A Novel Approach to Identify the Determinants of Online Review Helpfulness and Predict the Helpfulness Score Across Product Categories

Debasmita Dey(✉) and Pradeep Kumar

Indian Institute of Management Lucknow, Lucknow, UP, India
{fpml7006,pradeepkumar}@iiml.ac.in

Abstract. The proliferation in the number of available online reviews provides an excellent opportunity to use this accumulated enormous information of any product in a more strategic way to improve the quality of the product and services of the e-commerce company. Due to the non-uniform quality of online reviews, it is crucial to identify those helpful reviews from the pile of a large amount of low quality and low informative other reviews. This system will help the customers to form an unbiased opinion quickly by looking at its level of helpfulness. The e-commerce companies measure the helpfulness of a review using the number of votes it gets from other customers. This situation arises problems to newly-authored potentially helpful reviews due to lack of votes. Thus it is essential to have an automated process to estimate and predict helpfulness of any review. This paper identifies the essential characteristics of online reviews influencing the helpfulness of it. This study categorized all characteristics of reviews collected from previous literature in four main categories and then study the combined effect of the four aspects in predicting the helpfulness of a review. The product type (Search or Experience) acts as a control variable in the factors identification model of helpful prediction of a review. An analysis of total 14782 reviews from Amazon.com across five different product category shows the factors influencing the helpfulness of a review varies across product categories. Then a comparative study of two widely used machine learning, Artificial Neural Network and Multiple Adaptive Regression Spline are presented to predict the helpfulness of online review across five different categories and a better method of predicting helpfulness of online reviews are suggested based on the type of product. This study solves the starvation problem of potential newly-authored or infamous reviews without any manual votes along with high accuracy of helpfulness prediction.

Keywords: Online reviews · Helpfulness prediction · Artificial Neural Network · Multiple Adaptive Regression Splines

S. Madria et al. (Eds.): BDA 2019, LNCS 11932, pp. 365–388, 2019.
https://doi.org/10.1007/978-3-030-37188-3_21

1 Introduction

An increasing number of online reviews, e-commerce websites have offered customers their platforms to give their opinions and reviews of products, services, and the seller of the product. These reviews are given by those customers who have already bought or used the products from that e-commerce website or any other sources and are considered as a proxy for the product quality in offline word-of-mouth (WoM) communication. Online customer reviews are defined as "peer-generated product evaluations posted on company or third party websites" [17]. Ecommerce websites allow the customer to share their views about the products and services provided by them in two ways: (a) by numerical star ratings ranging 1 to 5, (b) by providing an area to write your opinion about the product. The availability of customer reviews on an e-commerce website has proved to make a better perception in the customer about its importance, social presence [1], and "stickiness" (time spent on a particular e-commerce site). Increasing the availability of online reviews also proliferate the opportunity of using it more strategically to improve the product and service of the e-commerce company. Now, the quality of these online reviews is not likely to be uniform, and this could range from an excellently vivid description and evaluation to spam with no value addition to make any decision about the purchase. The reviews which are helpful to customers in decision making generally lies under the heap of a large amount of low quality and low informative or fake (and spam) reviews and hence it is challenging for customers to identify and use those helpful reviews to form an unbiased opinion about any product or service.

To solve this problem, e-commerce websites allow the customer to vote for or against the review to show their support or disagreement with that particular review content. For example, "100 out of 140 people found the review helpful" shows the fact that 100 people apart from the reviewer found the review helpful to make their decision and rest 40 people did not find it helpful. The website also allows sorting all customer reviews as per helpfulness of it. However, there is found no such theoretically grounded specific explanation of what are the factors determining the helpfulness of a review and how they are being calculated to sort it. A massive portion of reviews contains no votes or have few votes, and this makes it more challenging to understand their helpfulness and social validity. As per Yang et al. [29], only 20% of the reviews from Amazon dataset [6] have more than five votes, and rest of them either do not have any votes or lesser than five votes. This situation arises problems to the newly authored potential reviews, which does not get the chance to be read by other customers due to lack of votes or infamous products. Therefore, it is essential to have an automated system to estimate helpfulness of any review instead of some manual process (Fig. 1).

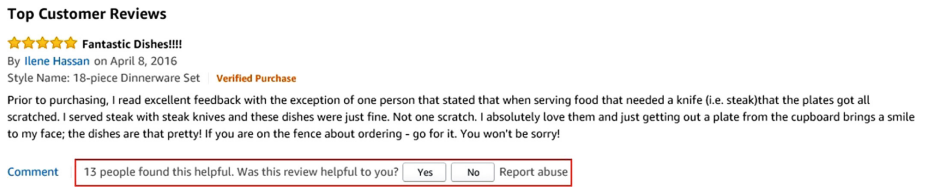


Fig. 1. An example of the structure of a review on Amazon.com.

The purpose of this study predicts a score denoting the helpfulness of a review automatically by identifying and analyzing linguistic, psychological, and peripheral factors of any review. Previous researches on this issue mostly address either linguistic determinants of psychological features affecting the helpfulness of any review. This study allows focusing on all possible aspect of factors possibly drive the helpfulness of online reviews together. The four identified an aspect of factors that determines helpfulness can be categorized into four types, e.g., Linguistic, Psychological, Text complexity, and peripheral cues. This study identifies the fact that considering only one type of factor to determine review helpfulness might neglect the other aspect of features, which can play a significant part in it. Therefore, apart from the confirmed variables (Rating, Positive emotion, Negative emotion etc.) from previous literature, this study considered not only considered the linguistic features (word count, word per sentence, adjective, etc.) but also the psychological (Analytical thinking, Tone, Authenticity, and Confidence) thought process of the reviewer.

The previous studies [12, 13, 17] also address the fact that product type act as a control variable, and for different types of products (Search good or Experience good), the factors driving helpfulness can differ. Therefore, this study considers this constraint and chooses five different categories of products to address whether and how the determinants of review helpfulness differ as per product type. For example, a highly analytical review may be more useful for cellphone category product customers than the grocery category of products. The five categories are chosen in such a way that it includes not only pure search type (cellphone, clothing) or experience type (beauty, grocery) of good but also products which do not have any physical presence (digital music) and can be considered both search and experience type.

Our research addresses three research questions. First, what are the linguistic and psychological features across different product categories and whether it is needed to study them separately or not? Second, what are the determinants driving the perceived helpfulness of an online review based on different product categories? Third, which method, among the two most used machine learning supervised methods, better predict the helpfulness of an online review?

To address these three research questions, a significantly large number of online reviews from Amazon.com are collected, and these reviews belong to five different categories of products (beauty, grocery, cellphone, clothing, and digital music). The linguistic and psychological variables of online reviews of each product category are extracted using R and Linguistic Inquiry and Word Count (LIWC) software. Then, to address our first research question, one-way ANOVA is performed across all five product categories on all the chosen variables (linguistics and linguistics, etc.) from literature. Next, to explore the factors determining the review helpfulness, the most widely adopted method, Linear regression (LR), is used. Due to the non-linear structure of the data, LR performed poorly and hence a wrapper built around Random Forest classification algorithm, Boruta, is performed, and a subset of variables are selected to predict the helpfulness of reviews of that particular product category reviews. Finally, two widely used machine learning algorithms, Artificial Neural Network (ANN), and Multiple Adaptive Regression Splines (MARS) are used to predict the helpfulness of online reviews using R 3.4.4 software. These two prediction methods are then compared in terms of their Mean Squared Error (MSE) to select the better performing model for each of five product categories.

In short, this study helps to process a large number of online reviews efficiently in an automated way even if reviews do not have any manually entered votes and generate a helpfulness score based on different types of characteristics of a review for each written review. Customers can quickly sort the reviews as per their helpfulness score to make a better purchase decision.

