Abstract

As online reviews provide essential information to guide customers in their prospective purchases. As more such reviews accumulate overtime, one would suspect that complete information becomes available about product and almost no customers should be disappointed in their purchases. Yet, we provide empirical evidence over a large dataset of reviews that negative reviews seem to be arriving at an accelerated rate later on in the lifetime of a product. To better understand this inconsistency, we frame the problem at hand as an information revelation problems. Using a novel approach, we then segment each review as an aggregation of aspects for which the reviewer provides weights, in line with how much she values those aspects, and corresponding experiences vis-à-vis those aspects, which range from positive to negative. We show that this segmentation better explains the review process and better explains the polarity of reviews.

Keywords

Online reviews, information inconsistency, review dimensions.

Introduction

Worldwide retail ecommerce sales surpassed $2.29 trillion in 2017\textsuperscript{1}, posting a 23.2\% year-over-year increase. The US share of that amount was more than $460 billion.\textsuperscript{2} Online sales channels, however, have their unique, sometime limiting characteristics, including the lack of product touch and feel. This makes user reviews an essential component in the success of online channels. As such, these reviews, the information embedded in them and the online review system have been hailed as the holy grail of online channels. In fact, Jeff Bezos, CEO of Amazon Inc., has been recently quoted saying “We don’t sell products, we sell information”\textsuperscript{3} when asked about his company’s business model.

It becomes therefore clear that information including user reviews play a central role in online channels. In this study, we investigate the nature of information revelation in online Amazon reviews by focusing on the text of those reviews, in two ways:

- First, we explore the pattern of information revelation over time. Specifically, we ask:
  - Within reviews of a specific numerical rating (namely 1-star and 5-star), do the topics discussed change over time?

\textsuperscript{1} According to an eMarketer study in 2017.
\textsuperscript{2} Amazon accounts for more than a third of this total, with yearly net sales revenue of $136 billion in 2016, according to a 2016 Forrester Research Inc. study.
\textsuperscript{3} In a 2013 interview with Harvard Business Review’s IdeaCast (hr.org/ideacast/2013/01/jeff-bezos-on-leading-for-the.html).
Is the nature of the evolution of the discussion different across numerical ratings (namely for 1-star and 5-star)?

- Second, we model the user experience of different aspects (or features or attributes) of a product and ask:
  - Does the user experience of these aspects change over time?
  - How much does the user experience over these aspects determine the numerical rating?

The first question is interesting in that it helps us understand, for example, whether recent reviews contain different information than older reviews, and also provides insight into what is actually being communicated in the text of the review. For example, if the discussion “converges” over time, in the sense that after some time, reviews report identical experiences of some product aspects, then focusing on recent reviews may be sufficient, and the arrival of reviews does not reveal new information; instead it ratifies the experience of previous reviewers. Alternatively, if the discussion keeps changing over time, reliance on recent reviews may not be sufficient unless recent reviews also summarize the experience of early reviewers. One simple reason why the information content could change is that experience with the longevity of a product’s usefulness or its durability takes time to experience. Finally, the pattern of information revelation across numerical ratings leads us ask if, for example, negative information emerges later or takes more time to emerge. Addressing this question contributes to an emerging stream of research that investigates the dynamics on online reviews (Godes and Silva 2012; Moe and Trusov 2011; Yong 2006).

The second question is important to both potential consumers of the product, the producer of a product, as well as producers of competitive products. For a consumer, understanding what aspects were important to previous users and what their experience was helps the consumer evaluate whether the review is useful or not. For example, if a product was rated favorably mostly because of a given product aspect A, and aspect A is not important to the potential consumer, then the review is not useful. A concrete example is battery life versus screen resolution of a cell phone: if reviewer A rated it highly because the battery life was long, but consumer B is only concerned about screen resolution, then the review is not useful in answering that question. For a producer, the implication is clear: an understanding of what aspects matter and what contributes to overall ratings (and presumably sales). The same applies to a competitor: an understanding of what is contributing to favorable ratings and what advantages or disadvantages their product has in that regard. The relationship between user experience of different aspects or attributes and the numerical rating also helps us understand what contributes to high or low rating. Since most products have multiple attributes, it is of keen interest to understand how user experience with specific attributes influences the overall numerical rating. Addressing this question contributes to a well-established stream of research that models products ones having multi-valued attributes and models customers as experiencing those attributes and making corresponding assessments (Archak et al. 2011; Chen et al. 2017; Chunhua et al. 2015; Decker and Trusov 2010; Roberts and Urban 1988).

