Information Systems and e-Business Management (2019) 17:261–284
https://doi.org/10.1007/s10257-019-00438-3

ORIGINAL ARTICLE

Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach

Satyendra Kumar Sharma[1] · Swapnajit Chakraborti[2] · Tanaya Jha[3]

Received: 11 January 2019 / Revised: 11 July 2019 / Accepted: 5 September 2019 /
Published online: 13 September 2019

Abstract

Prediction of customer demand is an important part of Supply Chain Management, as it helps to avoid over or under production and reduces delivery time. In the context of e-commerce, accurate prediction of customer demand, typically captured by sales volume, requires careful analysis of multiple factors, namely, type of product, country of purchase, price, discount rate, free delivery option, online review sentiment etc., and their interactions. For e-tailers such as, Amazon, this kind of prediction capability is also extremely important in order to manage the supply chain efficiently as well as ensure customer satisfaction. This study investigates the efficacy of various modeling techniques, namely, regression analysis, decision-tree analysis and artificial neural network, for predicting the sales of books at amazon.in, using various relevant factors and their interactions as predictor variables. Sentiment analysis is carried out to measure the polarity of online reviews, which are included as predictors in these models. The importance of each independent predictor variable, such as discount rate, review sentiment etc., is analyzed based on the outcome of each model to determine top significant predictors which can be controlled by the marketer to influence sales. In terms of accuracy of prediction, the artificial neural network model is found to perform better than the decision-tree based model. In addition, the regression analysis, with and without sentiment and interaction factors, generates comparable results. The comparative analysis of these models reveals several significant findings. Firstly, all three models confirm that review volume is the most important and significant predictor of sales of books at amazon.in. Secondly, discount rate, discount amount and average ratings have minimal or insignificant effect on sales prediction. Thirdly, both negative sentiment and positive sentiment of the reviews are individually significant predictors as per regression and decision-tree model, but they are not significant at all as per neural network model. This observation from the neural network model is contrary to the extant research which claims that both negative and positive sentiment are significant with the former having more influence in predicting sales. Finally, the interaction effects of review volume with negative and positive sentiment are also found to be significant predictors as per all three models. Hence, overall, out of various factors used for sales prediction of

Extended author information available on the last page of the article

books, review volume, negative sentiment, positive sentiment and their interactions are found to be the most significant ones across all models. The results of this study can be utilized by online sellers to accurately predict the sales volume by adjusting these significant factors, thereby managing the supply chain effectively.

Keywords E-commerce · Sentiment analysis · Neural network · Decision tree · Regression analysis · Predictive model

1 Introduction

Online shopping has gained tremendous popularity in India in the past decade. Easy access to internet, surge of less expensive smartphones, shopping apps for mobile phones, increase in awareness of the youth in rural areas and smaller cities and increase in prosperity have been major reasons behind the rising market share of online retail. Customers prefer online shopping for convenience, time-saving, home delivery, and wider range of products available (Social Media Report 2012; Jain and Kulhar 2015). Most businesses today feel the need for an online presence to increase their coverage and avoid the costs associated with setting up physical retail stores. E-commerce websites like Amazon, Taobao, Flipkart, eBay etc. enjoy a large customer base.

However, this domain is getting increasingly competitive and unpredictable, and customer expectations regarding price and quality have increased drastically. Businesses can no longer rely on cost advantages in manufacturing (Chong et al. 2009); they also need to focus on having efficient supply chain management with a better understanding of consumer demands and behavior (Chong and Zhou 2014).

Accurate forecasting of sales is an important part of efficient supply chain management, as it helps to avoid overproduction and underproduction of products. A better understanding of customer demand reduces risks such as the Bullwhip Effect, and significantly appreciates the effectiveness and efficiency of its supply chain (Chong and Zhou 2014). Consumers today make real-time decisions using the information available online in the e-commerce environment, and it is now possible to predict product sales and consumer demands by analyzing the data available online (Duan et al. 2008; Floyd et al. 2014).

Most major e-commerce websites now allow users to see reviews from other customers before making their purchasing decisions. Research shows that User Generated Content or UGC (Tang et al. 2014; Chong and Zhou 2014) has a significant impact on product sales. Findings of a survey state that 70% of consumers trust UGC for making purchase decisions (Social Media Report 2012). Another report found that four in five consumers reverse their purchase decisions due to negative online reviews (Online Influence Trend Tracker 2011). In a majority of cases, this data is freely available online and can be scraped easily with little cost.

The main purpose of this study is to determine whether factors like volume and sentiments of online reviews, discount offers, free delivery options, average rating of product, etc. can significantly influence product sales/demand on an e-commerce website.

Although the importance of volume and valence of reviews has been studied in literature (Cheung and Thadani 2012; Lu et al. 2013), not much research has been done in determining the significance of sentiments (which are qualitative) in online review text (Hu et al. 2014). Available studies do explore the influence of the quantitative features of UGC, such as numerical ratings and volume of reviews (Pang and Lee 2008; Zhu and Zhang 2010), but there is a scarcity of research as far as qualitative features like sentiments and pictures in reviews are concerned. Further investigation is needed to understand fully the complexity and variation in online feedbacks and reviews.

Further, in addition to online reviews, there are various other factors which may influence product sales in an e-commerce environment. For example, organizations can give discounts and respond to customer queries (Chong and Zhou 2014), or provide free delivery options. Differences in the nature of products may also cause sizeable difference in sales.

In addition to examining the predictive power of various factors, which may influence product sales on e-commerce websites, and the interplay between these variables, sentiment analysis of online reviews is also included in this study.

Further, most recent research that has been done in this field focuses solely on electronic product categories such as computers, televisions, cameras (Chong et al. 2016) and tablet computers (Schneider and Gupta 2016). These studies mention that research should be conducted for other product categories as well. Also, no study using this methodology has been done for the Indian market. This research focuses on addressing these gaps and therefore conducts the study based on sale of books in the context of the Indian market. In this context, it is worth mentioning that similar study is also possible for exploring demand/sales of services which are provided by e-commerce companies. However, in the current study that is kept out of consideration.

The data for this study is collected from amazon.in, which is a popular e-commerce website. Amazon is chosen because it provides a "Best Sellers Rank" for its products, which has been used as a proxy for product sales in various studies previously (Ghose and Ipeirotis 2006), and also because including the entire extent of online UGC for a product across the internet is not possible for us. Amazon is a major player in this segment in the Indian market and is used by large number of people; as such, it helps minimize the risk of bias and low reputation.

The paper is organized as follows: Sect. 2 contains theoretical background, Sect. 3 describes motivation for the research and its goal, Sect. 4 contains the data collection and preparation steps, Sect. 5 describes the research methodology, Sect. 6 contains the analysis and interpretation of results, Sect. 7 contains conclusions and managerial implications and Sect. 8 contains future work.

