NET Institute*

www.NETinst.org

Working Paper #16-09

September 2016

**The Effect of Word of Mouth on Sales:
New Answers from the Comprehensive Consumer Journey Data**

Xiao Liu
Stern School of Business, NYU

Dokyun Lee
Carnegie Mellon University

Kannan Srinivasan
Carnegie Mellon University

* The Networks, Electronic Commerce, and Telecommunications (“NET”) Institute, <http://www.NETinst.org>, is a non-profit institution devoted to research on network industries, electronic commerce, telecommunications, the Internet, “virtual networks” comprised of computers that share the same technical standard or operating system, and on network issues in general.

The Effect of Word of Mouth on Sales: New Answers from the Comprehensive Consumer Journey Data

Xiao Liu, Dokyun Lee, and Kannan Srinivasan*

September 30, 2016

Abstract

Online Word-of-Mouth has great impact on product sales. Although aggregate data suggests that customers read review text rather than relying only on summary statistics, little is known about consumers' review reading behavior and its impact on conversion at the granular level. To fill this research gap, we analyze a comprehensive dataset that tracks individual-level search, review reading, as well as purchase behaviors and achieve two objectives. First, we describe consumers' review reading behaviors. In contrast to what has been found with aggregate data, individual level consumer journey data shows that around 70% of the time, consumers do not read reviews in their online journeys; they are less likely to read reviews for products that are inexpensive and have many reviews. Second, we quantify the causal impact of quantity and content information of reviews read on sales. The identification relies on the variation in the reviews seen by consumers due to newly added reviews. To extract content information, we apply Deep Learning natural language processing models and identify six dimensions of content in the reviews. We find that aesthetics and price content in the reviews significantly affect conversion. Counterfactual simulation suggests that re-ordering review content can have the same effect as a 1.6% price cut for boosting conversion.

Keywords: Consumer Purchase Journey, Product Reviews, Review Content, Deep Learning

*Xiao Liu, Stern School of Business, New York University, xliu@stern.nyu.edu; Dokyun Lee, Tepper School of Business, Carnegie Mellon University, dokyun@cmu.edu; Kannan Srinivasan, Tepper School of Business, Carnegie Mellon University, kannans@cmu.edu. We gratefully acknowledge financial support from the NET Institute www.netinst.org.

1 Introduction

Product reviews are becoming an increasingly important part of the purchase journey for consumers. According to BrightLocal’s 2015 Local Consumer Review Survey, 92% of consumers read online reviews, compared to 88% in 2014. Similarly, 2013 Dimensional Research survey show that, an overwhelming 90% of customers say that their buying decisions are influenced by online reviews. Prior literature using aggregate online word-of-mouth data (Chevalier and Mayzlin 2006) has found a positive impact of review ratings on sales. However, much is left unknown about consumers’ review reading behavior and its impact on conversion at the granular level. Specifically, when do consumers read reviews text above and beyond the summary statistics? What content in the review text influences consumers’ final purchase?

In this paper, we tackle these research questions by means of the following. First, we provide topology of consumer product purchase journey, incorporating not only traditionally available click-stream and transactional data, but also their review reading behaviors. Second, we investigate what type of content in reviews actually matter in shifting consumer purchase decisions by identifying and content-coding product-related information that has been shown in previous literature to influence purchase decisions. We then come up with managerial implications on how E-commerce site could present review information to customers to improve conversion. Last, we demonstrate how to utilize recent advances in Deep Learning to content-code review text at scale.

Figure 1: Framework

Our research framework is given in Figure 1. We first look for conditions where reviews have no impact on conversion. Specifically, we examine three scenarios. The first is when, or for what type of products, reviews do not affect conversion. The second is for whom, or which consumers, reviews do not affect conversion. And the last is where, or on which device, reviews do not affect conversion. The answers to these questions give us the boundary conditions under which reviews influence conversion. As a result, we only focus on the right products, right consumers and right devices to examine the effect of reviews on conversion. In particular, we quantify the effect of reviews on conversion from three perspectives, *volume*, *valance*, and *variety*. For *volume*, we look at whether more reviews causes higher conversion. For *valance*, we compare the effects of positive and negative reviews. And for *variety*, we conduct content analysis using state-of-the-art Deep Learning models and find out which features of the reviews have the strongest impact on conversion.

In contrast to what has been found with aggregate data, individual level consumer journey data shows that around 70% of the time, consumers do not read reviews in their online journeys; They are less likely to read reviews for products that are inexpensive and have many reviews; When consumers do read reviews, the number of negative reviews they read has a stronger impact than the number of positive reviews. Further,

we apply Deep Learning natural language processing models to identify six dimensions of quality content in the reviews. We find that aesthetics and price content in the reviews significantly affect conversion. A counterfactual study suggests that re-ordering review content can have the same effect as a 1.6% price cut for boosting conversion.

The paper makes several substantive and methodological contributions. First, it leverages a comprehensive dataset to provide descriptive statistics of consumers’ review reading behaviors. The findings can help understand when, for whom and where reviews play a role in the consumer purchase decision journey. Second, we propose a novel identification strategy to quantify the causal impact of review reading on conversion. The exogenous variation of reviews comes from the newly posted reviews. Third, we conduct content analysis to extract distinct dimensions of price and quality information from the reviews. Instead of relying on surface-level measures like volume or valence, we dissect reviews to its core information, price and quality, which differentiates product reviews from other text documents like social media posts or news articles. The labeled review documents are of high precision and can be used for future studies by other researchers. Last, we introduce state-of-the-art Deep Learning techniques to the marketing literature. We demonstrate comparative advantages of Deep Learning models for automatic classification of unstructured data at scale and visualization of salient content.

The rest of the paper is organized as follows: First, we review relevant literature and identify the research gap. Second, we describe the data and exhibit model free evidence of consumers’ review reading behaviors. Third, we present the model and Deep Learning techniques to extract content features. Next, we discuss the results and counterfactual simulations. Finally, we summarize the managerial implications and limitations.

2 Literature Review

2.1 Effect of User Generated Content on Conversion

Research on the relationship between product reviews and sales has been proliferate in the past decades. Many researchers have found a positive relationship between product reviews and sales and attribute this to either a correlation or a causal relation (See Table 1 for a summary). On one hand, online reviews can reflect consumer preferences, therefore can be used to predict sales. This has been supported by Resnick and Zeckhauser (2002) for eBay products, Godes and Mayzlin (2004) for TV shows and Liu (2006), as well as Dellarocas et al. (2007) for movies. On the other hand, product reviews can directly influence consumer purchase decisions. For example, Chevalier and Mayzlin (2006) as well as Zhang and Dellarocas (2006) find that online ratings significantly influence sales in the book and movie industry separately. In contrast to these findings, Chen and Xie (2005) and Duan et al. (2008) find that product reviews do not have causal impact but merely serve as predictors for sales.

Table 1: Literature of The Effect of UGC on Demand

Study	Method	Data	Predictor or Causal
Resnick and Zeckhauser (2002)	Multiple regression	eBay 99	Predictor
Godes and Mayzlin (2004)	Multiple regression	TV Shows 99-00	Predictor
Chen and Xie (2005)	Multiple regression	Amazon Books 03	Causal
Chevalier and Mayzlin (2006)	Differences-in-differences	Books 03-04	Causal
Zhang and Dellarocas (2006)	Diffusion model	Movies 03-04	Causal
Duan et al. (2008)	Simultaneous system	Movies 03-04	Causal
Zhu and Zhang (2010)	Differences-in-differences	Video games 00-05	Causal

While these above findings are illuminating, they are all subject to the following limitations. First, they only focus on one product type or one product category. The fact that conflicting results have been found among Chevalier and Mayzlin (2006), Zhang and Dellarocas (2006), Chen and Xie (2005), and Duan et al. (2008) suggests that the effect between reviews and conversion may vary for different types and categories of products. It’s puzzling for which category and type of products should the effect exists. Zhu and Zhang (2010) give a first attempt to address this question and find that reviews are more influential for less popular products. But this work only looks at one product category, video games. We instead study more than 500 product categories and provide a systematic view on how product category moderate the relationship between reviews and demand. Second, all previous papers use aggregate measures of consumer response, hence overlook granular level behaviors. We instead take advantage of an individual level consumer panel to zoom in on consumers’ review reading behaviors. The quantified causal impact only applies to reviews “read” by consumers rather than all reviews posted. Finally, prior literature largely ignore the rich content information of reviews. Rather staying with the surface-level measures like volume, valence or variance, we extract multi-dimensional content features from the reviews and explain what content information makes some reviews more influential than others.

2.2 Unstructured Data Analysis

Recent reports by Oracle show that 80% of data enterprise and companies deal with are unstructured data such as text, images, and videos¹. However, the biggest challenge with using unstructured data is that the data require pre-processing and content-extraction, which are often not scalable. A large portion of this data is user-generated and in our case, we need to analyze the reviews generated by users on e-commerce sites to truly understand what type of information learning is happening during search and transactions. Taming the unstructured text data by content-coding and extracting specific information is the first step in this endeavor and now we describe the details on what content we profile and extract as well as natural language processing techniques we utilize.