This paper is organized in the following sections. Section 2 presents the previous literature reviews related to these topics. Section 3 presents the method of data collection. Section 4 presents the selection of various variables for this study from the literature. Section 5 gives a clear stepwise idea of the research methodology to solve our research questions. Section 6 presents the result of the experiments and the analysis of the results. Finally, the implication of this study and future scopes are concluded in Sect. 7.

2 Literature Review

Previous researches on the helpfulness of reviews were mainly addressed two typical questions: (i) finding out the critical factors influencing review helpfulness, and (ii) propose a suitable method to predict review helpfulness. The next two sections will address the studies focused on these two issues separately.

2.1 Important Factors Influencing Review Helpfulness

A customer review can have a set of different features, such as Numerical rating, Number of words used in that review, Polarity (Positive, Negative) of the review content, the readability of the text, Style of the writing, etc. Mudambi and Schuff [17] studied the influence of several words in a review, as well as the extremity of a review on the helpfulness of that review. They experimented with their hypothesis using Amazon.com review datasets. The results of this experiment (Tobit Regression) indicate a positive relationship between several words in a review and helpfulness of that review by considering the product type (search or experienced) as a moderator. Korfiatis et al. [10] studied the effect of readability of the review content with the helpfulness of it. In this paper, they measured readability in following four ways: Gunning Fog Index, Flesch reading scale, Coleman-Liau index, and Automated readability index and suggested that readability of review has a more significant impact on helpfulness than several words in it. Ghose and Ipeirotis [3] considered reviewer information, subjectivity, spelling error, and another six type of readability index to study the relationship of these factors with the helpfulness of review. Their study also shows a significant influence of these six readability indexes on review helpfulness. Krishnamoorthy [11], in his research, considered a set of linguistic characteristics (adjective terms, state and action verbs) and compared with other factors (readability, subjectivity). The results from his study showed that rather than considering only readability measure and numerical rating as essential factors, a hybrid model with some linguistic variables could better explain review helpfulness in terms of predictive accuracy. Ghose and Ipeirotis [3] and Forman [2] et al. also supported this explanation given by Krishnamoorthy, showing that a combination of subjective and objective features of review better explains the review helpfulness than considering any of them separately.

In our study, we adopted linear regression primarily to understand the relationship between the dependent (Helpfulness ratio) and an independent variable. Then Boruta algorithm [14] is applied to measure the variable importance as a better explanation of the dependent variable. The methods addressed in the literature did not take into the nonlinear data structure of the Amazon.com dataset. Thus, using linear regression or other methods with linearity assumption will not be appropriate to derive important factors of review helpfulness.

2.2 Prediction Methodology of Review Helpfulness

Mudambi and Schuff [17], Yin et al. [30], Yang et al. [29], Korfiatis et al. [10] and Forman et al. [2] defined helpfulness of a review as the ratio of helpful votes to total number of votes (Helpful votes + Unhelpful votes). They mostly adopted the commonly used method, Linear Regression (LR), to examine the critical factors and prediction of the helpfulness of reviews. Forman et al. [2] transformed the helpfulness

ratio in two-class, Helpful, and Unhelpful based on a threshold value and then predicted the helpfulness using linear regression.

Another commonly used method in this research is Support Vector Machine (SVM), as it handles both linear and nonlinear data. Kim et al. [9] applied Support Vector Regression (SVR) for review prediction. Hu and Chen [4] adopted M5P, SVR, and linear regression to measure their prediction performance and did a comparative analysis. Krishnamoorthy [11] used three techniques, e.g., Support vector classification, Random forest, and Naïve Bayes, to predict review helpfulness and then compared them to propose the best prediction model.

Khashei and Bijari [8] proposed a prediction model using Artificial Neural Network (ANN) due to its data-driven and self-adaptive features. Lee et al. [15] adopted a multilayer perceptron neural network (BPN) to predict review helpfulness and compared it with linear regression analysis.

In our study, we adopted three methods (linear regression, Multiple Adaptive Regression Splines, and ANN) and compared their results to find the best suitable method to predict helpfulness. Linear regression was chosen as it is a convenient method addressed in the literature. The other two methods (Multiple Adaptive Regression Splines and ANN) were selected due to their capability of handling nonlinearity in data and better accuracy in prediction.

3 Data Collection

We gathered data for this study from http://jmcauley.ucsd.edu/data/amazon/ [6] since 2005–2014. Amazon product reviews are collected category wise, and the categories are Beauty, Grocery, Cell phone, Clothing, and Digital music. These five product types were chosen in the study based on the following reasons:

- These five categories contain a large number of customer reviews to be analyzed and modeled for training and testing purposes.
- Based on the Nelson [18, 19] study, we included both search (Cell phone) and experience (beauty, grocery) type of products category.
- Our study addresses the fact that a product can exist along a continuum from simple search to pure experience type of product and hence, considers product categories involving mix (Clothing) of search and experience features.
- Digital products are the latest kind of products in the market with no physical presence. Thus, digital music category reviews are included in this study.

For each category mentioned above, all reviews are collected along with their respective numerical rating, Title, review text, the number of helpful and unhelpful votes. After preprocessing and cleansing of data, we excluded those reviews from analysis, which does not have a minimum of ten votes as minimal helpful votes can introduce biases in the model.

A total of 14782 reviews are finally collected after pre-processing to be analyzed after removing all reviews having lesser than ten votes (Table 1).

Table 1. Total number of reviews collected from each of five categories.

	Beauty	Grocery	Cell phone	Clothing	Digital music
Number of reviews	4139	2477	2442	3844	1880

The structure of the data collected is presented below (Fig. 2):

- Product ID: A2ENZ4FESUXXMT
- Reviewer ID: 1400501466
- Review Text: I was looking at expensive tablets that were more like mini notebook computers. I already have a high end notebook. I wanted something that was very portable and not combersome to take on trips, go to coffee shops and etc. I wanted the ability to get email, do limited surfing and read books. The Nook works flawlessly and the display is really nice. I have an N protocol router and the Nook is quick on the Net. I read some negative reviews here. They appear to be written by folks who want to take a $200 unit and turn it into a $500 unit with various apps and other applications. Here's a news flash for the naysayers. Go out and buy the $500 unit and quit complaining. If you want to read books, surf and get email, you'll like this unit.
- Rating: 5.0
- Review Time: 3 December 2012
- The number of votes on helpfulness: 9
- Total number of votes: 10

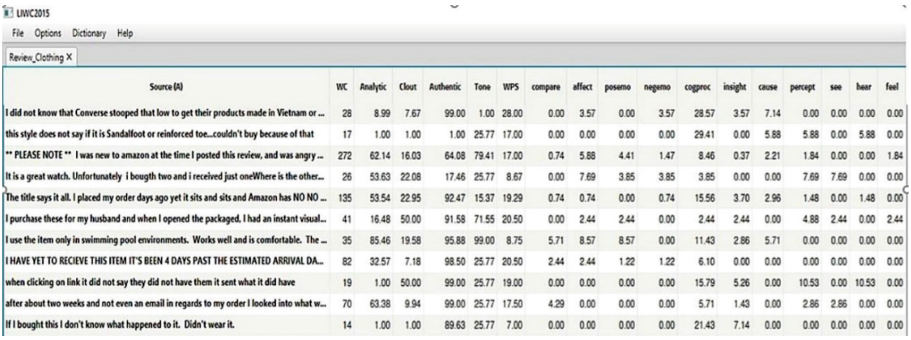
Fig. 2. An example of the structure of data collected from Amazon.com.

4 Variable Selection

Due to the unstructured form of review text apart from some explicit information (Numerical rating, Number of helpful votes, and several total votes), the review text is transformed into a standard structural format using LIWC (Linguistic Inquiry and Word Count) 2015 software.