2 Theoretical background and related work

2.1 Online promotional marketing: factors influencing online sales

With the advancement of technology, consumers have turned to new sources for access to information (Tang et al. 2014). Because there is more information

regarding products available, the number of factors that influence consumers' decisions has also increased (Floyd et al. 2014). Increase in competition also puts pressure on companies to attract customers and secure sales faster. Hence, analysis of factors that might affect product sales on online platforms is crucial to business success and, consequently, many of these variables have been studied extensively in literature. The extant research on these variables and their interactions is described below.

2.1.1 Discount value

The transaction utility theory suggests that higher discount offerings lead to higher sales (Lichtenstein et al. 1990), as discounts are a bargain for the customers. Gendall et al. (2006) reported that discounts lead to an immediate increase in short term sales; Faryabi et al. (2012) also reported a positive impact of discounts on store image and purchase intention for online sale of cell phones. The effect of discounts on product demand has been extensively studied in literature (McNeill 2013), but various contradictions have also emerged (Drozdenko and Jensen 2005). Studies show that while discount has a significant positive impact on online product sales (Gong et al. 2015), there may exist a saturation point for discount offerings (Gupta and Cooper 1992; Marshall and Leng 2002). As customers get used to frequent discounts, their expectations with respect to discount rates also increases. Also, customers belonging to different countries or cultures may display different attitudes to discount offers, as demonstrated by Marshall and Leng (2002).

2.1.2 Discount rate

The psychophysics-of-price heuristics theory states that the psychological utility derived from saving a certain amount of money through discount offerings decreases as the price of the product increases (Chen et al. 1998). Hence, there is a need to examine discount rate and discount value separately. If the price of the product is Indian Rupees (INR) 400, what has a greater impact on a customer's psychology, and consequently product demand, a discount of 10% or a discount of INR 40? (Chong and Zhou 2014).

2.1.3 Free delivery

Availability of free delivery is said to be beneficial in attracting customers for online platforms (Xu et al. 2017), and also may lead to an increase in consumer loyalty (Doern and Fey 2006). However, for e-commerce websites, the option of free delivery for fashion apparel and electronic goods is very common and expected by customers nowadays, and absence of this feature may lead to loss of customers. Hence this factor need not be considered for prediction of sales volume.

2.2 Online reviews: emerging driver of online sales

Online UGC is widely considered by customers to be a more reliable source of information than traditional advertising methods (Lee et al. 2008; Davis and Khazanchi 2008; Mudambi and Schuff 2010). It is also available more easily (Davis and Khazanchi 2008) and provides more detailed and balanced information (Floyd et al. 2014).

The effect of online UGC on product sales has been explored in the past (Archak et al. 2011). However, only valence, volume, credibility and dispersion rate of online reviews have been studied extensively (Lu et al. 2013). Reviews may influence users' experience and product prices (Li and Hitt 2010) also. Chevalier and Mayzlin (2006) found that volume and valence of online reviews have significant impact on sales of books on some e-commerce sites, while Lu et al. (2013) made the same observation for restaurant sales. Later research has also inspected the impact of user reviews on online sales, like Gaikar and Marakarkandy (2015) and Yao and Chen (2013) for movie sales prediction; Salehan and Kim (2016) and Chong et al. (2017) for electronic products like mobile phones, tablets, etc. Recent researches have also explored the genuineness of online reviews (Hu et al. 2012).

2.2.1 Online review valence

This variable represents the nature of the reviews, generally represented through the numerical rating of the review. The effect of this variable on product sales has been studied in the past, but the results are contradictory. While some studies assert that review valence has significant predictive power (Dellarocas et al. 2004; Cheung and Thadani 2012; Lu et al. 2013; Chevalier and Mayzlin 2006), others found no significant impact of review valence on sales (Davis and Khazanchi 2008). However, the impact of this variable may also differ according to platform, as demonstrated by Chevalier and Mayzlin (2006), or the impact may be indirect, as shown by Duan et al. (2008), who in his study found that review valence impacts box office sales indirectly by influencing volume of online reviews. Further, the predictive power of review valence may also depend on the category of products (Cui et al. 2012), and on qualitative features of the text (Ludwig et al. 2013). For example, the impact may be more for goods/services which can be evaluated only after experience (a.k.a. "experience" goods category), e.g. restaurant, hairdresser etc., compared to electronic products, which falls in "search" goods category, where, evaluation can be done prior to purchase purely based on technical aspects and specifications.

2.2.2 Online review volume

This variable also has been studied in literature (Duan et al. 2008; Lu et al. 2013; Davis and Khazanchi 2008), and is generally believed to have a significant impact on product sales through increasing awareness and popularity of the product (Yang et al. 2012). However, the predictive power varies according to product category, as demonstrated by Cui et al. (2012), who found that the impact is more for "experience" goods than for "search" products and is even more for electronic goods.

2.2.3 Percentage of negative online reviews

Studies show that negative reviews have a greater impact on purchasing decisions than positive reviews (Ito et al. 1998; Cheung and Thadani 2012; Lee et al. 2008). This may be attributed to the fact that positive reviews reflect good characteristics of a product, but negative reviews reflect bad experience or lack of faith in the brand's product quality. A case study carried out by Zhu and Zhang (2010) showed that negative reviews did more damage to the usage of software programs as compared to the benefit obtained by positive reviews, and Chevalier and Mayzlin (2006) demonstrated similar results for online sale of books on Amazon.com and Barnesandnoble.com.

2.2.4 Sentiment of review text

While review valence and volume are quantitative features, text sentiment is a qualitative feature. In the past, ratings or review valence was used to represent review sentiment (Chevalier and Mayzlin 2006), but that has proven to be an unsatisfactory assumption, as customers tend to consider review ratings as well as review text before making a decision (Chevalier and Mayzlin 2006). Further, valence may not always be consistent with text sentiment. Research shows that valence is quite bipolar in distribution and a majority of reviews tend to have either very high or very low ratings (Hu et al. 2014), while text sentiment may be mixed, as customers might be satisfied with some aspects of the product while being dissatisfied with others, and this complexity cannot be adequately captured through numerical ratings, which are generally on a range of 1–5. Studies indicate that emotional contagion occurs in online environments through text (Kramer et al. 2014; Hancock et al. 2008), and the emotions of customers have a significant impact on their purchasing decisions (Tsai 2001).

In view of the extant research, sentiment analysis of review text seems to be an important factor to accurately predict product sales. Further, not all reviews are of equal importance, as customers tend to go through only the most helpful reviews (Mudambi and Schuff 2010) which are also displayed on the first page, and hence these reviews have higher exposure and hold more importance over others. So, taking all reviews into account might dilute the effect of sentiment on product sales, as all reviews do not affect the customers' decisions equally.

2.3 Effects of interactions of factors on online sales

The amount of information available online has drastically increased, and customers are overwhelmed with choices in their decision-making (Chong 2013). Vendors have also caught onto this fact and are increasingly adopting multi-dimensional marketing approaches to increase their product sales. This is even more pronounced in the e-commerce business (Lu et al. 2013).