¹<http://www.oracle.com/us/solutions/ent-performance-bi/unlock-insights-1885278.pdf>

2.2.1 What Review Content Matters: Price and Quality

Price and quality of products has been the main drivers of economic transactions and consumers' purchase behavior in both online and offline. Thus, we look at how price and quality information within reviews influence consumer's purchase decision. While the price dimension of a product is unambiguous, the quality dimension requires a framework to define, identify, and content-code before we can measure the effect of reviews that contain these information on consumer purchase behavior. We take a seminal work by Garvin (1984) to identify and operationalize different dimensions of product quality found to influence purchase behavior. Garvin (1984) and Garvin (1987) introduced a set of quality dimensions that were aimed at helping organizations think about quality in terms of a multi-dimensional strategy. The eight dimensions proposed were: performance, features, reliability, conformance, durability, serviceability, aesthetics, and perceived quality. We closely follow Garvin's definitions of different dimensions to identify quality information in reviews. For some qualities that are conceptually close, we combine them. We describe each dimensions in Table 3.

2.2.2 Identifying Content in Large-scale Text Data

We have 74,958 reviews from 11,443 unique products. Since content coding 74,958 reviews manually is not feasible, we utilized recent advances in natural language processing and machine learning to extract relevant information we have profiled for all the reviews in our dataset. Natural Language Processing (sub-task includes text mining) refers to the process of extracting useful, meaningful, and nontrivial information from unstructured text (Dörre et al., 1999). Typical text mining tasks include concept extraction, sentiment analysis, etc. and over the last several years, there has been an increasing emphasis on utilizing text data to improve all aspects of business (Pang and Lee, 2008; Liu, 2012). For example, we not only want to learn the sentiment of consumer reviews, but also want to examine consumer review content to see what content are actually causing consumers to buy after they read product reviews on e-commerce sites. We briefly describe variety of text mining techniques to content-code our review data.

Extracting and content-coding natural language review data is divided into two parts. First, for a subset of our data, we obtain gold standard tags (defined as the best available from humans) and information extraction phrases from the human workers on Amazon Mechanical Turk ("Turkers"), a market place for online labor that ranges from simple data cleaning to complicated psychological studies. We use a survey instrument provided in Appendix A to obtain content that we have identified above, using price and Garvin's product quality dimension framework. To ensure high quality responses from the Turkers, we follow several best practices identified in literature (e.g., we obtain tags from at least N different Turkers, choosing only those who are located in the English speaking countries, have more than 100 completed tasks, and an approval rate more than 97%. We also include an attention-verification question among many others). After obtaining the gold standard tags to comprise our training dataset, in the second part, we build our automated text-mining model to content-code and extract information from all review data we have. Particularly, we use various Deep Learning (LeCun et al., 2015; Bengio et al., 2006) models for the supervised learning task.

2.3 Mobile and PC Purchase Behaviors

Several features of mobile devices make the shopping behavior on them distinctive from that on PCs. On one hand, the screens of mobile devices are limited in size which increase search cost (Ghose et al. 2012) and

conversion friction². On the other hand, portability of mobile devices give consumers more access to search and shop opportunities (Ghose et al. 2013, Daurer et al. 2015) therefore are routinely used for purchasing habitual products (Wang et al. 2015a). We follow this stream of literature and examine the differential effect of review reading behaviors on conversion on mobile devices versus PC.

3 Descriptions of Consumers' Review Reading Behaviors

The data come from a major online retailer in the United Kingdom. It is a panel data for 243,000 consumers over the course of two months in February and March 2015. The data tracks all consumers behaviors, including page views, impressions, used features, and transactions. Specifically, a page view is a single view of a product-specific or category-specific page. Whereas an impression is a single exposure to a product review.³ A used feature refers to a consumer's interaction with some web-features. For example, a consumer can click a page number button to go to the next page. Or the consumer can sort the reviews by time or ratings. The data include two broad product categories, Technology as well as Home and Garden. Each broad category consists of hundreds of well-defined sub-categories. For example, Pillowcases is a sub-category for Home and Garden while Printers is a sub-category for Technology. In total, there are 583 sub-categories. Please see Figure for some examples of the product sub-categories. Among all these product categories, consumers had around 2.5 million page views, 12.3 million review impressions, 500,000 used features and 30,000 transactions. These actions were taken on one of the two devices, PC or mobile phone.

Figure 2: Wordcloud of Product Sub-Categories

Note: We only include categories with more than 100 journeys. The font size indicates the number of journeys associated with the product sub-category.

Next we provide descriptive statistics to characterize consumers' review reading behaviors.

²Retailers, Listen Up: High Rates of Mobile Shopping Cart Abandonment Tied to Poor User Experience. Retrieved from <https://www.jumio.com/2013/05/retailers-listen-up-high-rates-of-mobile-shopping-cart-abandonment-tied-to-poor-user-experience-pr/>

³The data also include textual content like questions and answers. We don't differentiate questions and answers from reviews but refer all of them as reviews.

Figure 3: Journey Distribution

3.1 Consumer Decision Journey – When Not?

Review is only one step in the consumer decision journey. In order to accurately quantify the effect of reviews on conversion, we cannot overlook the entire decision journey. We define a consumer decision journey as the sequence of actions between search and final purchase for a certain product.⁴ In our data, search is reflected by a page view, either on a category information page or on a specific product's information page. Between search and purchase, the consumer might read reviews to gather more information of the product. Given this definition, the consumer decision journey can be classified to five types, as depicted in Figure 3.

Type 1 journey is the shortest where the consumer directly purchases a product without any search or review reading actions. The consumer knows exactly which product to buy and does not need to collect any information. This journey type takes 2% of the sample. Type 2 journey only has the search stage. Surprisingly, this journey type takes up 66% of all the journeys, which suggests a very high bounce rate. During type 2 journeys, consumers have relatively low intention to purchase therefore do not make effort to read reviews. Type 3 journey contains two steps, search and purchase, and happens 3% of the time. Reviews are not used during type 3 journeys. Type 4 journey also has two steps, search and reading review(s). Consumers in type 4 journeys make intensive effort to look for both product information provided by the retailer as well as user-generated reviews. Due to certain reasons, consumers drop out before the final purchase. This type 4 journey covers 27% of the sample. Lastly, type 5 journey is the longest that comprises all three steps, search, reading review(s) and purchase. It only involves 2% of all the journeys.

Looking across all five types of journeys we find that for 71.2% of the time, consumers do not read reviews (type 1, 2 and 3). However if we exclude the type 2 journey where consumers do not have strong intention to buy, then 85% of the time, consumers do read reviews. This suggests that reviews play an important role in the consumers decision journey when they are serious about purchasing.

⁴Technically, one journey is constrained to only one product sub-category. So when a consumer switches to search in a different product sub-category, a new journey starts.

Table 2: Characteristics of Different Types of Journeys

Type	1	2	3	4	5
	no search + purchase	search + no review + no purchase	search + no review + purchase	search + review + no purchase	search + review + purchase
avg price	12.45	22.28	25.00	48.71	41.93
# reviews	107.82	26.50	79.10	47.17	62.73
% recommend	89.88	90.22	88.67	76.25	90.78
avg rating	4.36	4.26	4.31	3.99	4.39

Table 2 provides more descriptive features of the five types of journeys, including the average price of products in the journey, the average number of reviews for products in the journey, the average percentage of consumers who recommend products in the journey and the average ratings of products in the journey. Here's a summary of the interesting findings. First, reviews play a role when the product is relatively more expensive. This is because for expensive products, consumer engagement is high due to the high stake. Moreover, reviews play a role when the number of reviews is relatively medium which suggests that the product is neither too popular or not popular at all. We think this happens because consumer uncertainty is low for most popular or dominating products while reviews are needed when the uncertainty is high.

3.2 Product Analysis – When Not?

To echo the above findings, we give concrete examples of the product categories for which consumers care about (Figure 4) or not care about reviews (Figure 5). For instance, floor care products, bed frames, mattresses are all relatively more expensive products with high quality variation. Consumers need to rely on reviews to assess the product quality and fit. In contrast, Pay as you go phones, laptop and PC accessories and Xbox One games are relatively cheap with known product features and generally guaranteed quality. Consumers have low incentive to read reviews before purchasing them.

Figure 4: Examples of Products that Consumers DO Read Reviews

You
 phone
 Accessories
 Headphones
 Laptop
 Mobile
 CD
 DVD
 Blankets
 Chairs
 Dictionaries
 Calculators
 Tools
 Kits
 DiY
 House
 Bubble
 Packs
 Moving
 Portable
 DVD
 Players
 Accessories
 Window
 Cleaners
 Audio
 Speakers
 Webbrowsers
 iPod
 and
 Apple
 iPhone
 and
 CD
 or
 radio
 Accessories
 In
 car
 DAB
 and
 CD
 or
 radio
 Accessories
 Garden
 fencing
 and
 trellis
 Barbecues
 tools
 covers
 and
 fuel
 Personal
 CD
 players
 and
 cassette
 players
 TV
 stands
 and
 wall
 brackets
 Electric
 blankets
 Office
 supplies
 PS3
 Wall
 art
 pictures
 and
 photo
 frames
 Xbox
 360
 Mobile
 phone
 accessories
 and
 attachments
 PS
 via
 Nintendo
 Wii
 U
 Shirtdriers
 Digital
 photo
 frames
 Hard
 drives
 USBs
 and
 data
 storage
 Pad
 and
 tablet
 covers
 and
 cases
 Nintendo
 3DS
 2DS
 and
 DS
 Home
 office
 Electrical
 accessories
 Telephones
 and
 cameras
 Coffee
 machines
 Decorating
 hand
 tools
 and
 accessories
 Cameras
 and
 camcorders
 Wallpaper
 and
 decorating
 PC
 games
 and
 accessories
 Telephones
 and
 cameras
 Accessories
 Radio
 Clocks
 Nintend

We also find that reviews might have heterogeneous effects on consumers. Figure 6 demonstrates that around 43% of consumers never read reviews for any products (in the sample period) while 10% of consumers always read reviews for all products. This consumer heterogeneity might be attributed to different searching cost and different purchase intentions.