LIWC is a text analysis software proposed by Pennebaker [22] to transform the unstructured text into approximately 90 output scores. This 90 output variable evaluates not only the structural and style feature if the text, but also the psychological thought process of reviewer [23–25, 27]. This tool is extensively used and validated by many articles and research papers [7, 20].

The LIWC output is shown below (Fig. 3):



LIWC2015

File Options Dictionary Help

Review_Clothing ×

Source (A)	WC	Analytic	Clout	Authentic	Tone	WPS	compare	affect	posemo	negemo	cogproc	insight	cause	percept	see	hear	feel
I did not know that Converse stooped that low to get their products made in Vietnam or ...	28	8.99	7.67	99.00	1.00	28.00	0.00	3.57	0.00	3.57	28.57	3.57	7.14	0.00	0.00	0.00	0.00
this style does not say if it is Sandalfoot or reinforced toe...couldn't buy because of that	17	1.00	1.00	1.00	25.77	17.00	0.00	0.00	0.00	0.00	29.41	0.00	5.88	5.88	0.00	5.88	0.00
** PLEASE NOTE ** I was new to amazon at the time I posted this review, and was angry ...	272	62.14	16.03	64.08	79.41	17.00	0.74	5.88	4.41	1.47	8.46	0.37	2.21	1.84	0.00	0.00	1.84
It is a great watch. Unfortunately I bought two and i received just oneWhere is the other...	26	53.63	22.08	17.46	25.77	8.67	0.00	7.69	3.85	3.85	3.85	0.00	0.00	7.69	7.69	0.00	0.00
The title says it all. I placed my order days ago yet it sits and sits and Amazon has NO NO ...	135	53.54	22.95	92.47	15.37	19.29	0.74	0.74	0.00	0.74	15.56	3.70	2.96	1.48	0.00	1.48	0.00
I purchase these for my husband and when I opened the packaged. I had an instant visual...	41	16.48	50.00	91.58	71.55	20.50	0.00	2.44	2.44	0.00	2.44	2.44	0.00	4.88	2.44	0.00	2.44
I use the item only in swimming pool environments. Works well and is comfortable. The ...	35	85.46	19.58	95.88	99.00	8.75	5.71	8.57	8.57	0.00	11.43	2.86	5.71	0.00	0.00	0.00	0.00
I HAVE YET TO RECIEVE THIS ITEM IT'S BEEN 4 DAYS PAST THE ESTIMATED ARRIVAL DA...	82	32.57	7.18	98.50	25.77	20.50	2.44	2.44	1.22	1.22	6.10	0.00	0.00	0.00	0.00	0.00	0.00
when clicking on link it did not say they did not have them it sent what it did have	19	1.00	50.00	99.00	25.77	19.00	0.00	0.00	0.00	0.00	15.79	5.26	0.00	10.53	0.00	10.53	0.00
after about two weeks and not even an email in regards to my order I looked into what w...	70	63.38	9.94	99.00	25.77	17.50	4.29	0.00	0.00	0.00	5.71	1.43	0.00	2.86	2.86	0.00	0.00
If I bought this I don't know what happened to it. Didn't wear it.	14	1.00	1.00	89.63	25.77	7.00	0.00	0.00	0.00	0.00	21.43	7.14	0.00	0.00	0.00	0.00	0.00

Fig. 3. An example of the output processed by LIWC software.

All variables are selected from previous literature studies and are categorized into four broad categories: (i) Linguistic, (ii) Psychological, (iii) Text complexity, and (iv) Peripheral cues.

The linguistic category of variables is based on the structure of sentence, punctuations, part of speech, polarity or tone of sentences, etc. Example: pronoun, article, preposition, auxiliary verb, word count, word per sentence, adjective.

The psychological category of variables focuses on feeling and thought processes using semantics. Example: comparative words (bigger, better), Analytic, Clout, Percept, Positive emotion, Tone, Negative emotion.

The text complexity category of variables includes those helping the review to understand or read easily or with difficulty. Example: Flesch reading ease index, Syllable, Dictionary word.

The peripheral ques category contains those variables which are independent of review text. Example: Rating, Time of the review posted.

In our study, the dependent variable (Helpfulness) is defined as the ratio of several helpful votes to the total votes (helpful +Unhelpful). For example, for a review where "100 people out of 150 found this review helpful", the helpfulness of the review will be (100/150) 0.67.

The independent variables are chosen from different kinds of literature listed below:

- Linguistic variables:
  - *Compare:* It is defined as the total number of comparative words (bigger, smaller, greater, etc.) used in the review.
  - *Pronoun:* It is measured by the number of pronouns in the text.
  - *Ppron*: It is defined as the number of personal pronouns (I, you, he, she, etc.) in the review and is calculated as the percentage of several pronouns.
  - *Article:* It is defined as the number of articles (a, an, the, etc.) mentioned in the text.
  - *Preposition:* It is measured by the total number of preposition in the review.
  - *Auxiliary verb:* It is defined as the total number of the auxiliary verb (might, must, could, etc.) used in the review.
  - *Adverb:* It is defined as the total number of adverb verb (very, slowly, quickly, etc.) used in the review.
  - *Adjective:* It is defined as the total number of adjectives (better, bright, dull, thick, etc.) used in the review.
  - *AllPunctuation:* It is defined as the total number of sentences with complete and with grammatically correct punctuations used in the review.
  - *I:* It is defined as the percentage number of occurrences of the word 'I' in the review.
- Psychological variables:
  - *Analytic:* It is known as the categorical dynamic index (CDI) and addresses the level of the formal, logical, and hierarchical thought process of the reviewer. A high score implies more formal, logical, and hierarchical thinking.
  - *Clout:* It is defined as the level of expertise or leadership in some context, or how much one is confident about his or her opinions. A higher clout score indicates a more professional and confident opinion, while a lower score indicates a tentative or humble style.
  - *Tone:* It defines the sentimental and emotional tone of the whole text. A score higher than 50 indicates a positive tone, and lower than 50 scores indicate tone with sadness or anxiety or hostility. The exact score of 50 indicates either a lack of emotion or ambivalence.
  - *Authentic:* It is defined as the level of honesty and disclosing thinking of the reviewer, i.e., expressing more personal, humble, and authentic opinions about something. A higher score indicates more honest and vulnerable thinking.
  - *Cogproc:* It is measured as the ratio words evoking the cognitive thought process (cause, know, etc.) of the thinker. A high cogproc score will indicate a more cognitive opinion rather than thorough normal senses.
  - *Percept:* It is measured as the ratio words evoking perceptual thought process ("look," "feeling," etc.) of the thinker. A high percept score will indicate that the opinion is generated and backed up by using the sensed of the reviewer rather than any cognitive information.
  - *Posemo:* It is measured by the ratio of positive emotion words to the total words in the text. It is identified as one of those confirmed factors determining review helpfulness [5, 11, 16, 19, 21, 24, 26].

  - *Negemo:* It is measured by the ratio of negative emotion words to the total words in the text. It is identified as one of those confirmed factors determining review helpfulness [5, 11, 16, 19, 21, 24, 26].
- Text complexity:
  - *WC:* It is defined as the total number of words in the review. It is a certain factor of review helpfulness studied in previous literature. It is used as a proxy for text complexity [15, 19].
  - WPS: It is defined as the number of words per sentence. It is used for sentence complexity [2, 3, 9].
  - *Sixltr:* It is defined as the number of words longer than six letters and is used as a proxy for word complexity.
  - *Dic:* It is defined as the percentage of target words captured by the LIWC dictionary.
  - *Flesch Kincaid Readability:* It is defined as the measure of difficulty in reading and understanding a text in English. A higher score indicates that the text piece is easy to read and understand.
- Peripheral Cues:
  - *Rating:* Rating is a numeric score (1 to 5) given by the customer. It is identified as a confirmatory factor of review helpfulness as studies in the literature [5, 9, 15].

