Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning

Abstract

How consumers use review content in their decision making has remained a black box due to the labor-intensive nature of extracting content from review text and the lack of review-reading behavior data. In this study, we overcome this challenge by applying deep-learning-based natural language processing on a comprehensive dataset that tracks individual-level review reading, searching, and purchasing behaviors on an e-commerce site to investigate how consumers use review content. We extract quality and price content from more than 500,000 reviews spanning nearly 600 product categories. We achieve two objectives. First, we describe consumers' review content reading behaviors. We find that although consumers do not read review content all the time, they do rely on review content for products that are expensive or of uncertain quality. Second, we quantify the causal impact of content information of read reviews on sales. We use a regression discontinuity in time design and leverage the variation in the review content seen by consumers due to newly added reviews. To extract content information, we use two deep learning models: a full deep learning model that predicts conversion directly and a partial deep learning model that identifies six theory-driven content dimensions. Across both models, we find that aesthetics and price content in the reviews significantly affect conversion across almost all product categories. Review content information has a higher impact on sales when the average rating is higher and the variance of ratings is lower. Consumers depend more on review content when the market is more competitive, immature, or when brand information is not easily accessible. A counterfactual simulation suggests that reordering reviews based on 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

1 Introduction

Consumers now live in a world where review ratings are often inflated and lose informational value. Uber disqualifies drivers who cannot maintain a 4.6-star (out of 5) average rating in Australia. More than half of Amazon reviews are 5-star[1]. The ubiquity of positive review ratings makes it impossible for consumers to use them as quality signals. Therefore, consumers increasingly read the review content when making purchase decisions. According to BrightLocal's 2015 Local Consumer Review Survey, 92% of the consumers surveyed acknowledge that they read online reviews. As a result, e-commerce platforms need to understand how review content impacts purchase behaviors in order to provide the most relevant content information.

And yet, despite the importance of product review content for business success, how consumers use the review content remains a black box. Prior studies on the impact of reviews are limited to easy-to-use volume, rating, and variance metrics that provide little information on the content of the reviews (Chevalier and Mayzlin, 2006). Those studies that do investigate the review content use simple sentiment or linguistic styles but not specific content in reviews regarding product quality and price. In this paper, we shed light on this black box by diving deep into the content of the reviews and measuring its impact on sales conversion.

In contrast to the studies using aggregate review metrics such as volume and valence, review content analysis poses four key challenges.

1. Data on review content reading behaviors are not available. When measuring the impact of review content on conversion, data on which reviews are read by consumers are indispensable. This is because even if it is safe to believe that consumers are aware of the total number of reviews and the average rating for the products they search, it is unreasonable to assume that consumers read all the content information of thousands of reviews. Therefore, without review reading data, the estimates for the review content effect would be biased. While researchers have access to consumer ratings as well as detailed reviews, the granular detail of how many and which reviews are actually read by consumers may be known to the firm but not to the researchers.

2. Content information in the reviews is represented by unstructured words and sentences. There is no guidance on which numerical variables to extract from reviews such that the variables can be relevant and insightful.

3. Traditional information retrieval and text mining methods are not applicable to cross-category analysis. Hundreds of product categories are present on e-commerce platforms such as Amazon or Walmart. A cross-category analysis on the impact of review content on purchase behavior has practical importance. However, prior studies on consumer content have been restricted to single or only few product categories. For cross-category analysis, researchers cannot depend on marketing experts to define useful content dimensions, as it would be too costly—if not impossible—to manually code content dimensions in hundreds of domains. Moreover, although existing natural language processing techniques can process large text corpora, they require extensive feature engineering, which is not scalable to multi-category content information. Therefore, a new framework and methodology for cross-category analysis is imperative.

4. Even if they know the relationship between review content and conversion, marketers still lack a practical strategy for using reviews to boost conversion.

In this paper, we tackle each of the research challenges.

1. We leverage a unique dataset to provide a topology of consumer product purchase journeys, incorporating not only traditionally available click-stream and transactional data but also consumers' review content reading behaviors. Although e-commerce companies, such as Amazon, track consumers' online actions

[1] See http://minimaxir.com/2014/06/reviewing-reviews/

using analytics tools such as Google Analytics, the data have not been available for academic research until recently. Fortunately, we have access to data that effectively measure the review content actually read by the consumers, and we leverage the data to provide insight on this concern. Specifically, we examine for what type of products, for how many consumers, and on which device review content does and does not affect conversion. This gives us the boundary conditions under which review content matters for conversion.

2. We open up the black box of how review content affects conversion by investigating what type of content in the reviews actually shifts consumer purchase decisions. Using a marketing theoretical framework, we identify and extract six distinct quality and price content dimensions from product reviews.

3. We use two deep learning models to tackle the problem of scalable information retrieval and quantify the impact of review content on conversion. The first model is a full deep learning model where content features and consumer/product characteristics are combined in a single deep learning framework to forecast conversion. We have developed the model to address the specific research issue we face. However, the proposed full deep learning model might have applications beyond this work. The second model is a partial deep learning model where we use deep learning natural language processing techniques for supervised learning and extract the six dimensions of quality and price information from the reviews. Then we pass these content dimensions to a traditional choice model to predict conversion. While the full model has better prediction performance and uses the unsupervised approach to portray salient content dimensions, the partial model provides asymptotics and valid statistical inference. The two models give qualitatively similar results. The deep learning architecture enables us to conduct the analysis on almost 600 product categories to establish generalizability. In comparison, most cross-category analyses in marketing deal with a far smaller set of categories, with most in the range of three to five categories.

4. We devise a new ranking algorithm for e-commerce sites to improve conversion and provide managerial implications for when review content plays a crucial role in consumer decision making.

Several interesting insights emerge from this effort. The numbering here corresponds to the research challenges outlined earlier.

1. Assuming that consumers read all the reviews would result in estimates that significantly understate the impact of review content on purchase conversion. We find that approximately 70% of the time, consumers do not read reviews in their online purchase journeys. They are likely to read reviews of products that are more expensive and about which they have many questions. It appears that as the number of reviews increases, consumers rely more on summary ratings. Importantly, assuming that consumers read all the reviews would lead to estimates that undervalue the effect of review content on conversion by an average of 20%.

2. Across numerous and diverse product categories, aesthetics and price content information are important dimensions. Given the enormous diversity of product categories, we believe it is highly significant that one key dimension of quality is important across virtually all categories. In contrast, conformance, durability, feature, and brand affect only a small number of categories. Moreover, we discover that review content information is more influential when the average rating is higher and the variance of rating is lower. In addition, consumers depend more on review content when the market is more competitive, immature, or when brand information is not easily accessible, implying that review content has a higher impact on sales when other sources of information are scarce.

3. Deep learning models outperform simple conventional natural language processing models. Deep learning models do not need time-consuming feature engineering and are scalable across different products. The comparative advantage of a deep learning model is that it allows us to sift through category-specific content features across a wide range of product categories without hand-coding features with human intervention or domain knowledge. In supervised deep learning, we merely need raw data (e.g., review text) with class labels (e.g., aesthetics information exists or not) without having to manually engineer features (e.g., sentence-level

attributes such as part-of-speech in text). This distinguishing aspect of deep learning makes it particularly useful in identifying polymorphic concepts of product quality in review texts, such as aesthetics, that vary across product categories in semantics. For example, words used to describe the aesthetics of a carpet may be different from words used to describe the aesthetics of a TV set. A simple "bag-of-words" model that relies on word frequency count may not have enough signal-to-noise ratio when we deal with many different product categories. Additionally, manually extracting domain specific features for all categories is often error-prone. Deep learning models outperform conventional natural language processing models by more than 27% in accurately labeling content information. 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 simple conventional machine learning models. Although the precision and recall of conventional text algorithms could become comparable to leading deep learning methods, conventional algorithms often involve very time-consuming engineering efforts of trial and error and are not scalable, in that the introduction of new product categories would require entirely new feature-engineering efforts with little or no economies of scale. By contrast, our approach is scalable, because the content-coding stays agnostic of domain-specific concepts (e.g., product categories) and can be readily applied to all types of textual data given that appropriate training data are available.

4. Firms can benefit significantly by simply reordering reviews based on content. In an interesting counterfactual, we find that simply reordering reviews based on their content will cause the same purchase conversion as that of a 1.6% price cut. Given the razor thin margins in e-commerce platforms, ordering reviews based on content is an economical way to achieve a high return on investment.

This paper makes several contributions to the marketing literature. Substantively, we open up the black box of how consumers use review content in their purchase journeys. Although many papers have studied the impact of online reviews, our paper is the first one to quantify the causal impact of specific review textual content on sales. The biggest differences between our paper and the extant literature are that we go beyond the easy-to-use metrics and analyze actual *content,* and we discover a *causal* relationship between review content and conversion. The second substantive contribution is that we study the effect of review content on conversion for a wide range of product categories, whereas prior works focus on only one or very few product categories. The cross-category analysis allows us to discover four dimensions of heterogeneity across product categories: competition, dynamics, brand, and rating. To our knowledge, no prior paper on product review has touched upon any of these four dimensions of heterogeneous effect. These substantive findings are highly valuable for e-commerce platforms and brands. On the methodology front, we propose a new deep learning model to directly predict sales. Instead of only using existing deep learning models from the computer science literature (the partial deep learning model), we create a full deep learning model that combines the convolutional neural network with consumer and product characteristics to model conversion in a joint framework. This full deep learning model is much more accurate and interpretable compared to extant methods.

The rest of the paper is organized as follows. In Section 2, we review relevant literature and identify the research gap. In Section 3, we describe the data and exhibit model-free evidence of consumers' review-reading behaviors. Then in Sections 4 and 5 we introduce the two deep learning models. The results and counterfactual simulations are presented in Section 6. Finally, we summarize the managerial implications and limitations.

2 Literature Review

Research on the relationship between product reviews and sales has been prolific in the past decades (See Table 1). Earlier studies have investigated simple and readily available summary statistics such as review vol-

ume, average rating, variance of rating, and how they influence consumer purchase decisions (e.g., Resnick and Zeckhauser 2002; Godes and Mayzlin 2004; Duan et al. 2008; Sun 2012. Refer to Table 1). However, as suggested in Chevalier and Mayzlin [2006], consumers do read review content beyond these summary statistics. Hence, the literature has evolved to also account for the textual aspect of the reviews. For example, Ludwig et al. [2013] extracted changes in affective content and linguistic style properties of customer book reviews to see how they influence conversion rate. A subsequent series of papers applied methods from natural language processing and extended the literature by looking at the effect of sentiment or lexical content of user-generated reviews on conversion (e.g., Schneider and Gupta 2016). These papers found that simplistic content features such as sentiment are correlated to conversion.

As such, recent papers now focus mostly on review text rather than summary statistics. Papers studying the impact of text on business outcomes face the challenge of structuring and extracting content. There are two approaches to extracting relevant content: One is to use human coders to identify information manually, and the other is to rely on natural language processing (NLP) and machine learning techniques to achieve scalability. Several studies have applied NLP to marketing problems (Eliashberg et al., 2007; Decker and Trusov, 2010; Lee and Bradlow, 2011; Archak et al. 2011; Netzer et al. 2012; Tirunillai and Tellis 2012; Liu et al. 2016; Schneider and Gupta 2016; Nam et al. 2017; Puranam et al. 2017; Lee et al. 2018). Here we focus on papers investigating the impact of consumer generated reviews on sales or market share. To emphasize scalability and low cost, a series of papers have proposed an unsupervised method to summarize and extract content from a massive amount of review data, to understand consumer preference and identify competitors using lexical-based semantic networks (Netzer et al. 2012), extract latent dimensions of consumer satisfaction (Tirunillai and Tellis 2014) using Latent Dirichlet Allocation (Blei et al. 2003), and better understand what consumers are saying by extending latent Dirichlet allocation (B¨uschken and Allenby 2016). Another set of papers, focusing on human-like accuracy (refer to Table 1), applied a supervised method to label the sentiment of reviews to see its correlation with conversion.

[Insert Table 1 about here]

And yet, when faced with the managerial question "What specific content in reviews do consumers care about?" the current literature lacks answers, for several reasons. First, most of these studies rely on text data from one or two product categories, which limits the generalizability of the findings. Second, extant papers lack unifying content dimensions that can be generalized to all product categories due to the limitation of the lexical-based approach. Third, the data requirement to see the causal impact of specific review content on conversion requires consumer review-reading behavior along with conversion data, which no previous studies had access to. This study fills this gap in the literature by first identifying specific conceptual content that consumers may care about (discussed in "Review Content Dimensions"), tagging the existence and sentiment of this content on reviews, and quantifying the causal impact of reading this specific content on consumer purchase behavior. Thus, the dimensions of products we identify can be extracted for all types of products, and the results from our study can effectively answer the managerial question "What specific content in reviews do consumers care about?" In effect, we go further than providing correlational empirical results to provide a basis for the relationship between specific content consumers read (which generalizes to all product categories) and the content's effect on final conversion outcome.

To do so, we obtain consumer purchase data tied to review-reading behavior (discussed in Section "Descriptions of Consumers' Review-Reading Behaviors") from an e-commerce platform with hundreds of product categories. A unique challenge with this dataset is that the review content for different product categories is immensely diverse and polymorphic, in that the same concept can be represented by different words. For example, the words used to describe the function of a watch may relate to "accurate" and "waterproof," whereas the words used to describe the function of a DSLR Camera may focus on "picture quality" and "resolution." Conventional NLP algorithms without extensive feature engineering would perform extremely

poorly when applied to our dataset. Instead, we rely on supervised deep learning (LeCun et al., 2015) that can be fed with raw data along with some human-tagged labels to automatically discover feature representations and deal with diverse content from a wide range of product categories. As a result, we can avoid time- and labor-intensive hand-coding, analyze platform-level review content data without domain knowledge in hundreds of areas, and approach near-human-level accuracy. We defer the discussion of deep learning models to later sections.

In summary, although many papers have studied the impact of online review text on sales, our paper is one of the first to quantify the causal impact of specific review textual content on sales across a wide range of product categories on dimensions that are generalizable across all types of products. In the process, we also demonstrate an application of deep learning methods suited for this particular task of polymorphic content extraction.

3 Descriptions of Consumers' Review-Reading Behaviors

Our data come from a major online retailer in the United Kingdom. The site regularly ranks as a top 5 e-commerce site in UK across different surveys and articles and employees more than 25,000 people. The data were gathered from a panel of 243,000 consumers over the course of two months in February and March 2015. The data track all consumer behaviors, including page views, impressions, used interactions, and transactions. For a more detailed description of how we combine several tables, please see Appendix E. We explain them one by one. First, a page view is a single view of a product information or category information page. On the category page, a consumer can see around 50 products with product images, average rating, and the total number of reviews for each product. On the product information page, a consumer can read more product descriptions. In Figure 1, we provide sample screenshots of the category information page (left panel) and product information page (middle panel). At the bottom of the product information page, a consumer can click a button to read detailed review content. A unique display feature of this website is that product reviews are presented in a group of five (right panel of Figure 1). If consumers want to read more reviews, they can click the page number below the reviews. Later, in Section 5.1, we explain how to leverage this design feature to achieve model identification.

Another user behavior is impression, which is a single exposure to a product review. The data also include textual content such as questions and answers. We don't differentiate questions and answers from reviews but refer to all of them as reviews. Through the impression data, we know precisely which review or reviews are read by consumers. The data provider is able to track and record (1) when a user clicks on review pages to view reviews, (2) whether each review content has appeared on user's browser, and (3) for how long the content was viewed on the user's browser accurately up to milliseconds. With these data, we assume that if a review appeared on a user's browser, the user has read the review. One caveat here is that we do not have eye-tracking data. Although the review reading behavior tracking technology used by this company is the most advanced in the e-commerce industry, we acknowledge the possibility of measurement error for the number of reviews read variable. In a robustness check, we assumed that the number of reviews read is proportional to the time review content appeared on the user's browser. The main result remain qualitatively unchanged. On average, consumers read 10 reviews.

The next user behavior is "used-interaction", which refers to a consumer's action of using 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. Lastly, transaction details are recorded including price paid and quantity purchased. For each of the above-mentioned user behaviors, we observe the time stamps, which we use to construct the consumer decision journey, to be explained in the next section.

[Insert Figure 1 about here]

The data cover two broad departments, Home and Garden and Technology. Each department consists of hundreds of well-defined product categories. For example, Pillowcases is a category within Home and Garden, while Printers is a category within Technology. In total, there are 583 categories. Please see Figure 2 for examples of product categories. Among all these product categories, consumers had around 2.5 million page views, 12.3 million review impressions, 500,000 used interactions, and 30,000 transactions. These actions were taken on one of two devices: PC or mobile phone. Tablets and mobile phones are clustered into the same group.

[Insert Figure 2 about here]

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

3.1 For What Products Do Consumers Read Reviews?

First, we describe for what products consumers do or do not read reviews. To do so, we need to put review reading in context, i.e., the consumer decision journey. We define a consumer decision journey as the sequence of actions between the search for and purchase of a certain product. One journey is constrained to only one product category. So when a consumer switches to search in a different product category, a new journey starts. In the case of no purchase, we assume that a journey ends when the consecutive session is more than one week away from the current session. 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 about the product. Given this definition, we construct 300,047 journeys, which can be classified into five types, as depicted in Figure 3.

A 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 makes up 2% of the sample. A Type 2 journey comprises only the search stage. Surprisingly, this journey type makes up 66% of all the journeys, which suggests a very high bounce rate. During Type 2 journeys, consumers have relatively low intention to purchase and therefore do not make an effort to read reviews. This might also be because consumers read reviews on other websites.[2] A Type 3 journey contains two steps, search and purchase, and happens 3% of the time. Reviews are not used during Type 3 journeys. A Type 4 journey also has two steps—search and reading reviews. Consumers in Type 4 journeys make an intensive effort to look for both product information provided by the retailer and user-generated reviews. But due to particular reasons, consumers drop out before the purchase. This Type 4 journey covers 27% of the sample. Lastly, a Type 5 journey is the longest, comprising all three steps—search, reading reviews, and purchase. It involves only 2% of all the journeys in our sample.

[Insert Figure 3 about here]

Looking across all five types of journeys, we find that for 71.2% of the time, consumers do not read reviews (journey Types 1, 2, and 3 in Figure 3). However, if we exclude the Type 2 journey, where consumers seem not to have a high intention to buy, then 85% of the time, consumers do read reviews.

Table in Figure 3 provides a descriptive analysis 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.

[2]Unfortunately, we have access to data from only one online retailer instead of the user-centric data across sites. We discuss this potential missing data problem in Appendix A.

We now summarize some 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 stakes (Laurent and Kapferer, 1985). Moreover, reviews play a role when the number of reviews is relatively modest, which suggests that the product is neither too popular nor unpopular. We think this happens because consumers have low uncertainty for the most popular, or dominating, products, so reading reviews is no longer critical.

To echo the above findings from our model results, as a preview we give concrete examples of the product categories for which consumers care about (Figure 4 left panel) or do not care about reviews (Figure 4 right panel). For instance, floor care products, bed frames, and mattresses are all relatively more expensive products with high quality variation. Consumers need to rely on reviews to assess the product quality and fit. These products are mostly classified as experience goods in the literature (Nelson, 1970). In contrast, pay-as-you-go phones, laptop and PC accessories, and clocks are relatively cheap, with known product features and guaranteed quality. Consumers have low incentive to read reviews before purchasing them. These products are classified as search goods in the literature (Nelson, 1970).

[Insert Figure 4 about here]

3.2 How Many Consumers Read Reviews?

We also find that reviews might have heterogeneous effects on consumers. Approximately 43% of consumers never read reviews for any products (in the sample period), 47% sometimes read, and 10% of consumers always read reviews for all products. This consumer heterogeneity might be attributed to different search cost and different purchase intentions. Unfortunately, we do not observe any demographic information to further explain the consumer heterogeneity.

3.3 On Which Devices Do Consumers Read Reviews?

There are also heterogeneous behavioral patterns of consumers using different devices to read reviews. Consumers are more likely to read reviews on PCs (92% of all the journeys) than on mobile devices (75% of all the journeys). This is consistent with the prior literature that finds that the smaller screen of the mobile device makes it less convenient than a PC to conduct in-depth searches. We further explore this effect in Section 6.

The above analysis describes that for certain products, for 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 products for which reviews matter, select the consumers who read reviews, and differentiate the effect on different devices. In the next two sections, we build models to do so. Section 4 introduces the full deep learning model, and Section 5 describes the partial deep learning model.

4 Full Deep Learning Model

We first build a full deep learning model that combines consumer characteristics, product attributes, and review content in a joint framework to predict conversion. As the model has many building blocks, we introduce the concepts sequentially. This section starts with the motivation of using deep learning to model text data by comparing classical natural language processing (NLP) models (Section 4.1) and deep learning

NLP models (Section 4.2). Then we present the intuition of using a convolutional neural network (CNN) model (Section 4.3) in our setting and some basic background information about neural network models (Section 4.4). Next we explain the structure (four layers, in Section 4.5) of the CNN model in detail and the estimation algorithms (Sections 4.6 and 4.7). We end (Section 4.8) by demonstrating how to interpret the results from the CNN model.