Frequency	Percentage
Sometimes	47%
Never	43%
Always	10%

Category	no review	review
mobile	10,000 (25%)	32,000 (75%)
pc	4,000 (8%)	56,000 (92%)

9

with the prior literature that finds that the smaller screen of the mobile device makes it less convenient to conduct in-depth search. We'll further test this effect in section 4 and 5.

The above analysis describes that for certain products, certain consumers and on certain devices, consumers do not pay attention to reviews in their online purchase journeys. So in order to quantify the effect of reviews on conversion, we need to select the right product that reviews matter, the right consumers who read reviews and the right device on which consumers read reviews. In the next section, we build models to do so.

4 Models to Quantify the Effect of Review Reading on Conversion

We propose two models. The first is the Cross-Sectional Model and the second is the Time-Series Model.

We adopt the random utility framework (Train 2009). As shown in equation (1) below, for consumer i , using device k considering product j at time t , the utility u_{ijkt} is determined by her intrinsic preference α_{ik} ⁵, observed consumer activities or characteristics vector $\vec{Z}_{it}$ (including total product searched and number of web-features used), product characteristics vector $\vec{X}_{jt}$ (including (log) price, average rating and cumulative number of reviews)⁶, other unobserved product characteristics ξ_j , review features vector for product j at time t $\vec{ReviewFeatures}_{jt}$ (details to be discussed later in Section 4.2), and an idiosyncratic shock ε_{ijkt} .

$$u_{ijkt} = \alpha_{ik} + \vec{\theta}_k \vec{Z}_{it} + \vec{\gamma}_k \vec{X}_{jt} + \xi_j + \vec{\beta}_k * \vec{ReviewFeatures}_{jt} + \varepsilon_{ijkt} \quad (1)$$

The intrinsic preference α_{ik} is related to factors such as income or willingness to purchase of consumer i 's and convenience of purchase using the device k . Unobserved product characteristics ξ_j are related to quality or popularity of the product. The shock term ε_{ijkt} is assumed to follow a Type I Extreme Value distribution. So the conversion rate has a closed-form formula. That is,

$$ConversionRate_{ijkt} = \frac{\exp(u_{ijkt})}{1 + \exp(u_{ijkt})}$$

Since our sample period is relatively short (only two months), we do not observe many repeated purchases from the same consumer. So we cannot obtain robust estimates of the consumer fixed effects. As a consequence, we make the assumption that consumers are homogeneous except for the observed characteristics. We think this assumption is reasonable because consumer purchase intention can be well-represented by her interactions with the web-features, like the total number of products searched and number of times to paginate or sort. Therefore we can eliminate the consumer fixed-effects and the regression becomes,

$$u_{ijkt} = \vec{\theta}_k \vec{Z}_{it} + \vec{\gamma}_k \vec{X}_{jt} + \xi_j + \vec{\beta}_k * \vec{ReviewFeatures}_{jt} + \varepsilon_{jkt}$$

Given the findings in Section 3.4, we hypothesize that there exists distinct consumer preferences on mobile devices versus PC devices. So we estimate the model using the mobile sample and PC sample separately⁷. Hence all the coefficients have a device subscript.

⁵We cannot separate the consumer from the device that she uses. So the intrinsic preference term has both the individual subscript i and the device subscript k .

⁶The product characteristics may vary overtime. For example, the e-commerce site performs the dynamic pricing strategy. From our conversation with the site managers, the pricing strategy is not targeted. So the concern of price endogeneity issue is eliminated.

⁷In the data, we observe less than 0.002% of the journeys that span across different devices. We eliminate these journeys from the sample used in the regressions.

4.1 Cross-Sectional Model vs. Time Series Model

For the cross-sectional analysis, we further assume that products are identical in terms of the unobserved characteristics ($\xi_j = \xi_{j'}$) so we can also eliminate ξ_j . The model specification hence becomes

$$u_{ikt} = \vec{\theta}_k \vec{Z}_{it} + \gamma_k X_{jt} + \vec{\beta}_k * \overrightarrow{ReviewFeatures}_{jt} + \varepsilon_{ikt} \quad (2)$$

The Cross-Sectional Model above aims to quantify the causal impact of the reviews on conversion. However, in reality, consumers' review reading behavior and purchase behavior are jointly determined, which leads to the endogeneity problem. In particular, when we observe that consumers who read more reviews are also more likely to make a purchase, it could be due to product quality rather than the effect of reviews. In other words, a high quality product will attract more consumers to read reviews and finally purchase it than a low quality one. Considering these, we need to look for exogenous changes in (the features of) reviews and control for product unobservables. This leads to our time series specification:

$$u_{ikt} = \vec{\theta}_k \vec{Z}_{it} + \xi_j + \vec{\beta}_k * \overrightarrow{ReviewFeatures}_{jt} + \varepsilon_{ikt} \quad (3)$$

We estimate this equation as a fixed effect model. Next, we discuss the identification strategy used in the time-series analysis.

4.1.1 Identification Strategy

Taking a first difference of equation 3, we have,

$$\Delta u_{ijk} = \vec{\theta}_k * \Delta \vec{Z}_i + \vec{\beta}_k * \Delta \overrightarrow{ReviewFeatures}_j + \varepsilon_{ijk} \quad (4)$$

So the identification of the parameters $\vec{\beta}_k$ comes from the within-product variations of review features. We take advantage of the fact that during the sample period, new reviews appear on the site which triggers the exogenous change of the reviews features. To illustrate, Figure 8 below shows the product review section on the webpage of a mattress. Up to June 28 2015, the product had two reviews. On July 1 2015, a new review was submitted which increased the total number of reviews to three. From a buyer's perspective, this change in number of reviews is exogenous. Since product characteristics remain unchanged, the mere change of the number of reviews as well as the associated review features allow us to identify the effect of reviews separately from unobserved product effects.

Figure 8: Example of Changing Number of Reviews

Although the Cross-Sectional Model suffers from endogeneity bias, we keep it because it allows us to use more data since during the sample period, the reviews of many products did not change. Our main insights of the causal impact of reviews on conversion come from the time-series analysis.

4.2 Review Features

In this subsection, we discuss what review features are considered in the models represented by equation (2) and equation (3). In a nutshell, we examine three types of features: *volume*, *valence*, and *variety*.

4.2.1 Volume Model

In the volume model, we calculate the marginal effect of number of reviews read on conversion. The element of review features in equation (2) and equation (3) becomes

$$\vec{\beta}_k * \overrightarrow{ReviewFeatures}_{jt} = \beta_k * \#Reviews_j + \vec{\lambda}_k \vec{C}_j \quad (5)$$

Different from the aggregate measure of cumulative number of reviews associated with the product, the $\#Reviews_{ijt}$ here measures the number of product j -related reviews read by the consumer i at time t . One endogeneity concern here is that consumers who have higher purchase intentions will read more reviews. Although we cannot totally remove this concern, it is alleviated by the exogenous change of more reviews appearing on the site. In the data we find that, when the number of reviews increases, (e.g. from n to $n + 1$ where $n \in \{0, 1, 2, 3, 4\}$), the number of reviews read by consumers also increases linearly. This implies that the number of reviews read by a consumer can be exogenously shifted by the number of reviews available on the website. We rely on this exogenous variation of number of reviews read for identification. We also include a list of control variables $\vec{C}_j$ to be explained later.

4.2.2 Valence Model

The Volume Model treats all reviews as the same, ignoring different sentiments of the reviews. However, past literature (Tirunillai and Tellis 2012) has suggested an asymmetric effect of positive versus negative reviews. As a result, we look at the effect of positive and negative reviews separately. So in the Valance Model, we change the element of reviews features in equation (2) and equation (3) to equation (6). We define a review as positive if its rating is 4 or 5 as to negative if its rating is 1 to 3. The effect of the number of positive reviews is captured by β_{ik}^p while the effect of the number of negative reviews is captured by β_{ik}^n .

$$\vec{\beta}_k * \overrightarrow{ReviewFeatures}_{jt} = \beta_k^p * \#PosReviews_{jt} + \beta_k^n * \#NegReviews_{jt} + \vec{\lambda}_k \vec{C}_j \quad (6)$$

Similar to what is explained before, the number of positive reviews read and the number of negative reviews read vary exogenously due to newly added reviews.