One step in studying this problem is to isolate the aspects of the product that are important to reviewers; these can, of course, vary by product. Rather than use keywords or other labels, we use topic modeling to isolate the aspects and to also model the discussion. This provides an automated way to identify what was important to the reviewers of a product. We then model the user experience by combining sentiment analysis with the topics. In doing so, our objective is to quantify whether the user experience was positive or negative for that aspect. We perform two tests: first, whether the topic modeling and user experience classifies reviews correctly, and second, how much the user experience of each aspect contributes to the overall rating. We then

In the next section, we provide a brief review of the relevant literature. We then describe our data and provide some information on the arrival process of reviews. After that, we present our model and approach followed by a preliminary analysis of the dynamics of topic evolution. We then present our main results, and summarize our conclusions.
Literature Review

While there appears to be consensus that individual reviews become more negative the longer the product is on the market (Godes and Silva 2012; Li and Hitt 2008), the literature provides varying rationales. The self-selection bias argument (Li and Hitt 2008) states that the more people value a product, the earlier they purchase it thus biasing up product ratings and consequently disappointing subsequent purchasers, leading to more negative reviews later on in the life of the product. The second argument, social influence (Moe and Trusov 2011), states that because reviewing is costly, requiring time and effort, subsequent reviewers worry that their reviews end up diluted by the large volume of positive reviews. They therefore become less motivated to post additional positive reviews for already highly rated products and end up posting rather critical reviews with lower ratings. The third argument (Godes and Silva 2012), building on the previous two, states that on one hand purchase errors may increase as more reviews arrive leading to lower ratings and, on the other hand, online reviewers seem to have become more negative over the years. In this research, we focus on the dynamics of the reviews textual content not the dynamics of the ordinal numbers.

Furthermore, there is a growing stream of research that breaks down reviews into information about distinct product attributes. Advances in natural language processing enabled the identification and extraction of attribute information from reviews (Archak et al. 2011; Decker and Trusov 2010). These developments have led to the finding that because product quality is often comprised of multiple dimensions, a multidimensional rating system is more suited to convey product information (Chen et al. 2017). We build on this literature to propose a novel approach to extract the product aspects, their weights and their experiences. This approach will be useful to study the prevalent assumption that that customers learn from reviews.

We use topic modeling to analyze the textual content of reviews, specifically Latent Dirichlet Allocation (LDA) (Blei et al. 2003). LDA assumes that each document follows a multinomial distribution over different topics (document-topic distribution) and that words in a document follow another multinomial distribution (topic-word distribution) with both distributions being estimated simultaneously. LDA has been applied to the text of online discussions and customer reviews to understand customer preferences and make strategic decisions about branding (Buschken and Allenby 2016; Puranam et al. 2017). McAuley and Leskovec (2013) use LDA to more accurately predict product ratings by harnessing the information present in review text and to identify useful and representative reviews.

Data and Review Arrival

We gathered data on 1,000 products listed on Amazon.com across four product categories as displayed in Table 1. The products tracked had variability along a variety of dimensions proposed in the literature: product involvement, product benefit, and frequency of purchase, as documented in the meta-analysis by Floyd et al. (2014) and the work by Gu et al. (2012); and search/evaluation efforts and risk reduction, as documented by (King et al. 2014). The products also vary in the number of reviews received, sales rank, price, and utility and hedonic qualities.

We collected information about the product as well as its reviews for two years between February 2014 and February 2016. The complete dataset consisted of a little over two million unique reviews posted on one thousand products.