4.1 Classical Natural Language Processing Model

The simplest way to represent content information in the reviews is using each word as a feature (independent or explanatory variable): the so-called bag-of-words representation. Let the entire corpus for all the reviews be $V = \{w_1, ..., w_N\}$, where w_i is the ith word in the corpus and N is the total number of words. The bag-of-words representation treats each word w_i as a single feature. And each review is represented by an N-dimensional vector, where the ith element in the vector denotes the frequency count (or binary variable denoting existence) of the word w_i. Similar to this bag-of-words representation for documents, traditional NLP represents each word sparsely, with a "one-hot vector" of size N, whereby word w_i is represented with a 0-vector of length N, except for position i, which is equal to 1.

These word features can then be used (combined with other explanatory variables) in a linear model (e.g., logistic regression) to predict conversion.

4.2 Comparison between Deep Learning-Based NLP and Classical NLP

Although the document (bag-of-words) or sparse word (one-hot vector) representations are easy to create, they suffer from several severe drawbacks. First, bag-of-words representation loses a tremendous amount of information from the ordering of words. That is, it ignores phrases and the syntactic structure of sentences. Although one can extend the bag-of-words representation to include n-grams (a sequence of n words), the number of features increases exponentially as the n-grams get longer (n is bigger). Second, it also treats features as entirely independent of each other. For example, under this framework, the word "like" is as dissimilar to "love" as it is dissimilar to "hate." Third, this representation creates a high dimensional sparse matrix, which increases computation time and storage requirement as well as reducing estimation efficiency. The high dimensionality also prevents it from being used in non-linear models, such as artificial neural networks (ANN).

To solve these problems, new deep learning-based natural language processing (NLP) models have been proposed (see the survey by Goldberg 2016). 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 provided by humans. For example, when the researcher is dealing with text data, sentence-level attributes such as part-of-speech, coreference resolution, and negation detection need to be hand-coded or explicitly extracted and entered into existing machine learning models. As a result, text mining algorithms often get confused or miss many natural language subtleties entirely if they are not explicitly encoded. In contrast, recent advancements in deep learning enable us to explore unstructured text data without ad hoc and error-prone feature engineering so that the entire process can be automated. Essentially, deep learning evolved from an already existing ANN machine learning technique to model high-level abstractions and patterns in data by using a network structure with multiple processing layers (thus "deep"), composed of multiple linear and non-linear transformations. Deep learning involves many improvements in techniques to overcome shortcomings in previous ANN model estimation LeCun

et al. 2015, Bengio et al. 2006. Using deep learning-based NLP models, we can let the data and sophisticated algorithms detect natural language subtleties instead of hand-coding the sentence-attributes to enter as X-variables (input variables) in typical classification models. In this way, the use of deep learning as a supervised learning content-coding algorithm is much more scalable and applicable in any domain and for commercial purposes. Specifically, in our case, in which we have review texts that could largely vary across product categories in semantics, we gain more accuracy with considerably less effort. Although conventional methods may ultimately achieve similar accuracy, they require time-consuming feature-engineering efforts (often many months) and lose economies of scale when a new dataset, such as new product categories in our case, is introduced.

Moreover, the deep learning-based NLP models aim to take into account interactions between words to better capture the semantic relations and syntactic structures of language. Deep learning-based NLP models use low-dimensional and dense vectors as representations. Words with similar meanings have similar values in the representation vectors. Furthermore, the dense vectors can be easily incorporated in ANN models to extract non-linear relationships.

4.3 Convolutional Neural Networks Framework and Intuition

Inspired by the works of Kalchbrenner et al. [2014] and Kim [2014], we propose the following convolutional neural network model to examine the impact of review content information on sales conversion. The model architecture is illustrated in Figure 5.

[Insert Figure 5 about here]

Before explaining the details of the model, we first describe the intuition behind convolutional neural network (CNN) natural language processing models. There are two important ideas. The first idea is that some local clues in sentences are more informative for predicting the outcome than others. For example, the exemplar review in Figure 5 says, "the washer is good looking and also very powerful." In this review, the local parts "good looking" and "very powerful" are more informative of the sentiment than the parts "the washer is" and "and also." The second idea behind CNN is that the local clues are informative regardless of their locations in the entire review document. So, if the review changes to "the washer is very powerful and also good looking," then "good looking" and "very powerful" are still the most informative parts.[3] By combining these two ideas, CNN models aim to use "filters" to identify the informative local clues from long sentences and then discard the position information of local clues to reduce the number of parameters in the model and avoid overfitting. Essentially, CNN models use a two-step approach: "convolution" and "pooling." "Convolution" applies a filter over each sliding window of the sentence to capture important local clues, whereas "Pooling" aggregates the outputs from the filters by creating a location-insensitive summary statistic. We show more details of convolution in Section 4.5.

4.4 Neural Network Basics

Above all, the CNN model we propose is a special case of the artificial neural network (ANN) models. The intuition behind ANN is to imitate the way the brain processes information. In an ANN, the basic

[3]This second idea behind CNN might not be correct in some settings. For example, the review, "the washer is not very good but very cheap", expresses a positive sentiment, whereas "the washer is very cheap but not very good", expresses a negative sentiment. This kind of more complicated semantic structure is not captured by CNN, and is a big limitation. Other deep learning NLP models, such as recurrent neural networks and recursive neural networks, are able to accommodate more complicated sentence structures. But they have other computational and data limitations that CNN avoids.

computation units are called neurons, which are interconnected to form hidden layers, and finally the whole network. Each neuron takes several inputs, multiplies them with the associated weights, sums them, and applies a non-linear activation function to deliver an output (we will explain weights and activation functions in Section 4.5.2.) Then the output becomes the input of the next layer of neurons. The layers in the neural network reflect information flow.

4.5 CNN Architecture

Our model architecture (Figure 5) has four layers. The first layer (the leftmost layer) is the word embeddings (to be explained next) of product reviews, and the second layer is the convolutional layer (to be described in Section 4.5.2). The third layer is the max-over-time pooling layer (to be defined in Section 4.5.3), appended with neurons of consumer and product characteristics information. And the final layer is our outcome: conversion. The squares in the figure denote neurons, and the lines denote the connections between inputs and outputs.

4.5.1 Layer 1: Word Embedding

In contrast to the sparse representation, content information is represented by low-dimensional dense vectors in neural networks. These vectors come from pre-trained word-embeddings. In our implementation, we use the word2vec embeddings published by Google (https://code.google.com/archive/p/word2vec/). These embeddings are trained on 100 billion words from the Google News dataset using the method created in Mikolov et al. [2013]. In a nutshell, the word2vec model takes the text corpus and transforms it into word vectors while preserving semantic distances between the words as much as possible; so, each word is represented by a 300-dimensional vector, and the word vectors carry desirable linguistic properties. For example, the vector-distance between similar words is higher than that between dissimilar words.

In Figure 5, the first layer is the word embeddings. Let us use one review as an example. Suppose n is the total number of words in the review and k is the dimensionality of the word embeddings ($k = 300$). We use $\overrightarrow{x}_i \in R^k$, a k-dimensional word vector to represent the ith word in the review, and we use $\overrightarrow{x}_{1:n} = \overrightarrow{x}_1 \oplus \overrightarrow{x}_2 \oplus \ldots \oplus \overrightarrow{x}_n$ to represent the entire review. Here, $\oplus$ is the concatenate operator which stacks all the n-word (column-) vectors to form a $nk \times 1$ vector. In Figure 5, we show a $n \times k$ matrix instead of a $nk \times 1$ vector to make the illustration easier to understand. In the model, we actually create a $nk \times 1$ vector to be further used in equation (1). This vector represents information in the entire review.

4.5.2 Layer 2: Convolution Operation

The next layer is the convolutional layer, which applies the convolution operation or filter to the word embeddings in the first layer. The convolution operation, or filter, is a one-dimensional vector of length h, applied to each sliding window of h words in the sentence. For example, in Figure 5, the green filter size is $h = 2$. The green filter can be written as $\overrightarrow{w} \in R^{hk}$. This is a $hk \times 1$ vector, where h is the window size and k is the dimensionality of the word embeddings in the previous layer ($k = 300$). The filter is first applied to the window of the first two words "the washer," then to the next two words, "washer is," then to the following two words, "is good," and so on. Let i be the current position and $\overrightarrow{x}_{i:i+h-1} \in R^{hk}$ be the window of words or n-grams that the filter is applied to. The output of the convolution operation is

$$c_i = f\left(\overrightarrow{w} \cdot \overrightarrow{x}_{i:i+h-1} + b\right), \tag{1}$$

where $\cdot$ denotes inner product, $b \in R$ is the bias parameter, and f is the activation function. The activation function is a non-linear function, which allows neural network models to incorporate non-linear relationships between input variables. We use the rectified linear units (ReLU) as the activation function (Goodfellow et al. 2016, p. 187). The ReLU function is defined as $f(x) = max(x, 0)$. The input to the activation function, $\overrightarrow{w} \cdot \overrightarrow{x}_{i:i+h-1} + b$ is a linear transformation of $\overrightarrow{x}_{i:i+h-1}$. If we use the analogy of a linear regression, we can consider $\overrightarrow{w}$ as the weights and b as the intercept term. Both weights ($\overrightarrow{w}$) and bias (b) are parameters to be estimated. In summary, the convolution operation first applies a linear transformation to the inputs using the weights and bias, and then a non-linear transformation using the activation function.

The filter is rolled over to each sliding window of h words for $i = 1, 2, \ldots$. So, the final output is a vector, $\overrightarrow{c} \in R^{n-h+1}$, called the feature map. Specifically,

$$\overrightarrow{c} = [c_1, c_2, \ldots, c_{n+h-1}].$$

In our setting, we try different window sizes: $h = 2, 3, 4, 5$ to incorporate bi-grams, tri-grams, 4-grams, and 5-grams.

4.5.3 Layer 3: Pooling

The third layer is the pooling layer. The pooling layer applies the max-over-time pooling operation to the feature map created in the previous layer (layer 2). The idea behind the max-over-time pooling is that we want to get the most salient information across all window positions, so

$$\hat{c} = max\{\overrightarrow{c}\}.$$

As mentioned in Section 4.3, the pooling layer will keep the informative local clues gathered by the filter in layer 2 but discard its position. In other words, $\hat{c}$ ignores which c_i is the biggest, but keeps the maximum value among all c_i's. To sum up, the outcome $\hat{c}$ is the representation of the entire review, and it captures the most indicative information in the review.

Each of the above-mentioned layers 1, 2, 3 extracts one feature from one filter. In reality, we can repeat the process and apply multiple filters with varying window sizes to create multiple features. Let the total number of filters be m. So, the penultimate layer becomes a vector of all the features (each feature corresponds to one filter) extracted from the text data, i.e.,

$$\overrightarrow{d} = [\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_m].$$

For instance, in Figure 5, the green filter ("the washer" and "very powerful") has a window size of 2, whereas the blue filter ("good looking and") uses a window size of 3. After the pooling layer, each filter creates one feature. The two features are concatenated to be passed to the next layer.

In our application, we use a total of 100 filters ($m = 100$) for each window size h ($h = 2, 3, 4, 5$). This results in 400 features, each representing a distinct content dimension. These features are not easily interpretable. In Section 4.8, we propose a solution to get around the interpretability issue.

4.5.4 Layer 4: Append and Output

In the last layer, the features extracted from the review text data ($\overrightarrow{d}_{ijt}$) are combined with variables of consumer characteristics ($\overrightarrow{Z_{it}}$, including total products searched and number of used interactions), observed product characteristics ($\overrightarrow{X_{jt}}$ including price, average rating, cumulative number of reviews, percentage of consumers recommended, number of questions and answers), unobserved product characteristics (ξ_j), and time fixed effects ($Weekend_t$, $Daytime_t$) to predict the final outcome: conversion. The activation function used in the last layer is the softmax function (Bishop 2006): $softmax(c) = \frac{e^{x_c}}{\sum e^{x_c}}$. In our case, this is equivalent to the logit choice probability function in discrete choice models (Train 2009), because we have a binary outcome of converting versus not converting.

Let y_{ijkt} denote conversion ($y_{ijkt} = 1$ implies purchase, and $y_{ijkt} = 0$ implies no purchase) for consumer i considering product j using device k at time t. Then the specification in the last layer of the neural network is

$$y_{ijkt} = softmax\left(\overrightarrow{w_k^y} \cdot \overrightarrow{d}_{ijt} + b_k^y + \overrightarrow{\theta_k}\overrightarrow{Z_{it}} + \overrightarrow{\gamma_k}\overrightarrow{X_{jt}} + \xi_j + Weekend_t + Daytime_t\right). \quad (2)$$

We append the consumer and product characteristics to text features in the last layer of the neural network to model conversion directly. In other words, this model creates a single, joint deep neural network model to forecast conversion. We call this the "full model." One thing to note is that the text features in this direct approach are not interpretable by face value. This is why, in Section 5, we introduce a two-step approach, or a "partial model." Later, we compare the full model and the partial model in terms of their accuracy, efficiency, and interpretability.

4.6 Training: Stochastic Gradient Descent – SGD

The estimation algorithm we use is stochastic gradient descent (SGD, see Goodfellow et al. 2016, for more details). It belongs to a broad class of gradient descent (Cauchy, 1847) algorithms. The key advantage of SGD is that it is scalable to a large amount of training data. The basic intuition is that the objective function of the machine learning problem is often the sum or the mean (expectation) of the objective functions applied to each observation. When the number of observations becomes enormous, it's too time consuming to calculate the objective function. So, instead of summing up all the observation-level objective functions, SGD uniformly samples a mini-batch of observations and uses only this subset to approximate the objective function. As a consequence, the estimation algorithm can speed up without much loss of accuracy. In practice, we use a mini-batch size of 50.

To calculate the gradient, we use the back-propagation method. The idea is that in the neural network model, information flows forward from the input layer to the hidden layers and then finally arrives at the output layer, allowing us to calculate the loss function (objective function). When calculating the gradient of the loss function with respect to a particular parameter (weights and biases in equation (1 and equation (2)), we can reverse this process and allow the information to flow backward through the layers of the network. The gradient calculation applies the chain rule to compute derivatives in each layer until the final derivative is obtained.

4.7 Regularization: Dropout

To prevent overfitting, we apply regularization to the model. Specifically, we use the dropout method (Srivastava et al. 2014). Intuitively, this method randomly drops $1-p$ percent of the neurons and keeps only the

rest p percent of the neurons in the training process, but in testing, all the neurons will remain for prediction. The mechanism behind this method is model averaging: By dropping subsets of neurons repeatedly, the model creates smaller neural networks, which are then averaged to avoid overfitting. We use a dropout rate of $p = 0.5$ in our practice. The dropout rate is a hyper-parameter. We tune all the hyper-parameters using cross-validation on the development set, a random sample of 10% of the data. Other hyper-parameters, including the number of filters and filter sizes, are chosen in a similar fashion.

4.8 Feature Interpretation

Deep learning models are not easily interpretable (Towell and Shavlik 1992). The weights and bias parameters as well as the neurons (including the features created from the filters) in the hidden layers of the model are not meaningful by themselves. To understand what content features affect conversion, we use the approach in Simonyan et al. [2013] to select the most "salient" n-grams. The intuition is that given the CNN model, we can compute the gradient of each input n-gram ($\overrightarrow{x}_{i:i+h-1}$ in equation (1)) by taking a derivative (using a single back propagation pass through) of the output function with respect to the n-gram. The magnitude of the derivative indicates which n-gram needs to be changed the least to affect the prediction outcome, conversion. For each review, we select the n-gram that has the highest derivative. After obtaining the "salient" n-grams, we run topic modeling (LDA, Blei et al. 2003) to cluster the n-grams to a handful of topics for ease of interpretation. Note that, different from most applications of LDA, in our case, each document contains only one n-gram.

Recall that in Layer 4 (Section 4.5.4) of the CNN model, we append the consumer and product characteristics to the text neurons, and the last activation function is softmax, which resembles a logistic regression. So, the coefficients ($\overrightarrow{\theta_k}$, $\overrightarrow{\gamma_k}$) in equation (2) can be directly interpreted. One caveat is that only the signs of these coefficients can directionally indicate the relationship between consumer/product characteristics (the input variables) and conversion (the output variable). However, because the model does not produce any confidence intervals, we cannot make any valid statistical inference, such as the significance of the coefficients.

5 Partial Deep Learning Model

As the review content features in the full deep learning model (Section 4) lack interpretability and cannot be used for statistical inference, in this section, we build a partial deep learning model where we use deep learning to extract interpretable content dimensions and then pass them to a classical choice model for inference. To better explain the model, we first elucidate the identification strategy in Section 5.1. Then Section 5.2 introduces the theory-driven content dimensions. Lastly, Section 5.3 lists the deep learning algorithms used for information retrieval and visualization.

5.1 Identification Strategy

To identify the effect of reviews read on conversion, we need exogenous variations of the reviews read by consumers. In our setting, the exogenous change of the review content comes from the timing of when new reviews are posted to the site. This allows us to use the Regression Discontinuity in Time (Lee and Lemieuxa 2010, Hausman and Rapson 2017) design (RDiT) to achieve identification.

To illustrate, Figure 6 shows an excerpt from the product review section for a clock. On October 13, 2017, the product had the first review, titled "Easy to read." On October 19, 2017, a new review with the title

"Best small clock ever" was submitted. From a buyer's perspective, this new review creates a discontinuous change in the pool of review content available to the buyers. Suppose each buyer reads only one review. Those buyers who visited the site before October 19 would read the old review, titled "Easy to read." But those consumers who visited the site after October 19 would read the new review, "Best small clock ever," instead. Comparing the two reviews, we find that though both reviews contain the same rating, the new review provides information about aesthetics, whereas the old review gives information about price. Before and after this review post, product characteristics stay unchanged, and other unobserved demand shocks (advertising, offline word-of-mouth, seasonality, etc.) change continuously in time. However, the review content dimensions read by consumers–in this example, aesthetics and price dimensions–have a discontinuous change when the new review arrives. So long as the timing of when new reviews are added is not correlated with the unobserved demand shocks, the timing of new reviews can be used to identify the effect of review content separately from unobserved demand shocks.

[Insert Figure 6 about here]

Essentially, using the terminology in the literature of regression discontinuity in time (Lee and Lemieuxa 2010, Hausman and Rapson 2017), we let the review post date be c (short for cutoff). For all the time $t > c$, the consumer is treated, and for all $t < c$, the consumer is not treated. In this setup, the running variable or the assignment variable is time. As long as the unobserved determinants of the outcome are continuous with respect to the running variable, identification is achieved with the inclusion of a flexible time trend.

Graphically, in Figure 7, we plot conversion rate (y-axis) against time (x-axis). In the four subplots, we overlay lines of best fit corresponding to polynomial time controls of order 1, 2, 3, and 4. The vertical line is the cutoff time when a new review with positive aesthetics information is posted. The identification strategy for other content dimensions is the same. We use the positive aesthetics dimension as an illustration. Each point in the plots is the average conversion rate in the corresponding bin, with bin size determined by the mimicking variance evenly spaced method. We use the Stata function rdplot in the package rdrobust for all the plots. As can be seen, the conversion rate becomes significantly higher after a new review with positive aesthetics information is posted. This example affirms our identification strategy. Following Lee and Lemieuxa [2010] and Hausman and Rapson [2017], we also test the robustness of the RDiT design. The details are presented in Appendix B.

We want to note two things related to the specific characteristics of our RDiT design. The first is that for different products, new reviews appear at different time points. This implies that our setting can be categorized as the Multi-Cutoff Regression Discontinuity Design (Cattaneo et al. 2016). We follow the conventional method in the literature by estimating a pooled regression discontinuity (RD) treatment effect (Brollo et al. 2013; Chay et al. 2005). The second is that we examine the effect of multiple review content dimensions (to be introduced in Section 5.2). Therefore, we consider each dimension change as a separate treatment. This setting resembles the "double discontinuity" framework in Card et al. [2007]. With the same identification argument in Card et al. [2007], we independently identify the effect of each review content dimension on conversion. We will further explain this after we introduce the multiple conditions in Section 5.2.