5 Research Methodology

The research methodology of this study is described here stepwise below:

Firstly, reviews collected were collected and cleaned, and the preprocessed to get a basic format as below to proceed further (Fig. 4):

helpful votes	total votes	Helpfulness	reviewText	overall	summary
24	24	1	I haven't been a big fan of Prada's fragrances over the years but absolutely fell in love with the sweetness and candy-like scent of this perfume! This smells like a sweet, decadent caramel with tones of vanilla and I'm not sure what else, but it smells great! Although, I must say that this seller is asking for WAY too high a price for this bottle! You could get the 2.7 oz bottle for around the same price at Neiman Marcus (the 1.7 oz for around $80 if you prefer a smaller bottle)!	5	Love the smell of this!
11	14	0.785714286	I bought a similar type of dispenser back in 1995, when I was outfitting my new apartment. Perhaps that was the Dispenser Classic I. That one lasted 11 years without a problem and I threw it out only because it looked old and I was moving. Flash forward to 2011. I decided to buy one for my condo. A day or two after I put in the bottles, I noticed shampoo from the dispenser dripping onto the faucet. It turns out that one of the bottles had a hairline crack in it. It was tiny but enough to cause a serious leak. I examined the bottles closely and realized how thin and cheap the plastic containers really are. Unbelievable. Now I have to decide whether to use Crazy Glue or go through the hassle of emptying the two other containers and sending the whole thing back to Amazon for a refund. If this is the Classic III, I wonder how bad IV will be. I'd be hesitate to order this again.	2	spensers Made of Cheap, Fragile Plastic - Beware!
17	17	1	This really does cover under eye circles and redness but a little goes a long way. Since it's a thick cream stick you should mix in some moisture cream on your hand and smooth it on sparingly. Too much will look cakey and get into wrinkles. Just a very thin layer is best.	5	Excellent
13	13	1	As long as this eyeliner is applied correctly there should be no peeling, fading or flaking. The only time I get some flaking or peeling is if I apply it too thickly. I put this on early in the morning (7am or so) and don't take it off until 11 or 12 pm. And in all that time it stayed where I put it and the color stayed true.I have horrible allergies and I also have tear producing issues...combine these problems and there is really nothing I could wear to line my eyes-until I found Maybelline Lineworks. My eyes can and will at times ooze tears all day.Sometimes it looks as if I'm having a crying jag. BUT even through all this, my liner never goes anywhere! I use Ponds to take it off at night. Easy peasy!The only thing I'm upset with is that Maybelline seems to be dis-continuing many of the colors and while I can understand it, I don't LIKE it.	5	Inexpensive Perfection

Fig. 4. A glimpse of input data after data cleaning and pre-processing.

In the second step, the reviews written in the English language is transformed into a various numeric score using LIWC dictionary. If any target word is matched with the dictionary word, then the corresponding variable's (out of those 90 variables) score is incremented by one. Figure 5 shows an example of this process.

File Options Dictionary Help

Clothing_Shoes_and_Jewelry_5 ×

Word	compare	affect	posemo	negemo	cogproc	insight	cause	percept	see	hear	feel
great		X	X								
deal											
before	X										
purchasing											
and											
even											
got											
free		X	X								
really					X						
do like		X	X								

Fig. 5. Cataloging target words using LIWC 2015 into different linguistic and psychological variables scores.

Next, the exploratory data analysis is performed on our five datasets to calculate the mean and standard deviations of reviews of all categories. To address our first research question, one-way ANOVA is performed on all the variables selected to test whether the product categories are significantly different or not.

In the next step, the determinants of the review helpfulness are explored using linear regression and the Boruta algorithm. Though linear regression is easy to understand and explainable, due to the nonlinear structure of data, it is not suitable to explore the relationship between determinants and the target variable. Hence, Boruta algorithm is performed to explore which independent variables affect the target variable.

In the final step, the helpfulness of online reviews is predicted using the two most widely used techniques for nonlinear data, i.e., Multiple Adaptive Regression Splines (MARS) and Artificial Neural Network (ANN). These two methods to predict online helpfulness of reviews are then compared to determine the suitable method to predict the helpfulness. The helpfulness is predicted considering 70% of training data, and then the mean squared error (MSE) is calculated for both to compare their results for each of five categories.

6 Results and Discussion

6.1 ANOVA Analysis Across Product Categories

Our first research question was whether review characteristics varied across different product categories and if the result is positive, then how they were different. The averages of review features were identified from the literature, and one-way ANOVA was performed to examine the differences, as presented in Table 2. The hypothesis for the ANOVA test is as follow:

H Null: The mean of each feature across five categories are same, i.e.,

$$\mu 1 = \mu 2 = \mu 3 = \mu 4 = \mu 5$$

H Alternative: Means of features across five categories are not all equal.

The ANOVA result shows that the p values for all identified features are less than 0.01, which indicates that all research variables are significantly different at the 99% confidence interval across the five product types. Thus, the null hypothesis is rejected.

The F critical value is 2.37, which is much lesser than the F value of each feature. Therefore, the ANOVA result indicates to the fact that review characteristics vary across different product categories, and so it should be analyzed separately.

The results of Table 2 and Fig. 6 can be interpreted as follows. Product reviews for the Digital Music category (average 260 words) are found to be the longest among the five product types based on WC, and approximately twice the average length of reviews for Clothing (123 words) and Grocery category (153 words). Moreover, WPS (27.48) and Analytic (72.22) for Digital Music are the highest. This means that the reviews for Digital Music are composed of lengthy and analytical sentences. The reason can be explained as reviewers may require more words to write reviews containing the personal experience and analytical expressions of the music for Digital Music category products. The level of Clout shows the highest score (54.67) for Music, but the lowest score (27.40) for Beauty category.

On the other hand, the Authentic scores showed the opposite results. Reviews for beauty have the highest Authentic scores (58.83), while the video has the lowest Authentic (23.70) scores. In other words, product reviews for Digital Music tend to be written expertly, whereas those of beauty is written authentically in a personal manner. Additionally, Tone (75.72) scores are highly positive in all five categories and are highest in the Clothing category pf products. This can be because reviewers express their personal experience of using and fitting of the product as per their product quality more elaborately than Cell phone category (64.33) products. The score of Percept for beauty Category (5.54) is found to be the highest among five categories as reviewers may use many sensory-based expressions such as “looked,” “heard,” or “feeling” for beauty, the quality of which is evaluated based on senses. The Flesch Kincaid score of Digital Music category (12.13) is highest as reviewers mostly write their personal feeling of that music is a very informal easy way to express emotion involved with it.

Table 2. Comparison of the average scores for review variables across five product categories.