2.3.1 Interactions between online reviews and discount rate

A study by Lu et al. (2013) examined how online UGC could affect marketing promotion strategies, but it focused on differences between product categories. The study also showed that for restaurant sales, promotions strategies become inefficient when UGC volume is high. In certain cases, discounts also may be perceived as indication of lack of quality (Marshall and Leng 2002).

2.3.2 Interactions between sentiments and valence and volume

The effect of sentiments and that of volume and valence on product sales have been individually studied in literature (Duan et al. 2008; Yu et al. 2012), however, the role of sentiments has not been studied extensively (Hu et al. 2014), and previous studies regarding the impact of review volume on online sales are inconsistent with their results (Chevalier and Mayzlin 2006; Dellarocas et al. 2004). In a study in Hu et al. (2014) found that online reviews for books have an indirect effect on product sales through sentiments. Cui et al. 2012 found that for electronics products, valence of reviews is a more important predictor.

The effect of review sentiment on valence and volume has not been studied in the past. For example, even if the product has a high rating, the sentiment of the most helpful reviews may affect product sales. Also, it is possible that customers do not trust the reviews when the review volume is very low. Hence, in these kinds of situations it is possible that the impact of review volume is affected by review sentiments.

Based on the preceding discussions in Sects. 2.1, 2.2 and 2.3, following factors and interaction effects emerge as important in the context of prediction of online book sales which are taken up for investigation:

- *Individual factors* Discount Rate, Price, Average Ratings, Sentiment (Positive/Negative), Review Volume. These are found to be significant predictors consistently across extant research and hence need validation in the context of prediction of online book sales.
- *Interaction effects* Interaction of Discount Rate with various aspects of online reviews, namely, Review Volume, Average Ratings, Positive Sentiment and Negative Sentiment. In addition, interaction of Review Volume with Positive Sentiment and Negative Sentiment need investigation. Extant research presents contradictory results on these interactions as predictors on various products/services investigated so far. Hence, they need a validation in the current context of online sales prediction for books.

3 Motivation and research goal

Predicting customer demand is an important part of Supply Chain Management, as it helps in avoiding both over-stocking and under-stocking of products. This is especially interesting for e-commerce, which is currently booming, and where most of the information which customers use to make purchasing decisions is easily and

freely available. As noted in Sect. 2, there are various factors which influence customers' decision making thereby affecting the sales volume. The UGC in the form of online reviews of products has also empowered the customers to not just depend on marketers' promotion for purchase decision making. For marketers or online sellers, it is extremely important to analyze the sentiment of buyers from these online reviews but due to its enormous size, sometimes it is left out and more focus is given to quantitative factors such as discount, ratings, review volume, etc. It is evident that there is a need to incorporate all such factors into a single framework/model so that a clearer understanding can be generated on significance of individual factors in predicting sales. Such models should also be able to handle the large volume of data as well as derive knowledge from it. Figure 1 shows a conceptual predictive model for sales with all influencing factors. This is further elaborated in Sects. 5 and 6.

This study aims to demonstrate the efficacy of various modeling techniques including regression analysis and advanced machine learning-based models, namely, decision tree and artificial neural networks in predicting sales on e-commerce platforms, and examine what variables are significant predictors for the sales of books on amazon.in.

The choice of books as the product category is driven by the fact that most current research in this domain has focused primarily on electronic products only. Moreover, lack of India-centric research in this domain is also another motivating factor for undertaking this subject for research.

4 Data collection and preparation

The dataset/corpus for this research is created using a sample of 1408 books (only paperback) from amazon.in. As mentioned earlier, books are chosen as product category because most recent research on sales prediction on e-commerce websites focuses on electronic product category (Chong et al. 2017; Schneider and Gupta 2016) and there is a scarcity of recent research on online book sales. The Amazon

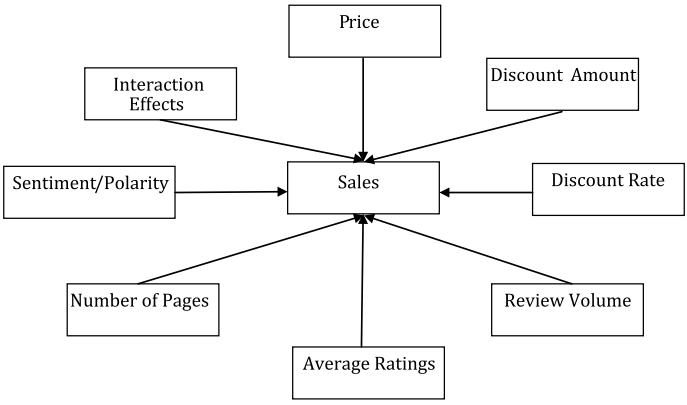

Fig. 1 Conceptual model for prediction of sales using relevant predictors

API is used to get a list of ISBNs (which are unique for a book on Amazon) along with "Requests" (Requests 2017) and "Beautiful Soup" (Beautiful Soup 2017) modules to scrape data for the independent variables, namely, current price, discount amount and rate, number of pages, average rating etc. as product characteristics. While "Beautiful Soup" is supported by Natural Language Toolkit (NLTK), which is a very reputed consortium, "Requests" is also a very popular tool as evident from various testimonials at its website. The data collection was done in the month of December 2017. Only paperback editions have been considered.

Since this study focuses on the importance of user reviews in predicting sales, the products with no reviews are excluded. Only reviews on amazon.in have been considered as the study is in the context of Indian users.

The product reviews on Amazon are grouped and displayed page by page. For sentiment analysis, the first two pages of the reviews (English reviews only) are collected, and both review title and review text are included. Approximately 16,000 reviews were collected during the month of December 2017. Only reviews of "Amazon Verified" customers were considered to reduce the effect of fake reviews. In order to extract sentiment-specific variables, the Sentiment Analysis API (2017) was used. This API handles all text pre-processing functions such as tokenization, stopword removal, stemming etc. before extracting sentiment. Appropriate sanity testing of this API was done before choosing this as a tool for the experiments.

The dependent variable in this study is Sales Rank or Best Sellers Rank, which is used as a proxy for sales as sales data is not available. The Sales Rank of the relevant 1408 book titles was collected in January, 2018. This is to ensure that the predicted variable comes later in time sequence than the predictor variables.

5 Research methodology

The primary goal of this research is to analyze the significance of various relevant factors while predicting the sales of books in amazon.in. The factor variables are chosen using extant research in the e-commerce domain and are gathered from amazon.in as explained in Sect. 4. The importance or significance of these factors is determined using two steps. First multiple types of predictive models are created from the data set/corpus gathered from amazon.in using Sales Rank as dependent variable and relevant factors as predictor/independent variables (ref Table 1). Next, the significance of the predictor variables generated by these predictive models is compared and interpreted.