4.2.3 Variety Model

Finally, we dig deeper into the content of the reviews. On top of the number of positive and negative reviews, we consider the “Quality” and “Price” information embedded in the review content. As discussed in Section 2.2.1, we content-code price and quality of product information in reviews. Specifically, we include six dimensions: “Aesthetics”, “Conformance”, “Durability”, “Feature”, “Perceived Quality” and “Price”. We consider these attributes to be the main focus of review content analysis.

$$\begin{aligned} \vec{\beta}_k * \overrightarrow{ReviewFeatures}_{jt} = & \beta_k^p * \Delta \#PosReviews_{jt} + \beta_k^n * \Delta \#NegReviews_{jt} + \vec{\lambda}_k \vec{C}_j \\ & + \beta_k^a \Delta S_Aesthetics_{jt} + \beta_k^c \Delta S_Conformance_{jt} + \beta_k^d \Delta S_Durability_{jt} \\ & + \beta_k^f \Delta S_Feature_{jt} + \beta_k^{pq} \Delta S_PerceivedQuality_{jt} + \beta_k^{pr} \Delta S_Price_{jt} \end{aligned} \quad (7)$$

We calculate the marginal effects of these information variables by replacing the element of reviews features in equation (2) and equation (3) with equation (7).

We also include a few control variables that have been found in the previous literature to influence conversion. These include “Length”, “Readability” and “Sales”. We explain the rational of using them one by one. First, “Length” measures the number of words in the reviews read. We include this variable because longer

reviews provide more detailed information which can strongly affect readers' decisions. Second, Ghose and Ipeirotis (2011) have shown that high readability of reviews are linked to increased sales. So we automatically calculate and control for the measure of readability using a widely used metric called the SMOG Index ("Simple Measure of Gobbledygook"). Higher values of SMOG implies a message is harder to read (Mc Laughlin 1969). Third, reviews sometimes contain sales information. For example, the sentence "really pleased with this cover, got it on sale. so even better. looks great" indicates that this consumer purchased the product on sale. We hypothesize that readers of this review will be negatively affected by it because the sale is temporary and the reader might not have access to the lower price any more.

4.3 Deep Learning

To extract the six dimensions of information from review texts, we use state-of-the-art Deep Learning natural language processing models. Deep Learning stems from Machine Learning, which employs computer science and statistics algorithms that can automatically learn patterns from data and make predictions. Conventional Machine Learning models are limited by their inability to process raw and unstructured data without the careful feature engineering inputted by humans. For example, when dealing with text data, the identification of sentence-level attributes such as part-of-speech, coreference resolution, negation detection, etc needed to be hand-coded or explicitly extracted and entered into existing machine learning models. As a result, text mining algorithms would often get confused or miss many natural language subtleties entirely - if not explicitly encoded. In contrast, recent advancements in Deep Learning enables us to explore unstructured text data without the ad-hoc and error-prone feature engineering so that the entire process can be automated. Essentially, Deep Learning evolved from an already existing machine learning technique called the artificial neural networks, but models high level abstractions and patterns in data by using a deep graph with multiple processing layers, composed of multiple linear and non-linear transformations. It involves many improvements in techniques to overcome shortcomings in previous artificial neural networks model estimation (LeCun et al., 2015; Bengio et al., 2006). Using Deep Learning based text mining, we can let the data and sophisticated algorithm detect natural language subtleties instead of hand-coding the sentence-attributes to enter as X-variables in typical classification models.

In our analysis, we utilize Deep Learning for two purposes. First, we use various Deep Learning models as supervised learning classifiers to identify pre-profiled price and quality content. Second, we use a visualization technique from the computer vision literature to highlight sentences that are topic-relevant. Next we explain these two applications in details.

4.3.1 Supervised Learning

As mentioned in section 2.2.1, we believe that the price and quality information of products embedded in the reviews are key drivers of consumer purchase. However, there is no prior work in machine learning or natural language processing that has identified useful content features that represents price and quality information. Instead of performing the ad-hoc, error-prone and time-consuming feature engineering, we rely on Deep Learning models to discover intricate textual structures in high-dimensional data to identify specific content in a large number of reviews.

We conduct supervised learning in multiple steps. First, we collect a labeled dataset of 5000 random reviews. To obtain this labeled set of data, we hire workers from Amazon Mechanical Turk (henceforth "AMT") to provide labels to these review. AMT is a crowd sourcing marketplace for simple tasks such as data collection, surveys, and text analysis. It has now been successfully leveraged in several academic papers for online data collection and classification. To content-code our reviews, we create a survey instrument comprising of a set

of binary yes/no questions we pose to workers (or “Turkers”) on AMT. For each review, we ask Turkers to identify whether each of the six dimensions of information exists in the text and what the associated valence of each dimension is. In other words, we ask Turkers to do both detection and sentiment analysis on reviews along each information dimensions. Table 3 below is a description of the six dimensions.

Table 3: Six Dimensions of Information in the Reviews

Dimension	Description
Aesthetics	The review talks about how a product looks, feels, sounds, tastes or smells.
Conformance	The review compares the performance of the product with per-existing standards or set expectations.
Durability	The review describes the experience with the durability or product malfunctioning or failing to work as per the customer’s satisfaction
Feature	The review talks about presence or absence of product features
Perceived Quality	The review talks about indirect measures of the quality of the product like the reputation of the brand
Price	The review contains content regarding price of the product

For example, a review that says “TV looks good but it’s too expensive” will be identified as having positive Aesthetics information and negative Price content. To ensure high quality responses from the Turkers, we follow several best practices identified in literature (e.g., we obtain tags from at least 5 different Turkers choosing only those who are from the U.S., have more than 100 completed tasks, and an approval rate more than 97%. Turkers also have to pass a short test to be qualified.) Please see the Appendix A for the final survey instrument and Appendix B for the complete list of strategies implemented to ensure output quality. Figure 16 in Appendix B presents the histogram of Cronbach’s Alphas, a commonly used inter-rater reliability measure, obtained for 5,000 reviews. The average Cronbach’s Alpha for our tagged reviews is 0.84 (median 0.88), well above typically acceptable thresholds of 0.7. About 84% of the reviews obtained an alpha higher than 0.7, and 90% higher than 0.6. For robustness, we replicated the study with only those messages with alphas above 0.7 (4,193 messages) and found that our results are qualitatively similar. At the end of the AMT step, approximately 800 distinct Turkers contributed to content-coding 5,000 messages. This constitutes the labeled dataset for the Deep Learning algorithm used in the next step.

Second, the labeled data is divided into a training set with 70% of the observations and a test set with the rest 30% of the observations. We then perform content detection and sentiment analysis by training various models on the training dataset and test the classification accuracy using the test dataset. This is a two-step process. For each review, we begin with detecting whether the content dimensions exists. If yes, a sentiment analysis is followed. For sentiment analysis, we median split the Likert scale and turn it into a binary classification problem of positive vs negative sentiment. We train the models, to be introduced later, separately for each of the six dimensions of information listed in Table 3, namely Aesthetics, Conformance, Durability, Feature, Perceived Quality, and Price. Third, we perform a prediction task to classify the rest of the 57,685 reviews so that each review will have twelve scores that indicate the existence and sentiment of each of the six content dimensions respectively.

We apply both conventional machine learning models and Deep Learning models for content detection and sentiment analysis. Before introducing the Deep Learning models, we first explain the intuition behind traditional machine learning models to perform sentiment analysis. To start off, in its essence, sentiment analysis is a text classification problem. Therefore any existing supervised learning method can be applied, e.g., Naïve Bayes classifiers, support vector machines (SVM) (Joachims, 1999; Shawe-Taylor and Cristianini, 2000). So in a first application, Pang, Lee and Vaithyanathan (2002) take this approach to classify movie reviews into two classes, positive and negative. It shows that using unigrams or bag-of-words as features in

classification performs quite well because sentiment words, such as “good” “bad”, are the most important indicators of sentiments. However, this “bag-of-words” (BoW) representation and other simple representations, such as part-of-speech, ignore order of words, syntactic or semantic relations between words. To address these problems, follow-up works propose many feature engineering techniques. But as mentioned before, these techniques are usually domain-specific and time-consuming. This motivates us to use Deep Learning models.

Here we introduce the Deep Learning models employed in our analysis and the rational of using each of them.

Figure 9: Recurrent Neural Networks - Long Short Term Memory.

from “Predicting polarities of tweets by composing word embeddings with long short-term memory,” by Wang et al. 2015, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Vol. 1, pp. 1343-1353). Copyright 2015 by the Proceedings.com. Adapted with permission.

Recurrent Neural Networks - Long Short Term Memory (LSTM) The first Deep Learning model we implement is a Long Short Term Memory Recurrent Neural Networks (Wang et al. 2015b) which works with *sequences* and can simulate interactions of words in the compositional process. As shown in Figure 9, in this model, to characterize sequences, each word is mapped to a vector through a Lookup-Table (LT) layer. For each hidden layer, its input comes from two sources, one is the current Lookup-Table layer activations and the other is the hidden layer’s activations one step back in time. The last hidden layer is considered as the representation of the whole sentence. The example in the Figure shows that the three words, “not”, “very” and “good” are first mapped to a vector through the LT layer. And the last hidden layer $h(t)$ represents the entire (sub)sentence “not very good” to be used for classifying Y , the outcome variable. This model excels in distinguishing negation because it tunes vector representations of sentiment words into polarity-representable ones. Therefore, it shows promising potential dealing with complex sentiment phrases.