We start by providing data on the pattern of arrival of positive and negative reviews over the lifetime of the product. Figure 1 shows the arrival rate of reviews for two representative products in our dataset of Amazon reviews. The X-axis shows the number of days from the time the first review was posted and the Y-axis shows the rating (1, 2, 3, 4, or 5). The bold black line depicts the cumulative average rating over time starting from the day the first review was posted. Each time a review is posted to the product we depict this review with a colored bubble; a 5 star review is depicted in dark blue on the 5 line of the Y-axis, a 4 star review light blue, a 3 star review in green, a 2 star review in pink and a 1 star review in red. The size of the “bubble” reflects the number of reviews for that given day. For both products, it is apparent that the 5 star ratings and the 1 star ratings, as well as the others, continue to be posted to the product at a steady rate long after the first review.
The sustained arrival of positive reviews (mainly 5 star and 4 star) is an expected finding, indicating that the review system is functioning as designed to ensure that prospective customers are well-informed and end up happy with their purchases. However, the sustained arrival of negative reviews (mainly 1 star) seems to contradict those conclusions. We therefore proceed to further validate the sustained arrival of extreme ratings, namely 5 star and 1 star. Figure 2 (left) shows the rate at which reviews arrive for all the products in our dataset that received a 1* rating. Each line corresponds to a product. The X-axis divides the calendar time between the first review and the last review for a product into 10 deciles. The Y-axis shows the cumulative proportion of reviews. If reviews were to arrive at a constant rate over time, the plotted line for a given product should look like the 45-degree line, shown in black. If most reviews arrived soon after the first review, the plot of the arrival rate would be a concave line. If most reviews came much after the first review, the plot of the arrival rate would be a convex line. Figure 2 (left) shows that for an overwhelming majority of products, the line has a convex shape, and so the rate of arrival of 1 star reviews increases over time. The pattern for 5-star reviews is shown in Figure 2 (right) and is essentially identical to 1 star reviews. This provides a conclusive evidence across all products in our dataset of the sustained arrival of extreme ratings over the lifetime of the products.
Dynamics of Information Revelation in Online Reviews

The persistent arrival of reviews, as well as the increasing rate of arrival over time, is in fact what prompted us to investigate the information content of reviews. This led us to analyze the dynamics of information revelation through topic modeling. We conduct key preprocessing steps on the Amazon review documents to build up the corpus for the LDA topic modeling. These include tokenization, stop word removal, stemming, and tf-idf.

**Topic Modeling:** We use LDA (Blei et al. 2003) to extract topics from the Amazon reviews. LDA is a generative probabilistic model. We are able to extract two matrices from the LDA model: the coverage of the topics in each document, and the word distribution for each topic. We view the comments in each day as one document and apply the variational expectation-maximization algorithm to obtain both matrices. Suppose we examine two days review comments on one mobile phone. In day 1 the reviewers are discussing the battery life of the mobile phone and in day 2 the reviewers are commenting on its screen resolution. Using LDA model we are able extract two topics battery and screen. There is a higher probability we see the words such as “battery”, “life”, “last” in the battery topic and the words such as “screen”, “size”, “resolution” in the screen topic. The coverage of the topic battery in day 1 comments is higher while the coverage of the topic screen is higher in day 2 comments. By doing so we can observe how topics are changing over time. The number of topics is the most critical parameter in the LDA model. To determine the number of topics, coherence of topics was used (Röder et al. 2015). After comparing the coherence value with the topic number from 5 to 50, we ended up using 10 topics.

**Dynamic Information Revelation:** We first handle the issue of multiple reviews on a given day. On each given day, all the 1* reviews are collapsed to form one document and all the 5* reviews are collapsed to form another document. Then all the documents from one product are used to conduct topic modeling, setting the number of topic to be 10. As a result, we get the topic distributions across different days for both 1* reviews and 5* reviews.

This helps us understand how the discussion evolves over time. So for each star rating, say the 1*, we have a matrix of size D by 10 where D is the total number of days for which the product has received reviews. The dth row provides the topic distribution for day d (multiplied by the total number of 1* reviews.) A standard Chi-squared test can be used to test whether the distribution of topics is constant over time.

**P-Values:** One artifact of the data is that there are many more 5* reviews than 1* reviews, and in fact, there can be multiple 5* reviews on a given day. On average, there is only one 1* review on a given day. This artifact makes variation in the day to day comparison less obvious, unless the reviews from two consecutive days are about the same topic. Because of this observation, we group the 1* reviews into groups of reviews by combining every 10 consecutive reviews into one group. Now, instead of testing for the daily topic distribution, we look at the topic distributions for every 10 reviews for the 1* reviews. Since there are many
more 5* reviews, the 5* reviews are not grouped. The log transformed p-values of standard Chi-squared tests for 1* reviews are calculated. The p-values for the 5* reviews are obtained similarly. We proceed in the following section to analyzing the results of this analysis.