[Insert Figure 7 about here]

Now we lay out the model specification. We apply the random utility framework in classical choice models (Train, 2009) and estimate the following specification:

$$u_{ijkt} = \overrightarrow{ReviewContent_{ijt}} \cdot \overrightarrow{\tau_k} + \sum_{n=1}^{3} \delta_{nk} t^n + \alpha_{ik} + \overrightarrow{\theta_k} \cdot \overrightarrow{Z_{it}} + \overrightarrow{\gamma_k} \cdot \overrightarrow{X_{jt}} + \xi_j + Weekend_t + Daytime_t + \varepsilon_{ijkt} \quad (3)$$

In equation (3) , u_{ijkt} denotes the utility for consumer i, using device k considering product j at the visit time t.[4] The term $\overrightarrow{ReviewContent_{ijt}} \cdot \overrightarrow{\tau_k}$ identifies the effect of review content on conversion. As we consider multiple review content dimensions (to be introduced in Section 5.2), we use the vector notation $\rightarrow$ for both $\overrightarrow{ReviewContent_{ijt}}$ and $\overrightarrow{\tau_k}$. Specifically, following Lee and Lemieuxa [2010], we let $\overrightarrow{ReviewContent_{ijt}} = \left\{d_{ijt}\right\}$ be a vector of treatment variables or assignment variables. For each review dimension d (to be introduced in Section 5.2), $d_{ijt} = 1$ if the consumer i's visit time t is *after* a new review for product j with content dimension d being posted to the site, and $d_{ijt} = 0$ if the consumer i's visit time t is *before* the review (for product j with content dimension d) posting time. In Section 5.2, we will explain the definition of $d_{ijt} = -1$. The coefficients in the vector $\overrightarrow{\tau_k}$ are our primary parameters of interest. To account for potential time-varying factors, we include flexible time polynomials with order up to 3, represented by the term $\sum_{n=1}^{3} \delta_{nk} t^n$. The optimal order of the polynomial is chosen using Akaike's criterion (Lee and Lemieuxa 2010).

As suggested by Hausman and Rapson [2017], we also add many covariates as control variables to account for potential discontinuous effects in time and consumer and product heterogeneity. These covariates include consumer i's intrinsic preference using device k,[5] α_{ik}, observed consumer activities or consumer characteristics vector $\overrightarrow{Z_{it}}$ (including total products searched and number of used interactions), product characteristics vector $\overrightarrow{X_{jt}}$ (including price, average rating, cumulative number of reviews, percentage of consumers recommended, number of questions and answers),[6] other unobserved product characteristics ξ_j, as well as weekend and time of day fixed effects $Weekend_t$ and $Daytime_t$. Finally, we add the idiosyncratic shock ε_{ijkt}. The intrinsic preference α_{ik} is related to factors such as income or consumer i's willingness to purchase and convenience of purchase using the device k. Unobserved product characteristics ξ_j are related to quality or popularity of the product. The weekend ($Weekend_t \in \{0,1\}$, when the visit time is on Saturday or Sunday, $Weekend_t = 1$) and time of day ($Daytime_t \in \{0,1\}$, when the visit time is between 6:00 and 18:00, $Daytime_t = 1$) fixed effects are used to control for unobserved discontinuous demand shocks, such as advertising, offline word-of-mouth, seasonality, etc. The shock term ε_{ijkt} is assumed to follow a Type I Extreme Value distribution. So, the conversion rate has a closed-form formula:

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

In equation (3), the time dimension t represents a consumer visit incidence at calendar time t. The time stamp granularity is one second.

Because 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 a consumer's purchase intention can be well-represented by that consumer's interactions with web features, such as 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

[4]The time dimension represents one incidence of a consumer's visit. It takes values of the actual calendar time, measured in seconds. For example, one observation in our data took place on 26FEB15:20:41:11. In equation (3), t for this observation would be 2,234,471 seconds, which is the time span between this calendar time and the starting time of our sample, 01FEB15:00:00:00.

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

[6]Product characteristics may vary over time. For example, the e-commerce site has a dynamic pricing strategy. From our conversation with the site managers, the pricing strategy is not targeted, so the price endogeneity issue is eliminated.

$$u_{ijkt} = \alpha_0 + \overrightarrow{ReviewContent_{ijt}} \cdot \overrightarrow{\tau}_k + \sum_{n=1}^{3} \delta_{nk} t^n + \overrightarrow{\theta_k} \overrightarrow{Z_{it}} + \overrightarrow{\gamma_k} \overrightarrow{X_{jt}} + \xi_j + Weekend_t + Daytime_t + \varepsilon_{ijkt}. \quad (4)$$

Note that the findings in Section 3.3 suggest that there exist distinct consumer preferences on mobile devices versus PC devices, so we assume that all the parameters are device-specific; hence, all the coefficients have a device subscript k. In other words, we estimate two sets of parameters from the model, for mobile and PC separately. All the parameters for PC and mobile are estimated in a joint model to achieve efficiency. In 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.

5.2 Review Content Dimensions

In this subsection, we discuss what review content dimensions are considered in the models represented by equation (4).

To start off, 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. We define a review as positive if its rating is 4 or 5, and as negative if its rating is 1 to 3. The effect of a positive review is captured by τ_k^{Pos}, and the effect of a negative review is captured by τ_k^{Neg}.

In addition to the sentiment of reviews, we consider the quality and price information embedded in the review content. Price and quality of products have been the main drivers of economic transactions and consumer purchase behavior both online and offline. Thus, we look at how price and quality information within reviews influences consumer purchase decisions. 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 its effect. We adopt a theory-driven approach and take a seminal work by Garvin [1984] to operationalize different dimensions of product quality found to influence purchase behavior. Garvin [1984] introduced a set of quality dimensions aimed at helping organizations think about quality. The eight dimensions proposed are performance, features, reliability, conformance, durability, serviceability, aesthetics, and brand (perceived quality). We closely follow Garvin's definitions of these different dimensions to identify quality information in reviews. We combine some qualities that are conceptually close. Specifically, we include six dimensions: "aesthetics," "conformance," "durability," "feature," "brand," and "price," and each has two valences, positive and negative. The coefficients are denoted as $\tau_k^{AestheticsP}$, $\tau_k^{AestheticsN}$, $\tau_k^{ConformanceP}$, $\tau_k^{ConformanceN}$, $\tau_k^{DurabilityP}$, $\tau_k^{DurabilityN}$, $\tau_k^{FeatureP}$, $\tau_k^{FeatureN}$, $\tau_k^{PerceivedQualityP}$, $\tau_k^{PerceivedQualityN}$, τ_k^{PriceP}, and τ_k^{PriceN}, respectively. We describe each dimension in Table 3. We consider these attributes to be the main focus of our review content analysis.

[Insert Table 3 about here]

$$\begin{aligned}
\overrightarrow{ReviewContent_{ijt}} \cdot \overrightarrow{\tau_k} &= Pos_{ijt} \cdot \tau_k^{Pos} + Neg_{ijt} \cdot \tau_k^{Neg} \quad (5)\\
&+ AestheticsP_{ijt} \cdot \tau_k^{AestheticsP} + AestheticsN_{ijt} \cdot \tau_k^{AestheticsN}\\
&+ ConformanceP_{ijt} \cdot \tau_k^{ConformanceP} + ConformanceN_{ijt} \cdot \tau_k^{ConformanceN}\\
&+ DurabilityP_{ijt} \cdot \tau_k^{DurabilityP} + DurabilityN_{ijt} \cdot \tau_k^{DurabilityN}\\
&+ FeatureP_{ijt} \cdot \tau_k^{FeatureP} + FeatureN_{ijt} \cdot \tau_k^{FeatureN}\\
&+ BrandP_{ijt} \cdot \tau_k^{BrandP} + BrandN_{ijt} \cdot \tau_k^{BrandN}\\
&+ PriceP_{ijt} \cdot \tau_k^{PriceP} + PriceN_{ijt} \cdot \tau_k^{PriceN}
\end{aligned}$$

We calculate the marginal effects of these content variables by replacing the element of reviews content dimensions in equation (4) with equation (5).

We now give more details on how to identify the effect of each content dimension separately, using the RDiT design. Recall that we find that each consumer reads only a limited number of reviews. As a consequence, a newly posted review will crowd out an old review. In the RDiT design, a new review post date is a treatment. Because of the crowd-out effect, each treatment involves two content dimensions. For instance, Figure 8 continues the example in Figure 6 and adds two more reviews published on October 21, 2017 and October 23, 2017. As mentioned before, the first review posted on October 13 contains price information, whereas the review posted on October 19 contains aesthetics information. In contrast, the third review posted on October 21 contains conformance information, whereas the fourth review posted on October 23 contains feature information. Under this setting, we treat October 21 and October 23 as two separate experiments. Imagine two consumers, A and B. A visits the website on October 22 to buy the product, and B visits on October 24. Both consumers read only two reviews. Consumer A's visit on October 20 is classified as treated for the first experiment but untreated for the second experiment, whereas consumer B's visit is classified as treated for the second experiment. More specifically, consumer A is considered as having read the review with conformance information (the review titled "A great buy!" posted on October 21) and the review with aesthetic information (the review titled "Best small clock ever" posted on October 19) but not the review with price information (the review titled "Easy to read" posted on October 13). So, the review with conformance information crowds out the review with price information. Similarly, consumer B is regarded as having read the review with feature information (the review titled "A Winner" posted on October 23) and the review with conformance information (the review titled "A great buy" posted on October 21) but not the review with aesthetic information (the review titled "Best small clock ever" posted on October 19). So, the review with feature information crowds out the review with aesthetic information.

In equation (5) above, consumer B's visit will be assigned with $FeaturePB_{jt} = 1$ and $AestheticsPB_{jt} = -1$, aiding the identification of the coefficients $\tau_k^{FeatureP}$ and $\tau_k^{AestheticsP}$. Note that $AestheticsPB_{jt} = -1$ means that time t, the consumer visit time, is after a new review for product j is posted to the site, crowding out an old review with the positive aesthetics content dimension.[7] We set $AestheticsPB_{jt} = -1$ because, compared with consumer A, consumer B reads not only one more review with feature information ($FeaturePB_{jt} = 1$) but also one fewer review with aesthetics information. By the same logic, consumer A's visit will be assigned with $ConformancePA_{jt} = 1$ and $PricePA_{jt} = -1$, helping identify the coefficients $\tau_k^{ConformanceP}$ and τ_k^{PriceP}. Note that if a review contains multiple content dimensions by itself, then a treatment can involve more than two dimensions. For example, if a newly posted review contains both positive durability and negative brand information, it crowds out an old review with positive feature information. Then three content dimensions, $DurabilityPB_{jt}$, $BrandNB_{jt}$, and $FeaturePB_{jt}$, are involved in this treatment.

The crowd-out effect is prominent due to the unique design feature of this website, illustrated in Figure 1. Recall that each review page contains only five reviews. A consumer who reads five reviews in total will not click the page button to read the reviews on the second page. When a new review is posted, the old review, which sits in the fifth position of the first review page, will be forced out to the second page.

[7] The crowding out assumption applies to all cases when the number of reviews read by a consumer is N where N is a positive integer. More specifically, the crowding out assumption also applies when a consumer reads multiple pages of reviews. The complete rule is $d_{ijt} =$
$\begin{cases} 0 & \text{consumer i read reviews of product j at time t before a new review with dimension d was posted} \\ 1 & \text{consumer i read reviews of product j at time t after a new review with dimension d was posted} \\ -1 & \text{consumer i read reviews of product j at time t bafter a new review crowded out an earlier review with dimension d} \end{cases}$
$d \in \{AestheticsP, AestheticsN, ConformanceP, ConformanceN, DurabilityP, DurabilityN, FeatureP, FeatureN, BrandP, BrandN, PriceP, PriceN\}$.
It implies that the value (of the content dimensions) will be 0 for all consumers who will be exposed to this page of reviews until a new review is posted.

Therefore, the consumer will not have a chance to read it. This design feature of the website strengthens our identification argument. Note that reviews are presented to a consumer in reverse chronological order if the consumer does not engage in any "used-interaction." This online retailer allows users to filter reviews by star ratings. For example, a consumer can filter to look at only one star reviews. When doing so, all the one star reviews are displayed to the consumer, also in reverse chronological order. Because the order is always reverse chronological, no matter whether a consumer filters or not, our identification is immune to consumers' filtering/sorting behavior.

[Insert Figure 8 about here]

We also include a few control variables that have been found to influence conversion in the previous literature. These include "length" and "readability." We explain the rationale of using them: first, "length" measures the number of words in the reviews read. We include this variable because longer reviews provide more detailed information that can strongly affect readers' decisions. Second, Ghose and Ipeirotis [2011] have shown that high readability of reviews is linked to increased sales; thus, we calculate and control for the measure of readability with a widely used metric called the SMOG Index ("Simple Measure of Gobbledygook"). Higher values of SMOG imply that a message is harder to read (Mc Laughlin, 1969).

5.3 Deep Learning Algorithms Used

In our partial deep learning model, we utilize deep learning for two purposes. First, we use various deep learning algorithms as supervised learning classifiers to extract predefined price and quality content dimensions. All the algorithms are scalable across different product categories. Second, we use a salient phrase visualization technique from the deep learning literature to highlight salient sentences that are topic-relevant. We adopt these deep learning algorithms from the computer science literature. Next, we briefly explain the intuitions behind the algorithms and refer readers to the original papers for technical details.

5.3.1 Supervised Learning

As mentioned in Section 5.2, we believe that the price and quality information of products embedded in the reviews are key drivers of consumer purchase. However, no prior work in machine learning or natural language processing has identified useful textual features that represent price and quality information. Instead of performing ad-hoc, error-prone, and time-consuming feature engineering, we rely on Deep Learning algorithms 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 5,000 random reviews. To obtain this labeled set of data, we hire workers from Amazon Mechanical Turk (AMT)[8] to provide labels to these reviews. To content-code our reviews, we create a survey instrument comprising 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 (Table 3) 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 dimension.

For example, a review that says "TV looks good but it's too expensive" is identified as having positive aesthetics information and negative price content. To ensure high-quality responses from the Turkers, we

[8] AMT is a crowdsourcing 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.

follow several best practices identified in the literature (e.g., we obtain tags from at least five different Turkers choosing only those who are from the U.S., and have more than 100 completed tasks and an approval rating of more than 97%. Turkers also have to pass a short test to be qualified.) Please see Appendix F for the final survey instrument and Appendix G for the complete list of strategies implemented to ensure output quality. 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. The entire data labeling process takes the Turkers 1.5 months and costs $7,230[9]. If we had labeled all 500,000 reviews manually, it would have taken 12.5 years and cost $723,000 (duration can be cut by either lowering the Turker quality requirement or paying more). In contrast, using the deep learning algorithms discussed later to label all the reviews takes only several hours.

Second, the labeled data are divided into a training set comprising 70% of the observations[10] and a test set with 30% of the observations. We then perform content detection and sentiment analysis by training various algorithms on the training dataset and test the classification accuracy using the test dataset. This is a two-step process. For each review, we begin by detecting whether the content dimensions exist. If yes, a sentiment analysis is followed. For sentiment analysis, we median split the Likert scale[11] and turn it into a binary classification problem of positive versus negative sentiment. We train the algorithms, to be introduced later, separately for each of the six dimensions of information listed in Table 3, namely aesthetics, conformance, durability, feature, brand, and price.

Third, we perform a prediction task to classify the rest of the nearly 500,000 reviews so that each review has 12 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 algorithms for content detection and sentiment analysis. Before introducing the deep learning algorithms, we first explain the intuition behind traditional machine learning models to perform sentiment analysis. In its essence, sentiment analysis is a text classification problem. Therefore, any existing supervised learning method can be applied, e.g., naive Bayes classifiers, support vector machines (SVM). So, in a first application, Pang et al. [2002] take this approach to classify movie reviews into two classes: positive and negative. This application shows that using unigrams or bags-of-words as features in classification performs well because sentiment words such as "good" and "bad" are the most important indicators of sentiments. However, this bag-of-words representation and other simple representations such as part-of-speech ignore the order of words and 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. Considering the fact that we need to extract content information from a large number of reviews of a wide range of products, we decide to use deep learning algorithms.

Here we briefly explain the intuition of the deep learning algorithms employed in our analysis and the rationale for using each of them. For more details on these algorithms, we refer readers to the original papers and Appendix C. The first deep learning algorithm we implement is long short-term memory **recurrent neural networks** (Wang et al., 2015), which works by taking word or character *sequences as inputs.* This algorithm excels in incorporating semantic relations between words in long sentences. The second deep learning algorithm is **recursive neural networks** (Socher et al., 2013). Instead of focusing on sequences as in the recurrent neural networks, recursive neural networks focuses on a more complicated sentence *parse-tree* structure that is aware of sentence syntactic context. Recursive neural networks is very powerful, but it requires a syntactic parse tree, which is not available in many settings. The last algorithm, **convolutional**

[9]One could achieve faster rate of collection by 1) raising pay rate, 2) lowering qualification requirement, and 3) having a better requester reputation.

[10]We perform a three-fold cross-validation on the training set.

[11]https://www.simplypsychology.org/likert-scale.html

neural networks (Kim, 2014), has a data-driven structure that does not rely on externally provided parse trees. This algorithm is similar to the one we presented in Section 4.3. The key difference is that the outcome variable here is the content dimensions or the associated sentiments instead of conversion.

5.3.2 Visualization to Extract Salient Sentences

In addition to using deep learning algorithms 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] that adapts visualization techniques from computer vision to automatically extract relevant sentences from labeled text data. This method allows us to identify the most salient sentence from each review. For example, if a review contains three sentences, and only the first sentence mentions aesthetics information. This method will highlight only the first sentence in the output. The intuition behind the method is that it can calculate saliency scores using gradient magnitudes, because the derivative indicates which words need to be changed the least to affect the score the most. We refer readers to Appendix D and the original paper for more details.

6 Results

We discuss all the empirical results in this section, starting with a performance comparison between the full deep learning model and the partial deep learning model. Next, we focus on the full deep learning model and provide qualitative insights. Subsequently, we focus on the partial deep learning model, present the effect of review content on conversion, the cross-category analysis, and a counterfactual simulation.

6.1 Comparison of the Full Deep Learning Model and the Partial Deep Learning Model

We first compare the prediction performance of the full deep learning model (Section 4) and the partial deep learning model (Section 5). The objective is to directly predict the sales conversion rate in the 1% holdout sample. We compare the models based on three metrics: hit rate (1-misclassification rate), precision, and recall. The results are reported in Table 4. The hit rate for the full model is 88.54%,[12] much higher than the 66.13% for the partial model. This implies that combining all input variables, text features, and consumer/product characteristics in a joint deep learning framework improves prediction accuracy because this is a more efficient way of utilizing information. The two-step approach in the partial model only extracts six dimensions of content dimensions from the reviews, omitting other potentially useful content information. We also compare the two deep learning models with alternative models used in the prior literature. Specifically, in Model 3, we use only review rating, volume, and variance (as well as other consumer and product characteristics) but no content features. In Model 4, we add simple content features currently used in the literature, including topics (or LDA, Blei et al. 2003),[13] subjectivity, and readability. In Model 5, we replace the unsupervised LDA model with the supervised seeded LDA model (Jagarlamudi et al. 2012). The comparison reveals that adding the review content features to the models significantly improves model accuracy. This result highlights the importance of studying the impact of review content above and beyond review volume and rating, as well as the simple content features used in the literature.

[12] Hit rate is measured on the balanced sample. Our sample is highly imbalanced because 96% of the journeys result in no conversion. We use the over-sampling approach to construct the balanced sample (Kotsiantis et al. 2006). The hit rates on the original, imbalanced sample are in the range of 96.0% and 98.9%, across various models.

[13] We use the distribution over topics estimated using LDA. We choose the number of topics as 20 based on Perplexity [Blei et al., 2003].

[Insert Table 4 about here]

6.2 Results of the Full Deep Learning Model

[Insert Table 5 about here]

As explained in Section 4.8, the coefficients for the consumer and product characteristics in the full deep learning model can directionally inform us how they affect conversion, though no asymptotic inference can be drawn. In Table 5, we present the model coefficients. We present only a subset of the model coefficients in the last layer. The coefficients for percentage of products recommended, number of questions, number of answers, readability, and length are available upon request.

The signs of the coefficients are consistent with our expectation: price negatively influences conversion. A higher total number of reviews and higher average rating positively affect conversion. Surprisingly, we find a negative effect of the number of products searched in the journey. We believe that this variable indicates the consumer's purchase intention, because if a consumer searches many products, that consumer must have a high willingness to buy. However, counterforces also exist. First, it could be that the consumer is in the early stage of conversion funnel. Or it could be competition. Recall that our dependent variable is the conversion of the product the reviews are associated with. If the consumer has a large consideration set, the consumer is less likely to purchase any single product because of competition. Thus, the coefficient unveils that the competition effect is stronger than the intention effect. Moreover, the number of used interactions, for example pagination or sorting, is found to have a positive association with conversion because it reflects a higher purchase intention. Expectedly, the number of positive and negative reviews read affects conversion in opposite directions.

Although the coefficients for the text features in the full deep learning model are not directly interpretable, we extract salient n-grams that affect conversion the most (see Section 5.3.2 for details). Table 6 presents the five topics gleaned from running topic modeling (LDA, Blei et al. 2003) on all the salient n-grams. For each topic, we report the top six n-grams with the highest probabilities. Based on the n-grams, we label the five topics as aesthetics, price, feature, favor, and easy-to-use. For example, the topic "aesthetics" is represented by n-grams such as "look good," "lovely looking," and "great color." The topic "price" contains n-grams such as "great value," "the money," "good price," etc. The topic "feature" represents "what it says," "fit to," "work well," etc. Among all the salient n-grams, "aesthetics" accounts for 36%, while "price" makes up 28%. These are the top two topics that influence conversion. Notice that using this data-driven full deep learning model, we uncover that content related to "easy-to-use" has a big impact on sales. This content dimension attracts consumers but may be bypassed by firms.