There are many techniques for creating predictive models, such as regression, decision tree (DT), artificial neural network (ANN) etc. Additionally, the choice of a specific technique for building a predictive model depends on the quality of result generated by that model on the testing set. Hence, the methodology followed in this research involves creating multiple predictive models, starting from the simplest one (namely, linear regression) and keep improving incrementally and gradually move towards more sophisticated machine learning based models, such as an artificial neural network, to observe the importance of factors and also improve quality of prediction.

Table 1 The input/predictor variables for the models

Variable	Min	Max	Mean	SD	Skewness
Current price	17	5083	371.333	406.510	4.812
Discount amount	0	2227.530	140.168	189.645	4.5
Discount rate	0	85	27.401	18.392	0.061
Review volume	1	4425	83.114	287.581	8.663
Average ratings	1	5	4.259	0.658	− 1.898
Number of pages	16	5232	400.239	303.361	4.046
Positive sentiment strength	1	5	2.765	0.569	− 0.342
Negative sentiment strength	− 4	− 1	− 1.409	0.473	− 1.939
Overall polarity (nominal)	NA	NA	NA	NA	NA

5.1 Identifying the independent variables of the model

As discussed in Sect. 2, review of the extant research revealed a set of appropriate quantitative factors, namely, current price, discount amount, review volume etc., which are suitable as independent variables of the predictive model. In addition to these variables, the sentiment-specific variables are also included as independent variables in order to improve the accuracy of the prediction.

5.1.1 Quantitative variables gathered from amazon.in

The following set of variables is created based on the data collected from amazon.in (ref Sect. 4). These are aligned with the factors noted in extant research on online sales and e-commerce.

- *Current price* The actual offer price of the book after discount, shown at amazon.in at the time of data collection.
- *Discount amount* The difference between original/printed price and the current price offered. Often this is shown by striking out the original price and showing the current/actual price at amazon.in against the book title.
- *Discount rate* Percentage of discount amount with respect to original price. This is calculated as ((Original Price − Current Price)/Original Price) * 100.
- *Review volume* Total number of reviews available for a book at the time of data collection. This is also shown at amazon.in against a book title.
- *Number of pages* Total number of pages taken up by all the reviews of a book assuming 10 reviews per page shown at amazon.in.
- *Average ratings* The average rating out of 5 shown for each book at amazon.in.

5.1.2 Extracting variables related to sentiment of product reviews

In addition to these variables, the sentiment of the online reviews needs to be captured as one of the factors because of its possible importance in present context of online sales.

Three additional variables are extracted from the review text scrapped from amazon.in. The first one is "Overall Polarity" of the review text, which can be classified as Positive, Negative, or Neutral. For this, the Sentiment Analysis API (2017) has been used. The API first determines the probability of the text being neutral; if it is greater than 0.5, the label is neutral. Otherwise, the label is "pos" (Positive) or "neg" (Negative), depending upon which probability is greater. For this purpose, the entire review text for a book title (ISBN) is considered as one unit. Hence this is an aggregate variable of all reviews for a book which reflects the overall polarity of its reviews.

The other two variables "Positive Sentiment Strength" and "Negative Sentiment Strength" are created using SentiStrength (2017), which is a freely available Java software package. It reports the positive sentiment strength and negative sentiment strength for short texts and is reportedly referenced by many peer-reviewed research publications using sentiment analysis. Positive Sentiment Strength can range from 1 (weakest) to 5 (strongest), while Negative Sentiment Strength can range from -1 (weakest) to -5 (most negative). For this purpose, each review is considered separately, and then the positive scores of all reviews are averaged to get the Positive Sentiment Strength for each product. The same is done to get Negative Sentiment Strength.

While measuring the Overall Polarity, all the reviews are taken as one unit because users often go through the first couple of pages of reviews and consider them together when forming an opinion about their product. Secondly, as mentioned earlier, the reviews visible on the first two pages of reviews are only taken for each book title because users do not go through all available reviews (Leino and Raiha 2007; Hu et al. 2014). Generally, the first few pages are enough, and taking all the reviews will dilute the importance of the most visible reviews while giving extra importance to old/unhelpful ones. While measuring the sentiment strengths, however, each review is taken as a separate entity because SentiStrength (2017) uses the word with the strongest emotion to determine the scores. If all the reviews are merged together, some reviews would dominate while others would be overshadowed. This is consistent with the approach taken by Hu et al. (2014).

Important data characteristics of the independent variables are shown in Table 1 below. Overall Polarity is the only nominal variable.

5.2 Building the predictive models

The model building phase starts with creation of simple models based on linear regression and then incrementally moving towards machine learning models with higher complexity such as decision tree and artificial neural network. Decision trees and neural networks are competitive classifiers. Neural networks are generally better at generalizing (Zhou et al. 2002), but decision tree results can be interpreted more easily (Zhou and Jiang 2004). Both of these models are explored in this study. SPSS Modeler 16.0 from IBM was used to create these two models while multiple linear regression models are created using Microsoft Excel.

The overall flow of the methodology of model building is outlined by the steps given below.

Step 1: Build a linear regression model as a basic prediction model for sales using the six independent variables mentioned in Sect. 5.1.1.
Step 2: Perform sentiment analysis of online reviews to extract three new variables (ref Sect. 5.1.2) that measure the sentiment strength and polarity.
Step 3: Incorporate these three new variables (ref step 2) into our regression model created in step 1. This means the regression model now has all 9 independent variables as shown in Table 1.
Step 4: Build a regression model by including interaction effects (6 new variables) formulated based on extant literature (ref Sect. 2.3). The total number of independent variables in this model is 15 (=9+6).
Step 5: Build a machine learning classifier, namely, a decision tree with the same input variables.
Step 6: Finally, an Artificial Neural Network based model is created for prediction.

5.2.1 Multiple linear regression based predictive model

Regression analysis is a statistical modeling procedure used to predict a dependent variable, using one or more independent variables. More accurately, it analyzes how the dependent variable's value changes as any of the independent variable values change, all else being constant. Regression analysis is a basic and widely used tool but is generally outperformed by more sophisticated machine learning models when the relationships between variables are complex (Santibanez et al. 2015).

For current research, regression modeling in Microsoft Excel is used to determine the relationship between sales rank and various predictor variables (ref Table 1). As noted earlier, for the current study, three regression models have been created and explored in sequence. The first model uses 6 variables as input, the second model uses 9 variables as input and the third model uses 15 variables as input.

5.2.2 Decision tree based predictive model

To improve the accuracy of results as well as interpretability, decision tree-based models are also created as part of this study. A decision tree is a supervised machine learning classifier which learns decision rules from the training data and uses them to build a predictive model. Many algorithms have been developed for learning of Decision Trees, like Iterative Dichotomiser (ID3), C4.5, C5.0, Classification & Regression Tree (CART), Chi square Automatic Interaction Detector (CHAID) etc. Decision trees are easy to use and interpret and give satisfactory results with a large set of data.

For this model, all predictor/input variables are scaled between 0 and 1, except for Overall Polarity which is a nominal variable. The output variable i.e. Sales Rank is mapped into 3 equi-frequency bins for classification purposes. The data set is also partitioned into training set (70%) and testing set (30%). The C5.0 decision tree algorithm is used for building the model for this study. The decision tree model is created using all 15 variables together as input.