Figure 10: Recursive Neural Networks

from “Recursive deep models for semantic compositionality over a sentiment treebank,” by Socher et al. 2015, Proceedings of the conference on empirical methods in natural language processing (EMNLP) vol. 1631, (, 2013), pp. 1642. Copyright 2015 by the Proceedings.com. Reprinted with permission.

Recursive Neural Networks The second Deep Learning model is Recursive Neural Networks (Socher et al. 2013). Instead of focusing on sequences as done in the Recurrent Neural Networks, Recursive Neural Networks focuses on a more complicated *tree* structure. It can take phrases of any length as its input. Then it represents a phrase through word vectors and a parse tree using the same tensor-based composition function. As shown in Figure 10, in this model, one needs to compute parent vectors in a bottom up fashion. For example, the parent node p_1 uses a compositionally function g and node vectors b and c as features for a classifier. This method can accurately capture the sentiment change and scope of negation. It also learns that the sentiment of phrases following the contrastive conjunction “but” dominates.

Figure 11: Convolutional Neural Networks

from “Convolutional Neural Networks for Sentence Classification,” by Kim 2014, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha... (, 2014), pp. 1746--1751. Copyright 2014 by the Proceedings.com. Adapted with permission.

Convolutional Neural Networks The Recursive Neural Networks is very powerful, but it requires a parse tree which is not available in many settings. The last model, Convolutional Neural Networks (Kim 2014), has internal input-dependent structure that does not rely on externally provided parse trees. As displayed in Figure 11, a sentence of length $n = 9$ is first represented by a $n \times k$ dimensional matrix where k is the word vector length. Then a convolution operation, or a filter, applying to a window of $h = 2$ words creates a new feature (the middle layer). The model then applies a max-over-time pooling operation to represent a particular filter. Finally, the features from multiple filters are pushed into the softmax layer to produce the output. Given this structure, the Convolutional Neural Networks model captures short or long-range semantic relations between words that do not necessarily correspond to the syntactic relations in a parse tree. We refer readers to the original papers for more details of these models and their estimation algorithms.

4.3.2 Visualization to Extract Salient Sentences

In addition to using Deep Learning models to classify reviews, we can use Deep Learning to visualize the most salient sentences in the reviews in order to gain a better understanding of the content information. We implement a method created by Denil et al. (2014). It adapts visualization techniques from computer vision to automatically extract relevant sentences from labeled text data. Essentially, it is a Convolutional Neural Networks (henceforth “CNN”) model that has a hierarchical structure divided into a sentence level and a document level. At the sentence level the model transforms embedding for the words in each sentence into an embedding for the entire sentence. While at the document level another CNN model transforms sentence embedding from the first level into a single embedding vector that represents the entire document. Figure 12 is schematic of the model. Specifically, at the bottom layer, word embedding are concatenated into columns to form a sentence matrix. For example, each word in the sentence “I bought it a week ago.” is represented

by a vector of length 5. Then these vectors are concatenated to form a 7×5 dimensional sentence matrix (7 denotes the number of words in the sentence, including punctuation). The sentence level CNN applies a cascade of operations (convolution, pooling, and nonlinearity) to transform the projected sentence matrix into an embedding for the sentence. The sentence embeddings are then concatenated into columns to form a document matrix (the middle layer in the figure). In the example, the sentence embeddings for the first sentence “I bought it a week ago.” until the last sentence “The found it really really funny.” are concatenated to form the document matrix. The document model then applies its own cascade of operations (convolution, pooling and nonlinearity) to form an embedding for the whole document which is fed into a final layer (softmax) for classification. After training this model, it can then be used to extract salient sentences. The first step in the extraction procedure is to create a saliency map for the document by assigning an importance score to each sentence. These saliency scores are calculated using gradient magnitudes because the derivative indicates which words need to be changed the least to affect the score the most. Then the following step is to rank sentences based on the saliency score and highlight the sentences with the highest score.

Figure 12: Using Convolutional Neural Networks to Extract Salient Sentences

from “Extraction of Salient Sentences from Labelled Documents,” by Denil et al. 2014. Copyright 2014 by University of Oxford. Adapted with permission.

5 Results

5.1 Natural Language Processing Model Comparison

Table 4: Model Comparison: Sentiment Analysis

Classifier/Accuracy %	Mean	Aesthetics	Conformance	Durability	Feature	Perceived Quality	Price
SVM + BoW	0.482	0.465	0.607	0.472	0.427	0.618	0.486
NB+ BoW	0.474	0.541	0.473	0.633	0.295	0.588	0.384
Recurrent-LSTM	0.680	0.624	0.622	0.606	0.808	0.626	0.796
Recursive	0.606	0.627	0.597	0.622	0.573	0.602	0.615
Convolutional	0.846	0.854	0.813	0.826	0.868	0.840	0.872

We now compare the performance of various conventional machine learning models and Deep Learning models. Table 4 shows the sentiment analysis accuracy of various models in the test sample in each information dimension. The counterparts in the information detection task is shown in Appendix C. The top two rows in Table 4 are for conventional machine learning models using the bag-of-word representation while

row 3 to 5 are for Deep Learning models. The conventional classifiers, Support Vector Machine (SVM) and Naive Bayes (NB) both have an average prediction accuracy of 48%. On the contrary, Deep Learning models generally have better performance, with Recurrent Neural Networks’s average accuracy of 68%, Recursive Neural Networks’s 61%, and Convolutional Neural Networks’s 85%. Dimension-wise, we have mixed results. Feature and Price have relatively higher accuracy than other dimensions for Recurrent Neural Networks and Convolutional Neural Networks, but not for Recursive Neural Networks.

To explain why Deep Learning Models have better prediction performance, we examine reviews that are correctly predicted by Deep Learning models but incorrectly predicted by conventional machine models. Table 5 illustrates some examples for each of the Deep Learning models.

Table 5: Examples of Reviews Correctly Classified by Deep Learning Models but Not Conventional Machine Learning Models

	Example 1	Example 2
Recurrent	The curtain is the least appealing	The toy is hardly surprising
Recursive	although the parts when they are spread out initially seem daunting. Looks great in our conservatory	It is good for the money but too flimsy
Convolutional	Without this battery, my phone is useless	The bed is not only comfortable but also pretty.

For Recurrent Neural Networks, it excels in distinguishing negation. So keywords like “least appealing” and “hardly surprising” are detected for expressing negative sentiments. Recursive Neural Networks, which relies on a tree structure to decipher syntactic relations, can discover that phrases following the contrastive conjunction “but” dominates the entire sentiment. For instance, it correctly pinpoints that a review that states “It is good for the money but too flimsy” conveys a negative sentiment about the aesthetics of the product. Last, a Convolutional Neural Network which captures local cues can recognize that sentences with many negative sentiment words can express positive sentiment semantically. For instance, although the review “Without this battery, my phone is useless” contains negative words like “without” and “useless”, the entire sentence delivers a positive message about the battery.

Given the advantages of Deep Learning models, we choose them to perform classification jobs. Considering the fact that Convolutional Neural Networks has the best prediction performance, in the rest of paper we report results generated from the Convolutional Neural Networks.

5.2 Visualize Saliency Sentences in Reviews

Next we show the effectiveness of the Deep Learning model to correctly detect distinct dimensions of information in the reviews. In Figure 13, for each of the six dimensions of information, we exhibit one example for both the positive and negative sentiment. The full text of the review is shown in black and the sentences selected by the CNN appear in color.

Figure 13: Salient Sentences for Six Dimensions of Information in Reviews

	Positive	Negative
Aesthetics	As expected worked brilliantly straight out of box. It is a 500GB as description and not 1TB as in title but that is still a good amount of storage. Graphics are fantastic.	This set is very poor quality, after using for one week only and then washing the fabric is all bobbly and feels rough, looks bad too. I won't be using it again!
Conformance	it does the job. would recommend for the price.good spin. although low temp wash has a smaller spin i just spin the clothes again on the 1200. my last beko machine lasted 7 years.family of 4.	I got these baskets to fit on shelves. but when I got them they were a lot smaller than the picture on line and so do not hold very much. Wasted my money!!!!
Durability	I have had this caddy for nearly 2 years and so far there has been no sign of rust. It looks fantastic in our en-suite and everything is now much tidier.	I don't recommend this product - it has broken after first use. Most likely the rubber seal.Tower Customer service is terrible, not helpful at all.Please chose another product and avoid dissapointment.
Feature	Extremely pleased with this mirror, great quality mirror - nice sharp reflection. We love the bevelled edge. Fully recommend.. Top stars	I have only used it twice, but I will be returning it. You pay for what you get I suppose. it takes ages, doesn't make the toast crispy and is hard to clean.
Perceived Quality	I purchased this camera after looking at different options and reading their reviews, this camera had very good reviews and I am glad I went with this Canon Powershot. It is comfortable to hold easy to follow the instructions ,and it has some great features,I even managed to send photos to my daughters Iphone by following the instruction manual which I did have to download but it was easy to follow. The photos that I have taken so far are very good so overall a very good choice.	I thought I would buy the genuine Weber cover for my new BBQ. This was a big mistake. The cover is too short to adequately cover it. There are no ties or fixings so the wind gets under it and lifts it off.Its now off again and my new BBQ is exposed to the elements. I think it needs to go back and Weber needs to revise this product
Price	excellent, made several meals in this cooker, meat is cooked so tender it drops off the bone. features good, easy too understand instructions. may be the best thing we have bought for the kitchen, and at its price it is an absolute steal	I thought this was rather flimsy and not very robust. I will be surprised if it lasts any length of time. A bit expensive for what it is.