[Insert Table 6 about here]

6.3 Results of the Partial Deep Learning Model

In the partial model, we leverage deep learning natural language processing techniques to extract six dimensions of content and their associated valence. Specifically, we use a recurrent neural network, a recursive neural network, and a convolutional neural network.

[Insert Table 7 about here]

We now compare the performance of various simple conventional machine learning models and deep learning algorithms, using three metrics: precision, recall, and F1 score (the harmonic average of precision and recall, $F1 = \frac{2*Precision*Recall}{Precision+Recall}$) [Jurafsky, 2000]. Table 7 shows the sentiment analysis accuracy of various models in the test sample for each information dimension. The counterparts in the information detection task are shown in Appendix H. Columns (1) and (2) in Table 7 are for conventional machine learning models using the bag-of-words representation, and columns (3) to (5) are for deep learning algorithms described in Section 5.3.1. The conventional classifiers, support vector machine (SVM) and naive Bayes (NB) have an average prediction accuracy (F1 score) of 26.3% and 67.7%, respectively. In contrast, deep learning algorithms generally have better performance, with the recurrent neural networks' average accuracy of 74.5%, recursive neural networks' 75.4%, and convolutional neural networks' 93.3%. This suggests that all deep learning algorithms outperform conventional natural language processing models. Dimension-wise, we have mixed results. Aesthetics 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 algorithms have better prediction performance, we examine reviews that are correctly predicted by deep learning algorithms but incorrectly predicted by conventional machine models. Table 8 illustrates some examples for each of the deep learning models.

[Insert Table 8 about here]

The recurrent neural networks model excels in distinguishing negation. For instance, keywords such as "least appealing" and "hardly acceptable" are detected for expressing negative sentiments. The recursive neural networks model, which relies on a tree structure to decipher syntactic relations, can discover that phrases following the contrastive conjunction "but" dominate the entire sentiment. For instance, the model correctly pinpoints that a review which states "It is good for the money but too flimsy" conveys a negative sentiment about the aesthetics of the product. Lastly, a CNN, 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 such as "without" and "useless," the entire sentence delivers a positive message about the battery.

Given the advantages of deep learning algorithms, we choose them to perform classification jobs. Considering the fact that CNN has the best prediction performance, in the rest of the paper, we report results generated from the CNN.

We also conduct a comprehensive comparison of different information sets. Below in table 9, we report the model comparison for nine different information sets (all using the partial deep learning framework with a balanced sample). Model 9 is our proposed model. As you can see, the hit rate of using reviews read is 66.13% while that of using all the reviews is only 55.88%, a 10% improvement. This highlights the informational value of using the reviews read instead of all the reviews available to predict sales conversion. The comparison also implies that using deep learning to carefully model review content can lead to a remarkable improvement in model performance compared to using only summary statistics of reviews (e.g. rating/volume/variance) or simple content matrics (e.g. topic/subjectivity/readability). The informational value of carefully sifted review content is even higher than that of shopper information.

[Insert Table 9 about here]

6.3.1 Visualize Salient Sentences in Reviews

Next we show the effectiveness of the deep learning algorithm to correctly detect distinct dimensions of information in the reviews. In Figure 9, for each of the six dimensions of content, 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.

The examples demonstrate that CNN can correctly locate the review fragment that corresponds to the particular information dimension. For example, in one review, the consumer said, "As expected worked brilliantly straight out of box. It is a 500 GB as description and not 1TB as in title but that is still a good amount of storage. Graphics are fantastic." The first sentence in this review indicates conformance, the second sentence describes feature, and the last sentence comments on aesthetics. The deep learning algorithm correctly detects that the third sentence is related to aesthetics because "Graphics are fantastic" are the words that can be changed the least to affect the gradient of the salience score the most, as explained in Section 5.3.2.

[Insert Figure 9 about here]

[Insert Figure 10 about here]

[Insert Figure 11 about here]

[Insert Figure 12 about here]

Another comparative advantage of deep learning models that is worth highlighting is their capability to sift features exclusive to various domains without researchers' domain knowledge to hand-pick features. Rather, deep learning models can accept raw data from any domain and automatically discover the representations pertinent to each domain. In our application, this comparative advantage of deep learning allows us to detect distinct review features appropriate for each product category. We use three examples to demonstrate that deep learning, or CNN in particular, can spot category-specific salient sentences which contain the six dimensions of information in the reviews. In Figures 10, 11, and 12, we exhibit examples of salient sentences in the Floorcare, TV, and Curtain categories, respectively. By comparing the three figures, we stress that even for information dimensions that involve broad representations, such as aesthetics or feature, deep learning is capable of identifying vastly diverse salient sentences across different categories.

For instance, in the Floorcare category, aesthetics can regard "smell," as shown in the review "the rooms have negative smell for few days" (Figure 10, row 2, column 3). On the other hand, in the TV category, aesthetics can mean "vibrations," as manifested by the review sentence "The vibrations are giving me headache" (Figure 11, row 2, column 3). Finally, in the Curtain category, aesthetics can reflect colors, as in the review "but they are the perfect colour for my kids room" (Figure 12, row 2, column 2). This distinguishing aspect of deep learning[14] makes it particularly useful for analyzing text data from a wide realm where extracting domain-specific features is time-consuming and error-prone.

6.3.2 Quantify the Effect of Review Content on Conversion

Next we present the measured causal impact of reviews on conversion. The results are presented in Table **??**. The summary statistics of the variables in this model are presented in Table 10. The last two columns indicate the source table and the key variable in the source table.

[Insert Table 10 about here]

[14]This ability is in part due to the semantic-aware word embedding (please see appendix W4, "word embedding") and in part due to the feature learning ability of neural networks.

We find that reviews containing favorable aesthetics and adverse price information can significantly affect conversion on both devices, whereas other dimensions, such as conformance, durability, features, or brand, are not prominent on PC. This finding echoes the result in the full deep learning model reported in Table 6, implying that the data-driven full deep learning model and the theory-driven partial deep learning model produce consistent results.

Moreover, we find negative price coefficients, which suggest that when price increases by $10, the odds ratio of conversion decreases by 3.3% on a mobile device and by 2.5% on a PC. This implies that price has a stronger effect on mobile than on PC. Besides, both the total number of reviews (available) and the average rating have significant positive impact on conversion, as expected. Similar to the result in Table 5, we find a significant negative effect of the number of products searched in the journey, indicating that the competition effect seems to be taking place. In contrast to Ghose and Ipeirotis [2011], we do not find readability to have a significant effect on conversion. Interestingly, the length of reviews read is found to have a significant but negative effect on conversion, supporting the finding in Chevalier and Mayzlin [2006].

As mentioned in the introduction, we are concerned that without the review-reading data, the effect of review content on conversion can be considerably underestimated. To illustrate this, in Table 12 we repeat the regressions in Table **??**, but assume that consumers read all reviews (or reviews posted before the consumer journey starts) instead of only a limited number of reviews. Under the RDiT design, this setting is equivalent to the one where each treatment involves only one content dimension, corresponding to the newly added review.

[Insert Table **??** about here]

Again, we use the example in Figure 8 as an illustration. For example, consumer A would be assumed to have read all the three reviews with conformance, aesthetics, and price information, and consumer B would be assumed to have read all the four reviews. Accordingly, the second experiment on October 23 relates only to the change in the feature dimension, but not the aesthetics dimension, because the review with aesthetics information is not crowded out. In other words, compared to consumer A, consumer B reads only one more review with feature information, but not one less review with aesthetics information. Therefore, for consumer B's visit, $FeaturePⁿ_{jt} = 1$ and $AestheticsP_{jt} = 0$. This visit helps identify only the coefficient $\tau_k^{FeatureP}$, but not $\tau_k^{AestheticsP}$. Similarly, the first experiment involves only the conformance dimension, but not the price dimension. As a result, consumer A's visit will be assigned with $ConformanceP_{jt} = 1$ and $PriceP_{jt} = 0$. So, consumer A's visit aids the identification of only $\tau_k^{ConformanceP}$, but not τ_k^{PriceP}.

[Insert Table 12 about here]

Table 12 shows the results with available reviews instead of read reviews. Comparing Table **??** and Table 12, we learn that the coefficients for review content measured without the review-reading data are on average underestimated by 20%. Importantly, the positive aesthetics content information no longer significantly impacts conversion. This highlights the importance of obtaining the review-reading data to correctly quantify the impact of review content on conversion.

6.3.3 Product Category Analysis

The above time-series analysis assumes that the effect of product reviews on conversion is homogeneous among product categories. This implies that, for instance, aesthetics information in reviews plays a similar role in the consumer purchase journey for "Lawnmowers" as for "Televisions." However, common sense suggests that the extent to which aesthetics information in reviews affects conversion should be greater for "Televisions" than for "Lawnmowers." To capture this heterogeneous effect of review information on

conversion across product categories, we modify the time-series model in equation (4) by allowing the coefficients to be category-specific. The new model specification becomes

$$u_{ijkt} = \overrightarrow{ReviewContent_{jt}} \cdot \overrightarrow{\tau}_{kc} + \sum_{n=1}^{3} \delta_{nkc} t^n + \overrightarrow{\theta_{kc}} \overrightarrow{Z_{it}} + \overrightarrow{\gamma_{kc}} \overrightarrow{X_{jt}} + \xi_c + Weekend_{tc} + Daytime_{tc} + \varepsilon_{ijkt}, \quad (6)$$

where the subscript c in coefficients $(\overrightarrow{\tau_{kc}}, \delta_{nkc}, \overrightarrow{\theta_{kc}}, \overrightarrow{\gamma_{kc}}, \xi_c, Weekend_{tc}, Daytime_{tc})$ indicates the product category to which product j belongs.

We estimate this model using a hierarchical Bayes framework.[15] Please see more model details in Appendix I.

Figures 13 and 14 present the effect of review content information on conversion estimated using the hierarchical Bayes model. Note that this is the subset of parameters that we are most interested in when comparing the heterogeneity across categories.

[Insert Figure 13 about here]

[Insert Figure 14 about here]

Consistent with the findings in Table **??**, in almost all the 37 categories,[16] favorable aesthetics and adverse price information have a significant impact on conversion.[17] A few exceptions occur for "Pet Supplies," "Lawnmowers and garden power tools," and "Barbecues and garden heating," where the effect of aesthetics is not statistically significant. This matches our intuition that in categories such as "Lawnmowers and garden power tools" and "Barbecues and garden heating," aesthetics should not be a crucial factor in consumers' decisions. As for conformance, we find that in most product categories, the effect is not significant. However, for "Sofas, armchairs and chairs," "Lighting," and "Floorcare," the effect turns out to be significantly positive. This might be because the quality uncertainty for products in these categories is relatively higher than that for other categories, and consumers' performance evaluation is often subjective, with a considerable variation. Therefore, conformance matters more in these categories than in others. A similar result is found for durability. Although in most product categories the effect of durability is not significant, for "Storage desks and filing," "Lawnmowers and garden power tools," "Heating and cooling," and "Beds," durability information in reviews plays a significant role in driving purchases.

The hierarchical Bayes analysis in this section enables us to investigate the effect of review content across categories. Next, we further decompose the difference among categories in four aspects: rating, competition, dynamic, and brand.

Rating

[Insert Figure 15[18] about here]

[15] Both the homogeneous model and the Hierarchical Bayes model are trained by pooling all the product categories to increase efficiency.

[16] There are 583 categories in total. We choose the top 37 sub-categories, with more than 220 journeys in each category, to report in Figure 17 and 18. The figures would look too messy if we were to report all the 583 categories.

[17] For aesthetics, conformance, durability, and brand, we present the coefficients for the positive sentiment. For price, we present the negative sentiment. The other coefficients are omitted for brevity.

[18] The numbers in the figure and the subsequent figures in this section come from the hierarchical Bayes estimates.

Is review content more relevant when the information in ratings is tenuous? This question is of great importance in a world with rating inflation. To investigate, we plot the effect of review content on conversion by the mean and standard deviation of ratings in Figure 15. In this figure, each point is one product category. The color of a point denotes the sum of the absolute values of the review content coefficients.[19] The darker (more purple, less yellow) the color, the higher the effect of content on conversion. The color pattern in the figure suggests that for categories where products have high ratings with low dispersion, review content has a high impact on conversion. This might be because the information is noisy, so consumers heavily rely on review content to compare products. Consistent with our expectation, the higher the variance of the rating, the less effective review content is. We also find that the higher the mean rating, the more important review content is. This might be because consumers tend to ignore review content if the rating is lower than some minimum threshold.

Competition Effect

Does review content have a higher impact on conversion in a more competitive market? On the one hand, this might be true because in a more competitive market, consumer choices might be based on small differences in quality or price. Under such conditions, review content may be more likely to provide the marginal push that determines final purchases. On the other hand, in a more competitive market, firms may compete to provide information to buyers so that reviews provide no additional value. We test these competing hypotheses in Figure 16. The x-axis in Figure 16 is the Herfindahl–Hirschman Index (h-index), which measures market concentration of each product category. The lower the h-index, the more competitive the market. The y-axis is the review content effect on conversion, measured as the sum of the absolute values of the review content coefficients in Table **??**. The downward sloping trend (slope = -1.475, p_value<0.0001) indicates that review content is more useful in a more competitive market, consistent with the former view.

[Insert Figure 16 about here]

Dynamic Effect

Next we explore the dynamic effect and answer the question, "Is review content more effective for early stage products?" This is intuitive because consumers have higher uncertainty with products in the early stage than with mature products. Hence, they are more likely to rely on review content to learn about the early stage products. In our data, products in different categories were added to the market at different time points. Figure 17 exhibits the histogram of the (product) category launch time. Although a large number of categories have existed since the inception of the platform, many categories debuted later, between 2009 and 2015.

[Insert Figure 17 about here]

To assess the dynamic effect, in Figure 18, we plot the effect of review content (measured by the sum of the absolute values of the review content coefficients) against the number of days since a category was launched on the e-commerce platform. The fitted regression line suggests a negative relationship (slope=-0.0002, p_value=0.045) between the review content effect and tenure. In other words, review content is more effective for product categories that are new to the site than for those categories that have existed for a long time.

[19] $|\tau_k^{Pos}|+|\tau_k^{Neg}|+|\tau_k^{AestheticsP}|+|\tau_k^{AestheticsN}|+|\tau_k^{ConformanceP}|+|\tau_k^{ConformanceN}|+|\tau_k^{DurabilityP}|+|\tau_k^{DurabilityN}|+|\tau_k^{FeatureP}|$
$+|\tau_k^{FeatureN}|+|\tau_k^{PerceivedQualityP}|+|\tau_k^{PerceivedQualityN}|+|\tau_k^{PriceP}|+|\tau_k^{PriceN}|$

[Insert Figure 18 about here]

Brand Effect

Finally, we hypothesize that consumers are less likely to be influenced by review content when brand information is easily accessible. This happens because a brand itself conveys rich information that can substitute for the information in review content. In our data, some products are "unbranded," or without brand names. To test this hypothesis, for each category, we compute the percentage of products with brand names and plot it against the effect of review content in Figure 19. The fitted regression line has a slope of -0.35 (p_value=0.03). This implies that in categories where most products have brand identity, consumers depend less on review content in their purchase decisions. This confirms that brand names and review content information can be substitutes for each other.

[Insert Figure 19 about here]

6.4 Counterfactual of Changing the Ranking Algorithm

After discovering the relative importance of different content information in the review texts, in this section, we propose a strategy that marketers can leverage to boost conversion rate: reordering reviews. Earlier in previous section, our results imply that consumers pay attention not only to the summary statistics of reviews (e.g., average rating, total number of reviews) but also to the actual content of reviews. Their conversion rate is influenced by the content information embedded in the reviews. For example, aesthetics information has a stronger positive impact on conversion than other dimensions. As a consequence, within the set of reviews with the same rating score, marketers can display the reviews with positive aesthetics before other reviews, to increase conversion.[20]

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.[21] 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 20a displays the histogram of the increase in the conversion rate odds ratio. The average increase in the odds ratio of the conversion rate is 44%, while the maximum is 143%. Given the average conversion rate of 3.9%, this implies after reordering reviews, the conversion rate can go up to an average of 5.52%, with the highest of 8.98%. This also indicates that on average, reordering reviews by presenting one more review with positive aesthetics information is as effective as a 1.6% price cut to increase the conversion rate odds ratio.[22]

[Insert Figure 20a about here]

[20] Similar practices have been undertaken by Amazon, which 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.

[21] The counterfactual assumes that the consumers review-reading behaviors will not be affected by this policy change. Please see our discussion of the endogenous review-reading behavior in Appendix J. Although in our setting, empirical evidence suggests that consumers' review-reading behaviors (e.g., reading sequence and how many reviews to read) will not change in the counterfactual scenario, this pattern might not hold in other settings or on other e-commerce websites. Future research should explicitly model consumers' endogenous reading behavior in order to get a robust estimate of the effect of re-ordering reviews.

[22] To obtain price elasticity, we run the model in equation (4) with logged average price. We find that decreasing price by 1% can lead to an increase in odds ratio by 28% on a PC.

We also consider another couterfactual where we reorder reviews such that consumers can read the most diversified content information, while keeping the number of reviews and average rating of the reviews unchanged. We think that presenting the most diversified information to consumers can help consumers become fully-informed efficiently. We use the Herfindahl-Hirschman Index (HHI) to measure content (un)diversification. Please see the equation below. Simply put, HHI measures how concentrated (the opposite of diversified) the content information is

$$HHI = \sum_{d\in\{aesthetics,conformance,durability,feature,brand,price\},s\in\{P,N\}} shares_{ds}^2$$
$$= \sum_{d\in\{aesthetics,conformance,durability,feature,brand,price\},s\in\{P,N\}} \left(\frac{\#reviews\,with\,ds}{total\,\#\,reviews\,read}\right)$$

The counterfactual result is presented in Figure 20b below. As shown, when presenting well-diversified content information to consumers, although the conversion rate decreases in some journeys, the majority, or almost two thirds (66.5%) of the journeys have a higher conversion rate odds ratio. This implies that providing valuable and impartial information to consumers can be a win-win for both the e-commerce platform and consumers.

[Insert Figure 20b about here]

7 Conclusions and Limitations

This paper studies the role of review content in consumer purchase journeys. We leverage a unique, granular dataset that tracks individual consumers' entire decision journeys, including review reading, search, and purchase. This allows us to discover for what (types of products), for whom (how many consumers), and where (on which device) consumers read review content, as well as what dimensions of review content have a causal impact on conversion. We find that favorable aesthetics information and adverse price information can significantly affect conversion on both mobile and PC and across a wide range of product categories. The impact of review content on conversion is stronger when the review rating is concentrated and inflated, in a more competitive market, for new products, and when brand information is not easily accessible.

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 of any product category. Second, based on our finding regarding the relative importance of review content dimensions, managers can incorporate reviews as a new marketing mix, by refining the ranking and information presentation algorithms to provide the most relevant reviews to consumers. Third, managers can collect real-time information about the consumer purchase journey, including device and reviews read, to predict final conversion more accurately.

This paper has several limitations. Currently, we look only at the effect of review-reading behaviors on conversion. Another interesting angle would be to examine the effect of reviews on consumer search behaviors. Questions such as "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, due to data limitation, we have not accounted for consumer heterogeneity when quantifying the causal impact of review reading on conversion. This consideration might be useful for marketers to design targeted review ranking and presentation algorithms. Furthermore, we acknowledge the possibility of measurement error for the variable "number of reviews read." Future studies may consider alternative methods to measure consumers' review reading behavior, such as eye-tracking.

Lastly, the counterfactual exercise assumes a partial equilibrium without taking into account other platforms' competitive responses and consumers' entry or exit decisions. Future research may investigate the long-run consequences of the proposed review ranking algorithm in a general equilibrium setting.

References

Nikolay Archak, Anindya Ghose, and Panagiotis G Ipeirotis. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8):1485–1509, 2011.

Yoshua Bengio, Holger Schwenk, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

Jonah Berger, Alan T Sorensen, and Scott J Rasmussen. Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*, 29(5):815–827, 2010.

Christopher M Bishop. Pattern recognition. *Machine Learning*, 128:1–58, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Fernanda Brollo, Tommaso Nannicini, Roberto Perotti, and Guido Tabellini. The political resource curse. *The American Economic Review*, 103(5):1759–1796, 2013.

Joachim B¨uschken and Greg M Allenby. Sentence-based text analysis for customer reviews. *Marketing Science*, 35(6):953–975, 2016.

David Card, Raj Chetty, and Andrea Weber. Cash-on-hand and competing models of intertemporal behavior. *The Quarterly journal of economics*, 122(4):1511–1560, 2007.

Matias D Cattaneo, Rocío Titiunik, and Luke Keele. Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4):1229–1248, 2016.

Augustin Cauchy. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

Kenneth Y Chay, Patrick J McEwan, and Miguel Urquiola. The central role of noise in evaluating interventions that use test scores to rank schools. *The American Economic Review*, 95(4):1237–1258, 2005.

Judith A Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354, 2006.

Pradeep K Chintagunta, Shyam Gopinath, and Sriram Venkataraman. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, 29(5):944–957, 2010.

Reinhold Decker and Michael Trusov. Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 27(4):293–307, 2010.

