Mitigating Online Product Rating Biases through the Discovery of Optimistic, Pessimistic, and Realistic Reviewers

Sunghoon Lim
Industrial and Manufacturing Engineering,
The Pennsylvania State University,
University Park, PA 16802
e-mail: slim@psu.edu

Conrad S. Tucker [1]
Mem. ASME
Engineering Design and
Industrial and Manufacturing Engineering,
The Pennsylvania State University,
University Park, PA 16802
e-mail: ctucker4@psu.edu

ABSTRACT

The authors of this work present a model that reduces product ratings biases that are a result of varying degrees of customers' optimism/pessimism. Recently, large-scale customer reviews and numerical product ratings have served as substantial criteria for new customers who make their purchasing decisions through electronic word-of-mouth. However, due to differences among reviewers' rating criteria, customer ratings are often biased. For example, a 3-star rating can be considered low for an optimistic reviewer. On the other hand, the same 3-star rating can be considered high for a pessimistic reviewer. Many existing studies of online customer reviews overlook the significance of reviewers' rating histories and tendencies. Considering reviewers' rating histories and tendencies is significant for identifying unbiased customer ratings and true product quality, because each reviewer has different criteria for buying and rating products. The proposed customer rating analysis model adjusts product ratings in order to provide customers with more objective and accurate feedback. The authors propose an unsupervised model aimed at mitigating customer ratings based on rating histories and tendencies, instead of human-labeled training data. A case study involving real-world customer rating data from an electronic commerce company is used to validate the method.

Keywords: Data-driven design, user generated data, electronic word-of-mouth, online review, customer rating

[1] Corresponding author

Conrad S. Tucker, MD-17-1247

1 INTRODUCTION

Customer feedback from those who have already purchased products is one of the significant factors for new customers to consider when making purchasing decisions [1-2]. Customer feedback can be extracted using the traditional sources, such as customer interviews, surveys, focus groups, and self-reports, that are predesigned for getting specific customer needs or preferences. Interviewed customers passively express their sentiment and experiences corresponding to predetermined questions or topics, but it is difficult to discover latent customer needs or preferences with predesigned topics or questions. In addition, customer interviews and surveys that require a sufficient number of customers are expensive and time consuming processes. However, due to reliable internet access, low-cost and large-scale online customer feedback, such as online customer reviews and ratings, is now available across the globe through electronic word-of-mouth (eWOM) in real-time [3-7]. Customers' textual reviews (i.e., qualitative feedback) and numerical star ratings (i.e., quantitative feedback) are significant criteria that determine if new customers will decide to purchase products through online commerce websites [8]. Customers can compare competing products based on online customer reviews and numerical ratings [9]. It is also known that higher ratings and more positive reviews correlate to higher product sales [10-11]. A 2012 Nielsen report found that online consumer reviews and ratings were the second most trusted source for customers, after recommendations from family and friends [12]. Most major electronic commerce companies (e.g., *Amazon.com*, *eBay*) provide star ratings in addition to online customer reviews [13]. For instance, star ratings on a scale of 1-5 can represent customers' opinions of a product from extremely negative (i.e., 1) to extremely positive (i.e., 5) [14].

While many studies of customer ratings exist, most overlook the effects of customers' rating histories and tendencies when analyzing online customer ratings. Online customer ratings are often biased, because different reviewers have different criteria when they buy and rate products. In this work, optimistic reviewers are defined as *biased* reviewers who have given relatively high ratings to all products that they have reviewed, including low-quality products. Pessimistic reviewers are defined as *biased* reviewers who have given relatively low ratings to all products, including high-quality products. Realistic reviewers are *unbiased* reviewers who have given high ratings to high-quality products and low ratings to low-quality products, respectively. Unreliable reviewers are defined as reviewers who have given high ratings to low-quality products and low ratings to high-quality products. A customer only having one review is classified as *not decided*, because only one review is not enough to determine his/her (optimistic/pessimistic) tendencies. However, because there is no ground truth for "true" product quality (i.e., the collective assessment of the product's perceived value across customers [15]), product sales rankings (i.e., the ordinal ranking of product sales within a product category (e.g., *Amazon Best Sellers*)) are used as an approximation of product quality instead [8]. Figure 1 indicates that optimistic and pessimistic reviewers are reviewers who have given high customer ratings to all products (the dotted blue line) and low customer ratings to all products (the dotted purple

line), respectively. Realistic reviewers have consistently given high ratings to high-ranked products and low rankings to low-ranked products in the product sales rankings (the solid green line). Unreliable reviewers have given high ratings to low-ranked products and low ratings to high-ranked products (the solid red line).

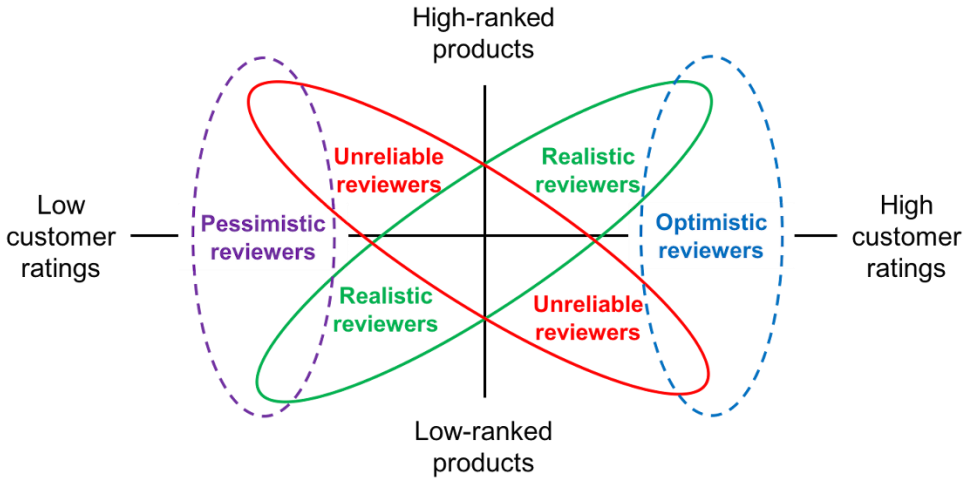

Fig. 1 Classification of optimistic/pessimistic/realistic/unreliable reviewers based on customer rating histories and product sales rankings

