Journal of Theoretical and Applied Information Technology

31st October 2023. Vol.101. No 20
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MULTI-DIMENSIONAL ASPECT LEVEL HELPFULNESS PREDICTION OF ONLINE CUSTOMER REVIEWS

ALA'A AHMAD KREASHAN[1], MOHAMMAD SAID EL-BASHIR[1] , ANAS JEBREEN[2] , ATYEH HUSAIN [*]

[1]Department of Computer Science, Al al-Bayt University, Jordan
[2]Department of Information Systems, Al al-Bayt University, Jordan
*Correspondence: anasjh@aabu.edu.jo

ABSTRACT

Estimating and predicting the reviews' helpfulness become essential for consumers and e-commerce systems that help access the proper reference through massive product reviews. Reviews' helpfulness is usually calculated based on perceived helpful votes. This study extends the prior review helpfulness studies. It aims to identify review characteristics that best represent the review helpfulness and then use them to predict accurate new helpfulness scores at the minimum possible error. Several natural language processing tools are configured and used to extract review characteristics from the Amazon.com dataset. Six review characteristics (i.e., review age, review aspects, review length, review polarity, review rating, and review subjectivity) that span the three main categories of the review elements were identified as the most influential for review helpfulness and proposed a multiple linear regression (MLR) model that makes use of such characteristics to predict review's helpfulness. The ability of the proposed model to predict the review helpfulness at minimum error was tested and compared with related prediction methods under various scenarios. The results show that combining the characteristics associated with the linguistic, content, and peripheral review elements improves the accuracy of helpfulness prediction, and the proposed MLR model predicts the most accurate helpfulness score at minimum error. The MLR model outperforms the SVM and DT methods by 17.68% and 1.74% in reducing MAE error and by 9.3% and 0.91% in reducing RMSE error, respectively. This study offers a novel contribution to the literature by illustrating the importance of incorporating the most influential review characteristics in the review helpfulness prediction and how it affects the predictive performance. This study extends ongoing studies on helpfulness prediction and provides notable implications for research and practice; e-commerce systems can have better organization and ranking of their reviews, and customers can efficiently access knowledge to make better purchase decisions.

Keywords: *Customer Review, Prediction, Review Helpfulness, Regression Analysis, Multiple Regression.*

1. INTRODUCTION

Customers can only test the products online after buying them. Accordingly, they need to know how the product feels; it is difficult to visualize its dimensions and size. The target item in online shopping may not be as original as the customer expects. Its size, color, and other specifications may also differ from what they expect. While this is possible in traditional shopping, it is not online shopping. Electronic word-of-mouth (eWOM) can be a promising solution for such limitations [1]; eWOM represents one or a set of statements provided by customers about a service or product that is available to other potential or actual consumers [2]. Thus, eWOM is used for information exchange in various forms, such as user-generated content, online reviews, and social media posts. As new multimedia and internet technologies evolve, eWOM has been identified as an essential topic in research, especially communication research, and addressed as a form of communication within Internet usage, social connection, m-commerce, and e-commerce [3].

The m-commerce and e-commerce systems usually present eWOMs as online reviews for buyers to help them make sound product selections and purchase decisions. Potential buyers often use the reviews written by previous buyers and published by online selling sites and forums as an essential and accessible source of information about the service or product to make the right purchase decision. Those forthcoming buyers collect information about a specific product or service because they need to learn more about it through the manufacturer and company descriptions.

Reviews allow buyers to share post-purchase experiences with other potential buyers. Consumers commonly resort to opinions of previous customers of the product, that is, reviews, as an alternative in online shopping to learn more about the target product and make sure that they make the right purchase decision. A product review summarizes knowledge and the result of previous buyers' experience of that product or service. Bearing this in mind, the m-commerce and e-commerce sites have provided a huge amount of product reviews, and the sheer volume of information published through these sites prevents potential buyers from determining the good and valuable information and feedback [4]. Indeed, it is difficult for the potential buyer to access and review all the reviews due to their huge number and disorganization, making acquiring knowledge from these reviews hard. Consequently, a need arose to arrange the reviews by their helpfulness as a potential consumer benefit and presenting the most helpful reviews for potential buyers directly.

In response, online selling platforms and e-commerce systems have included a service enabling the reader to assess and evaluate the review's helpfulness. In some platforms, this is achieved by asking the reader after each review: "Was this comment helpful to you?" [5]. This measure, or metric, has become commonly known as the helpfulness score. It is defined as the ratio of helpful votes to the total number of votes for a given review and is written as “n out of t potential buyers found this helpful” [6]. Potential buyers who read the review can vote and evaluate the ability of reviews to provide helpful information about the intended product(s). The evaluation is performed as a single vote (1) on the helpfulness of the particular review. The more votes obtained from buyers, the higher the review helpfulness score assigned.

The helpfulness score provided by its readers and employed by e-commerce systems is to guide other potential buyers directly to the most helpful reviews that best support product selection and the purchase decision. However, too many reviews in the e-commerce system were left without helpfulness score and not voted by any customer mainly for two reasons: the limited willingness of customers to spend time and effort to evaluate and vote the visited reviews, or customers are not able to see and vote such reviews due to their huge number. As a result, many potentially helpful reviews still need to vote and helpfulness score. Less helpful reviews, in comparison, which are reviewed, evaluated, and voted on, are considered valuable. Indeed, most customers may need help accessing and voting for many useful reviews because the most helpful reviews can be covered by less helpful reviews [7]. Because of this, actual helpful reviews may not be identified and presented to new potential buyers. Another potential voting issue is that review evaluation and helpfulness determination based on one human evaluation or few evaluations and votes can be criticized since it is subject to desires and subjective opinions that may not present adequate judgment. As the number of helpfulness votes acquired for a review grows, it gains a more robust evaluation, and its helpfulness value is determined based on varied viewpoints and aspects of the product.

In sum, most product reviews need a helpfulness value to be evaluated. Such reviews are marked as a “not helpful review.” This can be ascribed to several reasons: firstly, many reviews for each product in the e-commerce systems and the low willingness of potential buyers to spend time and effort reading, evaluating, and voting on them. Secondly, review evaluation and determination of helpfulness score based on the evaluation and voting of one person or few others can be unreliable. Consensus among many buyers about review helpfulness is required. Finally, adopting only one attribute of the review for determining its helpfulness is not as reliable and credible as the determination of review helpfulness based on information derived from different aspects of the product. The former approach cannot determine review helpfulness efficiently.