The examples demonstrate that CNN can correctly locate the review fragment that corresponds to the particular information dimension.

5.3 Review Effect

5.3.1 Cross-Sectional Analysis

We first show the results of the cross-sectional analysis, assuming that there are no product-specific fixed effects. In Table 6⁸, we present the estimates of the three models introduced in section 4.2 (equation 2, 3, 5, 6 and 7). For each model, we present the results for the mobile sample and the PC sample separately.

From the Volume Model we find that on mobile device, reading more reviews makes a consumer more likely to purchase the product. One additional review can boost the conversion rate odds ratio by 1.6%⁹. Note that our identification comes from the fact that more reviews arrive to the site overtime. Hence, it reflects that when more reviews are available and consumers read more reviews, their purchase likelihood increases. This result should not be interpreted as consumers who have higher purchase intention choose to read more reviews because we have already controlled for consumer purchase intention using the consumer activity data. We warn readers of the generalizability of this result. Due to the specific design setting on this website, consumers will simultaneously see reviews in groups of five. So this result is driven by variations of the number of reviews when the total number of reviews is relatively small. For example, when the total number of reviews increases from 3 to 4 or from 7 to 8. This being said, we still think that this scenario is

⁸The summary statistics of the variables in these models are presented in Appendix D

⁹Odds Ratio= $\exp(0.0155)-1 \approx 1.6\%$

Table 6: The Effect of Review Reading on Conversion

	Model 1: Volume		Model 2: Valence		Model 3: Variety	
	Mobile Est.(Std.)	PC Est.(Std.)	Mobile Est.(Std.)	PC Est.(Std.)	Mobile Est.(Std.)	PC Est.(Std.)
Log_Price	-0.296*** (0.0234)	-0.247*** (0.0243)	-0.300*** (0.0234)	-0.251*** (0.0243)	-0.297*** (0.0237)	-0.242*** (0.0246)
Total # of Reviews	0.000220*** (0.0000694)	0.000220*** (0.0000795)	0.000222*** (0.0000691)	0.000218*** (0.0000793)	0.000219*** (0.0000693)	0.000218*** (0.0000797)
Average Rating	0.519*** (0.0512)	0.533*** (0.0560)	0.411*** (0.0565)	0.409*** (0.0624)	0.388*** (0.0581)	0.389*** (0.0639)
# Products	-0.0586*** (0.00578)	-0.0405*** (0.00435)	-0.0582*** (0.00578)	-0.0403*** (0.00435)	-0.0587*** (0.00583)	-0.0424*** (0.00442)
# Used Features	0.120*** (0.0372)	0.284*** (0.0508)	0.113*** (0.0392)	0.291*** (0.0505)	0.112*** (0.0389)	0.287*** (0.0504)
# Reviews Read	0.0159*** (0.00490)	-0.0138* (0.00767)				
Low Readability	0.000567 (0.0128)	-0.00676 (0.0136)	0.00136 (0.0128)	-0.00654 (0.0136)	0.000820 (0.0128)	-0.00656 (0.0136)
Length	-0.00311* (0.00186)	-0.00350* (0.00200)	-0.00228 (0.00187)	-0.00249 (0.00200)	-0.00240 (0.00189)	-0.00242 (0.00203)
Sales Info	0.523 (0.340)	0.576 (0.375)	0.512 (0.340)	0.565 (0.374)	0.439 (0.343)	0.427 (0.376)
# Positive Reviews Read			0.0221*** (0.00540)	-0.00948 (0.00764)	0.0214*** (0.00537)	-0.00963 (0.00763)
# Negative Reviews Read			-0.0198* (0.0105)	-0.0563*** (0.0131)	-0.0137 (0.0108)	-0.0469*** (0.0135)
Review-Aesthetics					0.0452 (0.0774)	0.277*** (0.0824)
Review-Conformance					-0.0245 (0.156)	-0.0660 (0.166)
Review-Durability					0.0531 (0.116)	0.0195 (0.127)
Review-Feature					0.0949 (0.0695)	0.107 (0.0753)
Review-Perceived Quality					0.181 (0.224)	-0.388 (0.258)
Review-Price					0.117 (0.0855)	0.167* (0.0900)
Constant	-4.092*** (0.251)	-5.218*** (0.275)	-3.626*** (0.271)	-4.688*** (0.298)	-3.636*** (0.275)	-4.801*** (0.304)
Obs	37101	70680	37101	70680	37101	70680
BIC	15322.3	15631.3	15314.5	15624.3	15372.2	15670.7

quite consistent with the reality since consumers rarely read more than 10 reviews¹⁰. In contrast, its effect is negative on PC.

Moreover, we find negative price coefficients which suggest that when price increases by 1 percent, the odds ratio of conversion decreases by 34.6% on a mobile device and by 28% on PC. This implies that price has a stronger effect on mobile than on PC, consistent with the prior findings of XXX. Besides, both the total number of reviews and the average rating have significantly positive impact on conversion, as expected. Surprisingly, we find a significantly negative effect of the number of products searched in the journey. We believe that this variable indicates the purchase intention because if a consumer searches many products she must have a high willingness to buy. However, a counter force, competition, seems to take place. Recall that our dependent variable is the conversion of the product the reviews are associated with. If the consumer has a large consideration set, she is less likely to purchase any single product because of competition. So the regression coefficient picks up the competition effect more than the intention effect. The number of used features, for example pagination or sorting, is found to have a positive association with conversion because it reflects a higher purchase intention. Interestingly, the length of reviews read has a significant but negative effect on conversion. This echoes the finding in Chevalier and Mayzlin (2006). In contrast to Ghose and Ipeiritis (2011), we do not find Readability or Sales information to have significant effect on conversion.

For the Valence Model, new insights emerge. On a mobile device, when a consumer reads more positive reviews, her conversion rate gets higher. This effect becomes insignificant on PC, similar to the effect of the number of reviews read in the Volume Model. On the flip side, the effect of the number of negative reviews read has a significant negative effect on PC but not on mobile. Overall, the effect of the number of negative reviews is stronger than that of the positive reviews.

These puzzling opposite effects are replicated in the Variety model when we add more review text features. Remarkably, the review text features play a bigger role on PC than on mobile. Specifically, reviews containing favorable aesthetics and/or price information can significantly boost conversion, while other dimensions, like conformance, durability, feature or perceived quality are not prominent.

5.3.2 Time Series Analysis

The cross-sectional analysis is subject to the omitted variable bias where quality or popularity of the product is not controlled for. So the positive relationship between number of reviews read and conversion found in the volume model can be a result of high quality products both attracting consumer to read more reviews and generating higher sales. To account for this, we use the within-product but cross-time variation to pin down the coefficients for the review-related parameters. The results are presented in Table 7¹¹ including the three models in Section 4.2 on the mobile and PC sample.

Since there is too little variation over time, we cannot estimate the price coefficient or the coefficients for total number of reviews and average rating as we did in the cross-sectional analysis. Although the number of observations reduced in the time-series analysis because some products did not have new reviews during the sample period¹², most results remain qualitatively unchanged from the previous section. Again, when reading one more review on a mobile device, the odds ratio of consumer's conversion rate can increase by 1.6% ($\text{Odds Ratio} = \exp(0.0157) - 1 \approx 1.6\%$). The result does not hold on PC.

There are two exceptions compared with the results in the cross-sectional analysis. One is that in the Valence models, it shows that on mobile devices, both the number of positive reviews read and the number of negative

¹⁰According to a survey conducted by BrightLocal in 2014, 85% of consumers said they read up to 10 reviews. <http://searchengineland.com/88-consumers-trust-online-reviews-much-personal-recommendations-195803>

¹¹The summary statistics of the variables in these models are presented in Appendix E

¹²6390 out of 12974 products had new reviews appeared on the retailer's website during the sample period.