Figure 2 shows an example of the biased average star rating (i.e., 2.5 out of 5 stars) of the product P_X, along with the rating histories of two reviewers (i.e., Reviewer A and Reviewer B). Figure 2 indicates that Reviewer A might be categorized as an optimistic reviewer based on her product rating histories, because she has given 5-star ratings to all of the products, including the relatively low-ranked products (i.e., P_X). The second reviewer (i.e., Reviewer B) might be considered an optimistic reviewer as well, since she has given 5-star ratings to other products and has only given one 1-star rating to a relatively low-ranked product (i.e., P_X). Even though both reviewers (i.e., Reviewer A and Reviewer B) are considered optimistic reviewers, their ratings (i.e., 5 stars and 1 star, respectively) for the product significantly differ from each other. High ratings from optimistic reviewers (e.g., a 5-star rating from Reviewer A) can be considered overrated ratings. But low ratings from optimistic reviewers (e.g., a 1-star rating from Reviewer B) can be considered realistic ratings that express reviewers' substantial dissatisfaction. In the same manner, low ratings from pessimistic reviewers can be considered underrated ratings, but high ratings from pessimistic reviewers are considered realistic ratings that express reviewers' significant satisfaction. All ratings from realistic reviewers can be considered realistic, while all ratings from unreliable reviewers are considered unreliable and are disregarded.

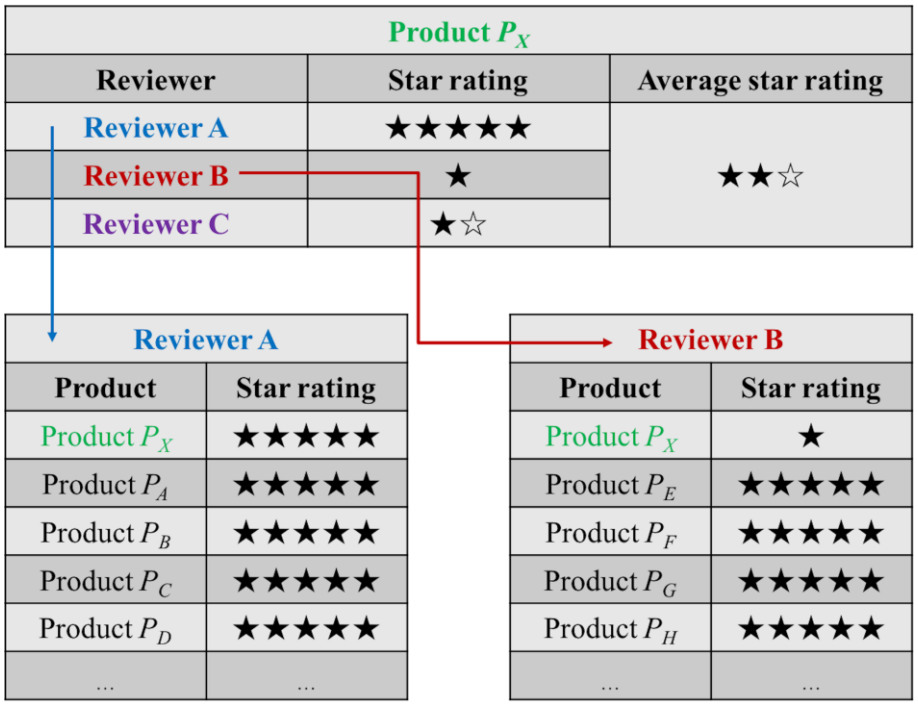

Fig. 2 An example of biased ratings, along with Reviewer A's and Reviewer B's product rating histories

Figure 2 indicates that the biased original average rating (i.e., 2.5 out of 5 stars) cannot represent the product's value, because, among the three reviewers, no one has actually rated the product near 2.5 stars. Hu et al. show that an overall distribution of customer ratings on *Amazon.com* is a *J*-shaped distribution (i.e., the solid blue line in Fig. 3), instead of a unimodal distribution (i.e., the dotted red line in Fig. 3), which is the distribution of controlled lab experimental results, in that everyone reviewed a randomly selected product (i.e., *Jason Mraz*'s *Mr. A-Z*) without purchasing it. This is because many customers tend to write reviews when they are significantly satisfied or significantly dissatisfied with their purchased products. The results indicate that current product ratings on a large number of electronic commerce websites do not represent true product quality, because the product ratings do not follow a normal distribution [15].

Fig. 3 The distributions of *Amazon.com*'s and the lab experiments' product star ratings

Therefore, in order to reduce the bias that exists among customers with varying degrees of optimism and pessimism, the authors of this work present a customer rating analysis method. In particular, an unsupervised model is presented to provide adjusted star ratings using customer rating histories and tendencies instead of human-labeled training data, because manually labeling training data is an expensive and subjective process [16] and there is no ground truth data for manually adjusting the original ratings. This model could prove helpful not only for customers' future purchasing decisions (from a customer's perspective) but also for predicting market sales (from an enterprise's perspective). The proposed method is especially useful in product design and development, because online product reviews and ratings can be used to predict emerging product trends and analyze customer requirements for new product development. In addition, biased product ratings can decrease accuracy in prediction and analysis [3]. A case study involving real customer rating data from an electronic commerce company (i.e., *Amazon.com*) validates that the adjusted star ratings better represent the product's true quality compared to the original star ratings.

The rest of the paper is organized as follows. This section provides an introduction and motivation for this work. Section 2 provides the literature review for this research. Section 3 explains the unsupervised numerical data-driven model, based on customer rating histories and tendencies, which provides the adjusted customer ratings. Section 4 introduces an application, while Section 5 provides the experimental results and discussion. Section 6 concludes the paper.

2 LITERATURE REVIEW

This section introduces works related to (1) online customer review analysis and (2) filtering and weighting online customer reviews and ratings.