In effect, relying on the characteristics and attributes of each review to determine its helpfulness score, instead of depending mainly on humans to read and evaluate the helpfulness of all reviews, constitutes a promising solution. It is important to use a supportive method to help predict helpfulness scores for reviews automatically based on analysis of their content. Due to the steady and significant increase in the number of reviews for each product, there is an immense need for new ways of sorting those reviews based on their importance for the buyers who can deal with them efficiently. The major research question in this study can be presented as the following:

- How to automatically predict review helpfulness that best represents human evaluation at the minimum error?

Helpfulness prediction are necessary and useful for manufacturers, customers, and sellers[8][9] since a good review discloses the advantages and disadvantages of products and influences the purchase intentions and decisions of customers[10]. The solution would help the consumers gain reasonable knowledge about the product to make an educated purchase decision. It will also assist the companies and manufacturers by making it possible for them to get accurate feedback on their products, which will, in turn, contribute to product improvement based on the needs and opinions of their customers.

Consequently, to answer the research question, this study aimed to identify review characteristics that best represent the review helpfulness and then use it to predict an accurate new helpfulness score at the minimum possible error. Six review characteristics (i.e., review age, review length, review aspects, review polarity, review rating, and review subjectivity) that span the three main categories of the review elements were identified as the most influential for review helpfulness, and a multiple linear regression (MLR) model that makes use of such characteristics to predict review’s helpfulness is proposed.

2. LITERATURE REVIEW AND RELATED WORK

Several works on the prediction of the review helpfulness are reviewed, analyzed, and presented in this section, and then some research gaps are outlined and discussed at the end. The contributions of related studies are broadly divided into three themes: (i) factors of the prediction, which are review characteristics, (ii) quality of prediction, and (iii) prediction methods. Prediction factors are characteristics of the review that greatly help the prediction of helpfulness. However, determination of the helpfulness of reviews based on their characteristics, which can be extracted from their metadata, is a key to prediction and the responsibility of the proposed model. The prediction quality is ascertained by performing several steps when pre-processing the reviews to use and adopt only the high-quality reviews to achieve better or more robust prediction results. The prediction method uses these factors for the high-quality reviews to assess new helpfulness scores automatically.

Several factors and review characteristics, like age, rating, aspects, polarity, subjectivity, and title, are frequently considered for review prediction. For example, Eslami et al. [11] investigated how review length, score, and argument frame can predict review helpfulness for products and services. In another example, Yang & Yao [12] adopted the similarity between the review title and its content to locate review helpfulness. In contrast, Huang et al. [13] used the similarity of the potential buyer and reviewing interest to assess review helpfulness. In other respects, Malik & Hussain [6] and Wang [14] highlighted that emotions play a significant role in analyzing and interpreting the reviews. Other researchers confirmed that negative and positive feelings in the review impact the review's helpfulness [10], [15].

However, according to Park [16], the various review characteristics are categorized into three main groups: linguistic features, review content characteristics, and other peripheral characteristics. The current study addressed these three groups of review characteristics. The linguistic aspect relates to the effects of grammar, text length, and readability on the perception of helpfulness. The content of reviews pertains to the semantic and sentiment features of the text. It assesses the effects of the online reviews' essential, stylistic, and semantic characteristics on their helpfulness. It also addresses the importance of embedded emotions, if any, in a review. On the other hand, the peripheral factors encompass the product rating score, review time, and reviewer's reputation. Among several potential factors that might related to the review helpfulness, only elements that are best related to review helpfulness need to be adopted in the prediction. The factors that have the highest effects on the reader’s evaluation of review helpfulness have been identified and selected in the light of the recommendations of several previous studies and Pearson’s Product-Moment Correlation Analysis, namely aspect, polarity, subjectivity, rate, age, and length.

Huang et al. [13] confirmed that product aspects are essential determinants of review helpfulness for that product. This issue has also been previously pinpointed by several studies, e.g.[1], [11], [12], [17]. In sentiment analysis, the polarity and subjectivity of a review are important factors for determining the helpfulness of a product review [6]. Several previous studies emphasized this importance, e.g., [14], [15], [18], [19]. Previous studies have provided evidence that product rate is an essential determinant of review helpfulness for the product [10], [11], [20], [21]. Length of review is a crucial factor in the determination of review helpfulness, as was reported in previous studies [9], [14], [22]–[24]. Age of the review was a variable investigated by many researchers, including [11], [13], [16], [25], [26]. These researchers found that the age of the review influences its helpfulness score.

Several steps can be applied to select and use the high-quality reviews only for training the model and obtaining more robust, accurate prediction results. Some researchers limit the minimum required total votes for a review for [2], [17], [27]. Others like Lee and Park [2], [16] determine a minimum length for a review for prediction. Ghose & Ipeirotis [28] limit the number of aspects mentioned in the reviews to be used for prediction. However, the number of persons who reviewed the product and voted on review helpfulness, the number of words that describe the product in the review, and the number of product aspects covered by the review were adopted simultaneously in this study to classify reviews before training the model and performing prediction.

Researchers used different approaches to predict helpfulness scores of reviews; some are based on machine-learning methods, whereas others used statistical methods. Some of the essential contributions of relevant previous studies have been reviewed to identify gaps in this area of research, align the herein proposed method to prior studies, and show how this study contributes to filling some of the gaps identified in the literature: Singh et al. [7] developed a model based on machine learning methods and utilize many textual factors to predict customers' reviews, including subjectivity, entropy, polarity, and ease of reading. When a review is posted, helpfulness values will be added automatically by the model so that the review can be viewed by other customers researching the product[7].

Regression analysis is quite helpful in predicting the helpfulness score of a review [16]. Multiple regression is widely adopted and frequently used because it is generally faster than other methods and can reveal how the explanatory variables affect the dependent variable. It has been widely employed for predicting the review helpfulness scores by many researchers, including [20], [27], [29]–[34], amongst others. Another reason that justified the adoption of multiple regression analysis in this study is that it allows for studying the combined effects of many independent variables simultaneously on the dependent variable[35]. Zhang and Tran [36] proposed a linear regression model to predict the helpfulness of online product reviews to help rank and classify the best ones. The regression model was compared with several machine-learning algorithms and proved its efficiency.