Chrysanthos Dellarocas, Xiaoquan Michael Zhang, and Neveen F Awad. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing*, 21 (4):23–45, 2007.

Misha Denil, Alban Demiraj, and Nando de Freitas. Extraction of salient sentences from labelled documents. Technical report, University of Oxford, 2014.

Vasant Dhar and Elaine A Chang. Does chatter matter? the impact of user-generated content on music sales. *Journal of Interactive Marketing*, 23(4):300–307, 2009.

Wenjing Duan, Bin Gu, and Andrew B Whinston. Do online reviews matter an empirical investigation of panel data. *Decision support systems*, 45(4):1007–1016, 2008.

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

David A Garvin. What does product quality really mean? *Sloan management review*, page 25, 1984.

Anindya Ghose and Panagiotis G Ipeirotis. 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, 2011.

David Godes and Dina Mayzlin. Using online conversations to study word-of-mouth communication. *Marketing science*, 23(4):545–560, 2004.

Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Catherine Hausman and David S Rapson. Regression discontinuity in time: Considerations for empirical applications. Technical report, National Bureau of Economic Research, 2017.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics, 2012.

Daniel Jurafsky. Speech and language processing: An introduction to natural language processing. *Computational linguistics, and speech recognition*, 2000.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014.

Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.

Gilles Laurent and Jean-Noel Kapferer. Measuring consumer involvement profiles. *Journal of marketing research*, pages 41–53, 1985.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

David S Lee and Thomas Lemieuxa. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

Dokyun Lee, Kartik Hosanagar, and Harikesh Nair. Advertising content and consumer engagement on social media: Evidence from facebook. *Management Science*, 2018.

Thomas Y Lee and Eric T Bradlow. Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48(5):881–894, 2011.

Xiao Liu, Param Vir Singh, and Kannan Srinivasan. A structured analysis of unstructured big data by leveraging cloud computing. *Marketing Science*, 35(3):363–388, 2016.

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

Stephan Ludwig, Ko De Ruyter, Mike Friedman, Elisabeth C Brüggen, Martin Wetzels, and Gerard Pfann. More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, 77(1):87–103, 2013.

Shawn Mankad, Hyunjeong "Spring" Han, Joel Goh, and Srinagesh Gavirneni. Understanding online hotel reviews through automated text analysis. *Service Science*, 8(2):124–138, 2016.

G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Wendy W Moe and Michael Trusov. The value of social dynamics in online product ratings forums. *Journal of Marketing Research*, 48(3):444–456, 2011.

Hyoryung Nam, Yogesh V Joshi, and PK Kannan. Harvesting brand information from social tags. *Journal of Marketing*, 2017.

Phillip Nelson. Information and consumer behavior. *Journal of political economy*, 78(2):311–329, 1970.

Oded Netzer, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko. Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3):521–543, 2012.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86, 2002.

Dinesh Puranam, Vishal Narayan, and Vrinda Kadiyali. The effect of calorie posting regulation on consumer opinion. *Marketing Science*, 2017.

Paul Resnick and Richard Zeckhauser. Trust among strangers in internet transactions. *The Economics of the Internet and E-commerce*, 11(2):23–25, 2002.

Matthew J Schneider and Sachin Gupta. Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, 32(2):243–256, 2016.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, volume 1631, page 1642, 2013.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1): 1929–1958, 2014.

Monic Sun. How does the variance of product ratings matter? *Management Science*, 58(4):696–707, 2012.

Seshadri Tirunillai and Gerard J Tellis. Does chatter really matter? dynamics of user-generated content and stock performance. *Marketing Science*, 31(2):198–215, 2012.

Seshadri Tirunillai and Gerard J Tellis. Mining marketing meaning from online chatter: Strategic brand analysis using latent dirichlet allocation. *Journal of Marketing Research*, 51(4):463–479, 2014.

Geoffrey Towell and Jude W Shavlik. Interpretation of artificial neural networks: Mapping knowledge-based neural networks into rules. In *Advances in neural information processing systems*, pages 977–984, 1992.

Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

Xin Wang, Yuanchao Liu, Chengjie Sun, Baoxun Wang, and Xiaolong Wang. 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*, volume 1, pages 1343–1353, 2015.

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

Tables and Figures

Table 1: Literature Review of the Effect of UGC (User-Generated Content) on Market and Conversion

Paper	UGC Variable	Text Mining Method	Products	# of Categories
Resnick and Zeckhauser 2002	Volume	N/A	Unknown	unknown
Godes and Mayzlin 2004	Volume, rating, variance	N/A	TV shows	1
Liu 2006	Volume, rating	N/A	Movies	1
Chevalier and Mayzlin 2006	Volume, rating	N/A	Books	1
Dellarocas et al. 2007	Volume, rating, variance	N/A	Movies	1
Duan et al. 2008	Volume, rating	N/A	Movies	1
Dhar and Chang 2009	Volume, rating	N/A	Music	1
Chintagunta et al. 2010	Rating	N/A	Movies	1
Berger et al. 2010	Rating, sentiment	N/A	Book	1
Zhu and Zhang 2010	Volume, rating	N/A	Games	1
Archak et al., 2011	Volume, rating, content dimensions	Feature engineering	Digital cameras, camcorders	2
Moe and Trusov 2011	Rating	N/A	Bath, fragrance, beauty	3
Sun 2012	Rating, variance	N/A	Books	1
Netzer et al. 2012	Lexical based	Lexical based semantic network	Cars	1
Ludwig et al. 2013	Affect	Keyword detection	Books	1
Tirunillai and Tellis 2014	Bag-of-words content	LDA	PC, phone, footwear, toys, data storage	5
Mankad et al. 2016	Sentiment	Lexical approach	Hotels	1
Schneider and Gupta 2016	Bag-of-words content	Bag-of-words random projection	Tablet computers	1
Liu et al. 2016	Volume, sentiment, n-grams	N-gram PCA	TV shows	1
This paper	Volume, sentiment, theory-driven content	Deep learning	Multiple categories	≈ 600

Table 2: Characteristics of Different Types of Journeys

Type	1	2	3	4	5
	no search + no review + 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 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 preexisting standards or set expectations.
Durability	The review describes the experience with the durability or product malfunctions or failing to work as per the customer's satisfaction.
Feature	The review talks about presence or absence of product features.
Brand	The review talks about indirect measures of the quality of the product such as the reputation of the brand.
Price	The review contains content regarding price of the product.

Table 4: Model Comparison: Full Deep Learning Model vs. Partial Deep Learning Model

	DV: Conversion	Balanced			Imbalanced		
		Hit Rate	Precision	Recall	Hit Rate	Precision	Recall
1	Full Deep Learning Model	88.54%	83.36%	96.48%	98.90%	20.08%	0.03%
2	Partial Deep Learning Approach	66.13%	63.43%	79.06%	97.23%	8.77%	2.35%
3	No Content (but with rating/volume/variance)	57.51%	59.13%	50.96%	96.02%	5.86%	1.50%
4	Simple Content Features (with rating/volume/variance/ topics /subjectivity/readability)	58.13%	56.90%	69.58%	96.28%	5.00%	1.66%
5	Simple Content Features (Seeded LDA)	58.05%	58.02%	60.55%	96.23%	5.81%	1.63%

Table 5: Coefficients in the Full Deep Learning Models: Consumer and Product Characteristics

	Mobile	PC
# Positive Reviews Read	0.039	0.06
# Negative Reviews Read	-0.088	-0.08
# Products Searched	-0.135	-0.084
# Used Interactions	0.442	0.256
Total # of Reviews	0.004	0.01
%Recommend	0.0007	0.0003
Rating Average	0.57	0.84
Rating Variance	-0.008	-0.001
# Questions	-0.057	0.009
# Answers	0.063	0.051
readability	0.752	-0.066
length	-0.0004	-0.0005
Log_Price	-0.787	-0.868
Obs	156,445	

Table 6: Topic Modeling of Salient N-grams

Aesthetics 36%	Price 28%	Feature 16%	Favor 12%	Easy-to-use 8%
look good	great value	what it says	highly recommend	easy to
looks good	the price	good quality	love this	to use
lovely looking	for the price	perfect for	good item	to assemble
look really nice	the money	fit to	very happy	to change
feel great	good price	the job	very pleased	set up
great color	good value	work well	quite nice	is fast

Table 7: Model Comparison: Sentiment Analysis

Classifier/Accuracy %		(1) SVM + BoW	(2) NB+ BoW	(3) Recurrent-LSTM	(4) Recursive	(5) Convolutional
Mean	Precision	0.634	0.769	0.740	0.710	0.955
	Recall	0.167	0.606	0.754	0.809	0.912
	F1	0.263	0.677	0.745	0.754	0.933
Aesthetics	Precision	0.698	0.810	0.774	0.705	0.976
	Recall	0.160	0.594	0.747	0.779	0.948
	F1	0.261	0.685	0.760	0.740	0.962
Conformance	Precision	0.545	0.738	0.728	0.693	0.884
	Recall	0.131	0.657	0.707	0.827	0.876
	F1	0.212	0.695	0.717	0.754	0.880
Durability	Precision	0.487	0.756	0.670	0.667	0.921
	Recall	0.125	0.671	0.857	0.936	0.862
	F1	0.199	0.711	0.752	0.779	0.891
Feature	Precision	0.667	0.781	0.750	0.716	0.981
	Recall	0.133	0.614	0.747	0.668	0.925
	F1	0.222	0.688	0.749	0.691	0.952
Brand	Precision	0.614	0.760	0.746	0.729	0.969
	Recall	0.197	0.533	0.697	0.810	0.916
	F1	0.298	0.627	0.720	0.767	0.942
Price	Precision	0.795	0.767	0.775	0.748	0.997
	Recall	0.256	0.570	0.769	0.837	0.947
	F1	0.388	0.654	0.772	0.790	0.971

Table 8: 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 carpet is hardly acceptable
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.

Table 9: Model Comparison: Different Information Sets

	DV: Conversion	Hit Rate	Precision	Recall
1	Product only	52.55%	52.32%	57.13%
2	Shopper only	53.16%	53.07%	58.07%
3	Reviews read only	55.28%	57.61%	56.07%
4	Product + shopper	53.76%	55.22%	59.79%
5	Product + shopper + reviews read no content (but with rating/volume/variance)	57.09%	58.96%	50.78%
6	Product + shopper + reviews read simple content (rating/volume/variance/subjectivity/readability/LDA)	57.87%	56.77%	68.76%
7	Product + shopper + reviews read simple content (rating/volume/variance/subjectivity/readability/seeded LDA)	57.98%	57.69%	59.73%
8	Product + shopper + all reviews deep learning	55.88%	57.82%	51.22%
9	Product + shopper + reviews read deep learning	66.13%	63.43%	79.06%

Table 10: Summary Statistics of Variables

Variable	Definition	Mobile					PC					Source Table	Key
		N	Mean	Std Dev	Min	Max	N	Mean	Std Dev	Min	Max		
Transaction[23]	Dummy variable indicating whether the journey ends with a transaction or not	156445	0.04	0.19	0	1						Transaction	userid
# Positive	number of positive reviews read by consumers in the journey	156445	1.42	2.06	0	125	156445	2.73	2.14	0	150	Impression	userid
# Negative	number of negative reviews read by consumers in the journey	156445	0.25	0.66	0	78	156445	0.48	0.85	0	87	Impression	userid
Review-Aesthetics P	average positive aesthetics score of reviews read in the journey	156445	0.09	0.16	0	1	156445	0.15	0.16	0	1	Impression	userid
Review-Aesthetics N	average negative aesthetics score of reviews read in the journey	156445	-0.05	0.11	-1	0	156445	-0.08	0.12	-1	0	Impression	userid
Review-Conformance P	average positive conformance score of reviews read in the journey	156445	0.05	0.11	0	1	156445	0.09	0.12	0	1	Impression	userid
Review-Conformance N	average negative conformance score of reviews read in the journey	156445	-0.12	0.18	-1	0	156445	-0.24	0.19	-1	0	Impression	userid
Review-Durability P	average positive durability score of reviews read in the journey	156445	0.04	0.11	0	1	156445	0.06	0.10	0	1	Impression	userid
Review-Durability N	average negative durability score of reviews read in the journey	156445	-0.10	0.16	-1	0	156445	-0.16	0.16	-1	0	Impression	userid
Review-Feature P	average positive feature score of reviews read in the journey	156445	0.01	0.07	0	1	156445	0.03	0.09	0	1	Impression	userid
Review-Feature N	average negative feature score of reviews read in the journey	156445	-0.02	0.08	-1	0	156445	-0.03	0.11	-1	0	Impression	userid
Review-Brand P	average positive brand score of reviews read in the journey	156445	0.03	0.09	0	1	156445	0.06	0.12	0	1	Impression	userid
Review-Brand N	average negative brand score of reviews read in the journey	156445	-0.04	0.11	-1	0	156445	-0.08	0.14	-1	0	Impression	userid
Review-Price P	average positive price score of reviews read in the journey	156445	0.01	0.06	0	1	156445	0.01	0.08	0	1	Impression	userid
Review-Price N	average negative price score of reviews read in the journey	156445	-0.01	0.06	-1	0	156445	-0.02	0.08	-1	0	Impression	userid
Time	difference (number of seconds) between the journey time and the sample start time	156445	871941	1463486	0	5096300	156445	1684110	1671912	0	5097562	Pageview	userid
# Products Searched	number of products searched in the journey	156445	1.75	4.48	0	89	156445	4.49	7.93	0	118	Pageview	userid
Used Features	number of web-features used in the journey	156445	0.34	0.48	0	2	156445	0.66	0.48	0	2	Usedfeature	userid
Total # Reviews	total number of reviews available for the product at the time of the journey	156445	62.60	215.73	0	4779	156445	114.37	265.27	0	4779	Review	productid
% Recommend	percentage of users that recommended this product at the time of the journey	156445	29.83	42.12	0	100	156445	57.42	42.78	0	100	Review	productid
Rating Average	average rating of reviews available for the product at the time of the journey	156445	1.46	2.05	0	5	156445	2.81	2.07	0	5	Review	productid
Rating Variance	variance of rating of reviews available for the product at the time of the journey	156445	0.46	0.80	0	8	156445	0.88	0.93	0	8	Review	productid
# Questions	total number of questions available for the product at the time of the journey	156445	5.26	18.16	0	276	156445	11.41	26.88	0	276	Question	productid
# Answers	total number of answers available for the product at the time of the journey	156445	6.07	21.45	0	329	156445	13.14	31.77	0	329	Answer	productid
Price	price of the product at the time of the journey	156445	27.29	68.60	0	2649.99	156445	54.39	91.00	0	2649.99	Transaction	userid
Readability	average readability score of reviews read in the journey	156445	2.88	4.07	0	15.74	156445	5.55	4.16	0	16.49	Impression	userid
Length	average length of reviews read in the journey	156445	10.89	16.55	0	87.40	156445	20.92	17.75	0	87.40	Impression	userid

[23] Because "transaction" is the dependent variable, we did not differentiate whether it occurred on PC or mobile. Instead, we differentiate all the independent variables based on the device (PC or mobile) because the coefficients on the right-hand side of equation (4) are device specific.

Table 11: The Effect of Review Content on Conversion: Read Reviews

Review Content Variables	Mobile Est.(Std.)	PC Est.(Std.)	Control Variables	Mobile Est.(Std.)	PC Est.(Std.)
# Positive Reviews Read	0.238***	0.0716	Time	3.43e-08	-5.71e-08
	(0.0551)	(0.0535)		(3.05e-08)	(3.26e-08)
# Negative Reviews Read	0.0951	-0.0493	Time^2	-1.19e-14	-2.57e-14*
	(0.0612)	(0.0609)		(9.61e-15)	(1.19e-14)
Review-Aesthetics P	0.0365*	0.0897***	Time^3	1.91e-21	3.32e-22
	(0.0183)	(0.0195)		(4.04e-21)	(4.56e-21)
Review-Aesthetics N	0.0529	0.0336	# Products Searched	-0.0600***	-0.0452***
	(0.0573)	(0.0626)		(0.00489)	(0.00375)
Review-Conformance P	-0.00493	-0.00103	Used Interactions	-0.987	0.880
	(0.0295)	(0.0318)		(0.558)	(0.452)
Review-Conformance N	-0.109*	-0.0157	Total # Reviews	0.000232***	0.000218***
	(0.0531)	(0.0555)		(0.0000541)	(0.0000609)
Review-Durability P	0.0807**	-0.0367	% Recommend	0.00842*	-0.0000245
	(0.0259)	(0.0291)		(0.00412)	(0.00416)
Review-Durability N	0.0103	0.0461	Rating Average	0.140	0.378***
	(0.0357)	(0.0380)		(0.110)	(0.112)
Review-Feature P	-0.0203	0.0128	Rating Variance	-0.0254	-0.00745
	(0.0255)	(0.0276)		(0.0313)	(0.0323)
Review-Feature N	-0.0791*	-0.000964	# Questions	0.00636	-0.00386
	(0.0360)	(0.0382)		(0.00495)	(0.00428)
Review-Brand P	-0.0362	-0.0697	# Answers	-0.00889*	0.000952
	(0.0359)	(0.0391)		(0.00422)	(0.00361)
Review-Brand N	0.00894	0.0284	Price	-0.00326***	-0.00248***
	(0.105)	(0.111)		(0.000276)	(0.000269)
Review-Price P	0.00545	0.0187	Readability	0.00416	-0.00370
	(0.0192)	(0.0202)		(0.00288)	(0.00308)
Review-Price N	-0.1381**	-0.1941**	Length	-0.000909*	-0.000862*
	(0.0586)	(0.0629)		(0.000375)	(0.000405)
			Intercept	2.325**	-5.887***
Product FE	Yes			(0.813)	(0.548)
Weekend FE	Yes		Daytime FE	Yes	
Obs	156445[24]				
BIC	50068.2				

Standard errors in parentheses * p<0.05, ** p<0.01, *** p<0.001

[24] See our explanation of the number of observations in Appendix K.

Table 12: The Effect of Review Content on Conversion: Available Reviews

Review Content Variables	Mobile Est.(Std.)	PC Est.(Std.)	Control Variables	Mobile Est.(Std.)	PC Est.(Std.)
# Positive Reviews Read	0.0299	0.00114	Time	3.27e-08	-5.78e-08
	(0.0172)	(0.0188)		(3.05e-08)	(3.25e-08)
# Negative Reviews Read	0.0308	0.00199	Time^2	-8.09e-15	-2.47e-14*
	(0.0173)	(0.0189)		(9.60e-15)	(1.19e-14)
Review-Aesthetics P	0.00777	0.0200	Time^3	2.14e-21	3.66e-22
	(0.0112)	(0.0125)		(4.00e-21)	(4.56e-21)
Review-Aesthetics N	-0.00626	-0.0607	# Products Searched	-0.0608***	-0.0448***
	(0.0403)	(0.0448)		(0.00491)	(0.00374)
Review-Conformance P	0.0267	-0.00487	Used Interactions	-0.179	1.033*
	(0.0189)	(0.0211)		(0.524)	(0.433)
Review-Conformance N	-0.0769*	-0.0160	Total # Reviews	0.000211**	0.000157*
	(0.0349)	(0.0384)		(0.0000648)	(0.0000761)
Review-Durability P	0.0658***	0.0337	% Recommend	0.00967*	0.00184
	(0.0179)	(0.0197)		(0.00400)	(0.00416)
Review-Durability N	0.0448	-0.0120	Rating Average	0.233*	0.450***
	(0.0238)	(0.0260)		(0.107)	(0.112)
Review-Feature P	-0.0172	0.00797	Rating Variance	-0.0383	-0.0203
	(0.0175)	(0.0190)		(0.0317)	(0.0328)
Review-Feature N	-0.0765**	0.0153	# Questions	0.00948	0.0000560
	(0.0243)	(0.0262)		(0.00518)	(0.00475)
Review-Brand P	-0.014	-0.0353	# Answers	-0.0108*	-0.00197
	(0.0216)	(0.0241)		(0.00446)	(0.00406)
Review-Brand N	0.0195	0.0108	Price	-0.00297***	-0.00236***
	(0.0671)	(0.0712)		(0.000281)	(0.000272)
Review-Price P	-0.00373	-0.0172	Readability	0.0000103	-0.000112
	(0.0110)	(0.0123)		(0.000169)	(0.000171)
Review-Price N	-0.109**	-0.128**	Length	-0.0000607**	-0.0000636**
	(0.0398)	(0.0440)		(0.0000223)	(0.0000239)
			Intercept	2.338**	-6.375***
Product FE	Yes			(0.793)	(0.528)
Weekend FE	Yes		Daytime FE	Yes	
Obs	156445				
BIC	42903.0				