		Beauty	Grocery	Cellphone	Clothing	Digital Music	F value	p-value
Rating	Mean	4.225	4.188	4.018	4.173	4.182	10.5	0.0
	SD	1.649	1.809	1.845	1.486	1.619		
WC	Mean	174.74	153.067	253.174	123.769	259.439	304.7	0.0
	SD	22070.1	18958.1	93185.413	12296.95	31688.998		
Analytic	Mean	47.811	59.194	63.488	50.829	72.223	491.2	0.0
	SD	518.890	547.656	487.714	613.524	397.078		
Clout	Mean	27.400	36.640	37.199	33.030	54.675	612.8	0.0
	SD	390.315	446.107	394.097	473.148	269.792		
Authentic	Mean	58.834	37.088	46.555	55.935	23.704	738.6	0.0
	SD	789.420	702.204	684.559	846.301	405.750		
WPS	Mean	20.000	20.437	24.439	18.263	27.481	75.0	0.0
	SD	155.647	195.541	1681.128	158.616	566.723		
Compare	Mean	3.172	3.089	2.836	2.989	3.152	12.2	0.0
	SD	4.186	4.834	3.965	5.195	2.592		
Posemo	Mean	4.026	4.671	3.804	5.176	4.547	130.0	0.0
	SD	6.037	7.685	6.859	11.309	3.849		
Negemo	Mean	1.042	1.177	1.134	0.996	1.436	41.8	0.0
	SD	1.422	1.864	1.504	1.772	1.971		
Cogproc	Mean	12.960	11.646	11.296	11.233	10.089	201.4	0.0
	SD	16.960	17.827	13.371	17.069	10.448		
Percept	Mean	5.541	4.513	4.613	3.841	4.654	161.9	0.0
	SD	11.075	9.562	8.270	9.545	3.801		
Tone	Mean	68.487	72.159	64.327	75.724	72.606	73.5	0.0
	SD	787.490	821.067	788.764	772.176	687.242		
Sixltr	Mean	14.635	15.711	15.182	13.213	16.047	162.7	0.0
	SD	22.145	27.198	22.380	23.451	19.888		
pronoun	Mean	16.614	13.778	13.730	15.517	11.546	516.0	0.0
	SD	18.643	21.445	19.703	21.922	15.737		
ppron	Mean	8.975	7.026	6.749	8.895	5.651	511.5	0.0
	SD	9.916	10.591	9.322	15.960	8.494		
article	Mean	6.297	6.762	8.399	7.437	7.799	273.1	0.0
	SD	6.533	7.210	8.209	9.708	5.002		
prep	Mean	11.660	11.983	12.239	11.431	12.361	45.6	0.0
	SD	8.395	10.621	7.944	12.428	6.269		
auxverb	Mean	8.532	8.326	8.285	8.987	7.526	87.1	0.0
	SD	7.644	9.305	7.076	9.929	5.633		
adverb	Mean	5.488	4.914	5.242	5.757	4.693	78.9	0.0
	SD	5.989	6.181	5.624	8.517	4.073		

(*continued*)

Table 2. (*continued*)

		Beauty	Grocery	Cellphone	Clothing	Digital Music	F value	p-value
adj	Mean	6.675	6.836	5.758	7.460	6.188	138.2	0.0
	SD	8.483	9.906	7.513	11.712	4.599		
AllPunc	Mean	15.931	17.457	15.693	17.005	23.219	329.4	0.0
	SD	37.399	69.431	53.109	79.893	72.967		
verb	Mean	16.226	14.001	14.358	15.946	12.895	387.6	0.0
	SD	13.688	16.025	11.414	15.490	10.974		
i	Mean	6.792	4.412	4.561	5.764	2.212	891.1	0.0
	SD	10.935	8.402	7.771	9.390	3.998		
Dic	Mean	85.479	83.038	81.691	85.385	78.655	533.4	0.0
	SD	31.163	48.898	45.262	33.732	43.154		
Flesch	Mean	8.589	9.149	9.885	7.376	12.132	197.9	0.0
	SD	31.562	29.980	54.283	27.541	74.447		

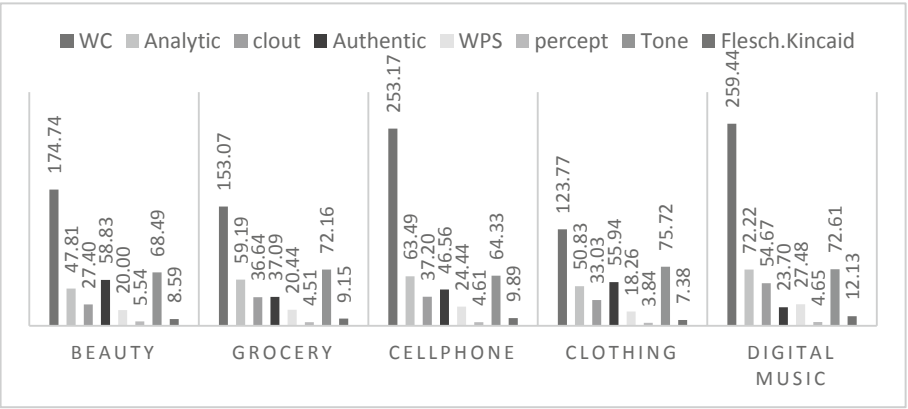


Fig. 6. Comparison of average scores of WC, Analytic, Clout, Authentic, WPS, Percept, Tome, and Flesch Kincaid score across five product categories.

In conclusion, as seen in the previous results, reviews for different product categories have different characteristics. Thus it would be necessary to analyze review helpfulness for each product category separately.

6.2 Factors Determining Review Helpfulness

Our second research question was to identify the determinant factors in the perceived helpfulness of reviews depending on their product category. To do so primarily, Linear regression (LR) is performed across product categories (Table 3).

Table 3. R-square value for each of five categories produced by Linear Regression analysis.

	Beauty	Grocery	Cell phone	Clothing	Digital Music
Multiple R square	0.13	0.21	0.11	0.05	0.47
Degrees of freedom	4113	2451	2416	3818	1854

From the above table, it is seen that the R-square value is very small except for the Digital Music category (0.47). The reason may be due to the nonlinear nature of the data set. To visualize the structure of five data sets, the high dimensional (25 dimensions) data is transferred to a lower dimension (2 dimensions using t-Distributed Stochastic Neighbor Embedding (t-SNE). t-Distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear method to reduce dimensionality for better visualization of data.

This algorithm works in the following way:

- Calculate the probability of similarity of points in high-dimensional space
- Calculating the probability of similarity of points in the corresponding low-dimensional (2D in this case) space.
- The similarity of data points is calculated as the conditional probability that a point X would select point Y as its neighbor if neighbors were chosen in proportion to their probability density under a Gaussian centered at X.
- The objective function is to minimize the difference between these similarities in high dimensional and low dimensional space to give a suitable representation of data in lower-dimensional space.

However, after this transformation, it is not possible to identify the input features and make any conclusions based on the output (Fig. 7).

Since the data is appeared to be nonlinear in shape, the Pearson correlation method cannot be used here to analyze the relationship between the dependent variable (Helpfulness Ratio) and Independent Variables.

Feature selection is a necessary procedure to reduce the high dimensional data into lower dimensions extracting the important variables among all variables. For this generally, Principal Component Analysis, Singular Value Decomposition, etc. methods are used. However, the primary assumption for the process mentioned above is that the data is linear. Also, these techniques do not consider feature values and target values. Therefore, using these methods shall not be applied in our data set.

The feature selection process can be categorized into three following process:

- Filter Methods: This method does not depend on the machine learning algorithm. Here, features are chosen based on various statistical tests for their correlation with the target variable. Example: Pearson Correlation, Spearman Correlation, Chi-squared test, Fisher's Score, etc.