2.1 Online Customer Review Analysis

Online customer reviews and numerical ratings express collective wisdom. When utilized efficiently, these ratings and reviews indicate future outcomes [17]. Analyzing customer reviews and ratings is essential for discovering customer requirements for market success [18]. Several researchers have developed theoretical approaches for customer review analysis. Dave et al. propose an automated method that distinguishes between positive and negative customer reviews [13]. Menon et al. present a vector space document representation method that derives customer requirements from consumer reviews and numerical ratings for new product development [19]. Rai employs text mining techniques to mine product attributes in order to rank attributes' importance in customer reviews [20]. Wong and Lam identify customer requirements from customer reviews on multiple auction websites using hidden Markov models and conditional random fields [21]. They also exploit a document object model to predict popular products on auction websites that contain customer required functions [22]. Wu and Huberman analyze the temporal evolution of large-scale customer reviews, discovering that latter reviews tend to show a large difference from previous ones, which in turn moderates the average review to the less extreme [23].

Online customer reviews and star ratings are used in various research areas, including data-driven product design. In order to enhance new product design processes, Tucker and Kim propose a method that utilizes customer review data to model and predict emerging product trends [24]. Liu et al. propose four feature categories that reflect designers' viewpoints. They develop a method that automatically evaluates the helpfulness of online customer reviews from a designer's perspective [25]. Ferguson et al. present an ergonomically-centered cue-phrase set for extracting useful information from online customer reviews, in order to inform designers for creating products that are universal and well-received by customers [26]. Online customer reviews and star ratings are also used to predict future product sales [27]. Chen et al. analyze how the social status of reviewers affect customer responses to consumer review information. They discover that highlighted customer reviews more strongly affect product sales than other reviews [11]. Numerical ratings, such as star ratings, are used to quantitatively review overall product quality [8]. For instance, McGlohon et al. analyze customer reviews and numerical ratings from different websites to measure true product quality [28].

Nevertheless, considerations of consumer rating histories and (optimistic/pessimistic) tendencies for customer rating analyses are still limited. Therefore, discovering unbiased customer ratings for various products remains uncertain. The method presented in this paper reflects customer rating histories and tendencies for customer rating analyses in order to obtain unbiased customer feedback.

2.2 Filtering and Weighting Online Customer Reviews and Ratings

While reliable reviews and ratings are important for customer review analyses, unreliable reviews, including opinion spam and fake reviews and ratings, might hinder accurate and objective customer review analyses. For example, conflicting aggregated numerical ratings decrease review credibility [29]. The problems associated with unreliable reviews and ratings have recently increased as well. For example, large numbers of paid fake reviewers have been recently detected [30]. The number of group and individual spammers has also recently increased [31-32].

Filtering and weighting online customer reviews and ratings are relatively new and emerging in customer review analyses. Online customer review filtering methods filter out customer reviews and ratings that disturb accurate analyses, including opinion spam and fake reviews. Weighting online customer reviews considers reliable or helpful reviews more significantly than normal reviews. Lim et al. present product scoring methods in order to identify spam reviewers using an *Amazon* review dataset [31]. Willemsen et al. analyze customer reviews and provide usefulness scores based on reviews from fellow customers [33]. Hu et al. test the underlying distribution of online customer ratings from *Amazon.com* and discover that the average score does not represent the product's true quality and that review filtering is necessary for measuring true quality [34]. A statistical method has been recently developed in order to investigate the writing style of reviewers and the effectiveness of manipulation through customer ratings, sentiments, and readability [2]. To mitigate these issues, *Amazon.com* provides the *Amazon Verified Purchase* label, which verifies that the person writing the review, purchased the product from *Amazon.com* [35].

Table 1 summarizes previous studies and this work. While previous studies have been widely applied to filtering and weighting online customer reviews and ratings, considering customer rating histories and tendencies remains ambiguous. This consideration is significant in customer review analyses, because overrated ratings by optimistic customers (or underrated ratings by pessimistic customers) can bias the results of customer review analyses and give biased information to future customers and designers seeking to mine the large-scale and abundantly-available product review data. The proposed work not only filters out and weights unreliable customer ratings but also adjusts customer ratings based on rating histories and tendencies.

Table 1 Summary of previous studies and this work

References	Filtering and weighting customer ratings	Adjusting customer ratings
[2,10,11,14,31,33]	√	
Ours	√	√

3 METHOD

Figure 4 outlines the steps involved in analyzing the original star ratings and providing adjusted star ratings. First, customer star ratings, along with each reviewer's rating histories and other available information, are extracted from online commerce websites. Then, online customer ratings are normalized to a default scale (i.e., a 1-5 scale) if they use a different scale. Filtering customer ratings filters out unverified reviews. A minimum distance classifier categorizes reviewers as realistic, optimistic, pessimistic, unreliable, or not decided, based on their rating histories and product sales ranks, instead of human-labeled training data. Finally, the proposed method maps optimistic/pessimistic customer ratings into realistic ratings in order to provide future customers with more objective and accurate information for their purchasing decisions.