Jerripothula et al.[37] proposed a feature-level rating system, which inputs, reviews, and reviews votes of customers and produces feature-level ratings. In terms of sentiment score accumulation and then finalization, "votes-aware cumulative rating" and "votes-aware final rating measures" were proposed as new rating measures [38]. Saumya et al. [39] developed a system that derives helpfulness score predictions from features of review text, customer question-answer feedback, and product description using gradient boosting regressor and random-forest classifier. This system assigned low/high classification to reviews based on the random-forest classifier. It did not calculate the helpfulness scores for low-quality reviews, arguing that they will not be relevant as they will not be among the top thousand reviews. This system provided placement of reviews only, with the high-quality reviews prominently displayed. These researchers concluded that including product description features and customer question-answer data improved the accuracy of the prediction of the helpfulness score. These researchers expanded the work of Singh et al. [7] by constructing an improved multiple regression prediction model that places the reviews in their suitable places in a list based on predicted review helpfulness scores. They aspired to build a model that employs a classifier to assign a quality rating to reviews and then only use high-quality reviews to construct an MLR model[39].

Lee and Choeh [2]used a model based on a neural network and validated their results using the 'amazon.com' dataset. Product type, textual characteristics, and review characteristics were used to determine the review helpfulness scores. Ghose and Ipeirotis [28] investigated the effect of Internet reviews on the sale of a product and perceived helpfulness by utilizing a random forest-based classifier to predict the usefulness of reviews and their impact on sales.

Review readability and subjectivity features and the relative effect of reviewer-related features were assessed. They found that the medium-length reviews were more helpful than others when they had fewer spelling errors and that the linguistic feature significantly impacted product sales. They employed Support Vector Machine (SVM) Regression to automatically assess the helpfulness of Internet reviews and used an 'amazon.com' dataset to evaluate the performance of their proposed method. They found that product rating, unigrams, and review length were critical determinants of the review helpfulness score. SVM and Decision Tree (DT) methods were adopted as prediction methods in the study of Ghose & Ipeirotis[28] and Park [16]. Malik and Hussain [6] studied the effect of emotion on helpfulness by using a deep-learning neural network model. Their findings indicate that positive emotion features are the best predictor upon consideration of an individual feature category. Nevertheless, the performance was compromised by visibility and emotion features. Yang & Yao [12] investigated the effects of reviews on customers' attitudes to products and product selection and sales. They also provided an analysis of the determinants of review helpfulness. Additionally, they examined the impacts of product type and review characteristics on the perceived review helpfulness. Data were collected from an online retailer. The study found that review length and valence positively impacted review helpfulness and that the type of product (i.e., experiential or utilitarian) remediates the effects of these two variables on perceived usefulness.

A growing body of literature reports on the review helpfulness prediction with varied performance results, but there needs to be more results to generalize. However, existing studies have few limitations. For example, few studies have examined the effect of various review characteristics that span several important categories, i.e., the linguistic features, review content characteristics, and other peripheral characteristics, on the helpfulness of reviews. It is necessary to consider the effect of different categories of review characteristics on the helpfulness of reviews. Another gap is that only some studies consider the quality of reviews in the dataset used for the prediction. Processing the reviews in the dataset and using only the highest quality ones that guarantee obtaining the most reliable prediction results is essential. Further comparative studies are required to extend the understanding of helpfulness prediction and examine the prediction performance with different environmental variables and test scenarios.

In this concept, it is believed that finding an appropriate set of review characteristics contributes significantly to accurate helpfulness prediction, and adopting a prediction method that best estimates review helpfulness score at minimum error, and with the use of the highest quality reviews can be a promising solution that fills research gaps. For instance, review age, review aspects, review length, review polarity, review rating, and review subjectivity, which span the linguistic, content, and other peripheral categories, are selected. The number of persons who reviewed the product and voted on the review, the number of words that describe the product in the review, and the number of product aspects covered by the review were the main variables proposed to select high-quality reviews. The actual helpfulness of the reviews chosen and the metadata of the adopted characteristics were used to train the MLR model to support better helpfulness prediction.

This study offers a novel contribution to the literature by illustrating the importance of incorporating the most influential review characteristics in the review helpfulness score and how it affects predictive performance. Several and multiple pre-processing steps to classify and select the highest quality reviews that guarantee to obtain the most reliable prediction results have been incorporated. A model based on regression analysis was formulated and trained to perform helpfulness prediction and find the best helpfulness scores at minimum error. Several natural language processing (NLP) tools, such as TextBlob and spaCy, are configured and combined to extract review characteristics from their text.

Table 1 summarizes the differences between the proposed method and methods presented in previous studies to review helpfulness score prediction. The comparison is based on the review characteristics, categories, and prediction methods.

3 THE PROPOSED SOLUTION AND TECHNIQUES (majority of this subsection affected)

To develop a solution and measure its ability to perform prediction at minimum error, several processes are performed using several natural language processing (NLP) tools and instruments, including (1) Data collection is the process of selecting a dataset of reviews; (2) data preprocessing, which is performed to extract review characteristics, identify the most important characteristics that contribute significantly to accurate helpfulness prediction, and lastly identify the quality reviews for training; (3) training process is aimed at performing predictive analysis to understand the voting behavior and determine prediction equation based on relation between the review characteristics and the actual helpfulness for training reviews; and (4) prediction, which includes estimating a new helpfulness score for each review using the prediction equation.

Table 1: Related Work Overview And Comparison (Current Study At The First Row)

Prediction factors						Prediction quality			Prediction method			Helpfulness variable: Ratio
Linguistic	Content			Other peripheral								
Length	Polarity	Subjectivity	Aspects	Rate	Age	Total votes≥ 20	Aspects	Length > 5	Regression	Classification	Ranking	
√	√	√	√	√	√	√	√	√	Multiple Linear regression (MLR)	√		√
√	√	√	√	√	√	√	√	√	Simple additive weighting (SAW) [40]	√		√
√	√			√	√			√	SVM, DT [16]	√		√
√				√	√	10			Linear [27]			√
√	√	√		√					Gradient boosting algorithm, tree-based [7]	√		√
√	√			√		10			Gradient boosting (GB) [17]	√		
√				√	√				Logistic regression [41]			
√				√	√	2		√	Neural network [2]			√
	√	√	√	√	√				Neural network [6]	√		√
√			√	√			√		SVM [28]	√	√	

The dataset obtained from Amazon.com consists of more than 100,000 reviews for 25,788 products. Mobile electronics are selected among several categories, such as electronics, furniture, toys, and cameras [42], [43]. During data preprocessing, the datasets are divided into training and target reviews. The training dataset consists of voted reviews, which include the actual helpfulness values provided by customers. By contrast, the target dataset, which represents the non-voted reviews, lacks the actual helpfulness scores. Consequently, for each review in the training dataset, the review textual content is extracted and processed by tokenizing their original text, segmenting it into sentences and words, and removing the punctuation marks and stop words. The review rating score (R_i), the number of helpful votes, and the total votes are directly available. All operations were performed based on NLP functions using RStudio, version 3.0 of the Python software via Jupyter Notebook, and two Python libraries: spaCy and TextBlob.