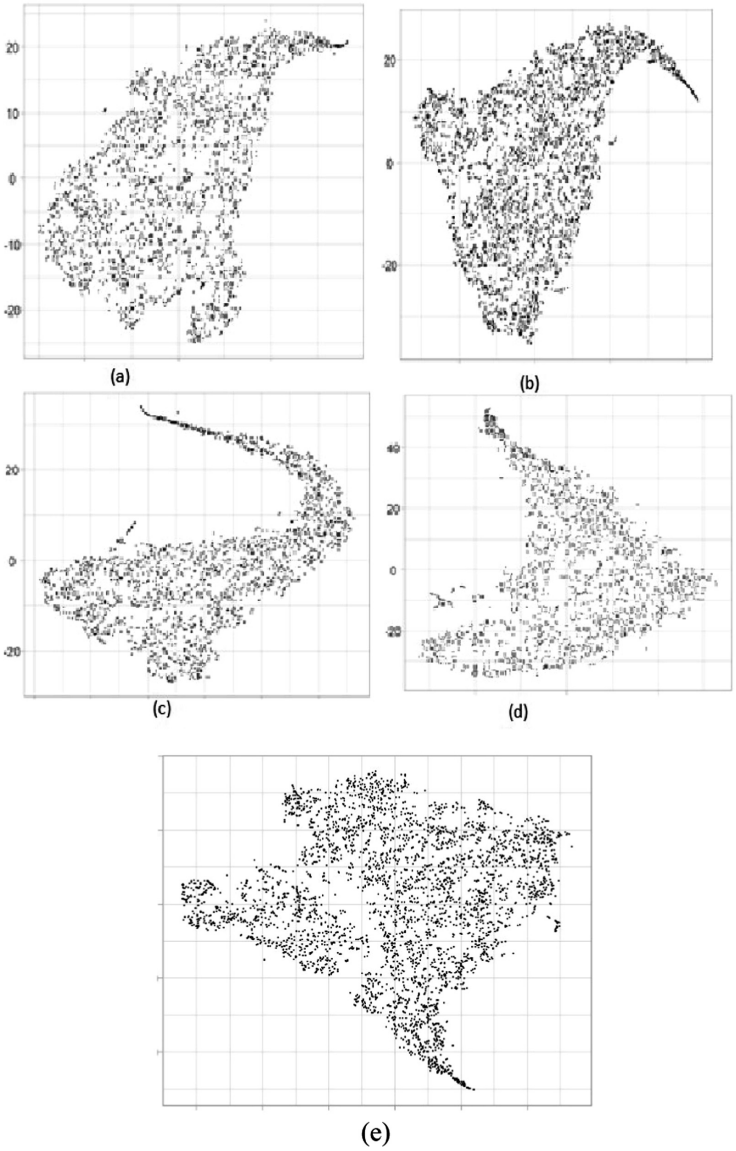


Fig. 7. A two-dimensional transformation of five datasets (a) grocery dataset, (b) beauty dataset, (c) cellphone dataset, (d) music dataset, and (e) clothing dataset.

- Wrapper Methods: This method of feature selection considers subsets of features that allow interaction with other variables by adding or removing features from that subset using a predictive model. Each subset is used to train the model and tested on a hold-out set. This is a computationally extensive algorithm but generally gives the best performing feature set for that model. Example: Recursive feature elimination, Sequential feature selection algorithms, Genetic algorithms, etc.

- Embedded Methods: Embedded feature selection method combines the advantages of both filter and wrapper selection methods and performs feature selection and classification together. This method of selection is not computationally extensive, like wrapper methods. Example: Lasso, Forward selection with Decision trees, Forward selection with Gram Schmidt, etc.

In our study, we used a wrapper built algorithm Boruta [14], which captures essential features with the target variable. Boruta is a wrapper build algorithm implemented in R package. In Boruta, Z score is used as a vital measure to consider the fluctuations of mean. This algorithm decides whether any feature is essential or not to predict the dependent variable. To do so, directly using Z score will not be ideal for measuring the importance of each variable as random fluctuations can mislead in this case. To handle this random fluctuations problem. For each variable, a corresponding shadow variable are defined whose values are assigned by shuffling the actual variables. And then the importance of shadow variables is used to decide the important variables.

Boruta algorithm ensures the randomness in the feature selection procedure and gives a better prediction on the importance of variables. Thus, in this paper, the Boruta algorithm is chosen for the feature selection procedure.

The final features identified important by Boruta algorithm is given below (Fig. 8):

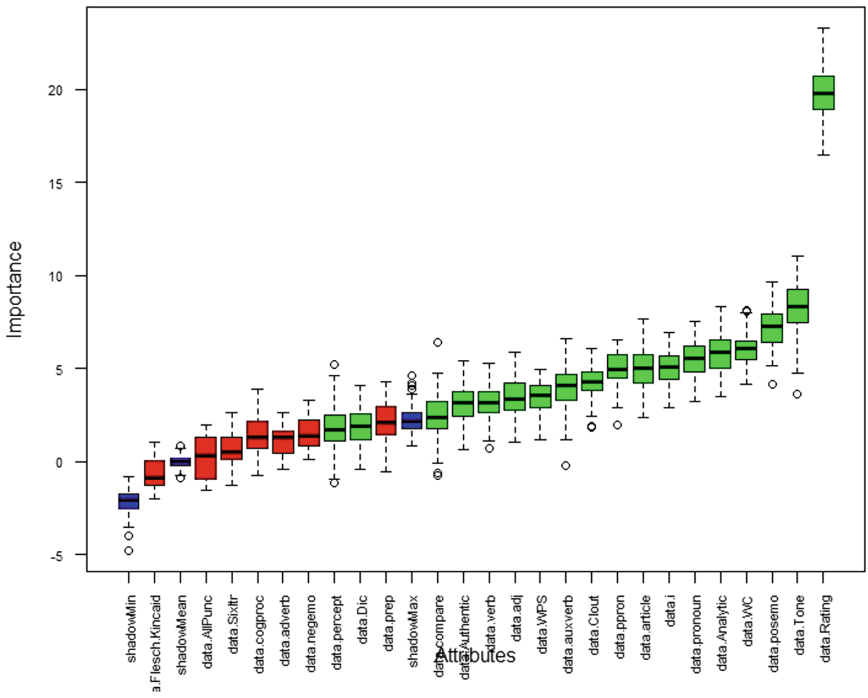


Fig. 8. Graphical plot of confirmed and rejected variables for the product category Clothing generated via Boruta algorithm.

From the above table, it is seen that for all five category product types, Rating, Analytic, WPS, posemo, Tone, Clout, pronoun, ppron, article, aux verb, adj, verb, I and Dic have a significant impact on review helpfulness. This implies these are the significant features to predict helpfulness scores where the rest of the variables only influence the helpfulness score of review specific to some product category. For example, Authentic variable is essential for all five categories of products. This implies a review comprising more honest and personal opinions with high involvement is perceived as more helpful to determine all five types of products. Compare variable is seen critical in Grocery, Cellphone, and Clothing category indicating usage of more comparative words (e.g., bigger, smaller, best, etc.) in the reviews. The negemo variables found vital features to predict the helpfulness score in Grocery and Cellphone category types of products. The percept variable is found necessary for Cellphone and Clothing category, which implies the usage of more perceptual words (e.g., feeling, hearing, etc.) in these product category reviews. The variable sixltr found crucial in only two product categories, namely Grocery and Digital Music, which implies the usage of more complex words (words longer than six letters). Adverb, Flesch Kincaid Readability, and All punc (Punctuation) are found prominent in only Digital Music category reviews. In other words, the usage of adverb words to express a more subjective view of the products in an easily readable manner with punctuation adequately used is an essential feature for Digital Music category products.

This concludes that with the conventional variables, for example, Rating, Word Count (WC), Word per Sentence (WPS), positive emotion (posemo), the other variables used as linguistic features of review, for example, Analytic, Clout, Tone, pronoun, personal pronoun, article, aux verb, adjective, verb, usage of I and Dictionary words are also have a significant impact on review helpfulness of any five categories of product reviews. The other variables (e.g., percept, negemo, compare, Sixltr, Flesch Kincaid Readability score, Allpunc, preposition, and adverb) also influence partially the target variables helpfulness for specific product categories.

The p-value for each of these Boruta results is 0.01, indicating the significance of the process of feature selection.