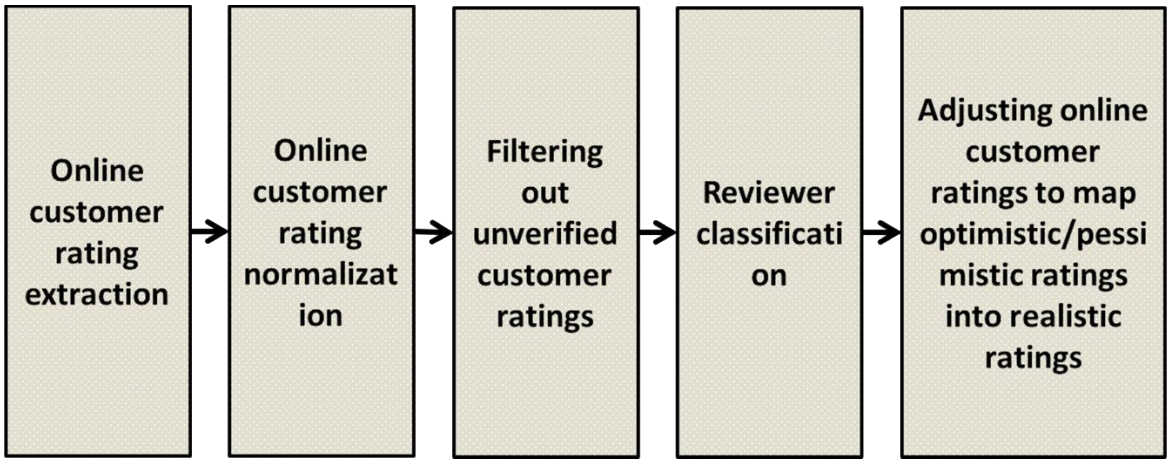

Fig. 4 Overview of the proposed method

3.1 Data Acquisition

Customer numerical ratings, reviewer information (e.g., user ID), product sales rankings, and other necessary information (e.g., ASIN, *Amazon.com*'s *Verified Purchase* marks) are extracted from online product data streams for each product. Each reviewer's historical data (i.e., each reviewer's product ratings and sales rankings of products they have reviewed in the past) are used as well. Web scraping applications (e.g., *Amazon Product Advertising* API) enable the scraping of website content in order to create a customer review database. A customized web scraper API (e.g., import.io) can be used to automatically extract product review data as well [7].

3.2 Online Customer Rating Normalization

Customer rating normalization is necessary because different electronic commerce companies provide different numerical customer rating systems [28]. A 1-5 scale is the most popular scale used on many electronic commerce websites. However, other scales

are also used for some companies (e.g., a 1-10 scale for *Redbox*). In this case, scales are proportionally converted to a 1-5 scale. For instance, 8 stars on a 1-10 scale can be approximately converted to 3.11 stars on a 1-5 scale $\left(\frac{8-1}{10-1} \cdot (5-1) \simeq 3.11\right)$. A 1-5 scale is set to a default scale, since most of major electric commerce companies use a 1-5 scale, such as *Amazon.com* and *eBay*. In addition, it is assumed that only an overall rating is used in this study, even when sub-ratings (e.g., *value*, *quality*), as well as overall ratings, are provided (e.g., *J. C. Penney*). Future work will use not only overall ratings but also sub-ratings based on product features, such as "price" and "quality", in order to discover significant product features for customer review analysis.

3.3 Filtering Customer Ratings Written by Unverified Reviewers

Filtering unverified reviewers helps remove opinion spammers and fake reviewers, which improves performance and reduces the bias caused by spams and fake reviews [35]. Some electronic commerce company websites provide verification marks for verified reviewers, such as *Amazon.com*'s *Verified Purchase* marks, which indicate reviewers who have purchased the product through *Amazon.com*. In this work, only verified users (as deemed by the electronic commerce company) are included in the model, hereby filtering all unverified users. The adjusted average ratings are calculated excluding unverified users' ratings, while the original average ratings are calculated including not only verified users' ratings but also unverified users' ratings.

3.4 Classification of Optimistic, Pessimistic, Realistic, and Unreliable Reviewers without Human-Labeled Training Data

Each reviewer who has rated the same product and has multiple reviews (i.e., at least two reviews) can be classified as an optimistic (O), a pessimistic (P), a realistic (R), or an unreliable (U) reviewer based on correlations between product rating histories and product sales rankings. A reviewer who has only one review is classified as not decided (N). Based on the definitions of an optimistic, a pessimistic, a realistic, and an unreliable reviewer (see Section 1), optimistic reviewers have relatively high averages and low standard deviations for their star ratings (the solid blue line in Fig. 5). Pessimistic reviewers have relatively low averages and low standard deviations (the solid purple line in Fig. 5). Realistic and unreliable reviewers tend to have medium averages (i.e., lower than optimistic reviewers' averages and higher than pessimistic reviewers' averages) and high standard deviations for their star ratings. Two different types of realistic and unreliable reviewers exist. One type is realistic or unreliable reviewers, who give various values (i.e., 1, 2, 3, 4, or 5 stars) for rating products (the solid green line in Fig. 5). The other type is realistic or unreliable reviewers, who give only extreme values (i.e., 1 or 5 stars) for rating products (the dotted green line in Fig. 5).

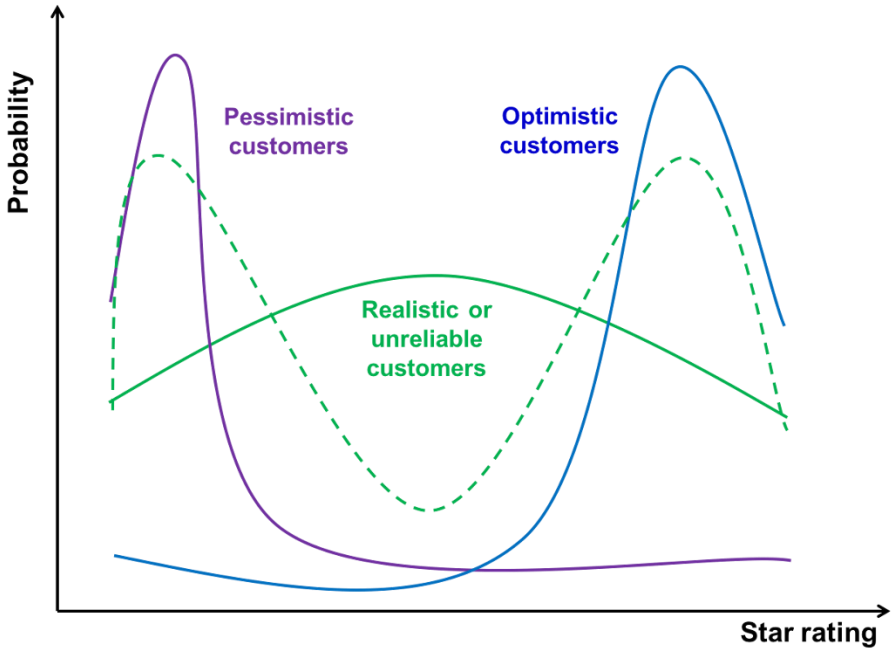

Fig. 5 Possible probability distributions of product ratings of optimistic, pessimistic, realistic, and unreliable reviewers