5.2.3 Artificial neural network based model

An Artificial Neural Network (ANN) is a computational system which mimics certain features of the human brain. It consists of interconnected artificial neurons organized in layers. The first layer takes input and the last layer produces the output. In between are hidden layers (one or more) which use a weighted function of the input values. An ANN learns these weights and consequently generates the output by learning in a supervised manner, that is, from a labeled data set.

ANNs can perform significantly better than traditional regression analysis in cases where the number of input variables is large and the relationships between variables is complex, for example in big data scenarios. Further, they are non-parametric, so the results depend solely on the learning of the ANN based on the real-world data provided instead of assumptions about variables.

For current research, a Multilayer Perceptron (MLP) model with back-propagation for ANN has been created using an ensemble of 10 component models. Each component model of ANN has 1 hidden layer with 5 nodes. The number of input nodes are 15 including 9 individual predictors and 6 interaction effects. The sigmoid activation function has been used for the network nodes. An ensemble of 10 components, instead of a single ANN model, is used for better accuracy. The overfit prediction is set at 30%. Only the output variable (predicted variable) i.e. Sales Rank is transformed using natural logarithm to address the issue of the wide range of values. The 15 input variables used in the ANN are normalized appropriately.

6 Results and interpretation

6.1 Analysis of multiple linear regression based model

6.1.1 Variable transformation

As noted earlier, the dependent variable used for this study is Sales Rank, also known as Best Sellers Rank, which is available on amazon.in and is used as a proxy for product sales, as actual sales data is not available. Since the variable Sales Rank ranged from 2 to 389,825, natural logarithm transform has been applied to reduce its variability. The predictor variables are also transformed using natural logarithm and are listed the Table 2 below. LN refers to natural logarithm of the variable. LN have been used to scale variables, as the range of most variables is quite big creating right-skewed distributions (ref Table 1).

6.1.2 Regression analysis with three models

Regression Analysis is carried out in stepwise fashion, first by considering only the singleton predictors (Table 2) without sentiment, next introducing the sentiment level predictors (Table 1, last three rows) and in the end, adding interaction effects as factors in addition to all the previous ones. The results of these three

Table 2 Input/predictor variables in the basic regression model (Step 1)

Variable	Min	Max	Mean	SD
LN (current price)	2.833	8.534	5.562	0.833
LN (discount amount)	0.00	7.709	3.941	1.921
LN (discount rate)	0.00	4.443	2.798	1.351
LN (review volume)	0.00	8.395	2.656	1.800
LN (number of pages)	2.773	8.563	5.767	0.683
Average ratings	1	5	4.259	0.658

Table 3 Regression results (MS Excel output)

Regression statistics		
Multiple R		0.708055
R square		0.501342
Adjusted R square		0.499207
Standard error		1.357702
Observations		1408
	Coefficients	*P* value
Intercept	9.122841	1.10E−57
ln(curr_price)	0.643128	*6.20E−14*
ln(disc_amt)	−0.15923	0.088075567
ln(disc_rate)	0.236044	0.07627463
ln(review_volume)	0.652	*5.58E−147*
ln(nop)	0.33616	*5.24E−07*
average_ratings	0.02444	0.663777605

levels of analysis are presented in the following paragraphs. MS Excel has been used for regression modeling.

Results of regression with the 6 input variables (Table 2) are displayed in Table 3. At 95% significance level, Current Price, Review Volume and Number of Pages ("nop") are significant (based on *p* value). The current price seems to have maximum effect based on coefficient value. The value of R-squared is 0.5. One similar research on some electronic item categories have reported R-squared in the range 0.18–0.38 (Ghose and Ipeirotis 2006), but no previous data is available on R-squared on prediction of Sales Rank for books.

The second regression-based predictive model uses 3 more input variables which are based on sentiment analysis. Details of these variable characteristics are displayed in Table 4 below.

Overall Polarity is a nominal variable with 3 values: 0 (Neutral), 1 (Positive Polarity) and 2 (Negative Polarity).

Table 4 Sentiment specific Input/Predictor variables

Variable	Min	Max	Mean	SD
Positive sentiment strength	1	5	2.765	0.569
Negative sentiment strength	−4	−1	−1.409	0.473

Table 5 Regression results (MS Excel output)

Regression statistics		
Multiple R		0.714676848
R square		0.510762997
Adjusted R square		0.507613402
Standard error		1.346258511
Observations		1408
	Coefficients	*P* value
pos_senti	0.198632539	*0.006329751*
neg_senti	−0.269830938	*0.000667931*
overall_polarity	−0.188711007	*0.029535947*

The results (only for the new sentiment specific variables) are displayed in Table 5. There is no change of significance for the previous 6 variables and hence they are not shown here.

At 95% significance level, all the three new variables (which are shown in Table 5) capturing various dimensions of sentiment, are found to be significant. However, R-squared has not increased much as compared to Table 3.

In the third regression-based model, the following 6 interaction effects are included along with 9 previous variables:

- Discount Rate * Review Volume
- Discount Rate * Average Ratings
- Review Volume * Positive Sentiment Strength
- Review Volume * Negative Sentiment Strength
- Discount Rate * Positive Sentiment Strength
- Discount Rate * Negative Sentiment Strength

Results (only for the new interaction variables) are displayed in Table 6. There is no change of significance for previous 9 variables. In Table 6, "rev" & "rev_vol" stands for Review Volume, "avg" stands for Average Ratings and "disc" stands for Discount Amount. Other variable names are self-explanatory and related to sentiment specific dimensions.

At 95% significance level, the interaction effects, namely, Discount Rate * Review Volume, Review Volume * Positive Sentiment Strength and Review Volume * Negative Sentiment Strength are significant (based on *p* value). However, the value of

Table 6 Regression results (MS Excel output)

Regression statistics		
Multiple R		0.723079238
R square		0.522843585
Adjusted R square		0.517330168
Standard error		1.331917191
Observations		1408
	Coefficients	*P* value
disc * rev	−0.600198927	*1.3E−102*
disc * avg	0.055090136	0.783823
disc * pos_senti	−0.004198911	0.070894
disc * neg_senti	−0.003995277	0.304932
rev_vol * pos_senti	−0.000955085	*0.000262*
rev_vol * neg_senti	−0.001445835	*0.003966*

R-squared has increased only slightly even after incorporating these interaction effects as compared to Table 5.

6.1.3 Overall analysis of regression models

Regression analysis seems to suggest that discount does not affect sales rank of books on amazon.in, and average rating of the product is not significant either (ref Table 3). This suggests that offers provided by e-commerce sites do not significantly affect book sales on amazon. Review Volume is significant, though it could be possible that higher sales lead to more reviews, and more reviews (indicating product popularity) lead to even more sales. Number of pages is again an indicator of review volume which is also found to be significant.