6.3 Prediction of Review Helpfulness Using Various Datamining Methods

The feature selection process gives us the critical variables affecting the target variable for each of the five category datasets. With the help of these variables, five different prediction models can be developed. In our study, we used Artificial Neural Network and Multiple Adaptive Regression Splines for prediction purposes and presented a comparative analysis of these two prediction models suggesting the best model choose for a specific category.

Table 4. The selected features across five categories using the Boruta algorithm.

	Beauty	Grocery	Cellphone	Clothing	Digital music
Rating	✓	✓	✓	✓	✓
WC	✓	✓	✓	✓	✓
Analytic	✓	✓	✓	✓	✓
Clout	✓	✓	✓	✓	✓
Authentic	✓	✓	✓	✓	✓
WPS	✓	✓	✓	✓	✓
compare		✓	✓	✓	
posemo	✓	✓	✓	✓	✓
negemo		✓	✓		✓
cogproc	✓	✓	✓		✓
percept			✓	✓	
Tone	✓	✓	✓	✓	✓
Sixltr		✓			✓
pronoun	✓	✓	✓	✓	✓
ppron	✓	✓	✓	✓	✓
article	✓	✓	✓	✓	✓
prep	✓	✓			✓
auxverb	✓	✓	✓	✓	✓
adverb					✓
adj	✓	✓	✓	✓	✓
AllPunc					✓
verb	✓	✓	✓	✓	✓
i	✓	✓	✓	✓	✓
Dic	✓	✓	✓	✓	✓
Flesch					✓

Multiple Adaptive Regression Splines (MARS) proposed by Friedman is a non-parametric statistical procedure to determine the relationship between a set of input variables and the dependent variable. This algorithm does not make any prior assumptions about the relationship between independent and dependent variables and works perfectly fine in both linear and nonlinear relationships. This flexibility in determining any relationship gives the idea of using MARS in this case (Fig. 9).

Artificial Neural Network (ANN) tries to model the human brain with the most straightforward definition where the building blocks are neurons. In multilayer artificial Neural Networks, each neuron is connected to others with some coefficients, and learning of the network is done by proper distribution of information through these connections. The capability of processing parallel and nonlinear behavior gives the reason use this algorithm in this paper (Fig. 10).

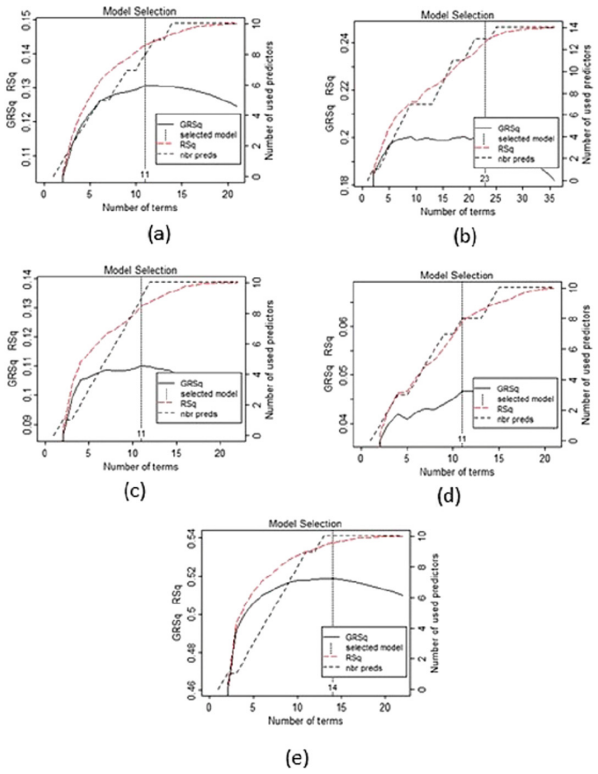


Fig. 9. Model summary capturing GCV R^2 (left-hand y-axis and solid black line) based on the number of terms retained (x-axis), which is based on the number of predictors used to make those terms (right-hand side y-axis) for (a) beauty dataset. (b) grocery dataset, (c) cellphone dataset, (d) Clothing dataset and (e) digital music dataset.

The selected features across five categories in Table 4 used to build five different models using MARS and ANN each to calculate the Mean Squared Error (MSE). The MSE gives the idea of the accuracy of the model across product categories. From the above Fig. 11, it is clear that Multiple Adaptive Regression Splines (MARS) produces lesser MSE than Artificial Neural Network (ANN) except the clothing category. Hence, it can be concluded that MARS gives more accurate results than the Artificial Neural Network for all categories except clothing product category. Therefore, it can be suggested that using the MARS algorithm would be better to predict review helpfulness in Beauty, Grocery, Cellphone and Digital Music category. The ANN produces lower MSE than MARS and hence can be used to predict the review helpfulness in case of clothing category.

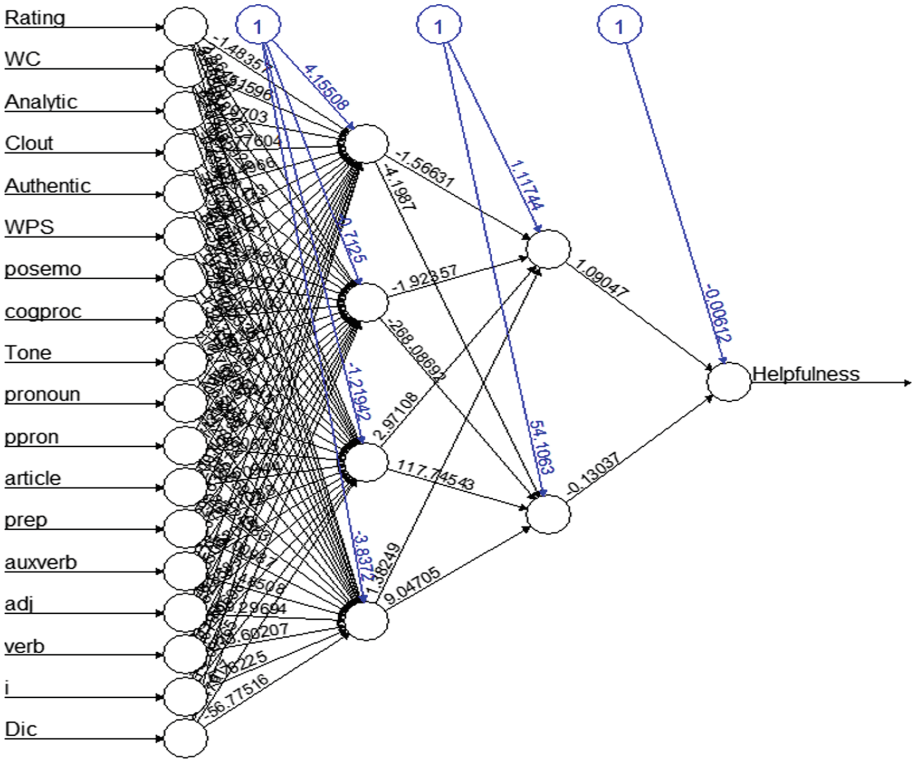


Fig. 10. Model summary for Artificial Neural Network capturing inputs and weights at each layer for grocery dataset.