However, the averages and standard deviations of customer ratings are not enough to classify optimistic, pessimistic, realistic, and unreliable reviewers. (1) It is very challenging to distinguish between realistic reviewers and unreliable reviewers only with the averages and standard deviations (see Fig. 5) and (2) many reviewers have only a few reviews, which are not enough for an analysis of averages and standard deviations (e.g., a customer having only two 5-star ratings). Therefore, product sales rankings (e.g., *Amazon Best Sellers*) are also used to classify optimistic, pessimistic, realistic, and unreliable reviewers. Product sales rankings of only the top T products listed in each category on the website (e.g., the top 100 bestsellers on *Amazon.com*), instead of all products, are considered, because the number of products in each category is not the same and often times, quite large. The products listed in the top T product lists are guaranteed high-quality products [36-37]. For example, it is not easy to estimate the quality of the 2,675th-ranked product in the *Music* category and the *Cell Phones* category on *Amazon.com*, respectively, because the total number of products in each category is not the same. On the other hand, the 3rd-ranked product that is listed on *Amazon Best Sellers* might be a guaranteed good product, regardless of the product categories. If a product has multiple ranks, a higher rank (i.e., a lower number) is used to represent the product. A normalized value of a product sales rank is used in order to (1) normalize the value to a range 0 to 1 and (2) give higher weightings to high-ranked products. I is defined as Eq. (1).

$$I = \frac{T - min\{a\ sales\ rank\ of\ the\ product\}}{T-1} \cdot \mathbf{1}_{\min\{a\ sales\ rank\ of\ the\ product\} \leq T} \quad (1)$$

Realistic reviewers' product ratings and I (i.e., the normalized value of a product sales rank) have high positive correlation coefficients (i.e., near 1), because high-ranked products have low numerical values of their product sales rankings. On the other hand, unreliable reviewers' product ratings and I have high negative correlation coefficients (i.e., near -1), because their ratings are assumed to be unreliable. Optimistic and pessimistic reviewers' product ratings and I have near-zero correlation coefficients, because they have given high and low ratings for all products, respectively. Table 2 shows an example of the values of the average, the standard deviation, and the correlation coefficient for four different types of reviewers (i.e., optimistic (C_1, C_5), pessimistic (C_2, C_6), realistic (C_3), and unreliable reviewers (C_4), respectively) who reviewed the same product P_Y. Table 3 illustrates how to calculate the average, standard deviation, and correlation coefficient values for the first reviewer (C_1) in Table 2. P_{11}, P_{12}, P_{13}, and P_{14} indicate the first, second, third, and fourth products that the first reviewer have purchased, respectively.

Table 2 An example of how to convert the original customer ratings to the adjusted customer ratings for each product

Product P_Y						
Reviewer (C_i)	**Average (C_{i1})**	**Standard deviation (C_{i2})**	**Correlation coefficient (C_{i3})**	**O/P/R/U/N**	**Original rating (R_{xi})**	**Adjusted rating (R'_{xi})**
1	4.75	0.43	0.03	O	5	$5 + \alpha \cdot (A_{XR} - 5)$
2	1.50	0.50	-0.08	P	2	$2 + \alpha \cdot (A_{XR} - 2)$
3	3.00	1.58	0.99	R	4	4
4	2.75	1.79	-0.94	U	3	–
5	4.85	0.27	0.03	O	1	1
6	1.47	0.32	0.19	P	5	5
7	–	–	–	N	4	4
…	…	…	…	…	…	…
Average product rating	–				4.3	3.9

Table 3 An example of how to calculate the average, standard deviation, and correlation coefficient values for the first reviewer (C_1)

Product	**P_{11}**	**P_{12}**	**P_{13}**	**P_{14}**	**Average**	**Standard deviation**
Product rating	4	5	5	5	4.75	0.43
min{sales rank}	67	11	99	85	–	
I	0.33	0.90	0.01	0.15		
Correlation coefficient	0.03					

Each reviewer can be classified as an optimistic, a pessimistic, a realistic, or an unreliable reviewer based on the values of the average, the standard deviation, and the correlation coefficient. Because human-labeled training data for classification are not used in this work, a minimum distance classifier, that is one of the existing classification algorithms, is applied based on extreme cases (i.e.,

templates or noise-free feature vectors) for each reviewer that can be assumed instead as follows [38]. Let $\boldsymbol{C_i}$ be a feature vector for a reviewer C_i in a three-dimensional space (Eq. (2)). For example, reviewer C_1 is expressed as $\boldsymbol{C_1}$ ={4.75, 0.43, 0.03}. The average rating has a range of 1 to 5, since a 1-5 scale is used as the default. By its definition, a correlation coefficient has a range of -1 to 1. Equation (3), which uses Popoviciu's inequality on variances, proves that the standard deviation has a range of 0 to 2 in this work [39]. Four (hypothetical) feature vectors, $\boldsymbol{M_1} = \{5, 0, 0\}$, $\boldsymbol{M_2} = \{1, 0, 0\}$, $\boldsymbol{M_3} = \{3, 2, 1\}$, and $\boldsymbol{M_4} = \{3, 2, -1\}$, are used as templates for indicating the *extreme* cases of an optimistic (O), a pessimistic (P), a realistic (R), and an unreliable (U) reviewer, respectively. For instance, an extremely optimistic reviewer (i.e., $\boldsymbol{M_1} = \{5, 0, 0\}$) can be a reviewer who have given only 5-star ratings. The reviewer has an average of 5, a standard deviation of 0, and a correlation coefficient of 0. On the other hand, an extremely unreliable reviewer (i.e., $\boldsymbol{M_4} = \{3, 2, -1\}$) can be a reviewer who has given 1-star ratings for half of the products that have the best qualities (i.e., I=1) and 5-star ratings for the other half of the products that have low qualities (i.e., I=0) yielding the maximum value of a standard deviation (i.e., 2) and the minimum value of a correlation coefficient (i.e., -1).