The overall polarity and sentiment strengths are all significant, even though Average Rating is not (ref Tables 3, 5). This reaffirms the hypothesis that users' perception is affected by the sentiment of the review text and not just the numerical rating, hence signifying the importance of sentiment analysis in this case. The coefficient for Negative Sentiment Strength is greater than that of Positive Sentiment Strength, reaffirming the results of Chevalier and Mayzlin (2006) that negative emotion has more effect than positive emotion on the user.

Discount Rate * Review Volume is a significant predictor (ref Table 6), although Discount Rate is insignificant. Hence, it can be concluded that discount offers are useful when review volume is high, that is, offering discounts on popular products may lead to increased sales. This contradicts Lu et al.'s (2013) findings for restaurant sales, where discount is insignificant when sales volume is high, but affirms Chong et al.'s (2016) results for electronic products.

Finally, the value of R-squared (approx. 0.5) has not improved much even after using 15 predictors including sentiment factors and interaction effects.

6.2 Analysis of decision tree based model

Results of decision tree-based model/classifier generated using SPSS Modeler 16.0 from IBM, are displayed in Table 7. For this machine learning based model, the training set comprises 982 instances while the testing set had 426 instances (total 1408 books). The input parameters are as in Table 1 including the interaction effects described in Sect. 6.1.2. The accuracy of prediction of this model is 66.9%, which means that 66.9% of instances of the testing set were correctly classified by the model.

Predictor importance of the input variables for Decision Tree model is displayed in Fig. 2.

It is observed from Fig. 2 that Review Volume is the most important predictor. Negative Sentiment Strength is more important than Positive Sentiment Strength, and once again the Interaction of Review Volume with Negative Sentiment Strength is much more important than its interaction with Positive Sentiment Strength, suggesting greater impact of negative sentiment. The interaction between Discount Rate and Average Ratings is more important than both Discount Rate and Average Ratings.

The interaction between Discount Rate and Positive Sentiment Strength is less important than both variables individually, suggesting that discount offers do not

Table 7 Results of decision tree classification

Partition	Testing set (number of instances)	Percentage
Correct	285	66.9
Wrong	141	33.1
Total	426	100

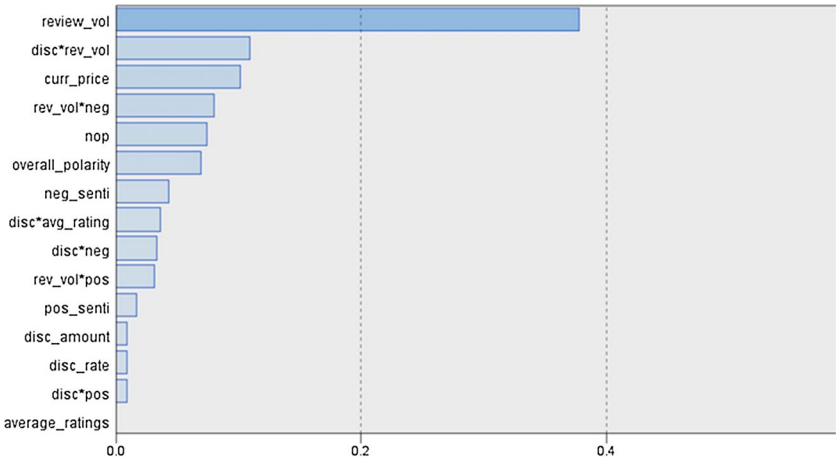

Fig. 2 Predictor importance for decision tree

affect sales when the product has good reviews. However, the interaction between Discount Rate and Negative Sentiment Strength is more important than Discount Rate, suggesting that discount offers may be useful for increasing sales when product reviews are negative.

The accuracy for this classifier is 66.9%, but, considering that Sales Rank has been mapped to 3 bins (refer 5.2.2) only, rather than predicting the actual value, it is not satisfactory. Hence, the ANN based model is explored next as neural networks can be used to predict continuous variables.

6.3 Analysis of artificial neural network based model

The ANN based model can predict the Sales Rank just like multiple regression because it can predict continuous variables. The accuracy of the ANN based model is displayed in Fig. 3 below. A single network had an accuracy of 60.9%, (Fig. 3, "Reference Model") but an ensemble model (Fig. 3, "Ensemble") created using 100 ANN models formed with boosting, increased the accuracy to 70.8%. The model is generated using SPSS Modeler 16.0 from IBM.

Predictor importance of input variables for the ANN based model is given in Table 8 below:

The results generated from the ANN based model are different from the regression as well as decision tree-based models. Positive Sentiment Strength, Negative Sentiment Strength, Number of Pages and Average Ratings are not important at all (Predictor importance = 0.0, Table 8) and Overall Polarity slightly more important (Predictor importance = 0.02). This suggests that the sentiment of review text is not very important, nor is review valence. The most important attribute of online UGC seems to be review volume (Predictor importance = 0.17), which is generally interpreted as an indicator of product popularity.

The interaction between Review Volume and Sentiment Strengths (both Positive and Negative) seems to be important, and so is the interaction between Discount

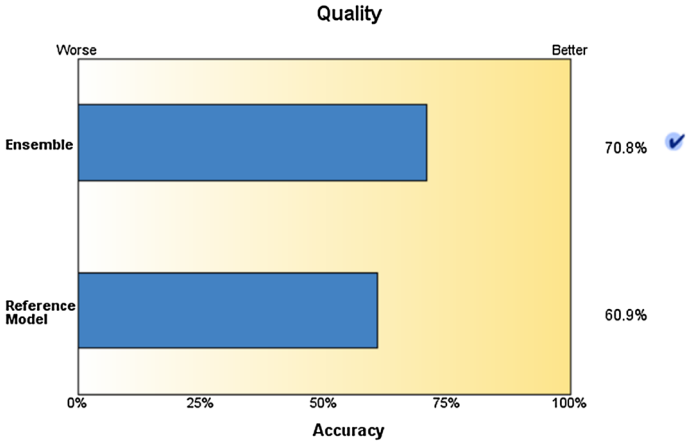

Fig. 3 Accuracy of neural networks

Table 8 Predictor importance for neural networks

Variable	Predictor importance
Current price	0.09
Discount amount	0.01
Discount rate	0.06
Review volume	0.17
Average rating	0.0
Number of pages	0.0
Positive sentiment strength	0.0
Negative sentiment strength	0.0
Overall polarity	0.02
Discount rate * review volume	0.16
Discount rate * average ratings	0.06
Review volume * positive sentiment strength	0.17
Review volume * negative sentiment strength	0.16
Discount rate * positive sentiment strength	0.06
Discount rate * negative sentiment strength	0.04

Rate and Review Volume. Discount rate appears to be a much more important predictor than Discount Amount, even though the two variables are highly correlated.

There seems to be no distinction between Positive Sentiment Strength and Negative Sentiment Strength, thus contradicting the findings of Chevalier and Mayzlin (2006) that negative sentiment has more impact on user perception. Further, Discount Rate*Positive Sentiment Strength is more important than Discount Rate*Negative Sentiment Strength, suggesting that discount offers for products with good reviews possibly lead to a greater increase in sales.