The Pearson’s chi-squared test for the reviews of a specific star rating is conducted as follows. For each product, we split the reviews into first half and second half and create a K by 2 table where the (k,1)th element is the count of how many of the first half comments are about topic k and (k,2)th element for the second half. This observed count is denoted by O_{k1}. By assuming that the distribution across topics is the same for the first half and the second half, we all also obtained the expected count E_{k1}. Then a Pearson’s chi-squared test test for the distance between the observed value and the expected value defined as $\chi^2 = \sum_{k1} \frac{(O_{k1} - E_{k1})^2}{E_{k1}}$. The above test statistics is compared with a $\chi^2$ distribution with $df = k - 1$ to calculate the p-value.

**Analysis, Results and Implications**

**Does the information content change within a numerical rating over time?**

The first question we posed is whether the information content changes over time. To answer this, we split our sample into two subsets: reviews arriving in the first half of the time period and those in the second half. We statistically test whether the topic distributions are different for the first half and second half. The negative of the logarithm transformed p-values are given in Figure 2. Products are sorted based on their categories and different categories are coded using different colors. Overall, the significant p-values imply that 28.1% of the 1* reviews change topic distributions over time and 90% of the 5* reviews change topic distributions over time. This provides insight into the first question we posed: the information in 5* reviews changes over time, but later reviews do not have different information than earlier reviews for 1* ratings.

![Figure 2](image-url)

*Figure 2. p-values for the test that the topics do not change over time.*
Is the pattern of information revelation different across numerical ratings?

The previous subsection compares the topic distributions between the first half and the second half. In this subsection we analyze the topic distribution evolution at a finer scale by looking at the daily topic distribution to gain insight into our second question: the pattern of information revelation.

For a large majority of the products in our sample, we find small p-values for 5* reviews, indicating strong evidence to support the hypothesis that the topic distributions of 5* reviews are changing over time. On the other hand, for a large majority of the products in our sample, the p-values for 1* reviews are close to one, indicating there is not strong enough evidence to say the topic distributions for 1* reviews are changing over time.

At an even more granular level, we can also calculate the actual level of variation between different days. This is measured by the similarity metric, defined next. For each day, we calculate the observed number of reviews for each topic. This is obtained directly from the topic modeling. We also calculate the expected number of reviews for each topic assuming same distribution over time. For day t and topic k, let us use $O_{kt}$ to denote the observed number of reviews and $E_{kt}$ to denote the expected number of reviews. Then we can calculate the ratio between the $\min(E_{kt}, O_{kt})$ and $\max(E_{kt}, O_{kt})$. A ratio close to 1 indicates small deviation from the constant distribution assumption and a ratio close to 0 indicates large deviation from the constant distribution assumption. So we define the similarity metric as

$$\sum_{kt} \frac{\min(E_{kt}, O_{kt})}{\max(E_{kt}, O_{kt})} / KT$$

A similarity close to 1 suggests similar topic distributions over time.

The similarity results are as follows. An overwhelming majority of the similarity measures for the 1* reviews are between 0.4 and 0.6 while those for the 5* reviews are between 0.1 and 0.3. In other words, the topic distributions for 1* reviews are more stable. This indicates that the recent 1* reviews are more informative compared to the 5* reviews. Thus, by looking at the recent 1* reviews will provide a fair summary of what’s covered in the overall 1* reviews. This is not so for 5* reviews.

Overall, the information content of positive reviews varies a lot more than that of negative reviews. Recent reviews do not summarize information in past reviews. In particular, there is less similarity in the day to day discussions for 5* reviews compared to 1* reviews. The recent 1* reviews are more informative compared to the 5* reviews.

How does experience vis-à-vis product aspects evolve over time?

To answer this question, we first define what we mean by experience. For sentence $s$ within a review $i$, let $w_k^i(s)$ be the normalized cosine distance between the sentence and the topic distribution for topic $k$ obtained from LDA. The weights $w_k^i(s)$ is normalized such that $\sum_k w_k^i(s) = 1$. It measures the relevance of sentence $s$ with all the topics. We calculate the polarity of each sentence by using the sentiment function in R package “sentimentr”. Then $e^i(s)$ provides an experience score based on sentence $s$ in review $i$.