Standard errors in parentheses * p<0.05, ** p<0.01, *** p<0.001

Figure 1: Sample Screenshots of the Webpages

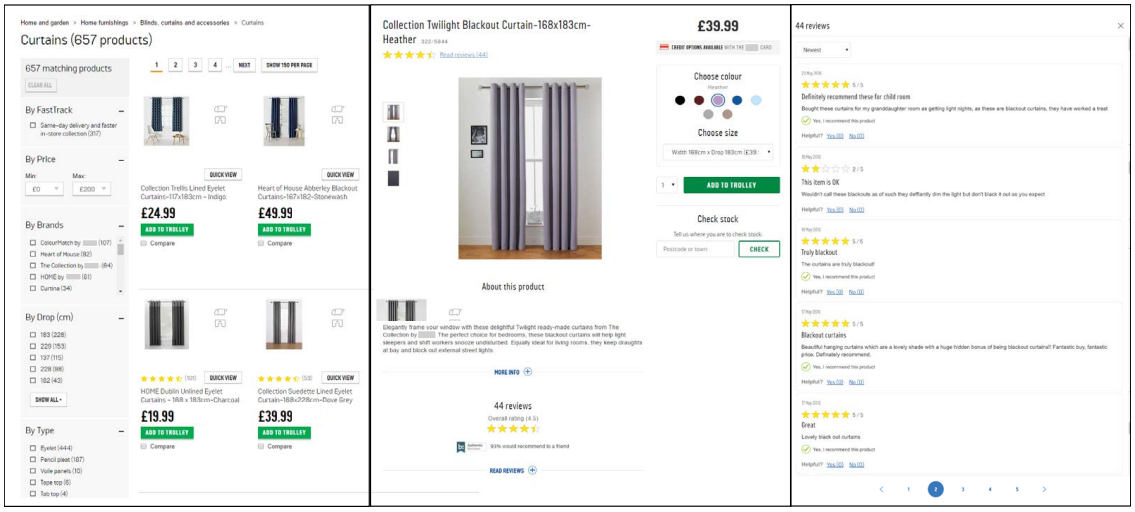


Figure 2: Word Cloud of Product Categories

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

Figure 3: Journey Distribution

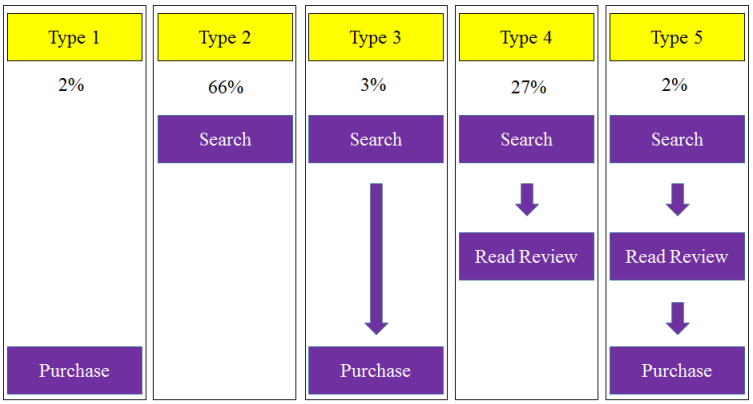
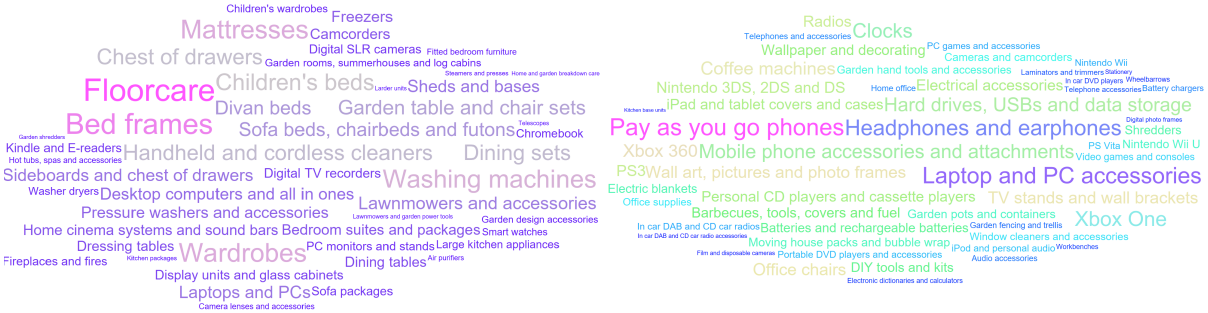
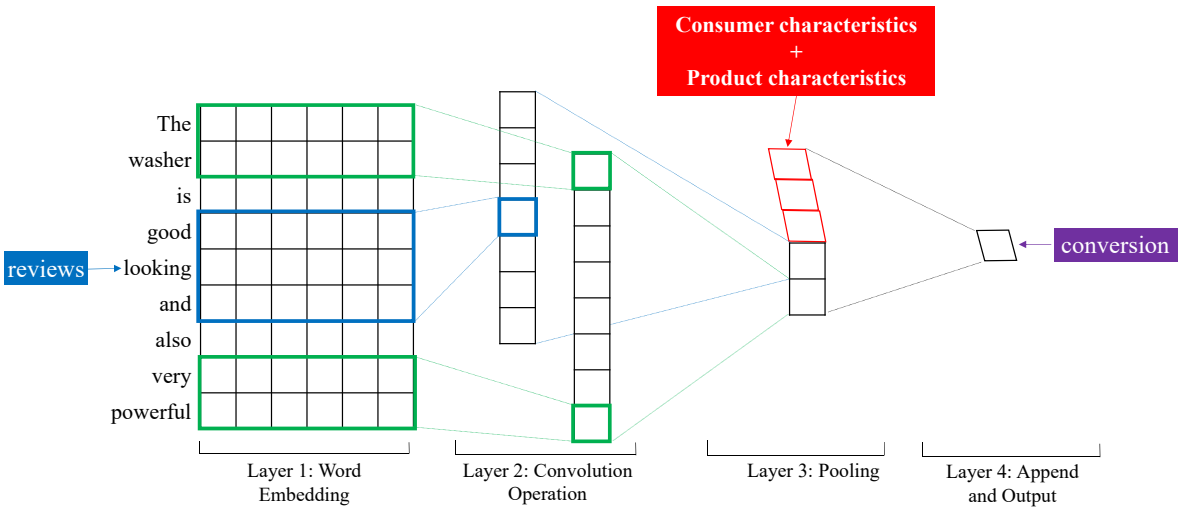


Figure 4: Examples of Products About Which Consumers Do (left) vs. Don't (right) Read Reviews



Note: The font size indicates the number of journeys associated with the product category.

Figure 5: Convolutional Neural Network Model for Conversion



Note. The squares in the figure denote neurons, and the lines denote the connections between inputs and outputs.

Figure 6: Example of Changing Review Content

Time 1

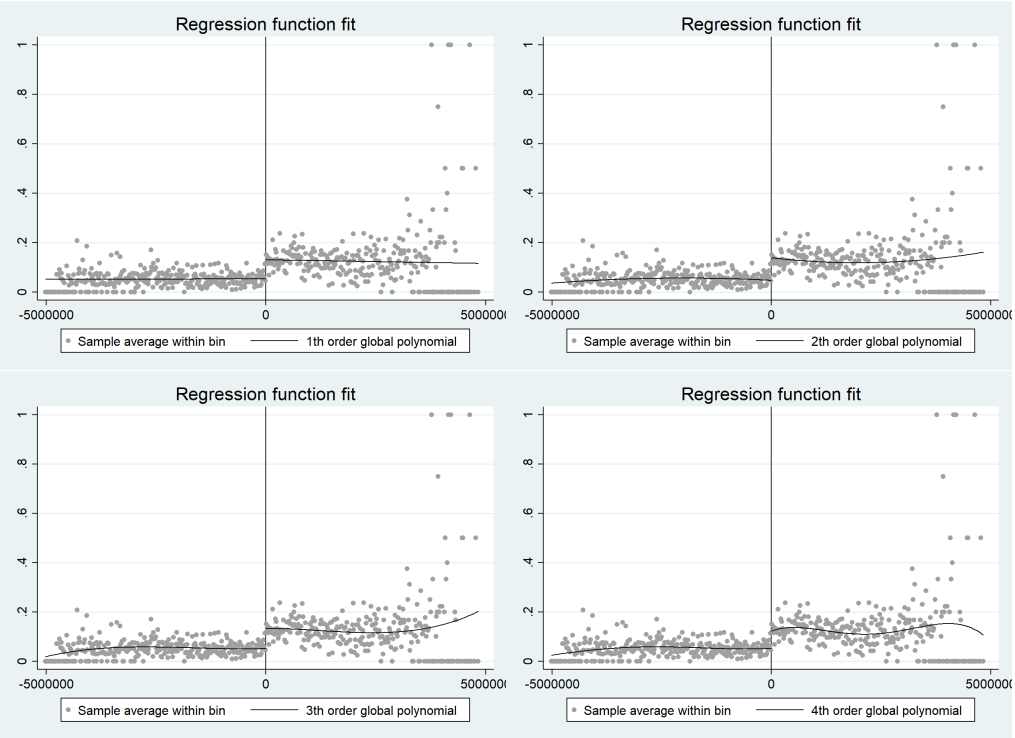
★★★★★ **Easy to read**
By Robert B. Costanzo on October 13, 2017
Color: White | Verified Purchase
Very simple to use. Easy to read. Best price. I bought another one.

Time 2

★★★★★ **Best small clock ever**
By mickey on October 19, 2017
Color: White | Verified Purchase
Perfect size. White face makes it easy to see. Keeps accurate time. Can use it anywhere as there is no cord to worry about. Perfect clock for stationary use, like a desk, or for a travel clock. Highly recommend.

★★★★★ **Easy to read**
By Robert B. Costanzo on October 13, 2017
Color: White | Verified Purchase
Very simple to use. Easy to read. Best price. I bought another one.

Figure 7: Regression Discontinuity In Time



Note: This figure plots the average conversion rate in each time bin. The vertical line denotes the cutoff time when a new review with positive aesthetics information is posted. The products included in these plots experienced only one review change during the sample period.

Figure 8: Illustration of Multiple Treatments

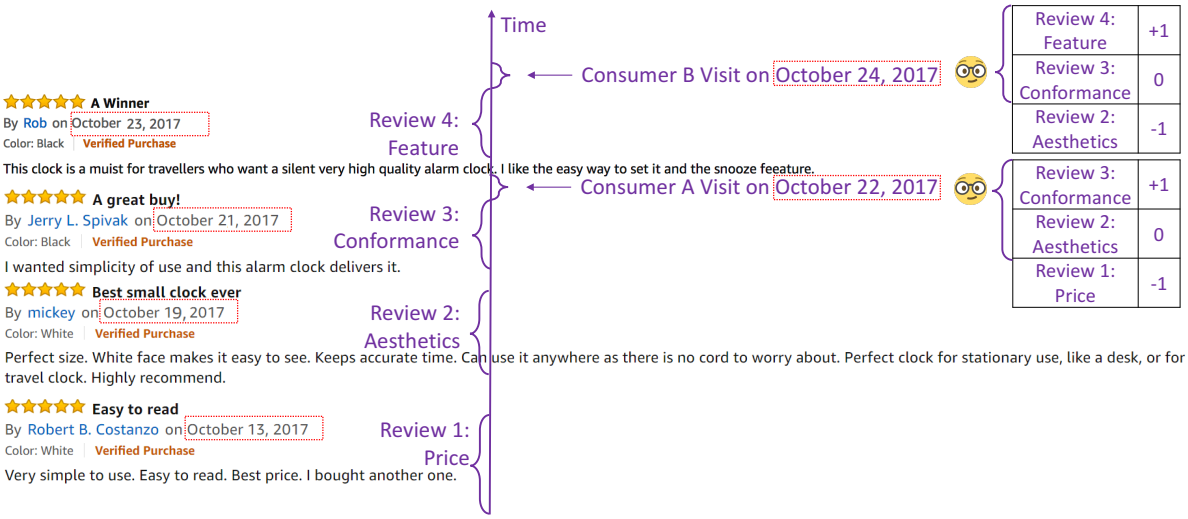


Figure 9: Salient Sentences for Six Dimensions of Information in Reviews

	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, it 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.
Brand	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 I,phone 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.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.

Figure 10: Salient Sentences for Six Dimensions of Information in Reviews: Floorcare

	Positive	Negative
Aesthetics	I found this product dealt with animal hairs extremely efficiently - I can definitely confirm it has made a great difference to the carpets, they are much brighter and cleaner looking.	Excellent product but one day after carpet hav been washed, the rooms have negative smell for few days.
Conformance	Arrived as promised, exceptionally well packed. The steam mop even outdoes it's TV commercial and is exactly what I needed for both hard floors and carpeted areas	Initially delighted with this machine. It is easy to put together & easy to use. The tanks click into place easily. It successfully removes ingrained dust from carpets. They are fresher & brighter after use. It gets out dust that even a Dyson cannot reach. But it will not remove stains that you could easily scrub out and it is not suitable for homes with pets. The instructions say hoover well before use which we did but the brush heads still get entangled and eventually stuck with pet hairs.
Durability	We bought it about 6 months ago and never got round to writing a review. I am about to buy another for my mum and dad and still think this looks the best one for the price. I am sure some of the more expensive ones are fab but I don't think you can go wrong with this for the money. I have a fairly big dining room and kitchen that are both tiled. We have 2 dogs who seem to shed a whole coat overnight, so I use this al the timet. It's quick, has great suction and I really can't fault it.	Was very happy with it the day it arrived, was really powerful, picked up everything! Now 3-4 weeks later it has no power at all and picks up nothing so taking it back tomorrow. Not happy!
Feature	This product is well worth getting.Gets all grime off wooden floors, and makes the room look spotless.	Wouldnt recommend this as a vacuum for all around the house, does not pick up great and useless picking up pet hair, also the tubes come apart when using so only best to use as an occasional or a back up vacuum.
Brand	Great vac for the money, fairly lightweight and performs as well as my previous (Dyson), much cheaper too. I would definitely recommend it.	I expect a good product with a brand name like Hoover. The lead is so short it continually pulls out past the red warning tape. The soft pipe is attached to the body by an angled rigid inset. This means that when something gets stuck in this bend there is no way to extract it unless you break the pipe off. This is an appalling design fault, when you vacuum occasionally something unintended will go up the pipe and need to be removed. Really basic avoidable problem. 0/10.
Price	When I first got this vac I was very happy with it, it was a good price and had really impressive suction.Unfortunately now I find that you cannot get a replacement filter :(	Great Hoover and works well only two things it let it down for me was the head it came with and the tools are on the tube should have left them on machine apart from that does the job and got it on sale so won't complain as much.lol

Figure 11: Salient Sentences for Six Dimensions of Information in Reviews: Television

	Positive	Negative
Aesthetics	Easy to setupInternal browser.Cheap price.Great connectivity and shape.Excellent picture.Definitely recommend it.	Sounds awful. you get a vibration on speech and some music has a quarter to half volume. I'm taking mine back. The vibrations are giving me headaches :(Thought i could live with it but its really really bad
Conformance	Product was as expected stylish, easy to install, ideal for room, plus please keep the standard up staff and service are excellent and professional at home base store Newmarket.	Quality of Pic very good. It says in the Spec that their is a USB PORT this is not true. I was disappointed.
Durability	Bought this tv a few weeks ago and i'm delighted with it. The picture quality is excellent, its not too heavy at all and was easy to set up.I know the previous reviews have a mixed feeling towards the sound quality, the sound is good to my ears but i understand if you've had a tv with better speakers before then you will notice the difference.Sky and Playstation 4 work perfectly with the HDMI.Overall a great TV for the price, wish i had saved up more for the 50inch one	unit failed within 24 hours of purchase. Technical response was appallingly bad and despite being advised that the problem was'probably' the remote control and that we should return it - this was from UMC UK - and it would be replaced, 14 days later it has not been replaced. Eventually, Argos direct, a woman called Lyndsey, sorted us out for a return/refund which went through without a problem.
Feature	Love the touch on/off button. Sick of the power buttons being at the side or back of other TV's. First LG and really happy.	brilliant picture the only downsides are the built in speakers are tinny, the remote control could be more responsive and internet use is a bit slow - connect it to external speakers though and it's well worth the money
Brand	Love the touch on/off button. Sick of the power buttons being at the side or back of other TV's. First LG and really happy. download but it was easy to follow. The photos that I have taken so far are very good so overall a very good choice.	we wanted a larger screen than our 40\ LG Led TV that we already have,so went for this one. you need a sound bar as the quality of the sound through the TV is below average. picture quality is also below par with our other LG"
Price	Easy to setupInternal browser.Cheap price.Great connectivity and shape.Excellent picture.Definitely recommend it.	I works we'll but there's a surprise when you open the box! You'll notice a port for connection of a power supply,but its not in the box! If yo want it / need it you'll have to pay extra.Now OK inmost instances you probably won't eed one. But if like me you wanted to connect a tablet which won't supply its own power to run the HDMI switch you need the additional power supply. I thought this a bit cheeky as this switch is really expensive already.

Figure 12: Salient Sentences for Six Dimensions of Information in Reviews: Curtain

	Positive	Negative
Aesthetics	Iv bought these in the past before - they are very thin - no lining at all - but they are the perfect colour for my kids rooms.Buy with black out blinds :)	I wouldn't say this is bubblegum pink - it is more like a lilac colour. In my daughter's pink bedroom these do not look pink. That is my only criticism - otherwise this is an ok product.
Conformance	Bought a pair similar to these in purple. They were easy to hang and looked quite elegant so I decided to buy these black ones. Just like the purple ones, they hang nicely and look elegant. They do actually look like silk which is a bonus so I am very happy with my purchase.	This blind does not fit into the window which it is designed to .my window is the right measurement so the blind indicated those measurement that I why I decided to buy it but was so disappointed when it was over 2.5cm to small
Durability	Bright and colourful iv washed them twice now and the colours are still purfect.... Really brightens up our sons room :)	This blind is very cheap and you get what you pay for.Actual blind okay but inner tube was made from cardboard so unsuitable for our bathroom as I felt it would deteriorate. Also the colour was quite dark and imposing in a small ish room.Ended up taking it back for a refund and ordered a made to measure blind that was much sturdier and had a metal inner tube.
Feature	Bought this blind to replace my old bathroom one. Didn't have to mess about cutting it to size as can fit on outside of window. Easy to fit, use and splash proof which is great for a bathroom. Would definitely buy again	These should be lined or at least thicker material, these are so thin you can see through them. If I didn't loose the receipt I would have returned these. Very poor quality.
Brand	Bought a pair similar to these in purple. They were easy to hang and looked quite elegant so I decided to buy these black ones. Just like the purple ones, they hang nicely and look elegant. They do actually look like silk which is a bonus so I am very happy with my purchase.	I bought this blind after reading reviews whilst the blind does look nice I was a little disappointed as it was an inch shorter than the size stated therefore had to return.
Price	This is just what you expect, I bought 2 the colour matches with no shade differences and I was impressed with the colour range. Just what I want plain cheap and inexpensive.	This blind is very cheap and you get what you pay for.Actual blind okay but inner tube was made from cardboard so unsuitable for our bathroom as I felt it would deteriorate. Also the colour was quite dark and imposing in a small ish room.Ended up taking it back for a refund and ordered a made to measure blind that was much sturdier and had a metal inner tube.