Among several review characteristics obtained and investigated for review helpfulness, the related characteristics (i.e., review age, review aspects, review length, review polarity, review rating, and review subjectivity) have been extracted and adopted for prediction based on Pearson's Product-Moment Correlation Analysis. However, such characteristics were not directly available and were calculated as follows: Review age (G_i) was calculated by counting the number of days starting from the review posting date. Review length (L_i) was calculated by measuring the words in the textual review content. Review polarity score (P_i), which ranges from −1 to 1, and review subjectivity score (S_i), which runs from 0 to 1, are calculated using TextBlob [44]. TextBlob is a Python-based library for performing NLP functions, proven to be an effective sentiment analysis tool [44], [45]. It provides a sentiment score for a text based on a Naïve Bayes analyzer [45]. The lower the polarity score, the more negative the review sentiment is. The higher the subjectivity score, the more feeling is expressed in the review. The aspects for review (A_i) are extracted also using the TextBlob and spaCy Python libraries. Lastly, the actual review helpfulness (H_i) was calculated by the ratio of helpfulness to total votes [34].

Pearson's Product-Moment Correlation Analysis is conducted and identified the six review characteristics (i.e., review age, review aspects, review length, review polarity, review rating, and review subjectivity) as the most influential characteristics for accurate helpfulness prediction. The obtained correlation coefficients for these characteristics indicate statistically significant correlations with the actual review helpfulness and are recommended as prediction parameters. To improve the reliability and quality of the results, training reviews are classified based on the number of individuals who voted for review helpfulness, the number of words that describe the product or service in the review, and the number of product aspects mentioned in the review. Only high-quality reviews containing valuable information for training (i.e., reviews with more than 20 votes, reviews with five words or more, and reviews that include at least 1 product aspect) are adopted. The final sample training dataset comprises 1,180 reliable, non-biased reviews. Table 2 shows the summary statistics of the sample training reviews with the results of the correlation analysis.

Table 2. Descriptive Statistics For Dataset (N = 1,180) And The Correlation Analysis Results

Review Characteristics	Description	Correlation (P value)	Instrument	Range	Mean	SD
Age (G_i)	Lifetime of the review from the time of its publication to the time of its use	$p < 0.001$	days count calculation	1950 - 3312	2312	615
Aspects (A_i)	Product features exist in the review text	$p < 0.001$	spaCy and TextBlob Python libraries	1 - 69	3	6.9
Length (L_i)	Number of words in a review	$p < 0.001$	Word count calculation	6 - 2654	231.6	271.9
Polarity (P_i)	Emotions of consumers about usage experiences of a product range from negative (−1) to positive (1) score	$p < 0.01$	TextBlob Python library	−0.875 - 1	0.17	0.175
Rating (R_i)	Numerical assessment of a consumer about a product quality range from negative (1) to positive (5)	$p < 0.001$	Directly avalable	1 - 5	3.59	1.618
Subjectivity (S_i)	The presence of subjective feelings in the review	$p < 0.05$	TextBlob Python library	0 - 1	0.509	0.139
Actual helpfulness (H_i)	The ratio of helpful votes to total votes		Ratio calculation	0 - 1	0.828	0.241

Finding precise and fine-grained predicted helpfulness is critical to performing this study efficiently and obtaining reliable results. Thus, regression analysis is employed as a machine-learning technique during the training process to perform predictive analysis and find predicted helpfulness (Hi_{new}), which represents the estimated helpfulness score for review (i). Regression analysis is a machine learning technique commonly used to conduct predictive analysis, investigate the relationship between the research variables, and efficiently obtain a predicted value based on their correlation [35]. For instance, a multiple linear regression (MLR) model is formulated and applied to predict a new helpfulness score (H_{inew}) resulting from the relationship between the adopted six review characteristics with the actual helpfulness score (H_i) of the training reviews [46]. The solution involves training the MLR model to understand and learn the voting behavior in training reviews, determine prediction equation parameters, and then estimate the predicted helpfulness score for any review during the prediction process, as shown in Eq. (1).

$$H_{inew} = b+(W_G*G_i)+(W_A*A_i)+(W_L*L_i)+(W_P*P_i)+(W_R*R_i)+(W_S*S_i) \quad (1)$$

Where H_{inew} is the new predicted helpfulness score, G_i, A_i, L_i, P_i, R_i, and S_i are the six identified review characteristics of the review (i). W_G, W_A, W_L, W_P, W_R, and W_S are the slope coefficients. MLR-related parameters represent corresponding weights for each review characteristic, respectively, and b represents an intercept MLR-related parameter. However, figure 1 illustrates the proposed solution framework that summarizes the overall data collection, preprocessing, learning, and prediction processes.

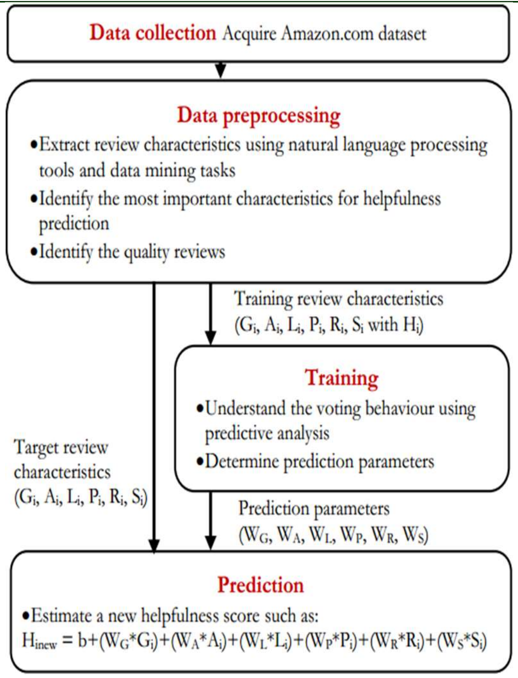
Data collection Acquire Amazon.com dataset
Data preprocessing
•Extract review characteristics using natural language processing tools and data mining tasks
•Identify the most important characteristics for helpfulness prediction
•Identify the quality reviews
Training review characteristics (G_i, A_i, L_i, P_i, R_i, S_i with H_i)
Training
•Understand the voting behaviour using predictive analysis
•Determine prediction parameters
Target review characteristics (G_i, A_i, L_i, P_i, R_i, S_i)
Prediction parameters (W_G, W_A, W_L, W_P, W_R, W_S)
Prediction
•Estimate a new helpfulness score such as:
$H_{inew} = b+(W_G*G_i)+(W_A*A_i)+(W_L*L_i)+(W_P*P_i)+(W_R*R_i)+(W_S*S_i)$

Figure 1: Overall View Of The Proposed Prediction Model

4 PERFORMANCE EVALUATION AND RESULTS

To measure and compare the performance of the proposed model, helpfulness prediction of training reviews was performed using several methods about several performance metrics, i.e., Mean Absolute Error (MAE) and Root Mean-Squared Error (RMSE). MAE and RMSE are common performance metrics that measure prediction accuracy [47]. Several performance tests were conducted according to the multiple evaluation metrics, environmental variables with their levels, and several related methods.