For brevity, we present results on the evolution of the experience over the ten topics for one product: Apple Certified Lightning to USB Cable - 3 Feet (0.9 Meters) – Black, Product ID: B009SYZ8OC. Figure 3 shows the charts for experience evolution over time for this product for each of the 10 topics across reviews. The left side of the Figure 3 tracks, for 1* ratings only, the evolution of the derived experiences (ranging from -1 to +1 on the y-axis) over the product lifetime (days since first review on the x-axis). The right side of Figure 3 tracks the evolution of those experiences for 5* ratings.

---

4 https://cran.r-project.org/web/packages/sentimentr/README.html
Figure 3. Plots of the evolution of user experiences over time for 1* ratings (first two columns of charts on the left) and for 5* ratings (last two columns of charts on the right).

For this product, the plots for 5* ratings (right side of Figure 3) show that most experience with the topics is positive (overwhelmingly above the zero line in all charts on the right), and that there is significant positive experience with multiple aspects of the product; See for example four representative charts identified with a blue star on the right panel. This is consistent with the variation in topics across time we presented in the previous section. It is also interesting to see that the same topics do not appear as prominently, with either positive or negative experience, in the 1* plots. It is also interesting to see that one aspect, indicated by an arrow on the left side of Figure 3, in the 1* ratings seems to have generated the most negative experience; this aspect does not figure prominently in the 5* reviews. This is also consistent with the plot of p-values in the previous section.

The plots confirm that experience with multiple aspects of a product matter, that there can be both positive and negative experience with the same product. We turn next to the question of whether these experiences generate the overall rating.

How does experience vis-à-vis product aspects explain overall review rating?

To answer this question, we conduct the following logistic regression:

$$\text{logit}(p(r_i = 5)) = \beta_0 + \beta_1 \sum_s w_i(s)e_i(s)/S_i + \cdots + \beta_k \sum_s w_k(s)e_k(s)/S_i,$$

where logit(p) = p/(1-a) and $S_i$ is the total number of sentences in review i. This lets us quantify whether the experiences with different aspects of the product can predict the rating. We present the results for the same product as above, and to predict probability of a 5* rating.

The output of the logistic regression is provided in Table 2. In Table 2, each row corresponds to the estimate of one coefficient. For example, row 2 corresponds to the estimate of the intercept; row 3 corresponds to the estimate of the coefficient for topic 1, etc. A p value less than 0.05 indicates the effect is statistically significant. In particular, we observe that the significant coefficients are all positive and significant, which implies that a positive experience has a positive impact on the rating. The coefficients for different topics differ significantly, which suggests different topics have different level of impact on the rating.
Table 2. Output from the logistic regression for whether the review rating is 5* or 1*.

We further calculated the correlation between the observed and predicted values to be 0.76. Both indicate that our model is able to capture the relationship between experience with different aspects of a product and the numerical rating.

Our research has a variety of implications in relation to customers’ perception of information presented in reviews. We initially show statistically the phenomenon of continuous arrival of reviews of extreme review ratings (both one and five stars). In doing so, we join an emerging body of literature that emphasizes the role of disagreement in reviews (Nagle and Riedl 2017). We then analyze the informational content of these extreme ratings and show that five-star reviews

Conclusion

For a large online retailer platform like Amazon, the number of reviews for a popular product is usually in thousands or even tens of thousands. As a result, it is almost not possible for a potential customer to go over all the reviews. Designing a system to better inform customers about their prospective purchases becomes of utmost importance. By answering the questions we posed at the beginning, we have shown that the information content for 5* reviews has a lot more variability over time, while that is not the case for 1* reviews. Our results suggest that it may be much more useful to summarize the main topics being discussed by rating and the user experience with those topics rather than emphasize recent reviews or the most frequent words that appear in reviews. Online platforms hosting reviews have started to recognize that by implementing a variety of review featuring approaches. For example, Steam, a digital distribution platform for multiplayer gaming, video streaming and social networking, adjusted it review featuring by displaying the ratio of positive to negative ratings by reviewers. Similarly, BestBuy breaks down reviews into themes and allows easy access to navigate to those reviews.

---

The Echo dot has already garnered more than 100 thousand reviews online a year and a half after it has been on the Amazon platform.
REFERENCES