A scaled Euclidean distance between two vectors (i.e., $\boldsymbol{C_i}$ and $\boldsymbol{C_j}$) is defined as Eq. (4), because the average, the standard deviation, and the correlation coefficient have different ranges (i.e., a range of 1 to 5 for the average, a range of 0 to 2 for the standard deviation, a range of -1 to 1 for the correlation coefficient). A minimum distance classifier classifies each reviewer C_i as Eq. (5). Figure 6 shows an example of applying a minimum distance classifier.

$$\boldsymbol{C_i} = \{C_{i1}, C_{i2}, C_{i3}\} = \{average\ rating, standard\ deviation\ of\ ratings, correlation\ coefficient\} \tag{2}$$

$$0 \le \sigma(rating) = \sqrt{VAR(rating)} \le \sqrt{\frac{(\max(rating) - \min(rating))^2}{4}} = \sqrt{\frac{(5-1)^2}{4}} = 2 \tag{3}$$

$$D(\boldsymbol{C_i}, \boldsymbol{C_j}) = \sqrt{\left(\frac{C_{i1} - C_{j1}}{5-1}\right)^2 + \left(\frac{C_{i2} - C_{j2}}{2-0}\right)^2 + \left(\frac{C_{i3} - C_{j3}}{1-(-1)}\right)^2} = \sqrt{\left(\frac{C_{i1} - C_{j1}}{4}\right)^2 + \left(\frac{C_{i2} - C_{j2}}{2}\right)^2 + \left(\frac{C_{i3} - C_{j3}}{2}\right)^2} \tag{4}$$

$$\begin{cases} C_i \in O & if \ \arg\min_{j\in\{1,2,3,4\}} \{D(\boldsymbol{C_i}, \boldsymbol{M_j})\} = 1 \\ C_i \in P & if \ \arg\min_{j\in\{1,2,3,4\}} \{D(\boldsymbol{C_i}, \boldsymbol{M_j})\} = 2 \\ C_i \in R & if \ \arg\min_{j\in\{1,2,3,4\}} \{D(\boldsymbol{C_i}, \boldsymbol{M_j})\} = 3 \\ C_i \in U & if \ \arg\min_{j\in\{1,2,3,4\}} \{D(\boldsymbol{C_i}, \boldsymbol{M_j})\} = 4 \end{cases} \tag{5}$$

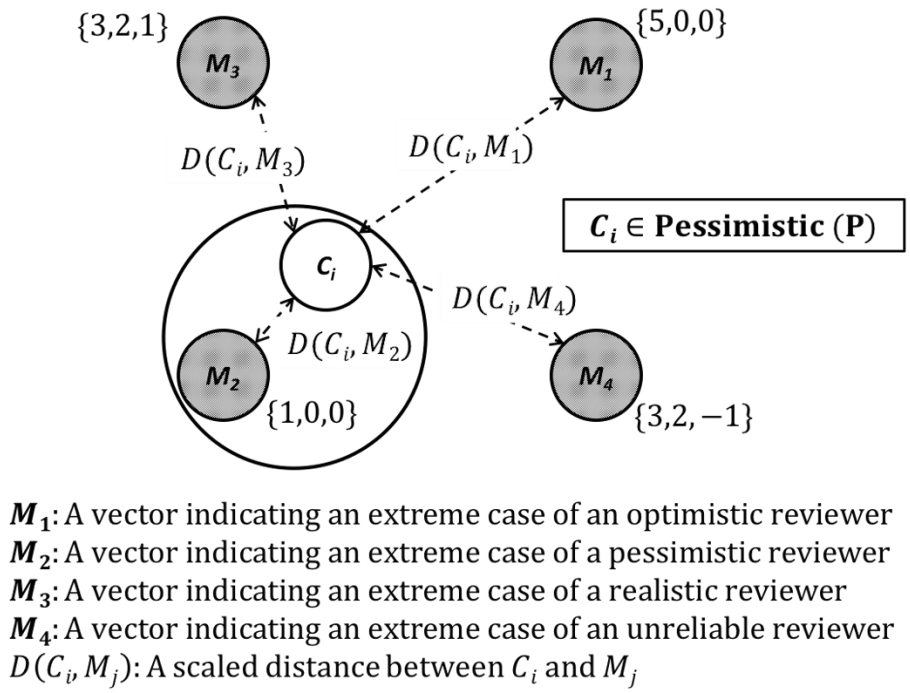

Fig. 6 An example of applying a minimum distance classifier to classify C_i

3.5 Adjusting Customer Ratings Written by Optimistic/Pessimistic Reviewers

The original product ratings by realistic reviewers (R) are not adjusted, since their product ratings are considered already realistic and represent true product qualities (see Section 1). The original product ratings by reviewers classified as not decided (N) are not adjusted as well, since only one review is not enough to determine his/her rating tendencies. Low ratings by optimistic reviewers (O) and high ratings by pessimistic reviewers (P) are not adjusted, since they are assumed to be realistic ratings as well (see Section 1). On the other hand, the original product ratings by unreliable reviewers (U) are filtered out and not considered in further analysis. High product ratings by optimistic reviewers (O) that are overrated and low product ratings of pessimistic reviewers (P) that are underrated are converted to the adjusted customer ratings. Let A_{XR} be the average value of product ratings by only realistic reviewers (R) for a product P_X, which is assumed as a breakpoint to distinguish between overrated ratings and realistic ratings by optimistic reviewers (or distinguish between underrated ratings and realistic ratings by pessimistic reviewers) in this work. Table 2 illustrates an example of how to convert customers' original ratings to the proposed adjusted customer ratings for each product. α is defined as an adjusting factor that has a range of 0 to 1. If α is set to 1, high ratings by optimistic reviewers and low ratings by pessimistic reviewers are converted to A_{XR}, and an adjusting effect is fully applied (e.g., $5 + \alpha \cdot (A_{XR} - 5) = 5 + 1 \cdot (A_{XR} - 5) = A_{XR}$). On the other hand, if α is set to 0, product ratings by optimistic and pessimistic reviewers are not changed, and an adjusting effect is not applied (e.g., $5 + \alpha \cdot (A_{XR} - 5) = 5 + 0 \cdot$

$(A_{XR} - 5) = 5)$. In this case, all customer ratings are not changed, except unreliable reviewers' ratings, which are filtered out. α is set to 1 as the default, but different values of α (e.g., 0.5) can be used for different online commerce websites and different products. Let R_{Xi} and R'_{Xi} be the original rating and the adjusted rating by a reviewer C_i for a product P_X, respectively. Equation (6) indicates how to calculate the proposed adjusted ratings for a reviewer C_i.

$$R'_{Xi} = \begin{cases} R_{Xi}, & C_i \in O, R_{Xi} < A_{XR} \\ R_{Xi} + \alpha \cdot (A_{XR} - R_{Xi}), & C_i \in O, R_{Xi} \geq A_{XR} \\ R_{Xi} + \alpha \cdot (A_{XR} - R_{Xi}), & C_i \in P, R_{Xi} < A_{XR} \\ R_{Xi}, & C_i \in P, R_{Xi} \geq A_{XR} \\ R_{Xi}, & C_i \in \{R \cup N\} \\ (fliltered\ out), & C_i \in U \end{cases} \quad (6)$$

Algorithm 1 summarizes the steps of the proposed model in order to provide the adjusted star ratings for a product P_X.