The first group of tests intended to evaluate the performance of the proposed method in predicting closer review helpfulness values to the actual helpfulness values for all reviews in the training dataset compared to the simple additive weighting (SAW) method. SAW is selected in this test to demonstrate the efficiency (show the importance) of applying the learning concept that results in variable weights for prediction factors when training the MLR model instead of treating them equally as in SAW. Results in Figure 2 show that predicted helpfulness values by the proposed model are closer to the actual helpfulness values, generating the best review helpfulness prediction. This indicates that the proposed model is valid for efficiently predicting the helpfulness score for non-voted reviews.

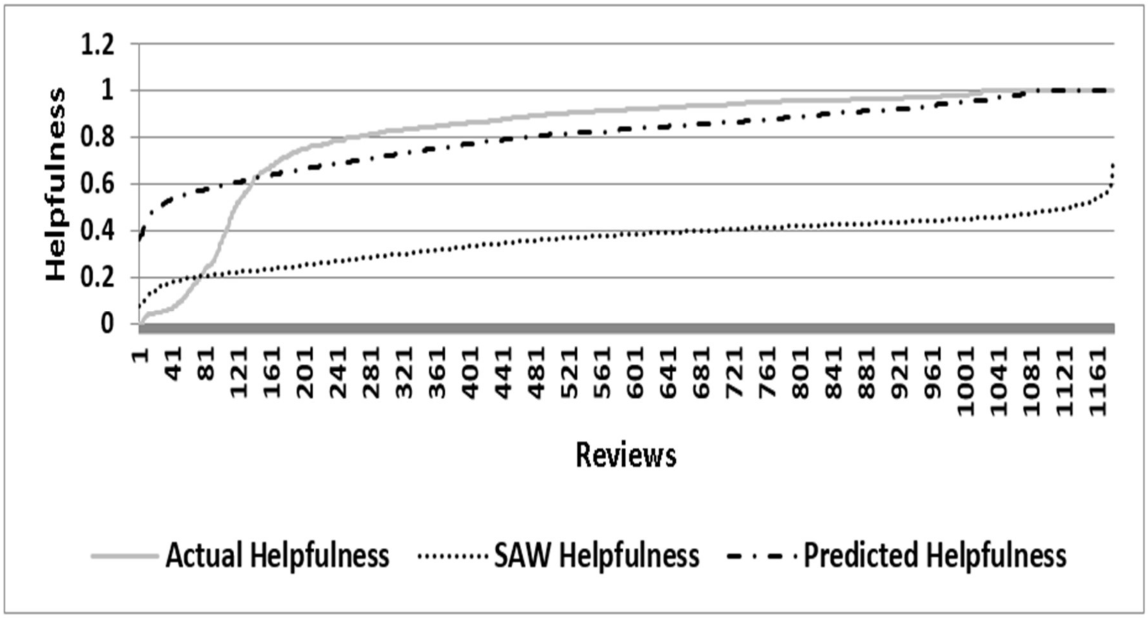

Figure 2: Actual Helpfulness Versus Predicted Helpfulness

The second group of tests intended to assess the performance of the proposed method in predicting review helpfulness based on the MAE and RMSE at three variable levels, namely: total vote ranges (low, medium, high), number of reviews per product (low, medium, high), and at all reviews in the training dataset about Support Vector Machine (SVM) and Decision Tree (DT) methods [16]. Adopting these variables allows more performance investigation for the proposed solution. Figure 3 depicts the MAE values result when predicting helpfulness at three variable levels: a: all reviews in the training dataset, different total vote ranges, and c: different numbers of reviews. The results showed that the proposed MLR model performs better in predicting the review helpfulness score at minimum error than other methods at all vote ranges, at all categories of numbers of reviews, and at the overall dataset level. Less error indicates close predicted helpfulness scores to the actual helpfulness scores provided by users. In other words, the number of humans who voted for a review and the number of reviews of a particular product (small, medium, or large) does not affect the efficiency of the herein-proposed MLR model in predicting the review helpfulness score. It can obtain the result at minimum error than other methods. This suggests that the MLR model works efficiently, regardless of the number of votes, the reviews per product, and the overall reviews in the dataset.

Figure 4 presents the RMSE values for three variable levels: a: all reviews in the training dataset, different total vote ranges, and c: different numbers of reviews. The results in this figure support that the proposed MLR model performs better in predicting helpfulness at minimum RMSE than other methods at all vote ranges, at all categories of numbers of reviews, and at the overall dataset level. Indeed, the number of humans who voted for a review and the number of reviews of a particular product (small, medium, or large) does not affect the efficiency of the proposed model in predicting the review helpfulness score. It can obtain better scores regardless of the changes in such environmental variables.