Figure 13: Review Information Effect by Product Category on Mobile Devices

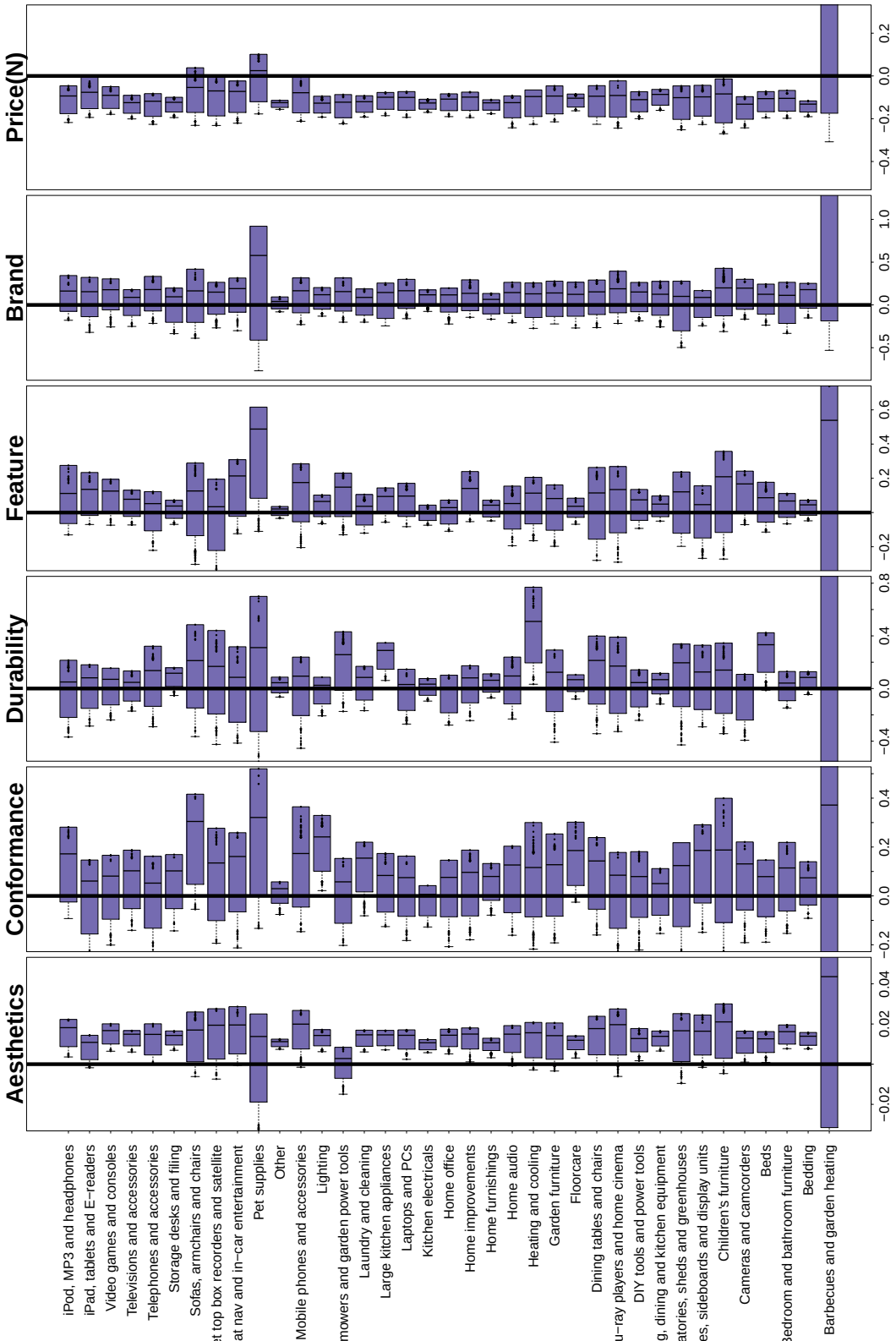


Figure 14: Review Information Effect by Product Category on PC

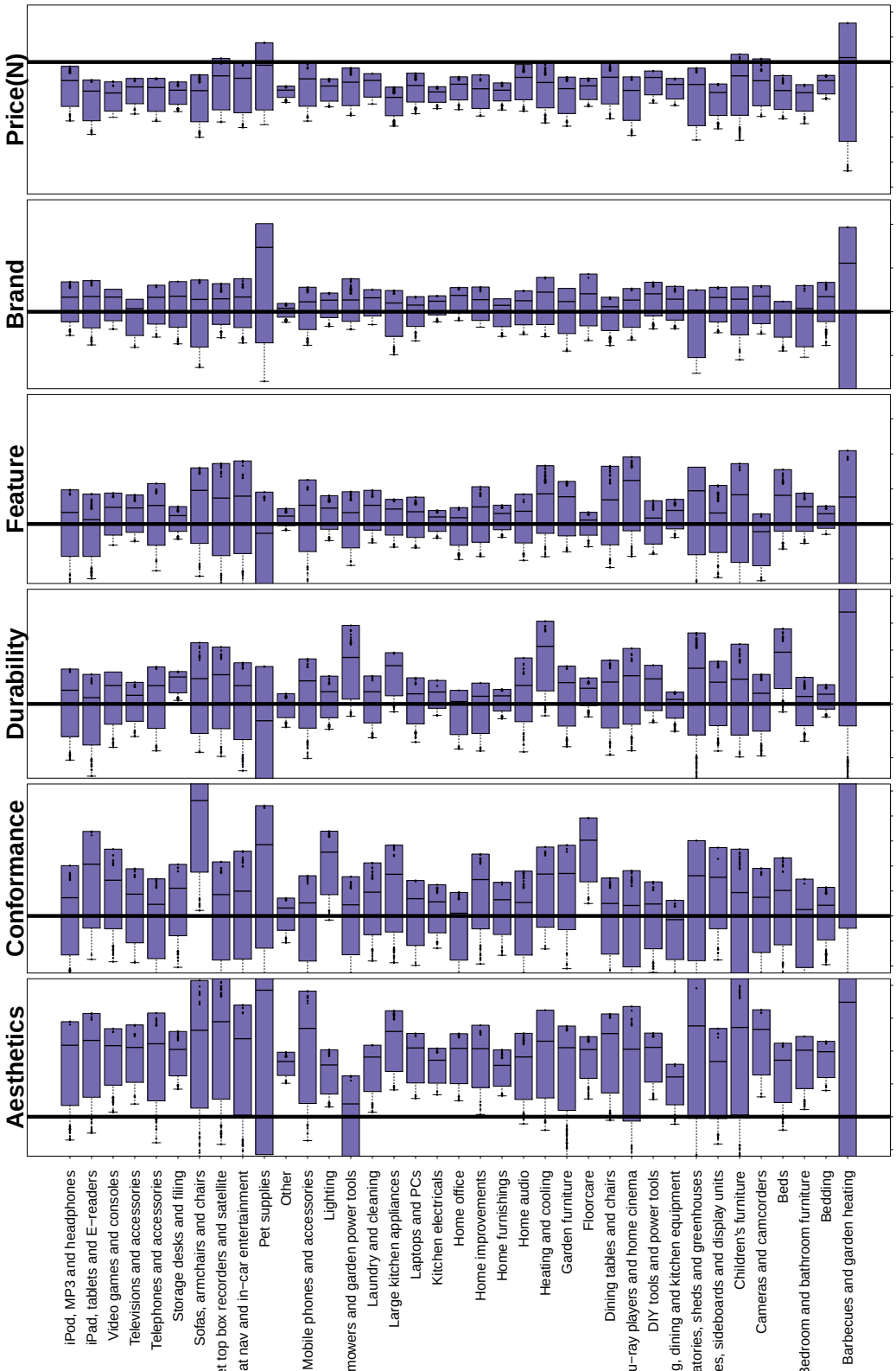


Figure 15: Effect of Content by Mean and Variance of Review Rating

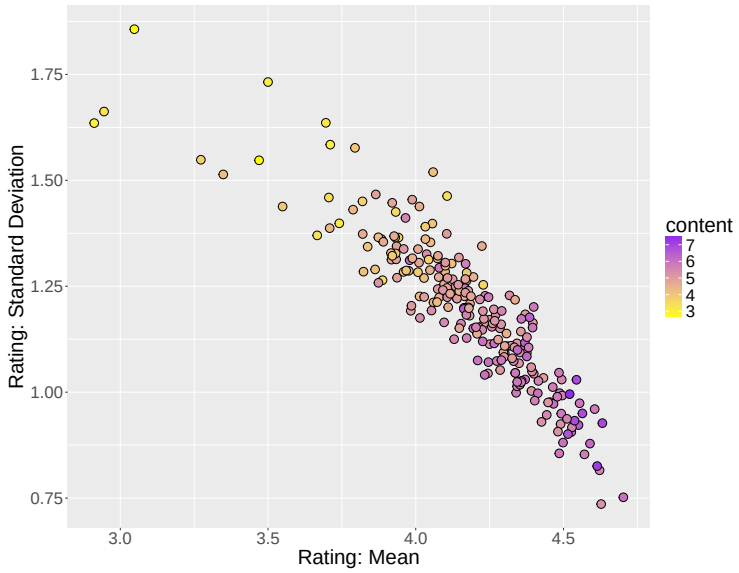


Figure 16: Effect of Content by Concentration

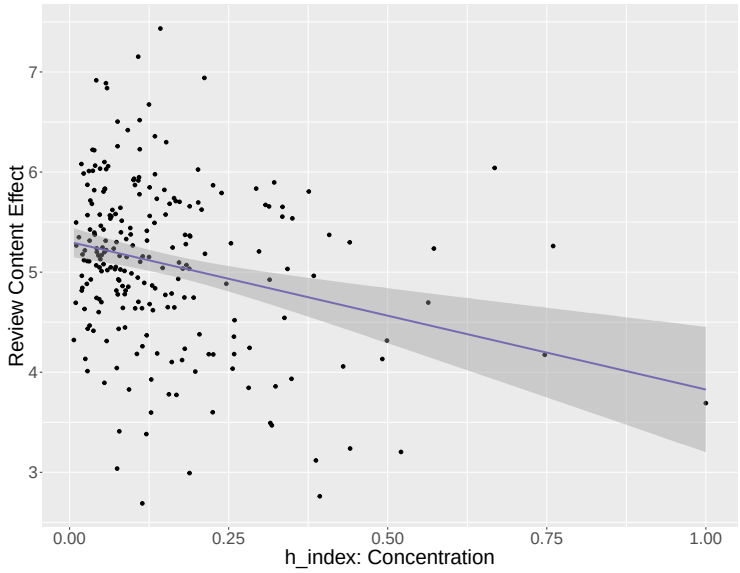


Figure 17: Histogram of Product Category Launch Time

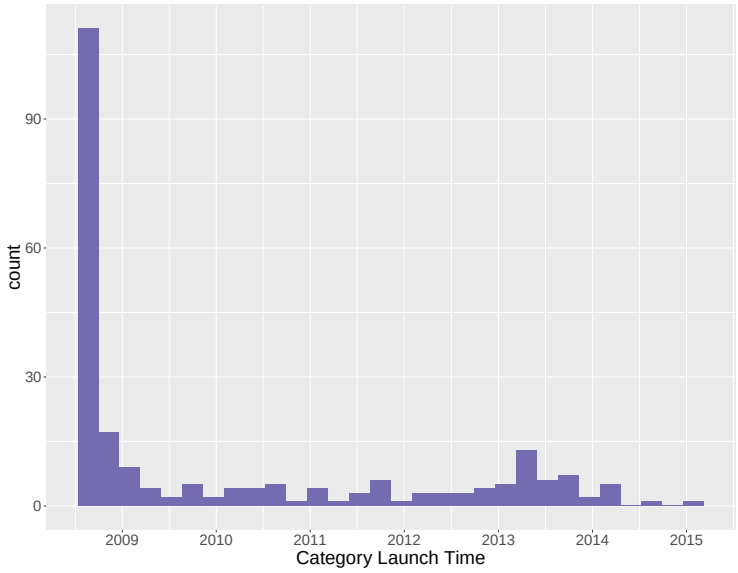


Figure 18: Effect of Content by Number of Days on Site

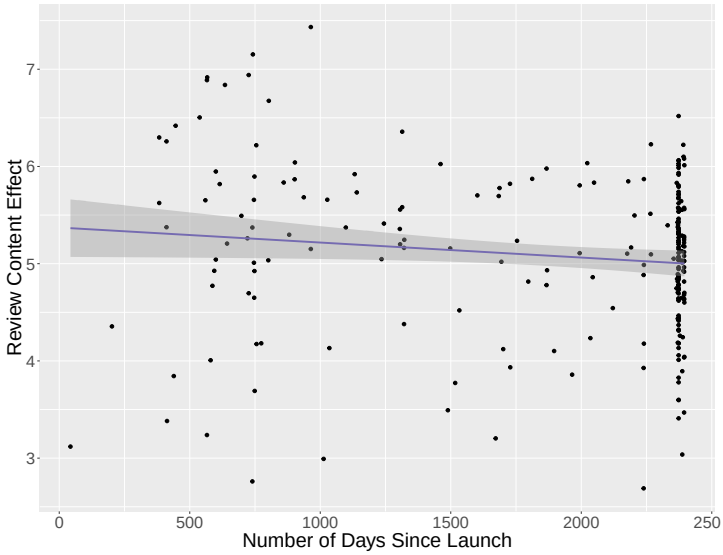


Figure 19: Effect of Content by Percentage of Branded Products

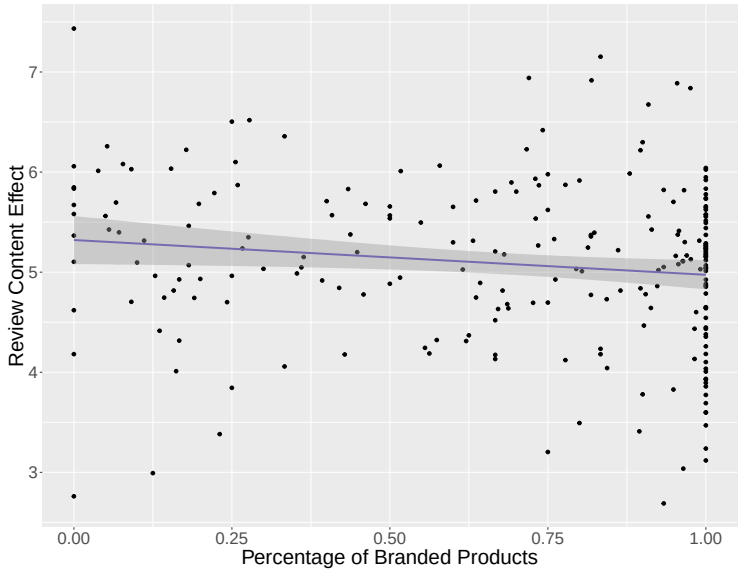
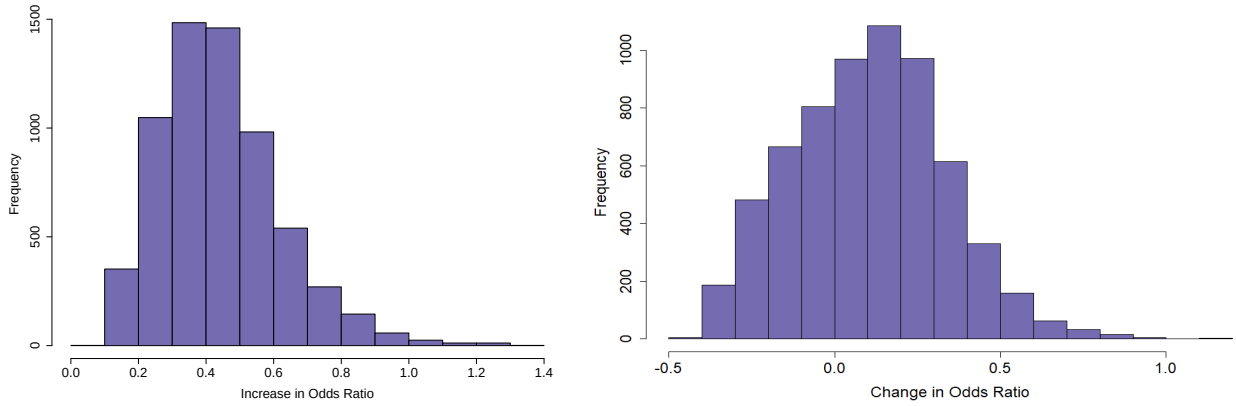


Figure 20: Counterfactuals

(a) Histogram of Increase in Odds Ratio by Reordering Reviews (b) Histogram of Increase in Odds Ratio by Diversifying Content



Note: The products included in (a) experienced only one new review with positive aesthetics information.

Appendix

A Missing Data Concern

We have access to data from only one online retailer instead of the user-centric data across all e-commerce sites. It is possible that consumers gather information and undertake transactions at many other sources during their purchase journey. We believe several facts could mitigate the concern that the competition among different retailers is missing in our analysis. First, this online retailer is a leading e-commerce site in the UK with a very high market share.[25] It is well known for its large-volume consumer reviews. The users in our sample all have a loyalty membership account with the company, so it is less likely for these loyal customers to conduct comparison-shopping across sites. Second, the ultimate goal of this paper is to identify the impact of review content on sales. We use the regression discontinuity in time (RDIT) identification strategy. This identification strategy implies that even if consumers shop or read reviews on multiple e-commerce sites, as long as the consumers' behaviors outside of the focal site are not systematically different before and after a new review is posted to the focal website, the estimated impact of review content on conversion will be unbiased. Because we cannot think of any strong argument for the systematic difference, we believe that our identification strategy is immune to the competitive effects. Last, although Section 3 may provide an incomplete profile of consumers' entire online decision journeys, this section only provides descriptive analysis of the data. The missing data problem will not affect our estimate of the impact of review content on conversion because we eliminate from the regressions the journeys in which consumers do not read reviews.

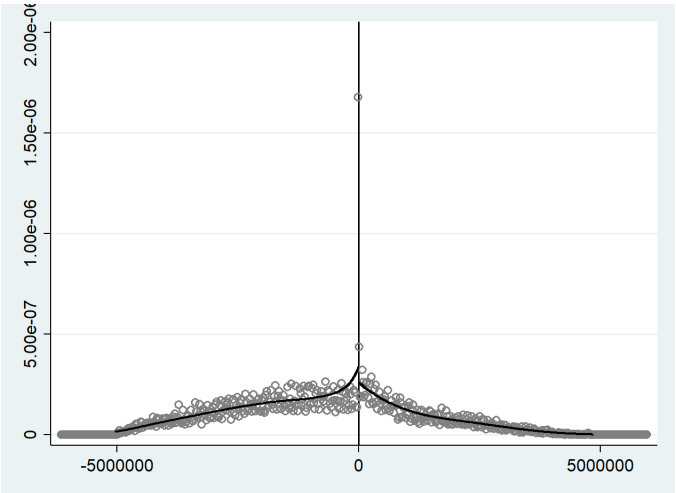
B Robustness Checks for the Regression Discontinuity in Time Design

Because we have multiple treatment variables for all the content dimensions, below we provide only one example for the positive aesthetics content dimension. Robustness checks for other content dimensions are available from the authors upon request.

We first assess the possibility of manipulation of the assignment variable, new review post time, by showing its distribution in Figure 21. The underlying assumption that generates the local random assignment result is that each consumer has imprecise control over the assignment variable. We can test this by checking whether the aggregate distribution of the assignment variable is discontinuous. Figure 21 shows no evidence of discontinuity at the cutoff point 0.

[25] Due to the non-disclosure agreement, we cannot reveal the identity of the company or its major financial statistics.

Figure 21: Density of the Assignment Variable: Time



Moreover, we plot a parallel RD estimated on control variables to demonstrate continuity. In Figure 22, we create the regression discontinuity plot for one of the covariates, price. In contrast to the discontinuous jump for conversion rate in Figure 7, price is continuously distributed before and after the cutoff point. This confirms the validity of the assumption that there is no precise manipulation or sorting of the assignment variable.

Figure 22: Parallel Regression Discontinuity: Price by Time

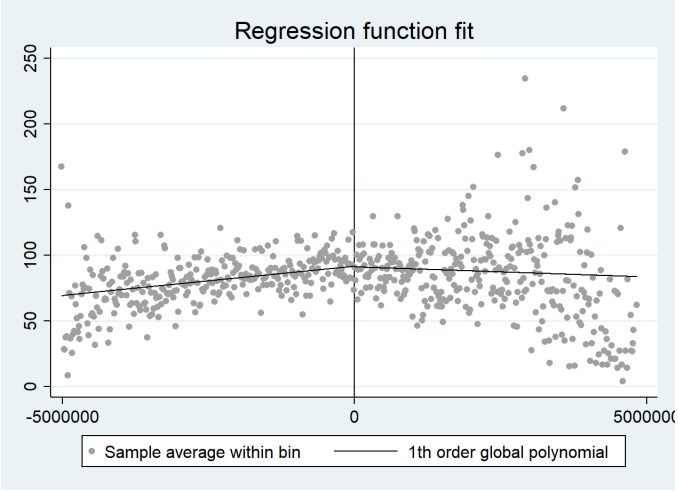
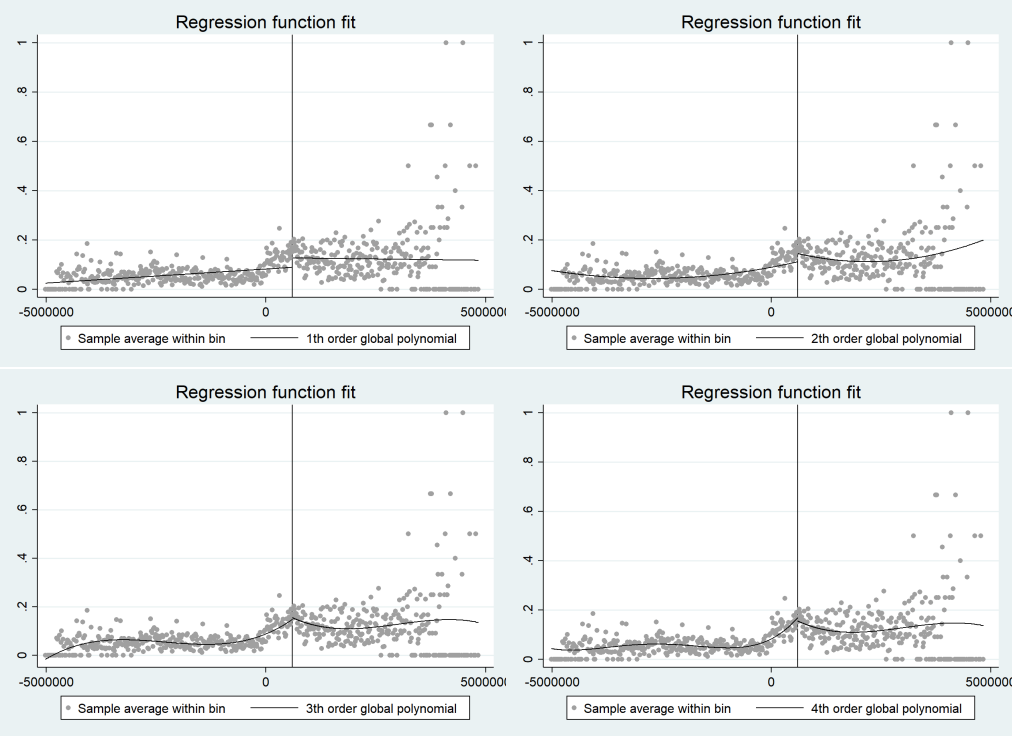


Figure 23: Placebo Test



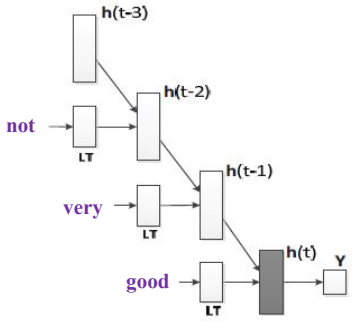
We also conduct a placebo test by estimating a parallel RD on a different date other than the new review (with positive aesthetics information) post date. The idea of the placebo test is that if RDiT does not work, we should observe non-zero and statistically significant jumps at other discretionary cutoff points. We would then conclude that there is something wrong with RDiT. In Figure 23, we present the regression discontinuity plots with polynomial orders 1 to 4 when the cutoff point is set at 600,000 seconds after the cutoff time stamp. The plots suggest that there is no discontinuity in conversion rate at a different cutoff point other than zero, which reassures the validity of our RDiT design.