7 Conclusion and implications

In summary, the comparative analysis of three modeling techniques, namely, regression analysis, decision tree and ANN, presents the importance of various predictors and their interactions in the context of book sales at amazon.in. Several conclusions can be drawn by comparing the importance of various predictors as reported by three models.

Firstly, as discount amount and discount rate both are reported as not significant (based on p values) by the regression analysis and as less important (in the range 0.01–0.06) by decision tree and ANN models, only discounting prices of books on online marketplaces may not boost sales. Hence book sales at amazon.in, which is a very popular online marketplace, shows a contradictory behavior as compared to an earlier study on impact of discounting on online product sales (Gong et al. 2015). Nevertheless, ANN also reported discount rate as a more important predictor than discount amount, which can be useful for managers.

Secondly, interaction of discount rate with review volume is found to be an important predictor by all three models and hence demonstrates the effect of review volume, which is detected as an important stand-alone predictor by all three models, in the context of book sale at amazon.in.

One important goal of this research was to investigate the effects of review sentiment parameters computed from review text, namely, overall polarity, positive sentiment and negative sentiment, on book sales. In this regard, although regression analysis demonstrates the importance of these factors (Coefficient range, 0.18–0.26), both the advanced machine learning models, i.e. decision tree and ANN, did not project these factors as important (Importance range, 0.0–0.05) at all. This is quite opposite to various extant research (ref Sect. 2.2) which has found that sentiment parameters do affect sales. The findings for decision tree and ANN are quite interesting and seems to originate from the fact that there is a concern among buyers about fake review content and hence they do not care much about the sentiment of these reviews. Another possible reason may be that once the customer is satisfied with the current price of the book (which is an important factor as reported by all three models), the review sentiments become unnecessary. In this regard, the average rating, which, in a way, also captures sentiment, is not found important at all by all three models, highlighting the fact that numeric ratings do not affect sales of books.

Although the sentiment related parameters are not important as stand-alone factors, interaction of positive and negative sentiment parameters with review volume is found to be an important factor by all three models. This shows that review volume has the capacity to moderate discount rate, positive sentiment and negative sentiment. Similarly, interaction of positive and negative sentiments with discount rate is also found to be important by all three models.

Thus, the comparative analysis of three models, namely, regression analysis, decision tree and ANN, helps in triangulation of results regarding importance of predictors and interaction effects. The implications of these observations for business managers/online sellers for books on amazon.in are as follows:

- Higher discount rates might be more important than discount amounts, so it might be better to offer a discount of INR 70 on a book with an M.R.P. of INR 200 than on one with an M.R.P. on INR 1000, as it would translate to a higher discount rate on the former. Also, in the websites, instead of showing the reduced price, the discount rate can be shown. But only discounting, be it by amount or rate, may not be very effective in boosting sales.
- The interaction of Review Volume and Discount Rate is one of the most important predictors of Sales Rank. This suggests that offering discounts for books with large number of reviews may be more beneficial. Secondly, marketers should nudge buyers to write reviews to increase review volume.
- By similar logic, discount offers may yield better results (in terms of better Sales Rank) when offered for products with good reviews.
- Possibly because books are not high value products, and there is a concern about fake reviews, negative reviews are not affecting the sales significantly. Hence, managers need not be too worried about negative reviews unless their volume is large.

8 Future scope of research

There is scope of future research in multiple areas. Firstly, the use of Sales Rank as a proxy for sales is only an approximation of actual sales, which is necessitated by the fact that actual sales data is not available. Availability of actual historical sales data would greatly improve the confidence on the results.

Secondly, the current study has examined only books on amazon.in; further studies could use the same methodology for different product categories on different e-commerce websites to compare the results. In addition, there is scope for similar studies for various services as well and not just products.

Thirdly, the data sample used for building models uses information only from 1408 paperback book titles. Future experiments can increase the sample size and consider various types of books, namely, fiction, non-fiction, various types of book editions, namely hard-bound, kindle version etc. Thus, further studies can focus on a larger dataset and apply more sophisticated models based on deep learning. This will help in understanding the online sales prediction more accurately and develop guidelines for marketers.

Fourthly, the study can be conducted at different snapshots of time to understand whether the findings are similar or not. The time lag between reviews and sales rank can also be varied to estimate its effect more accurately on the quality of prediction.

Finally, other predictor variables such as genre, reviewer characteristics etc. can also be included in the prediction model. Country or culture specific studies can also be done to find out if the results vary or not.

It is evident from this research that interaction effects are going to play an important role in predicting sales of online products. As machine learning based models, such as ANN, are known to handle these complexities better than popular models such as regression analysis, it would be advisable for marketers to adopt these techniques for modeling sales. Hence this approach can be applied to other products as well as services to predict sales more accurately and understand the importance of interaction effects better for controlling the relevant market parameters.

References

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. Manag Sci 57(8):1485–1509

Beautiful Soup (2017). Beautiful soup documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc. Accessed 1 Dec 2017

Chen S-FS, Monroe KB, Lou Y-C (1998) The effects of framing price promotion messages on consumers' perceptions and purchase intentions. J Retail 74(3):353–372

Cheung CM, Thadani DR (2012) The impact of electronic word-of-mouth communication: a literature analysis and integrative model. Decis Support Syst 54(1):461–470

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: online book reviews. J Mark Res 43(3):345–354

Chong AYL (2013) Predicting m-commerce adoption determinants: a neural network approach. Expert Syst Appl 40(2):523–530