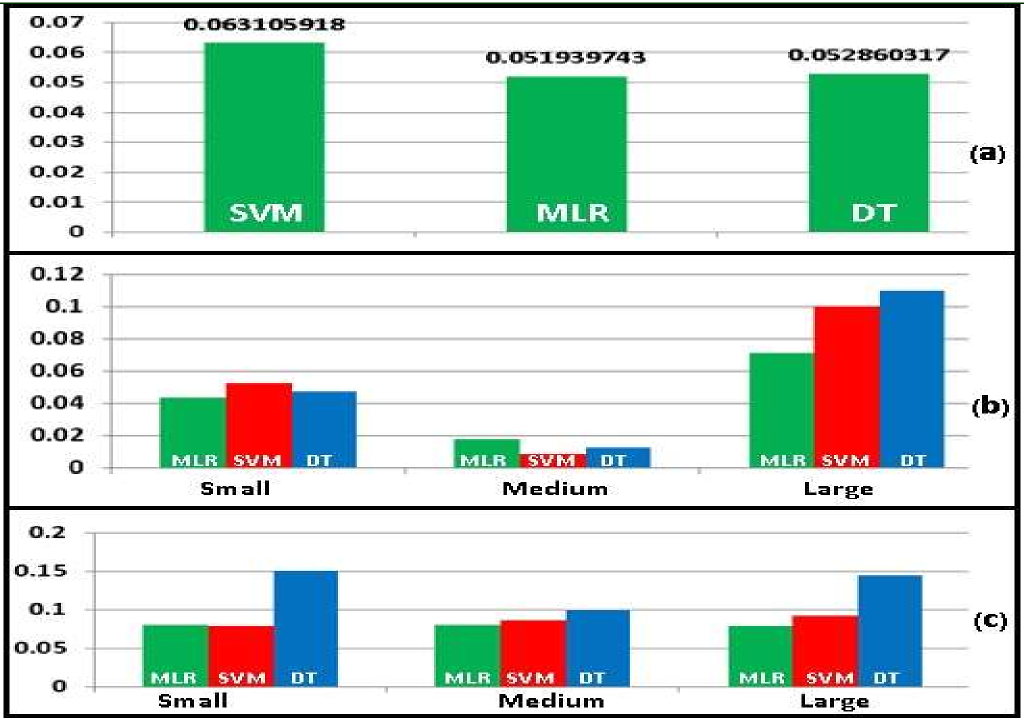

Figure 3: The MAE Values For (A) All Datasets, (B) Different Total Vote Ranges, (C) Different Numbers Of Reviews

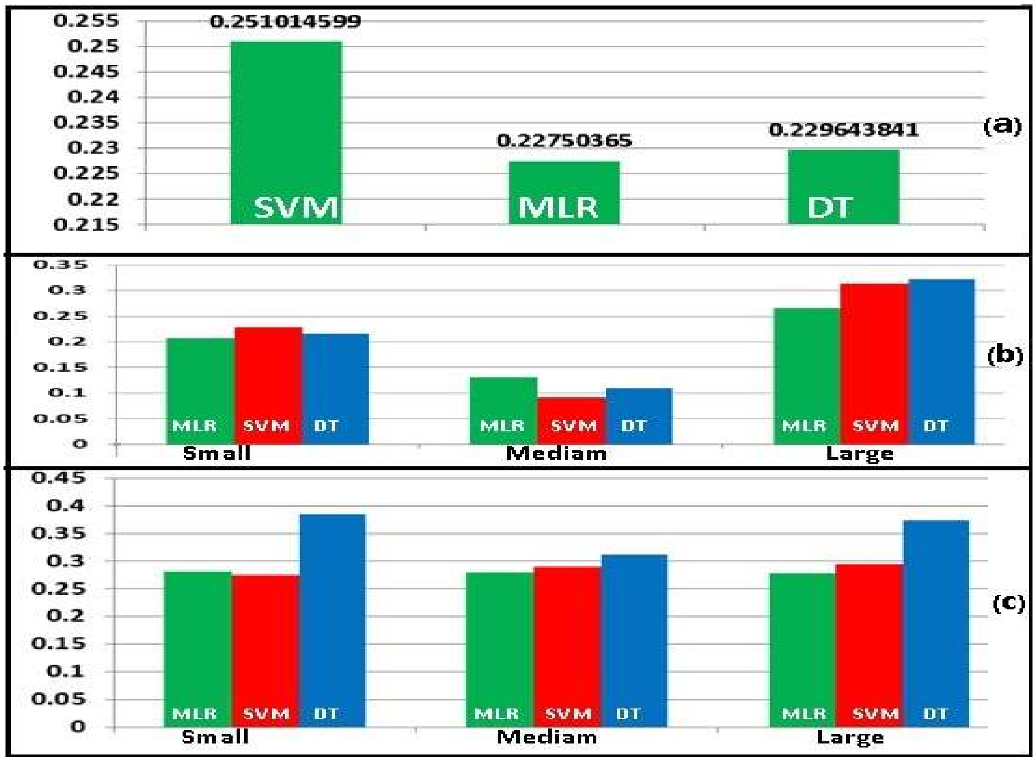

Figure 4: The RMSE Values For (A) All Datasets, (B) Different Total Vote Ranges, (C) Different Numbers Of Reviews

5 DISCUSSION

This study aimed to identify review characteristics that best represent the review helpfulness and then use them to predict an accurate new helpfulness score at the minimum possible error. Six review characteristics (i.e., review age, review aspects, review length, review polarity, review rating, and review subjectivity) that span the three main categories of the review elements were identified as the most influential for review helpfulness, and a multiple linear regression (MLR) model that makes use of such characteristics to predict review's helpfulness is proposed. The ability of the proposed model to predict the review helpfulness at minimum error was tested and compared with related prediction methods. The results show that combining and adopting the review characteristics associated with the linguistic, content, and peripheral review elements for multiple regression prediction improves the accuracy of helpfulness prediction and minimizes the error. Compared to the SVM and DT methods, the proposed MLR model performed better by 17.68% and 1.74% in reducing MAE error and 9.3% and 0.91% in reducing RMSE error on real-life amazon.com review datasets, respectively.

Additional analysis was performed on different test scenarios and environmental variables (i.e., the number of reviews for a particular product and votes for each review). The results showed that the proposed model that uses regression analysis with the selected review characteristics delivers the best performance in all scenarios. The obtained results demonstrated the ability of the MLR model to estimate accurate helpfulness values that best human evaluation at minimum error for all available reviews. Unlike the SAW method [40], which assigns fixed weights for the review characteristics, the proposed MLR model assesses appropriate parameters for the review characteristics based on their existing linear relationship with the actual helpfulness score. Therefore, the predicted helpfulness scores were remarkably close to the actual helpfulness. The experimental tests also revealed that the MLR model outperformed the other methods and predicted helpfulness at minimum MAE and RMSE errors. These results align with Zhang and Tran [36], which confirmed that multiple Linear regression achieved high predictive performance compared to other methods and estimated the best review helpfulness scores. However, contrary to our results, the findings of Park [16] indicated that the SVM method predicts review helpfulness most accurately. O'Mahony and Smyth [48] also showed that the DT method accurately predicts review helpfulness. This can be explained by the different datasets involved, other review characteristics adopted for prediction, and different environmental variables and test scenarios. That is to say, the number of reviews for a particular product (small, medium, or large) and the number of votes do not affect the efficiency of the proposed MLR model in predicting review helpfulness, and the MLR performs the prediction efficiently, regardless of the environmental changes.