Algorithm 1: The proposed unsupervised model that maps optimistic/pessimistic customer ratings into realistic ratings for a product P_X

STEP 1 Extract customer ratings and reviewer IDs for P_X, reviewers' rating histories, and product sales rankings
STEP 2 Normalize customer ratings to a 1-5 scale, if they use a different rating scale
STEP 3 Filter out unverified reviewers
STEP 4 Create a feature vector $\boldsymbol{C_i} = \{C_{i1}, C_{i2}, C_{i3}\}$ for each customer C_i who has rated P_X
STEP 5 If $\arg\min_{j \in \{1,2,3,4\}} \{D(\boldsymbol{C_i}, \boldsymbol{M_j})\} = 1$ and $R_{Xi} \geq A_{XR}$,
$R'_{Xi} = R_{Xi} + \alpha \cdot (A_{XR} - R_{Xi})$
Else if $\arg\min_{j \in \{1,2,3,4\}} \{D(\boldsymbol{C_i}, \boldsymbol{M_j})\} = 2$ and $R_{Xi} < A_{XR}$, $R'_{Xi} = R_{Xi} + \alpha \cdot (A_{XR} - R_{Xi})$
Else if $\arg\min_{j \in \{1,2,3,4\}} \{D(\boldsymbol{C_i}, \boldsymbol{M_j})\} = 4$, filter out C_i and remove R_{Xi}
Otherwise, $R'_{Xi} = R_{Xi}$
STEP 6 If all the values of R'_{Xi} for all reviewers who have reviewed a product P_X are obtained, go to STEP 7
Otherwise, go to STEP 4 and repeat the process
STEP 7 Calculate and return the average of R_{Xi} and R'_{Xi}, respectively, and stop

3.6 The Complexity of the Proposed Adjusted Ratings

This section provides the algorithmic complexity of the proposed unsupervised model. Big-O notation, that gives the asymptotic upper bound on execution time but is not necessarily related to running time for every input combination, is used for the complexity [40]. Suppose that the maximum number of reviewers for each product and the maximum number of product ratings that each reviewer has written are C and D, respectively. Each calculation of the average, the standard deviation, and the correlation coefficient involves O(D). O(1) is required to calculate a scaled Euclidean distance between a feature vector of each reviewer and a template (i.e., M_1, M_2, M_3, or M_4) and classify each reviewer an optimistic, a pessimistic, a realistic, or an unreliable reviewer, respectively (see Eqs. (4) and (5)). Calculating the average adjusted rating for each product involves O(CD), because each adjusted rating requires O(D) and the

calculation of the average of all reviewers' adjusted ratings involves O(C) (see Table 2). Therefore, the proposed unsupervised model is a polynomial time algorithm and the execution time for presenting the adjusted rating for each product is directly proportional to CD (i.e., O(CD)).

3.7 Validation of the Proposed Adjusted Ratings

Given the absence of ground truth data for product ratings (i.e., the challenge of following up with each customer to determine whether they feel as though their review was biased), two kinds of approximated validations are used in this research. The first approximated validations are based on the correlation coefficients (*Corr*) between I (see Eq. (1)) and (1) products' original average star ratings and (2) products' adjusted average star ratings, respectively. Table 4 is an example of product sales rankings that do not significantly correlate to the products' original average ratings. For instance, the first product on the list has the lowest average rating among the *best six products* (i.e., deemed best sellers on *Amazon.com*).

Table 4 An example of product sales rankings (i.e., *Amazon Best Sellers*) and product ratings

Best Sellers in Over-Ear Headphones		
–	Product sales ranking	Product average rating
Product A	1	4.3
Product B	2	4.4
Product C	3	4.8
Product D	4	4.5
Product E	5	4.7
Product F	6	4.6
…	…	…

The null ($H1_0$) and alternative ($H1_a$) hypotheses are developed to provide the approximated validations as below. The Z-test can be useful to test hypotheses, but the Z-test requires normally distributed values. However, the correlation coefficients of the star ratings that follow a *J*-shaped distribution do not follow a normal distribution. Fisher's *z*-transformation (Eq. (7)) is therefore used for transforming the correlation coefficients that are not normally distributed to normal distributions. The test statistic for Fisher's Z-test (Eq. (8)), that are generally used to assess the significance of the difference between two correlation coefficients, are applied to test the hypotheses [41]. A *p*-value that is defined as the conditional probability of finding the observed or more extreme results when the null hypothesis is true is used to test hypotheses. A small *p*-value provides strong evidence that the null hypothesis is not true. The *p*-values can be obtained by the values of Z in Eq. (8) as well.

$H1_0$: *Corr*(*I*, adjusted average ratings) ≤ *Corr*(*I*, original average ratings)

$H1_a$: *Corr*(*I*, adjusted average ratings) > *Corr*(*I*, original average ratings)

$$z \coloneqq \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{7}$$

$$Z = \frac{z_a - z_o}{\sqrt{\frac{2}{n_p - 3}}} \tag{8}$$

where:

r: the correlation coefficient

z: the transformed value of the correlation coefficient

z_o: the transformed value of the correlation coefficient of the original star ratings

z_a: the transformed value of the correlation coefficient of the adjusted star ratings

n_p: the total number of products

Z: the test statistic for Fisher's Z-test

If the correlation coefficients of the adjusted star ratings are significantly higher than the correlation coefficients of the original star ratings and $H1_0$ is rejected, the adjusted average ratings (e.g., 3.9 in Table 2) can be used as an indication of product popularity in the market instead of the original average ratings (e.g., 4.3 in Table 2). Table 5 indicates an example of the correlation coefficients between *I* and (1) the original average ratings and (2) the adjusted average ratings, respectively, where α is 1. Table 5 also shows the *p*-value.