Chong AYL, Zhou L (2014) Demand chain management: relationships between external antecedents, web-based integration and service innovation performance. Int J Prod Econ 154:48–58
Chong AYL, Ooi KB, Sohal A (2009) The relationship between supply chain factors and adoption of e-collaboration tools: an empirical examination. Int J Prod Econ 122(1):150–160
Chong AYL, Li B, Ngai EW, Ch'ng E, Lee F (2016) Predicting online product sales via online reviews, sentiments, and promotion strategies: a big data architecture and neural network approach. Int J Oper Prod Manag 36(4):358–383
Chong AYL, Ch'ng E, Liu MJ, Li B (2017) Predicting consumer product demands via big data: the roles of online promotional marketing and online reviews. Int J Prod Res 55(17):5142–5156
Cui G, Lui HK, Guo X (2012) The effect of online consumer reviews on new product sales. Int J Electron Commer 17(1):39–58
Davis A, Khazanchi D (2008) An empirical study of online word of mouth as a predictor for multi-product category e-commerce sales. Electron Mark 18(2):130–141
Dellarocas CN, Awad N, Zhang X (2004) Using online reviews as a proxy of word-of-mouth for motion picture revenue forecasting. SSRN Electron J. http://www.ssrn.com/abstract=620821
Doern RR, Fey CF (2006) E-commerce developments and strategies for value creation: the case of Russia. J World Bus 41:315–327
Drozdenko R, Jensen M (2005) Risk and maximum acceptable discount levels. J Prod Brand Manag 14(4):264–270
Duan W, Gu B, Whinston AB (2008) The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry. J Retail 84(2):233–242
Faryabi M, Sadeghzadeh K, Saed M (2012) The effect of price discounts and store image on consumer's purchase intention in online shopping context case study: Nokia and HTC. J Bus Stud Q 4(1):197
Floyd K, Freling R, Alhoqail S, Cho HY, Freling T (2014) How online product reviews affect retail sales: a meta-analysis. J Retail 90(2):217–232
Gaikar D, Marakarkandy B (2015) Product sales prediction based on sentiment analysis using Twitter data. Int J Comput Sci Inf Technol 6(3):2303–2313
Gendall P, Hoek J, Pope T, Young K (2006) Message framing effects on price discounting. J Prod Brand Manag 15(7):458–465
Ghose A, Ipeirotis P (2006) Designing ranking systems for consumer reviews. The impact of review subjectivity on product sales and review quality. In: Proceedings of the 16th annual workshop on information technology and systems. http://pages.stern.nyu.edu/~aghose/wits2006.pdf. Accessed 1 June 2017
Gong J, Smith MD, Telang R (2015) Substitution or promotion? the impact of price discounts on cross-channel sales of digital movies. J Retail 91(2):343–357
Gupta S, Cooper LG (1992) The discounting of discounts and promotion thresholds. J Consum Res 19:401–411
Hancock JT, Gee K, Ciaccio K, Lin JM-H (2008) I'm sad you're sad: emotional contagion in CMC. In: Proceedings of the 2008 ACM conference on computer supported cooperative work. ACM, pp 295–298
Hu N, Bose I, Koh NS, Liu L (2012) Manipulation of online reviews: an analysis of ratings, readability, and sentiments. Decis Support Syst 57:42–53
Hu N, Koh NS, Reddy SK (2014) Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. Decis Support Syst 57:42–53
Ito TA, Larsen JT, Smith NK, Cacioppo JT (1998) Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. J Personal Soc Psychol 75(4):887
Jain R, Kulhar M (2015) Growth drivers of online shopping in small cities of India. Int J Adv Res Comput Sci Manage Stud 3(9):80–87
Kramer ADI, Guillory JE, Hancock JT (2014) Experimental evidence of massivescale emotional contagion through social networks. Proc Natl Acad Sci 111(24):8788–8790
Lee J, Park DH, Han I (2008) The effect of negative online consumer reviews on product attitude: an information processing view. Electron Commer Res Appl 7(3):341–352
Leino J, Raiha K (2007) Case Amazon: ratings and reviews as part of recommendations. In: RecSys. ACM, pp 137–140
Li X, Hitt LM (2010) Price effects in online product reviews: an analytical model and empirical analysis. MIS Q 34(4):809–831
Lichtenstein DR, Netemeyer RG, Burton S (1990) Distinguishing coupon proneness from value consciousness: an acquisition-transaction utility theory perspective. J Mark 54(3):54–67
Lu X, Ba S, Huang L, Feng Y (2013) Promotional marketing or word-of-mouth? Evidence from online restaurant reviews. Inf Syst Res 24(3):596–612

Ludwig S, Ruyter K, Friedman M, Bruggen EC, Wetzels M, Pfann G (2013) More than words: the influence of affective content and linguistic style matches in online reviews on conversion rates. J Mark 77:87–103
Marshall R, Leng SB (2002) Price threshold and discount saturation point in Singapore. J Prod Brand Manag 11(3):147–159
McNeill L (2013) Sales promotion in Asia: successful strategies for Singapore and Malaysia. Asia Pac J Mark Logist 25:45–69
Mudambi S, Schuff D (2010) What makes a helpful review? A study of customer reviews on Amazon.com. MIS Q 34(1):185–200
NLTK. Natural language toolkit documentation. https://www.nltk.org/doc
Online Influence Trend Tracker Report (2011). http://conecomm.com
Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135
Requests (2017) Requests: HTTP for humans. http://docs.python-requests.org/en/master
Salehan M, Kim DJ (2016) Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics. Decis Support Syst 81:30–40
Santibanez SF, Kloft M, Lakes T (2015) Performance analysis of machine learning algorithms for regression of spatial variables. A case study in the real estate industry
Schneider MJ, Gupta S (2016) Forecasting sales of new and existing products using consumer reviews: a random projections approach. Int J Forecast 32(2):243–256
Sentiment API (2017) Natural language processing APIs. https://www.text-processing.com/docs/sentiment.html. Accessed 15 Dec 2017
SentiStrength. (2017). SentiStrength. http://sentistrength.wlv.ac.uk. Accessed 15 Dec 2017
Social Media Report 2012: social media comes of age (2012). http://www.nielsen.com
Tang T, Fang E, Wang F (2014) Is neutral really neutral? The effects of neutral user-generated content on product sales. J Mark 78(4):41–58
Tsai W-C (2001) Determinants and consequences of employee displayed positive emotions. J Manag 27(4):497–512
Xu N, Bai SZ, Wan X (2017) Adding pay-on-delivery to pay-to-order: the value of two payment schemes to online sellers. Electron Commer Res Appl 21:27–37
Yang J, Kim W, Amblee N, Jeong J (2012) The heterogeneous effect of WOM on product sales: why the effect of WOM valence is mixed? Eur J Mark 46(11/12):1523–1538
Yao R, Chen J (2013) Predicting movie sales revenue using online reviews. In: 2013 IEEE international conference on granular computing (GrC). IEEE, pp 396–401)
Yu X, Liu Y, Huang X, An A (2012) Mining online reviews for predicting sales performance: a case study in the movie domain. IEEE Trans Knowl Data Eng 24(4):720–734
Zhou ZH, Jiang Y (2004) NeC4. 5: neural ensemble based C4. 5. IEEE Trans Knowl Data Eng 16(6):770–773
Zhou ZH, Wu J, Tang W (2002) Ensembling neural networks: many could be better than all. Artif Intell 137(1–2):239–263
Zhu F, Zhang X (2010) Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics. J Mark 74(2):133–148

Affiliations

Satyendra Kumar Sharma[1] · Swapnajit Chakraborti[2] · Tanaya Jha[3]

✉ Swapnajit Chakraborti
swapnajit.chakraborti@spjimr.org

Satyendra Kumar Sharma
satyendrasharma@pilani.bits-pilani.ac.in

Tanaya Jha
f2013304@pilani.bits-pilani.ac.in

[1] Department of Management, Birla Institute of Technology and Science (BITS Pilani), Pilani, Rajasthan 333031, India

[2] Information Management, S. P. Jain Institute of Management and Research (SPJIMR), Munshi Nagar, Andheri (W), Mumbai, India

[3] Department of Computer Science, Birla Institute of Technology and Science (BITS Pilani), Pilani, Rajasthan 333031, India