6 CONCLUSIONS AND FUTURE WORKS

This study attempted to provide an accurate helpfulness prediction for online reviews, which help improve the online shopping experience; e-commerce systems strengthen their business, manufacturers improve their products by getting precise feedback, and individuals make better purchase decisions. To this end, this study designed a review helpfulness prediction method based on multiple regression that is anticipated to compensate for any errors that will make the high-quality reviews go unnoticed, whether due to the human factor or an error relating to the rating system. The method proposed in this study evaluates every posted review accurately. The predicted review helpfulness scores differ slightly from the actual human voting of review helpfulness. Moreover, the proposed MLR model was tested extensively using mean absolute error and root-mean-squared error performance metrics. In conclusion, the results of this study confirm that the returned outputs of the proposed model are almost identical to human work, which further demonstrates the predictive performance at minimum error in various situations and environmental changes. This study highlights the importance of identifying factors that best represent helpfulness votes for accurate helpfulness prediction and performing multi-dimensional aspect-level helpfulness prediction at the minimum error.

This study has several limitations, which may represent opportunities for future research. First, we needed to verify the reliability and validity of the extracted review characteristics. For example, the product aspects extracted from the review text need to be tested and validated to ensure the reliability of the results. Second, the target reviews in this study are limited to the search products and other product categories. Their reviews, such as experience products, are recommended, and additional analysis according to product categories would be more meaningful. Lastly, this study was mainly performed using only the Amazon.com dataset, while other datasets from various e-business domains would lead to more general results.

REFERENCES

[1] M. S. I. Malik and A. Hussain, “An analysis of review content and reviewer variables that contribute to review helpfulness,” *Inf. Process. Manag.*, vol. 54, no. 1, pp. 88–104, 2018, doi: 10.1016/j.ipm.2017.09.004.

[2] S. Lee and J. Y. Choeh, “The impact of online review helpfulness and word of mouth communication on box office performance predictions,” *Humanit. Soc. Sci. Commun.*, vol. 7, no. 1, 2020, doi: 10.1057/s41599-020-00578-9.

[3] T. Sun, S. Youn, G. Wu, and M. Kuntaraporn, “Online word-of-mouth (or mouse): An exploration of its antecedents and consequences,” *J. Comput. Commun.*, vol. 11, no. 4, pp. 1104–1127, 2006, doi: 10.1111/j.1083-6101.2006.00310.x.

[4] N. Amblee and T. Bui, “Freeware downloads: An empirical investigation into the impact of expert and user reviews on demand for digital goods,” *Assoc. Inf. Syst. - 13th Am. Conf. Inf. Syst. AMCIS 2007 Reach. New Height.*, vol. 3, pp. 1609–1620, 2007.

[5] Y. Hong, J. Lu, J. Yao, Q. Zhu, and G. Zhou, “What reviews are satisfactory: Novel features for automatic helpfulness voting,” *SIGIR’12 - Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 495–504, 2012, doi: 10.1145/2348283.2348351.

[6] M. S. I. Malik and A. Hussain, “Helpfulness of product reviews as a function of discrete positive and negative emotions,” *Comput. Human Behav.*, vol. 73, pp. 290–302, Mar. 2017, doi: 10.1016/j.chb.2017.03.053.

[7] J. P. Singh, S. Irani, N. P. Rana, Y. K. Dwivedi, S. Saumya, and P. Kumar Roy, “Predicting the ‘helpfulness’ of online consumer reviews,” *J. Bus. Res.*, vol. 70, pp. 346–355, Mar. 2017, doi: 10.1016/j.jbusres.2016.08.008.

[8] Husain, A. J. A., Alsharo, M., AlRababah, S. A., & Jaradat, M.-I. R. Content-rating consistency of online product review and its impact on helpfulness: A fine-grained level sentiment analysis. *Interdisciplinary Journal of Information, Knowledge, and Management*, *18,* 2023.

[9] Y. Kang and L. Zhou, “Longer is better? A case study of product review helpfulness prediction,” *AMCIS 2016 Surfing IT Innov. Wave - 22nd Am. Conf. Inf. Syst.*, 2016.

[10] G. Ren and T. Hong, “Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews,” *Inf. Process. Manag.*, vol. 56, no. 4, pp. 1425–1438, Mar. 2019, doi: 10.1016/j.ipm.2018.04.003.

[11] S. P. Eslami, M. Ghasemaghaei, and K. Hassanein, “Which online reviews do consumers find most helpful? A multi-method investigation,” *Decis. Support Syst.*, vol. 113, pp. 32–42, Mar. 2018, doi: 10.1016/j.dss.2018.06.012.

[12] Y. Zhou, S. Yang, Y. Li, Y. Chen, J. Yao, and A. Qazi, “Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining,” *Inf. Process. Manag.*, vol. 57, no. 2, p. 102179, 2020, doi: 10.1016/j.ipm.2019.102179.

[13] C. Huang, W. Jiang, J. Wu, and G. Wang, “Personalized review recommendation based on users’ aspect sentiment,” *ACM Trans. Internet Technol.*, vol. 20, no. 4, 2020, doi: 10.1145/3414841.

[14] X. Wang, L. (Rebecca) Tang, and E. Kim, “More than words: Do emotional content and linguistic style matching matter on restaurant review helpfulness?” *Int. J. Hosp. Manag.*, vol. 77, pp. 438–447, Mar. 2019, doi: 10.1016/j.ijhm.2018.08.007.

[15] M. Lee, M. Jeong, and J. Lee, “Roles of negative emotions in customers’ perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach,” *Int. J. Contemp. Hosp. Manag.*, vol. 29, no. 2, pp. 762–783, 2017, doi: 10.1108/IJCHM-10-2015-0626.

[16] Y. J. Park, “Predicting the helpfulness of online customer reviews across different product types,” *Sustain.*, vol. 10, no. 6, 2018, doi: 10.3390/su10061735.

[17] S. Saumya, J. P. Singh, and Y. K. Dwivedi, “Predicting the helpfulness score of online reviews using convolutional neural network,” *Soft Comput.*, vol. 24, no. 15, pp. 10989–11005, 2020, doi: 10.1007/s00500-019-03851-5.

[18] H. Chen and D. Zimbra, “AI and opinion mining,” *IEEE Intell. Syst.*, vol. 25, no. 3, pp. 74–76, 2010, doi: 10.1109/MIS.2010.75.

[19] M. Salehan and D. J. Kim, “Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics,” *Decis. Support Syst.*, vol. 81, pp. 30–40, Mar. 2016, doi: 10.1016/j.dss.2015.10.006.