Table 7: The Effect of Change in Review Reading on Change in Conversion

	Model 1: Volume		Model 2: Valence		Model 3: Variety	
	Mobile Est.(Std.)	PC Est.(Std.)	Mobile Est.(Std.)	PC Est.(Std.)	Mobile Est.(Std.)	PC Est.(Std.)
# Products	-0.0684*** (0.00682)	-0.0468*** (0.00507)	-0.0670*** (0.00679)	-0.0460*** (0.00506)	-0.0691*** (0.00685)	-0.0481*** (0.00513)
# Used Features	0.105** (0.0408)	0.265*** (0.0541)	0.105** (0.0428)	0.281*** (0.0535)	0.0994** (0.0423)	0.272*** (0.0536)
# Reviews Read	0.0157*** (0.00544)	-0.0147* (0.00821)				
Low Readability	-0.0185 (0.0147)	-0.0168 (0.0154)	-0.0190 (0.0146)	-0.0186 (0.0153)	-0.0196 (0.0146)	-0.0178 (0.0153)
Length	-0.0105*** (0.00210)	-0.00915*** (0.00220)	-0.00844*** (0.00211)	-0.00690*** (0.00222)	-0.00793*** (0.00214)	-0.00622*** (0.00224)
Sales Information	0.295 (0.420)	0.644 (0.419)	0.260 (0.420)	0.625 (0.419)	0.0401 (0.422)	0.408 (0.420)
# Positive Reviews Read			0.0233*** (0.00589)	-0.00925 (0.00808)	0.0220*** (0.00582)	-0.00998 (0.00812)
# Negative Reviews Read			-0.0421*** (0.0113)	-0.0743*** (0.0132)	-0.0239** (0.0118)	-0.0553*** (0.0141)
Review-Aesthetics					0.239*** (0.0905)	0.365*** (0.0951)
Review-Conformance					-0.0473 (0.182)	0.0437 (0.187)
Review-Durability					0.262* (0.134)	0.182 (0.144)
Review-Feature					0.153* (0.0789)	0.142* (0.0839)
Review-Perceived Quality					-0.0896 (0.251)	-0.468* (0.276)
Review-Price					0.293*** (0.0973)	0.325*** (0.101)
Constant	-2.533*** (0.133)	-3.526*** (0.141)	-2.561*** (0.133)	-3.561*** (0.141)	-2.804*** (0.145)	-3.839*** (0.154)
Product Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Obs	30251	57867	30251	57867	30251	57867
BIC	13538.3	13714.6	13504.5	13686.5	13538.4	13716.7

reviews read have significant effect on conversion. Expectedly, positive reviews improves conversion while negative reviews dampens conversion. The second exception is that aesthetics information in reviews also can drive up conversion on a mobile device. Note that, in these two exceptions, the model coefficients become significant in the time-series model but are not significant in the cross-section model in spite of fewer observations in the time-series samples. Since the time-series analysis corrects for the endogeneity bias, our conclusions lean toward the findings in the time-series analysis. Note that the effects of the content information are still stronger on PC than on mobile. This is consistent with our model-free evidence in section 3.4 that consumer pay less attention to review content information on a mobile device than on PC.

5.4 Counterfactual of Changing the Ranking Algorithm

After discovering the relative importance of different content information in the review texts, we propose a strategy that marketers can leverage to boost conversion rate: re-ordering reviews. Our results in Section 5.3.1 implies that consumers not only pay attention to the summary statistics of reviews (e.g., average rating, total number of reviews) but also the actual content of reviews. Their conversion rate is influenced by the content information embedded in the reviews. For example, since aesthetics and price information have stronger positive impact on conversion than other dimensions, within the set of reviews with the same rating score, marketers can display the reviews with positive aesthetics and price information before other reviews.¹³

We implement a counterfactual scenario where for each product, we randomly select an associated review that contains positive aesthetics information and move it from a lower position to the set of reviews read by each consumer. We then calculate the conversion rate odds ratio for each product and the increase in conversion rate ratio compared to what is observed in the data. Figure 14 displays the histogram of the increase in conversion rate odds ratio. The average increase in odds ratio of the conversion rate is 44% while the maximum is 143%. Recall that in section 5.3.1 we find that increasing price by one percent can lead to an increase in odds ratio by 28% on PC. This indicates that on average, re-ordering reviews by presenting one more review with positive aesthetics information is as effective as a 1.6 percent price cut to increase conversion rate odds ratio.

Figure 14: Histogram of Increase in Odds Ratio by Re-ordering Reviews

¹³Similar practices have been undertaken by Amazon who changed its algorithm to determine which top reviews to display. See <http://www.geekwire.com/2015/amazon-changes-its-influential-formula-for-calculating-product-ratings/> for more details.

6 Conclusions and Limitations

The paper studies the role of reviews reading behaviors in the consumer purchase journeys. We leverage a unique granular-level dataset that tracks individual consumers' entire decision journey, including review reading, search, and purchase. This allows us to discover when (for what types of products), for whom (which consumers) and where (on which device) consumer read reviews as well as what features (volume, valance, and variety) of reviews have a causal impact on conversion.

The results can assist managers in multiple ways. First, managers can implement the Deep Learning models to automatically extract price and quality information from reviews. Second, based on our finding regarding the relative importance of review features, managers can incorporate reviews as a new marketing mix by refining the ranking and information presentation algorithms to provide most relevant reviews and content to consumers. Third, managers can collect real time information of the consumer purchase journey, including device and reviews read to predict final conversion more accurately.

The paper has several limitations. Currently, we only look at the effect of review reading behaviors on conversion. Another interesting angle is to examine the effect of reviews on consumer search behaviors. Questions like “will reading consistent reviews reduce consumer search?” or “will reading negative reviews before positive reviews drive consumers to increase the consumer consideration set” invite more investigation. Moreover, we have not accounted for consumer heterogeneity when quantifying the causal impact of review reading on conversion. This might be useful for marketers to design targeted review ranking and presentation algorithms.

References

- Bengio, Y., H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain: 2006, ‘Neural probabilistic language models’. In: *Innovations in Machine Learning*. Springer, pp. 137–186.
- Chen, Y. and J. Xie: 2005, ‘Third-party product review and firm marketing strategy’. *Marketing Science* **24**(2), 218–240.
- Chevalier, J. A. and D. Mayzlin: 2006, ‘The effect of word of mouth on sales: Online book reviews’. *Journal of marketing research* **43**(3), 345–354.
- Daurer, S., D. Molitor, M. Spann, and P. Manchanda: 2015, ‘Consumer Search Behavior on the Mobile Internet: An Empirical Analysis’. *Available at SSRN 2603242*.
- Dellarocas, C., X. M. Zhang, and N. F. Awad: 2007, ‘Exploring the value of online product reviews in forecasting sales: The case of motion pictures’. *Journal of Interactive marketing* **21**(4), 23–45.
- Denil, M., A. Demiraj, and N. de Freitas: 2014, ‘Extraction of Salient Sentences from Labelled Documents’. Technical report, University of Oxford.
- Dörre, J., P. Gerstl, and R. Seiffert: 1999, ‘Text mining: finding nuggets in mountains of textual data’. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 398–401.
- Duan, W., B. Gu, and A. B. Whinston: 2008, ‘Do online reviews matter An empirical investigation of panel data’. *Decision support systems* **45**(4), 1007–1016.
- Garvin, D. A.: 1984, ‘What Does Product Quality Really Mean?’. *Sloan management review* p. 25.

- Garvin, D. A.: 1987, 'Competing on the 8 dimensions of quality'. *Harvard business review* **65**(6), 101–109.
- Ghose, A., A. Goldfarb, and S. P. Han: 2012, 'How is the mobile Internet different? Search costs and local activities'. *Information Systems Research* **24**(3), 613–631.
- Ghose, A., S. Han, and K. Xu: 2013, 'Mobile commerce in the new tablet economy'. In: *Proceedings of the 34th International Conference on Information Systems (ICIS)*. pp. 1–18.
- Ghose, A. and P. G. Ipeirotis: 2011, 'Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics'. *Knowledge and Data Engineering, IEEE Transactions on* **23**(10), 1498–1512.
- Godes, D. and D. Mayzlin: 2004, 'Using online conversations to study word-of-mouth communication'. *Marketing science* **23**(4), 545–560.
- Kim, Y.: 2014, 'Convolutional Neural Networks for Sentence Classification'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar; A meeting of SIGDAT, a Special Interest Group of the ACL*. pp. 1746–1751.
- LeCun, Y., Y. Bengio, and G. Hinton: 2015, 'Deep learning'. *Nature* **521**(7553), 436–444.
- Liu, B.: 2012, 'Sentiment analysis and opinion mining'. *Synthesis lectures on human language technologies* **5**(1), 1–167.
- Liu, Y.: 2006, 'Word of mouth for movies: Its dynamics and impact on box office revenue'. *Journal of marketing* **70**(3), 74–89.
- Mc Laughlin, G. H.: 1969, 'SMOG grading-a new readability formula'. *Journal of reading* **12**(8), 639–646.
- Pang, B. and L. Lee: 2008, 'Opinion mining and sentiment analysis'. *Foundations and trends in information retrieval* **2**(1-2), 1–135.
- Resnick, P. and R. Zeckhauser: 2002, 'Trust among strangers in internet transactions: Empirical analysis of ebays reputation system'. *The Economics of the Internet and E-commerce* **11**(2), 23–25.
- Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts: 2013, 'Recursive deep models for semantic compositionality over a sentiment treebank'. In: *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Vol. 1631. p. 1642.
- Tirunillai, S. and G. J. Tellis: 2012, 'Does chatter really matter? Dynamics of user-generated content and stock performance'. *Marketing Science* **31**(2), 198–215.
- Train, K. E.: 2009, *Discrete choice methods with simulation*. Cambridge university press.
- Wang, R. J.-H., E. C. Malthouse, and L. Krishnamurthi: 2015a, 'On the go: how mobile shopping affects customer purchase behavior'. *Journal of Retailing* **91**(2), 217–234.
- Wang, X., Y. Liu, C. Sun, B. Wang, and X. Wang: 2015b, 'Predicting polarities of tweets by composing word embeddings with long short-term memory'. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 1. pp. 1343–1353.
- Zhang, X. and C. Dellarocas: 2006, 'The Lord Of The Ratings: Is A Movie's Fate is Influenced by Reviews?'. *ICIS 2006 Proceedings* p. 117.