C Three Deep Learning Algorithms

For each example, consider a phrase that may appear in a review text: "but not very good."

Recurrent Neural Networks – Long Short-Term Memory (LSTM)

Figure 24: 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.

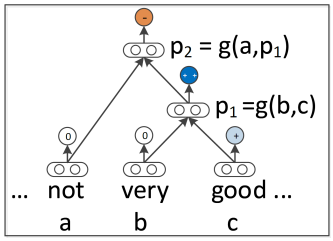
The first deep learning algorithm we implement is long short-term memory recurrent neural networks (Wang et al., 2015), which works by taking word or character *sequences as inputs,* and can simulate interactions of words in the sentence compositional process. The main idea in this deep learning algorithm is that the algorithm, in a sense, has a memory of words that came before a current word and thus does better than a simple bag-of-words model that ignores word positions and sequences. As shown in Figure 24, in this algorithm, to characterize sentence sequence, each word is mapped to a vector through a lookup-table (LT) layer, which adds value by staying dynamically tunable based on processed data. For each hidden layer, its input comes from two sources: One is the current word (taken and vectorized through lookup-table layer activations), and the other is the hidden layer's activation one step back in time, which incorporates information from previous phrases (i.e., previous phrase memory). The last hidden layer is considered as the representation of the whole sentence. The example in Figure 24 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 algorithm excels in distinguishing negation because it tunes vector representations of sentiment words into valence-polarity-representable ones. Therefore, it shows promising potential dealing with complex sentiment phrases.

Recursive Neural Networks

The second deep learning algorithm is recursive neural networks (Socher et al., 2013). Instead of focusing on sequences as in the recurrent neural networks, the recursive neural networks algorithm focuses on a more complicated sentence *parse-tree* structure that is aware of sentence syntactic context. Intuitively, this algorithm improves on the bag-of-words approach by acknowledging that sentences consist of several syntactic phrases which may vary in sentiment and, when put together to compose a sentence, naturally evolve sentence-level sentiment. This algorithm works to label the sentiment of sentences by labeling the sentiments for each separable syntactic phrase and combining them via recursive neural networks. As shown in Figure 25, in this algorithm, one needs to compute parent phrase representation-vectors in a bottom-up fashion. At the bottom level, the word "not" is classified as neutral (denoted by 0, in white), "very" is classified as neutral, and "good" is classified as (somewhat) positive (denoted by +, in blue). In the middle layer, the phrase "very good" is classified as very positive (denoted by ++, in dark blue). And in the top layer, the entire phrase "not very good" is classified as negative (denoted by -, in orange). From the

model perspective, the classifier for the parent node p_1, or phrase "very good," uses a specific and clever compositional function g and node vectors b and c as features. Similarly, the classifier for the top parent node p_2 uses the same composition function g and node vectors a and p_1 as features. Given this unique composition process, this method can accurately capture the sentiment change (from positive to negative and vice versa) and scope of negation (somewhat negative or very negative). This algorithm can also discern that the sentiment of phrases following the contrastive conjunction "but" in "but not very good" dominates and may be more informative for sentence-level sentiment.

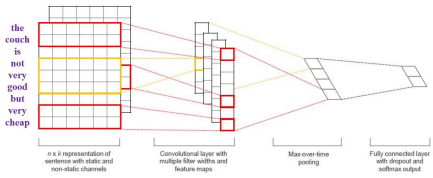
Figure 25: Recursive Neural Networks



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.

Convolutional Neural Networks

Figure 26: Convolutional Neural Networks



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

The recursive neural networks algorithm is very powerful, but it requires a parse tree, which is not available in many settings. A parse tree represents the syntactic structure in a sentence using a tree model. For more details, see https://en.wikipedia.org/wiki/Parse_tree. The last algorithm, convolutional neural networks (Kim, 2014, Figure 26), has a data-driven structure that does not rely on externally provided parse trees. This algorithm is similar to the one we presented in Section 4.3. The key difference is that the outcome variable here is the content dimensions or the associated sentiments instead of conversion.

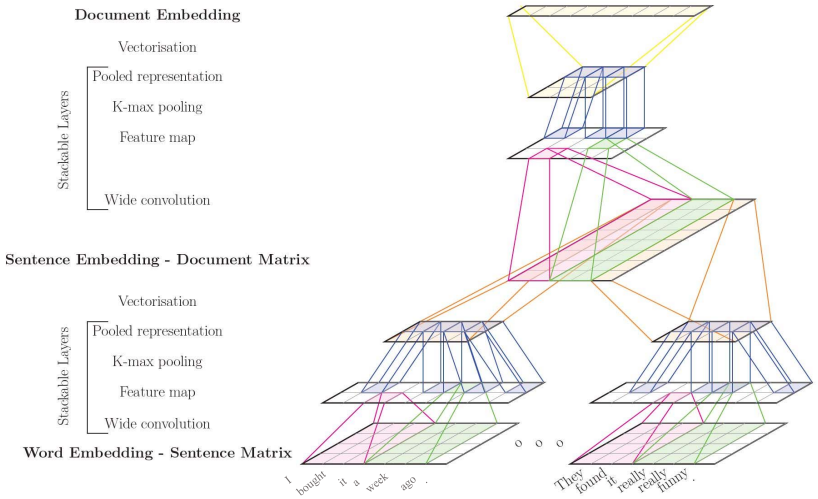
D Salience Extraction Method

We implement a method created by Denil et al. [2014] that adapts visualization techniques from computer vision to automatically extract relevant sentences from labeled text data. In essence, it is a CNN that has a hierarchical structure divided into a sentence level and a document level.

At the sentence level, the algorithm transforms embeddings for the words in each sentence into an embedding for the entire sentence. At the document level, another CNN transforms sentence embeddings from the

first level into a single embedding vector that represents the entire document. Figure 27 is a schematic of the algorithm. Specifically, at the bottom layer, word embeddings 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, "They 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 this algorithm is trained, it can then be used to extract salient sentences by identifying sentence locations that have the highest amount of influence to loss function. 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. The following step is to rank sentences based on the saliency score and highlight the sentences with the highest score.

Figure 27: 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.

E Summary Statistics of Variables in the Regressions

Our data consist of seven tables: four user behavior tables and three domain tables. The four user behavior tables are pageview, review impression, usedfeature[26], and transaction, as described in Section 3. The three domain tables store information of product reviews, questions, and answers. Here is the description of each table.

[26]Referred to as used-interaction in Section 3.

Table 13: Data Structure

Table Name	Description
pageview	a single product or category page view for a customer
impression	a single exposure to a product review
usedfeature	a single user engagement with a product review
transaction	a single product purchase made by a customer
review	a single review of a product
question	a single question related to a product
answer	a single answer related to a product

All the seven tables are used to construct the consumer decision journey. The four behavioral tables are linked to the journey using the key variable "userid." The three domain tables are linked to the journey using the "productid" variable. After constructing the journeys, we can calculate the variables used in the regression equation (4) for each journey.

F Survey Instrument to Content-Code Review Content

Figure 28: **Survey Instrument Shown to Amazon Mechanical Turkers for Identifying Review Contents**

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:
a. "This overpriced junk broke after using twice!"
b. "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:
a. "The screen size is quite small at 3.5 inches"
b. "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:
a. "These earbuds broke after 3 months of regular usage"
b. "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:
a. "The product does exactly what they says it would do...hydrating my dry skin."
b. "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:
a. "The lamp's sleek appearance is pleasing and I got many complements."
b. "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:
a. "Have been using HP ink for 5 yrs and think it's the best on the market!"
b. "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.

G 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 were recorded. We obtained the final classification by a majority-voting rule.

2. We restricted the quality of Turkers included in our study to only those with at least 100 reported completed tasks and 97% or better reported task-approval rates.

3. We used only Turkers from the US 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 include only pages located in the US).

4. We created a sample test, and only those who passed this test, in addition to possessing the 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 until almost all queried Turkers stated that no questions were confusing.

6. To filter out participants who were not paying attention, we included an attention question that asked the Turkers to click a certain input. Responses from Turkers who failed the verification test were 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 tagged more than 20 messages, a couple of tagged samples were 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 managed to pass time-filtering requirements. We concluded that these were automated programs. These results were dropped, and the Turkers were "hard blocked" from the survey, via the blocking option provided in AMT.

Figure 29: **Cronbach's Alphas for 5,000 Reviews**

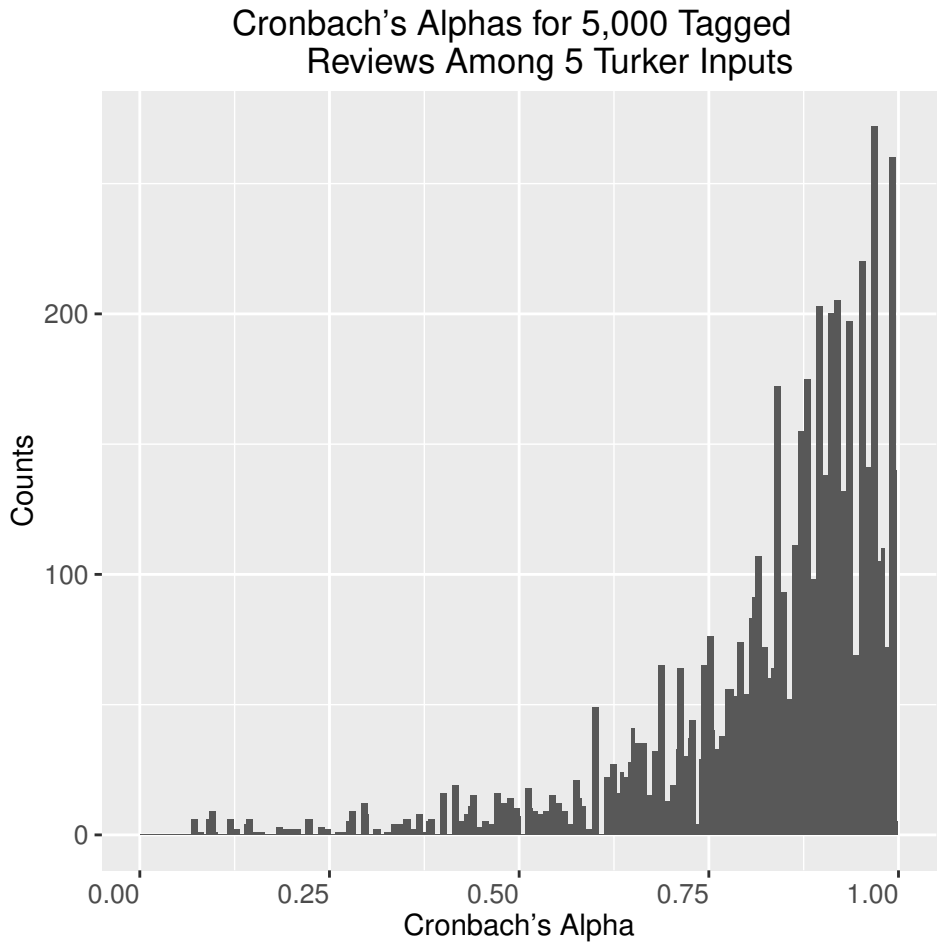


Figure 29 presents the histogram of Cronbach's alphas, a commonly used inter-rater reliability measure, obtained for 5,000 reviews. The average Cronbach's alpha for our tagged reviews is 0.84 (median 0.88), well above typically acceptable thresholds of 0.7. About 84% of the reviews obtained an alpha 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 were qualitatively similar.

H Model Comparison for Information Detection

Table 14: Model Comparison: Information Detection

Classifier/Accuracy %		SVM + BoW	NB+ BoW	Recurrent-LSTM	Convolutional
Mean	Precision	0.358	0.484	0.496	0.955
	Recall	0.098	0.563	0.483	0.946
	F1	0.151	0.472	0.475	0.950
Aesthetics	Precision	0.304	0.531	0.492	0.917
	Recall	0.057	0.623	0.516	0.863
	F1	0.097	0.574	0.504	0.889
Conformance	Precision	0.211	0.134	0.235	0.971
	Recall	0.061	0.455	0.162	0.992
	F1	0.094	0.207	0.192	0.981
Durability	Precision	0.350	0.333	0.373	0.910
	Recall	0.064	0.550	0.648	0.921
	F1	0.109	0.415	0.473	0.916
Feature	Precision	0.922	0.966	0.921	0.999
	Recall	0.327	0.570	0.791	0.951
	F1	0.482	0.717	0.851	0.974
Brand	Precision	0.059	0.082	0.094	0.983
	Recall	0.026	0.474	0.176	1.000
	F1	0.036	0.140	0.122	0.992
Price	Precision	0.304	0.858	0.861	0.948
	Recall	0.051	0.708	0.604	0.952
	F1	0.088	0.776	0.710	0.950

Note that Socher et al. [2013] is suitable only for sentiment analysis, so we could not perform information detection for the recursive neural networks model.

I Hierarchical Bayes Model

Let the total number of product categories be C and the total number of coefficients be NIV. If we concatenate the coefficients in equation (6) to a $C \times NIV$ matrix B where each row is a vector of coefficients associated with a particular category, i.e., $B_c = \left[\overrightarrow{\tau_{kc}}, \delta_{nkc}, \overrightarrow{\theta_{kc}}, \overrightarrow{\gamma_{kc}}, \xi_c, Weekend_{tc}, Daytime_{tc}\right]$, then the multivariate regression takes the form

$$\underbrace{B}_{C\times NIV} = \underbrace{\Delta}_{C\times NIV} + U. \tag{7}$$
$$where\, U \sim N\left(0, V_\beta\right)$$

The $C \times NIV$ matrix Δ contains the mean value of the NIV coefficients for each category. The error term U is assumed to be distributed normally with mean zero and covariance matrix V_β. To complete the model formulation, we set the following priors for Δ and V_β:

$$vec\left(\Delta | V_\beta\right) \sim N\left(vec\left(\overline{\Delta}\right), V_\beta \otimes A^{-1}\right)$$
$$V_\beta \sim IW\left(v, V\right).$$

They are the natural conjugate priors for the multivariate regression in equation (7).

J Endogenous Review-Reading Behaviors

In reality, consumers endogenously decide which reviews to read and how many reviews to read. We do not model this endogenous process explicitly. However, the endogenous review-reading behavior will not affect the counterfactual result for the following reasons.

First of all, regarding which reviews to read, our assumption is that consumers use a top down fashion, i.e., reading reviews from the top of the page to the bottom of the page. This assumption is supported by previous eye tracking studies. The retailer in our study posts reviews in the reverse chronological order. Therefore, consumers read the most recent reviews before the old reviews. Importantly, consumers can also engage in "used-interaction;" i.e., this online retailer allows users to filter reviews by star ratings. For example, a consumer can filter to look at only one star reviews. When doing so, all the one star reviews are displayed to the consumer, also in reverse chronological order. Because the order is always reverse chronological, no matter whether a consumer filters or not, our identification is immune to consumers' filtering/sorting behavior. In other words, if a consumer in our data used the filter to select only one star reviews, in the counterfactual, we also let this consumer read only one star reviews. Even though the star rating remains unchanged, we can change the content of reviews read by consumers in the counterfactual. In the observed data, a consumer may read only a one star review about "conformance." In the counterfactual, we re-order the reviews by content, so this consumer will read a one star review about "aesthetics" instead. Because the counterfactual does not alter consumers' review-reading behavior with respect to "which reviews to read," this problem is eliminated.

Second, regarding how many reviews to read, we provide additional evidence that, in our data, the content of the first review is uncorrelated with the number of reviews read. In other words, we find that no matter whether a consumer read a review about "conformance" at the beginning or a review about "aesthetics" at the beginning, he will read the same number of reviews in both scenarios. We test this hypothesis-whether the number of reviews read is uncorrelated with the content of the first review, because this is what we did in the counterfactual. On page 28, we state that "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 first position in the set of reviews read by each consumer." We acknowledge that if we did other counterfactuals, we would probably need to more carefully model the endogenous review-reading behavior. Now we present the data evidence. In Table 15 below, we show the distribution of "the number of reviews read" by the six content dimensions: aesthetics, conformance, durability, feature, brand, and price. For the distribution statistics, we present number of observations (N, column 2), mean, standard deviation, and median of "the number of reviews read."

Table 15: Distribution of Number of Reviews Read by Content in the First Review Read

	N	Mean	Std	Median
Aesthetics	25105	11.1265485	7.74282639	10
Conformance	9275	11.1926685	7.35463469	10
Durability	18521	11.007559	7.46181403	10
Feature	66011	11.1631243	7.65780554	10
Brand	2844	11.4388186	7.58973413	10
Price	28746	11.2147081	7.57472158	10

The table shows that, the mean, standard deviation, and median of the "number of reviews read" are all very similar across scenarios. We also perform a two sample Kolmogorov–Smirnov test to see whether the samples are drawn from the same distribution under different scenarios. The table below shows the p values of the pair-wise Kolmogorov–Smirnov tests. As you can see, across the fifteen pairs, only two are

significantly different from each other. This confirms our hypothesis that the number of reviews read is uncorrelated with the content of the first review.

Table 16: P_values of the Pair-wise Kolmogorov–Smirnov Tests

	Aesthetics	Conformance	Durability	Feature	Brand	Price
Aesthetics	X	X	X	X	X	X
Conformance	0.0759	X	X	X	X	X
Durability	0.3313	0.0533	X	X	X	X
Feature	0.1450	0.4484	0.0334	X	X	X
Brand	0.3435	0.8766	0.0584	0.4784	X	X
Price	0.0076	0.9686	0.0030	0.4842	0.4170	X

Therefore, although our model does not explicitly account for the endogenous review-reading behavior, after carefully checking the data evidence, we find that this does not limit our capability to answer the counterfactual question: how would re-ordering the reviews affect sales conversion?

K Unit of Analysis

The unit of analysis in the regression is journey-product. We use only the type 4 and type 5 journeys in the regression. In fact, we only use a subset of the type 4 and type 5 journeys in which the focal products had a least one new review posted during the sample period. We have to narrow down to this subset for the purpose of identification. The intuition is that for a focal product, if no new review was posted during the sample period, then there is no variation in the review content read by consumers, so we cannot identify the effect of review content on sales. This subset leaves us with 58282 journeys. Importantly, our unit of analysis is not a journey, but a journey and product combination. In some journeys, the consumer read reviews of multiple products (an average of 2.7) before purchasing, so we looked at the impact of reviews for each product separately. This explains why we have 156,445 observations in the regression.

We now explain why we choose the unit of analysis at the journey-product level. Our assumption is that the reviews a consumer read affects the purchase likelihood only for the focal product but not for the other products considered in the journey. Specifically, if the review content for product A is favorable, then the likelihood of purchasing product A will increase. But the review content of product A won't affect the purchase likelihood of product B, which is also in the journey. This specification rules out the competition effect. As you can see from the conversion rate formula on page 17 ($ConversionRate_{ijkt} = (exp(u_{ijkt}))/(1+exp(u_{ijkt}))$), this is a binomial logit function instead of a multinomial logit function. We have two justifications for this binomial logit specification. First of all, in a robustness check, we also tested the multinomial specification; the signs of the coefficients are consistent with the current specification, but some coefficients became insignificant due to the smaller sample size. Note that the sample size of the multinomial specification is equal to the number of journeys, which is about 1/3 of the number of observations in the current specification. Essentially, allowing for competition dramatically reduces the number of observations in the regression, which reduces the power of the analysis. However, the insights we obtain from the two specifications are very consistent. Therefore, we make a tradeoff to use the binomial logit function. Second, the binomial logit function requires a much shorter computing time because we do not have to pool the reviews for all the products in the CNN model. The pooling makes the convolutional operator in the second layer of CNN much slower. So above all, due to the concerns for economic insights, statistical power, and computational burden, we choose to use journey-product as the unit of analysis instead of journey itself.