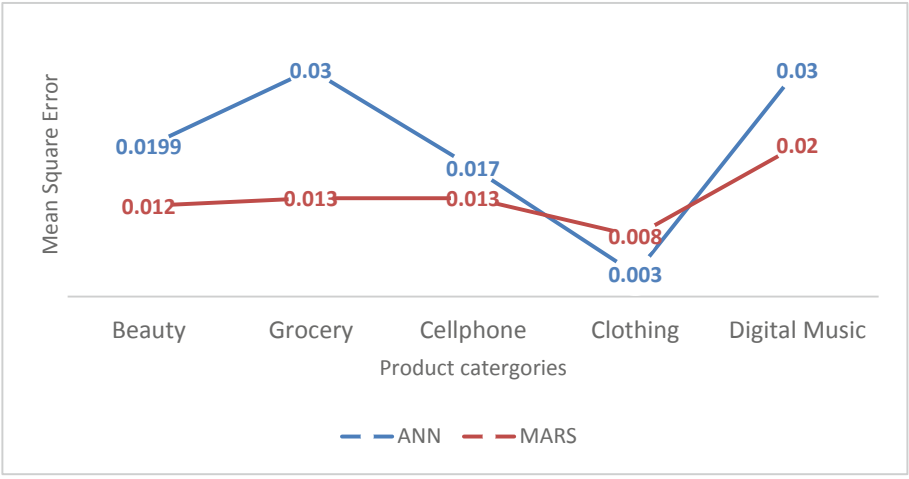


Fig. 11. Comparisons of mean square error: Multiple Adaptive Regression Splines Vs. Artificial Neural Network.

7 Conclusion

This study addresses three research questions. First, along with the conventional explicit variable (Rating), this paper explored several psychological and linguistic features from directly the product reviews across five different categories and examined whether these features are different for a different category using ANOVA analysis. The review of the Digital music category found to have the highest word count and written more analytically with maximum criticism. The authenticity of this category is least among all five categories.

On the other hand, Beauty category reviews are found to have the highest authenticity score but the lowest clout score indicating low expertise of reviewer. Also, it contained the most comparison words to describe the quality of the products. The ANOVA result shows that there are significant differences in review features among five different categories of product reviews at a 99% confidence interval. Secondly, the critical variables influencing the target variable (Helpfulness ratio) are explored using boruta algorithm. It is found that for all five category product types, Rating, Analytic, WPS, posemo, Tone, Clout, pronoun, ppron, article, aux verb, adj, verb, I, and Dic have a significant impact on review helpfulness. The other variables (e.g., percept, negemo, compare, Sixltr, Flesch Kincaid Readability score, Allpunc, preposition, and adverb) also influence partially the target variables helpfulness for specific product categories.

Finally, among two extensively used machine learning algorithms, the better method for review helpfulness prediction is determined. The result shows that both Multiple Adaptive Regression Splines and Artificial Neural Network performs very well as their Mean squared error is less than 5%. However, the MSE of the MARS algorithm is much lower than ANN except for the clothing category. Hence, the MARS algorithm should be used to predict review helpfulness for Beauty, Cellphone, Grocery, and Digital Music category reviews, and ANN should be used to predict the helpfulness score of Clothing category reviews.

This paper also solves the cold start problem, which arises when reviews do not receive any manual votes but have the potential to be a helpful review. Due to starvation, most of the reviews generally do not get the chance to get voted all the time. This prediction method will solve this problem by identifying important psychological, linguistic, and explanatory variables and producing a helpfulness score in the absence of any manual vote.

References

1. Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: online book reviews. J. Mark. Res. **43**, 345–354 (2006)
2. Forman, C., Ghose, A., Wiesenfeld, B.: Examining the relationship between reviews and sales: the role of reviewer identity disclosure in electronic markets. Inf. Syst. Res. **19**, 291–313 (2008)

3. Ghose, A., Ipeirotis, P.G.: Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. IEEE Trans. Knowl. Data Eng. **23**, 1498–1512 (2011)
4. Hu, Y.H., Chen, K.: Predicting hotel review helpfulness: the impact of review visibility, and interaction between hotel stars and review ratings. Int. J. Inf. Manag. **36**, 929–944 (2016)
5. Hu, N., Koh, N.S., Reddy, S.K.: 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 (2014)
6. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: SIGIR (2015)
7. Kacewicz, E., Pennebaker, J.W., Davis, M., Jeon, M., Graesser, A.C.: Pronoun use reflects standings in social hierarchies. J. Lang. Soc. Psychol. **33**, 125–143 (2013)
8. Khashei, M., Bijari, M.: An artificial neural network (p, d, q) model for time series forecasting. Expert Syst. Appl. **37**(1), 479–489 (2010)
9. Kim, S.M., Pantel, P., Chklovski, T., Pennacchiotti, M.: Automatically assessing review helpfulness. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006, pp. 423–430 (2006)
10. Korfiatis, N., Garcia-Bariocanal, E., Sanchez-Alonso, S.: Evaluating content quality, and helpfulness of online product reviews: the interplay of review helpfulness vs. review content. Electron. Commer. Res. Appl. **11**, 205–217 (2012)
11. Krishnamoorthy, S.: Linguistic features for review helpfulness prediction. Expert Syst. Appl. **42**, 3751–3759 (2015)
12. Kuan, K.K., Hui, K.L., Prasarnphanich, P., Lai, H.Y.: What makes a review voted? An empirical investigation of review voting in online review systems. J. Assoc. Inf. Syst. **16**, 48–71 (2015)
13. Kumar, N., Benbasat, I.: The influence of recommendations on consumer reviews on evaluations of websites. Inf. Syst. Res. **17**(4), 425–439 (2006)
14. Kursa, M., Rudnicki, W.: Feature selection with the Boruta package. J. Stat. Softw. **36**(11), 1–13. http://dx.doi.org/10.18637/jss.v036.i11
15. Lee, S., Choeh, J.Y.: Predicting the helpfulness of online reviews using multilayer perceptron neural networks. Expert Syst. Appl. **41**(6), 3041–3046 (2014)
16. McAuley, J., Leskovec, J.: Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM Conference on Recommender Systems, RecSys', Hong Kong, China, 12–16 October 2013, pp. 165–172 (2013)
17. Mudambi, S.M., Schuff, D.: What makes a helpful online review? A study of customer reviews on Amazon.com. MIS Q. **34**, 185–200 (2010)
18. Nelson, P.: Information and consumer behavior. J. Polit. Econ. **78**(20), 311–329 (1970)
19. Nelson, P.: Advertising as information. J. Polit. Econ. **81**(4), 729–754 (1974)
20. Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M.: Lying words: predicting deception from linguistic style. Pers. Soc. Psychol. Bull. **29**, 665–675 (2003)
21. Pan, Y., Zhang, J.Q.: Born unequal: a study of the helpfulness of user-generated product reviews. J. Retail. **87**, 598–612 (2011)
22. Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic inquiry and word count (LIWC2007), LIWC, Austin, TX, USA (2007). http://www.liwc.net. Accessed 27 Apr 2018
23. Pennebaker, J.W., Francis, M.E.: Cognitive, emotional, and language processes in disclosure. Cogn. Emot. **10**, 601–626 (1996)
24. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015. http://hdl.handle.net/2152/31333. Accessed 27 Apr 2018

25. Pennebaker, J.W., Chung, C.K., Frazee, J., Lavergne, G.M., Beaver, D.I.: When small words foretell academic success: the case of college admissions essays. PLoS ONE **9**, e115844 (2014)
26. Sen, S., Lerman, D.: Why are you telling me this? An examination into negative consumer reviews on the web. J. Interact. Mark. **21**, 76–94 (2007)
27. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**, 24–54 (2010)
28. Willemsen, L.M., Neijens, P.C., Bronner, F., De Ridder, J.A.: "Highly recommended!" The content characteristics and perceived usefulness of online consumer reviews. J. Comput. Mediat. Commun. **17**, 19–38 (2011)
29. Yang, Y., Yan, Y., Qiu, M., Bao, F.: Semantic analysis and helpfulness prediction of text for online product reviews. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015, pp. 38–44 (2015)
30. Yin, D., Bond, S., Zhang, H.: Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. MIS Q. **38**, 539–560 (2014)