Reading Between the Stars: Understanding the Effect of Online Customer Reviews on Product Demand

Many studies have examined star ratings in customer reviews and established that they are a relevant source of quality information. We examine the sentiment of text reviews and study how the interplay between sentiment and star ratings relates to product demand. To fully exploit the abundance of online reviews, we implemented a novel algorithm that reads the text of customer reviews and assesses its sentiment. Using the U.S. Automobile market data from 2002-2013, we find that sentiment and star ratings both have a decreasingly positive effect on product demand and that their interaction effect on demand suggest they act as complements, not substitutes.

When faced with a product whose quality is either unknowable ex-ante or difficult to assess, consumers tend to rely on signals and look to others for information about quality. In the days of traditional media, these sources of information were friends, neighbors, colleagues, and endorsements from celebrities or experts. With the rise of social media and the proliferation of online customer reviews, consumers can do the same at a much larger scale and take advantage of the wisdom generated by crowds of product reviewers. Indeed, many studies have shown that reviews have an influence on product demand (Chevalier and Mayzlin 2006, Liu 2006, Dellarocas et al. 2007, Ghose and Ipeirotis 2011).

Although reviews are comprised of both quantitative ratings (i.e., star ratings) and text, previous studies relating product reviews and product demand are primarily focused on the effects of the quantitative ratings. This is not entirely surprising because not only are star ratings readily available but also they are presumably the most salient feature of product reviews. Salience matters due to an interesting tradeoff in harnessing wisdom from crowds. More reviews not only mean that the crowd’s average estimate of quality is closer to the product’s true quality, but more reviews also mean that the cost of consuming this information is increasing.

While the ratings can be simply averaged and displayed as a number, text reviews cannot be averaged and require time and effort from the reader to process all the text. Chetty et al. (2009) showed that consumers overlook less salient information when purchasing consumables but suggest that this may not hold for big ticket items. Yet, Luca and Smith (2013) found that consumers still rely on coarse information for big ticket items such as college education. They show that even an incremental increase in cost of information processing can have an impact on consumer choice: university rankings had a smaller effect on application rate when universities were listed by alphabetical order versus when they were listed by ascending order of rankings.
Assuming similar saliency of star ratings and text reviews, prior studies have de facto assumed that text and ratings are substitutes.

Some studies, however, found that text reviews have a direct impact on product sales. Chevalier and Mayzlin (2006) found that the length of text reviews had a significantly positive effect on sales. Ghose and Ipeirotis (2011) showed that while controlling for average ratings, readability and subjectivity of the text still had a significantly positive impact on sales. These findings can be interpreted in two ways: 1) consumers read text reviews to judge the validity of the quality signal contained in ratings, or 2) text reviews contain information that is not captured in ratings. Other studies that have mined text reviews found that reviews reveal consumer's relative preference for different product features (Archak et al. 2011) as well which brands are associated together in consumers' minds (Netzer et al 2012). In particular, Archak et al. (2011) extracted opinions about frequently mentioned product features and found that feature specific opinions have a significant effect on sales above and beyond the effects of average ratings. The Archak et al. (2011) findings suggest that text reviews and ratings may contain different information and challenges the assumption that, in aggregate, opinions about individual product features should be captured by average ratings.

We also argue that star ratings and text reviews contain different information and therefore their interplay may have a significant effect on product demand. Drawing on dual-process theory in psychology, we argue that star ratings are the result from intuitive cognitive processing, System 1, whereas text reviews are the result from rational cognitive processing, System 2, (Kahneman and Frederick 2005). This conceptual model is consistent with the way customer ratings and reviews are typically solicited. When writing a review, customers are first asked to rate the product. Star ratings capture the reviewer's gut feeling about a product (system 1 thinking). Only after the star ratings have been assigned, reviewers are asked to comment further. Reviewers, still anchored on the rating they have just provided, reflect and try to rationalize their decision by offering their opinions about the product as well as accounts of their experience (system 2 thinking).

In order to understand how the interplay between quantitative and textual reviews affects product demand, we examine how the average sentiment of text reviews and the aggregated ratings relate to product demand. We also study how these effects interact to examine whether text reviews and star ratings are complements or substitutes.

An important challenge to carry out our study concerns measuring sentiment of text
reviews. We conduct sentiment analysis, via a supervised machine learning (ML) algorithm, on the entire body of text reviews available in our data to extract a measure that is comparable to star ratings. We first breakdown the body of the text reviews into sentence fragments by using an unsupervised ML algorithm to detect sentence boundaries and further split the fragments that contain transition words indicating contrast such that each fragment has uniform sentiment. Then, we randomly sample 6,000 fragments (0.8% of the data), each of which is hand coded by at least 10 mTurk workers to form a consensus. We conduct 5-fold cross validation, where one fifth of the hand coded sentiments are used as the test set and the rest are used to train the algorithm. We use Support Vector Machine (SVM) with Radial Basis Function (RBF) Kernel to classify each segment as negative or not negative and use Platt scaling to get probability estimates.

We empirically analyze the relationship among sentiment, ratings, and product sales (controlling for all other measurable factors that may contribute to product demand) by using an IV logit model with longitudinal data aggregated at the model-year. We find that text and rating are complements, confirming that not only do ratings and text capture different quality information but also that they have a synergistic effect on product sales.

References