Zhu, F. and X. Zhang: 2010, 'Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics'. *Journal of marketing* **74**(2), 133–148.

Appendix

A Survey Instrument to Content-Code Review Content

CONTENT DESCRIPTION

1. Price: Any content regarding the price of the item that's under review. The consumer may find the price too high, too low, or just right.

Example reviews with price content:

- "This overpriced junk broke after using twice!"
- "Fair price given it was less than 20 dollars."

2. Performance and Feature: This dimension involves observable and measurable attributes of the product. These include the products' primary characteristics that can be measured and compared. For example, if the product is an iPhone, the performance and feature attributes would include topics like screen size, weight, image and video resolution, camera megapixel, etc. If the product is a curtain, the performance could include topics about fabric feelings, size, laundry requirement, thickness, whether it blocks light etc.

Example reviews with performance and feature content:

- "The screen size is quite small at 3.5 inches"
- "The Aveeno Lotion's smell was great"

3. Reliability and Durability: Reliability reflects the probability of a product malfunctioning or failing to work as per the customers' satisfaction. For example, if a customer purchases a camera and finds operating defects within a short period of time, the product is ranked lower on the reliability. Durability measures the product life. Products may have high durability or lifespan (e.g., well built camera lens) or have low durability and lifespan (e.g., poorly made camera lenses which are fragile).

Example reviews with reliability and durability content:

- "These earbuds broke after 3 months of regular usage"
- "These new nokia phones are built like bricks! After we are gone, nokia will remain"

4. Conformance: This dimension reflects the degree to which the product's design and operating characteristics meet established standards. This dimension is perceived as the amount of divergence of the product feature specifications from an ideal or accepted standard. For example, if an automobile promises noise-free operation, but customers find that the car is actually quite noisy, then they would rank the car low on conformance.

Example reviews with conformance content:

- "The product does exactly what they says it would do...hydrating my dry skin."
- "The jacket wasn't rainproof as advertised!"

5. Aesthetics: This is a subjective measure. The aesthetic dimension captures how a product looks, feels, sounds, tastes, or smells, and is clearly a matter of personal judgment and a reflection of individual preference. For example, a person using an iPhone might feel that the phone has a "decent look and feel." This purely reflects the customer's own aesthetic preferences, as other customers might have differing opinions on what a "decent" look and feel might entail.

Example reviews with aesthetics content:

- "The lamp's sleek appearance is pleasing and I got many complements."
- "The color of the jean was not what I was looking for. It looks so cheap!"

6. Perceived Quality: Consumers do not always have complete information about a product or service's attributes, and hence, indirect measures may be their only basis for comparing brands. A leading source of perceived product quality is reputation of the brand. For example, consumers might prefer a new line of shoes purely because it comes from a leading shoe manufacturer that has a proven record of good quality e.g. Nike, Adidas etc.

Example reviews with perceived quality content:

- "Have been using HP ink for 5 yrs and think it's the best on the market!"
- "What's up with Samsung lately? The TVs are overpriced for what they offer!"

QUESTIONS

1. **[Price]** This review contains content regarding **pricing of product**

YES/NO

If you answered yes above, judge if the sentiment regarding this specific content is negative or positive. If answered no, then select Not Applicable.

Sentiment in Likert Scale 1 (Strongly Negative) 7 (Strongly Positive)

We exclude identical answer parts for other questions for brevity

2. **[Performance and Feature]** This review talks about **presence or absence of product features and performances**
3. **[Reliability and Durability]** This review describes the experience with the **durability or reliability or product malfunctioning or failing** to work as per the customer's satisfaction.
4. **[Conformance]** This review **compares** the performance of the product **with pre-existing standards or set expectations or as advertised**.
5. **[Aesthetics]** This review talks about **how a product looks, feels, sounds, tastes, or smells**, and is clearly a matter of personal judgment and a reflection of individual preference.
6. **[Perceived quality]** This review talks about **indirect measures of the quality of the product** like the reputation of the product brand or based on a history of past purchases.

B Amazon Mechanical Turk Strategies and Cronbach’s Alpha

Following best-practices in the literature, we employ the following strategies to improve the quality of classification by the Turkers in our study.

1. For each message, at least 5 different Turkers’ inputs are recorded. We obtain the final classification by a majority-voting rule.
2. We restrict the quality of Turkers included in our study to comprise only those with at least 100 reported completed tasks and 97% or better reported task-approval rates.
3. We use only Turkers from the US so as to filter out those potentially not proficient in English, and to closely match the user-base from our data (recall, our data has been filtered to only include pages located in the US).
4. We created a sample test and only those who passed this test, in addition to above qualifications, were allowed to work.
5. We refined our survey instrument through an iterative series of about 10 pilot studies, in which we asked Turkers to identify confusing or unclear questions. In each iteration, we asked 10-30 Turkers to identify confusing questions and the reasons they found those questions confusing. We refined the survey in this manner till almost all queried Turkers stated no questions were confusing.
6. To filter out participants who were not paying attention, we included an attention question that asks the Turkers to click certain input. Responses from Turkers that failed the verification test are dropped from the data.
7. On average, we found that review tagging took about 4 minutes and it typically took at least 30 seconds or more to completely read the tagging questions. We defined less than 30 seconds to be too short, and discarded any review tags with completion times shorter than that duration to filter out inattentive Turkers and automated programs (“bots”).
8. Once a Turker tags more than 20 messages, a couple of tagged samples are randomly picked and manually examined for quality and performance. This process identified dozens of high-volume Turkers who completed all surveys with seemingly random answers but manage to pass time filtering requirements. We concluded these were automated programs. These results were dropped, and the Turkers “hard blocked” from the survey, via the blocking option provided in AMT.

Figure 16: **Cronbach's Alphas for 5,000 Reviews**

C Model Comparison for Information Detection

Table 8: Model Comparison: Information Detection

Classifier/Accuracy %	Mean	Aesthetics	Conformance	Durability	Feature	Perceived Quality	Price
SVM + BoW	0.731	0.749	0.723	0.698	0.714	0.757	0.747
NB+ BoW	0.781	0.784	0.813	0.756	0.779	0.816	0.737
Recurrent-LSTM	0.855	0.802	0.864	0.796	0.910	0.922	0.834
Convolutional	0.872	0.832	0.816	0.858	0.895	0.911	0.920

Note, Socher et al. (2013) is only suitable for sentiment analysis. So we could not perform information detection for the Recursive Neural Networks model.

D Summary Statistics of Variables in the Cross-Sectional Regressions

Table 9: Summary Statistics of Variables for Cross-Sectional Analysis

Variable	N	Mean	Std Dev	Minimum	Maximum
Transaction	113725	0.03	0.18	0	1
Price	110946	87.68	111.23	0	2749.99
Total # of Reviews	113611	129.82	268.71	0.14	4779
Average Rating	113611	4.22	0.62	0.5	5
# Products	113725	6.81	9.09	1	118
# Used Features	113725	2.15	1.10	1	5
# Reviews Read	113725	9.78	7.56	0	195
Low Readability	113725	7.97	6.84	0	150
Length	113725	1.82	2.97	0	87
Sales Information	109480	32.35	15.30	2	105
# Positive Reviews Read	109348	8.60	2.07	3	25.58
# Negative Reviews Read	109480	0.01	0.06	0	1
Review-Aesthetics	109480	0.24	0.31	-1	1
Review-Conformance	109480	0.04	0.16	-1	1
Review-Durability	109480	0.01	0.23	-1	1
Review-Feature	109480	0.59	0.43	-1	1
Review-Perceived Quality	109480	0.02	0.10	-1	1
Review-Price	109480	0.28	0.28	-1	1

E Summary Statistics of Variables in the Time-Series Regressions

Table 10: Summary Statistics of Variables for Time Series Analysis

Variable	N	Mean	Std Dev	Minimum	Maximum
Transaction	90499	0.04	0.19	0	1
# Products	90499	6.40	8.47	1	118
# Used Features	90499	2.25	1.11	1	5
# Reviews Read	90499	10.53	7.65	0	195
Low Readability	90499	8.63	6.97	0	150
Length	90499	1.90	3.02	0	87
Sales Information	88170	32.09	14.46	2	100
# Positive Reviews Read	88118	8.58	1.96	3	21.17
# Negative Reviews Read	88170	0.01	0.06	0	1
Review-Aesthetics	88170	0.22	0.29	-1	1
Review-Conformance	88170	0.04	0.15	-1	1
Review-Durability	88170	0.01	0.21	-1	1
Review-Feature	88170	0.60	0.40	-1	1
Review-Perceived Quality	88170	0.03	0.11	-1	1
Review-Price	88170	0.28	0.27	-1	1