[20] C. Forman, A. Ghose, and B. Wiesenfeld, “Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets,” *SSRN Electron. J.*, 2011, doi: 10.2139/ssrn.1026893.

[21] J. McAuley and J. Leskovec, “Hidden factors and hidden topics: Understanding rating dimensions with review text,” *RecSys 2013 - Proc. 7th ACM Conf. Recomm. Syst.*, pp. 165–172, 2013, doi 10.1145/2507157.2507163.

[22] Q. Cao, W. Duan, and Q. Gan, “Exploring determinants of voting for the ‘helpfulness’ of online user reviews: A text mining approach,” *Decis. Support Syst.*, vol. 50, no. 2, pp. 511–521, 2011, doi: 10.1016/j.dss.2010.11.009.

[23] C. Hall, “How and When Review Length and Emotional Intensity Influence Review Helpfulness : Empirical Evidence from,” pp. 1–16, 2014.

[24] M. Siering and J. Muntermann, “What Drives the Helpfulness of Online Product Reviews ? From Stars to Facts and Emotions,” *11th Int. Conf. Wirtschaftsinformatik*, no. March, pp. 103–118, 2013.

[25] H. Hong, D. Xu, G. A. Wang, and W. Fan, “Understanding the determinants of online review helpfulness: A meta-analytic investigation,” *Decis. Support Syst.*, vol. 102, pp. 1–11, 2017, doi: 10.1016/j.dss.2017.06.007.

[26] Y. Zhou and S. Yang, “Roles of Review Numerical and Textual Characteristics on Review Helpfulness Across Three Different Types of Reviews,” *IEEE Access*, vol. 7, pp. 27769–27780, 2019, doi: 10.1109/ACCESS.2019.2901472.

[27] J. Otterbacher, “Helpfulness in online communities: A measure of message quality,” *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 955–964, 2009, doi: 10.1145/1518701.1518848.

[28] A. Ghose and P. G. Ipeirotis, “Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics,” *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 10, pp. 1498–1512, 2011, doi: 10.1109/TKDE.2010.188.

[29] U. of Missouri, D. Yin, S. D. Bond, G. I. of Technology, H. Zhang, and G. I. of Technology, “Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews,” *MIS Q.*, vol. 38, no. 2, pp. 539–560, Mar. 2014, doi: 10.25300/MISQ/2014/38.2.10.

[30] Y. Yang, M. Qiu, Y. Yan, and F. S. Bao, “Semantic analysis and helpfulness prediction of text for online product reviews,” in *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 2015, vol. 2, pp. 38–44, doi: 10.3115/v1/p15-2007.

[31] N. Korfiatis, E. García-Bariocanal, and S. Sánchez-Alonso, “Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content,” *Electron. Commer. Res. Appl.*, vol. 11, no. 3, pp. 205–217, Mar. 2012, doi: 10.1016/j.elerap.2011.10.003.

[32] J. A. Chevalier and D. Mayzlin, “The Effect of Word of Mouth on Sales: Online Book Reviews,” *J. Mark. Res.*, vol. 43, no. 3, pp. 345–354, Mar. 2006, doi: 10.1509/jmkr.43.3.345.

[33] K. Park, Y.-J., & Kim, “Impact of Semantic Characteristics on Perceived Helpfulness of Online Reviews,” *J. Intell. Inf. Syst.*, vol. 23, no. 3, pp. 29–44, 2017.

[34] S. M. Mudambi and D. Schuff, “What makes a helpful online review? A study of customer reviews on amazon.com,” *MIS Q. Manag. Inf. Syst.*, vol. 34, no. 1, pp. 185–200, 2010, doi: 10.2307/20721420.

[35] D. Maulud and A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” *J. Appl. Sci. Technol. Trends*, vol. 1, no. 4, pp. 140–147, 2020, doi: 10.38094/jastt1457.

[36] R. Zhang and T. Tran, “Helpful or unhelpful: a linear approach for ranking product reviews,” *J. Electron. Commer. Res.*, vol. 11, no. 3, pp. 220–230, 2010.

[37] K. R. Jerripothula, A. Rai, K. Garg, and Y. S. Rautela, “Feature-Level Rating System Using Customer Reviews and Review Votes,” *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 5, pp. 1210–1219, 2020, doi: 10.1109/TCSS.2020.3010807.

[38] K. R. Jerripothula, A. Rai, K. Garg, and Y. S. Rautela, "Feature-level Rating System using Customer Reviews and Review Votes," pp. 1–10.

[39] S. Saumya, J. P. Singh, A. M. Baabdullah, N. P. Rana, and Y. K. Dwivedi, "Ranking online consumer reviews," *Electron. Commer. Res. Appl.*, vol. 29, pp. 78–89, 2018, doi: 10.1016/j.elerap.2018.03.008.

[40] A. J. A. Husain, "New satisfying tool for problem solving in group decision-Support system," *Appl. Math. Sci.*, vol. 6, no. 109–112, pp. 5403–5410, 2012.

[41] Y. Pan and J. Q. Zhang, "Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews," *J. Retail.*, vol. 87, no. 4, pp. 598–612, 2011, doi: 10.1016/j.jretai.2011.05.002.

[42] R. He and J. McAuley, "VBPR: Visual Bayesian personalized ranking from implicit feedback," *30th AAAI Conf. Artif. Intell. AAAI 2016*, pp. 144–150, 2016.

[43] Amazon Customer Reviews Dataset. (n.d.). https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-dataset?select=amazon_reviews_us_Mobile_Electronics_v1_00.tsv

[44] S. Mundra, A. Dhingra, A. Kapur, and D. Joshi, "Prediction of a movie's success using data mining techniques," *Smart Innov. Syst. Technol.*, vol. 106, pp. 219–227, 2019, doi: 10.1007/978-981-13-1742-2_22.

[45] H. Gauba, P. Kumar, P. P. Roy, P. Singh, D. P. Dogra, and B. Raman, "Prediction of advertisement preference by fusing EEG response and sentiment analysis," *Neural Networks*, vol. 92, pp. 77–88, 2017, doi: 10.1016/j.neunet.2017.01.013.

[46] P. B. Palmer and D. G. O'Connell, "Research Corner: Regression Analysis for Prediction: Understanding the Process," *Cardiopulm. Phys. Ther. J.*, vol. 20, no. 3, pp. 23–26, 2009, doi: 10.1097/01823246-200920030-00004.

[47] T. Chai and R. R. Draxler, "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)," *Geosci. Model Dev.*, vol. 7, pp. 1525–1534, 2014.