Table 5 An example of the correlation coefficients between *I* (i.e., the normalized values of product sales rankings) and (1) the original average ratings and (2) the adjusted average ratings, respectively

Product	*I*	Original average rating	Adjusted average rating
P_1	1	4.3	3.9
P_2	0.22	2.7	3.6
…	…	…	…
Correlation coefficient		0.45	0.81
***p*-value**		≃ 0.000	

The second approximated validations are based on the correlation coefficients between the average sentiment scores of reviewers and (1) reviewers' original average star ratings and (2) reviewers' adjusted average star ratings, respectively. The null ($H2_0$) and alternative ($H2_a$) hypotheses are presented to provide the approximated validations as below. Fisher's *z*-transformation (Eq. (7)), the test statistic for Fisher's Z-test (Eq. (9)), and the *p*-values are used to test the hypotheses as well.

$H2_0$: *Corr*(adjusted average ratings, average sentiment scores of reviewers) ≤ *Corr*(original average ratings, average sentiment scores of reviewers)

$H2_a$: *Corr*(adjusted average ratings, average sentiment scores of reviewers) > *Corr*(original average ratings, average sentiment scores of reviewers)

$$Z = \frac{z_a - z_o}{\sqrt{\frac{2}{n_r - 3}}} \tag{9}$$

where:

z_o: the transformed value of the correlation coefficient of the original star ratings

z_a: the transformed value of the correlation coefficient of the adjusted star ratings

n_r: the total number of reviewers

Z: the test statistic for Fisher's Z-test

Because labeled training data is not used in this work, existing trained sentiment classifiers (e.g., *The Natural Language Toolkit* (*NLTK*) [42]) are exploited for calculating reviewers' sentiment scores based on their textual review histories. All textual reviews by each reviewer are used as inputs and the outputs are reviewers' sentiment scores. Sentiment scores on a scale of 0-1 can represent reviewers' sentiments for products, from extremely negative (i.e., 0) to extremely positive (i.e., 1). Textual reviews are only used for this validation and are unnecessary for providing the adjusted customer ratings.

If the correlation coefficients of the adjusted star ratings are significantly higher than the correlation coefficients of the original star ratings and $H2_0$ is rejected, the adjusted average ratings can be used as an indication of unbiased customer feedback instead of the original average ratings. Table 6 shows an example of the correlation coefficients between reviewers' sentiment scores and (1) the original average ratings and (2) the adjusted average ratings, respectively.

Table 6 An example of the correlation coefficients between the average sentiment scores of reviewers and (1) the original average ratings and (2) the adjusted average ratings, respectively

Reviewer	Average sentiment score	Original average rating	Adjusted average rating
C_1	0.8	4.5	4.6
C_2	0.1	2.9	2.5
…	…	…	…
Correlation coefficient		0.40	0.77
***p*-value**		$\simeq 0.000$	

4 APPLICATION

This section introduces a case study involving real-world customer rating data from an electronic commerce company (i.e., *Amazon.com*) in order to validate the proposed research, which better represents true product quality. Experiments are conducted on a 2.5 GHz *Intel Core i7* with 16GB RAM using *JAVA Development Kit* 8.0. *Amazon* customer rating data, provided by McAuley et al., are also used [43-44]. 1,689,188 reviews and 27,446 reviewers from the *Electronics* category on *Amazon.com* are used in this case study. Customer ratings, textual reviews, and reviewer IDs, along with information for *Verified Purchase* marks and product sales ranks, are extracted for each product. Each reviewer's historical data (i.e., each reviewer's product ratings, textual reviews, and sales rankings of products they have reviewed in the past) are used as well. T is set to 100 based on *Amazom.com*'s lists of the top 100 best sellers. Different values of α (i.e., 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1) are used in order to identify the effect of the different values of α. 0.05 is used as a significance level for the p-values in this study, because 0.05 is the most common value for the hypothesis tests.

In this case study, a baseline is defined as the original average ratings of all reviewers that are currently used on most electric commerce websites (e.g., *Amazon.com*). The results of the proposed adjusted star ratings are compared to a baseline for products in the *Electronics* category. Other existing methods are not used for comparison, because the proposed method is the only method (to the best of our knowledge) that directly adjusts original customer star ratings without human-labeled training data. *NLTK* is applied as the existing trained sentiment classifier in order to provide reviewer's average sentiment scores for validation [42].

5 RESULTS AND DISCUSSION

Among totals of 3,799 products in the *Electronics*, 739 products are listed in *Amazon Best Sellers* (i.e., the top 100 best sellers) that are used in this case study. Table 7 and Fig. 7 show the results of the correlation coefficients and the p-values for the correlation coefficients of product sales rankings in the *Electronics* category. Table 8 and Fig. 8 indicate the results of the correlation coefficients and the p-values for the correlation coefficients of reviewers' sentiment scores from the *Electronics* category. For instance, the p-value

of the adjusted ratings with α=0.5 (i.e., 0.0146) is obtained using the values of the correlation coefficient of the original ratings (i.e., 0.6867) and the correlation coefficient of the adjusted ratings with α=0.5 (i.e., 0.7423) in Table 7.

According to Table 7 and Fig. 7, the adjusted ratings and product sales rankings provide significantly higher correlation coefficients than the original ratings (i.e., baseline) and product sales rankings, regardless of the values of α, because all the *p*-values are less than 0.05 (i.e., a significance level in this study). Table 8 and Fig. 8 also indicate that the correlation coefficients of the adjusted ratings and reviewers' sentiment scores are significantly higher than the correlation coefficients of the original ratings and reviewers' sentiment scores, regardless of the values of α, since all the *p*-values are significantly less than 0.05. The research hypotheses $H1_0$ and $H2_0$ are rejected for all the values of α. It is therefore concluded that the proposed adjusted ratings outperform the original ratings.

Tables 7 and 8 also show that the correlation coefficients of the adjusted ratings, where α is 0, have larger values than the correlation coefficients of the original ratings in both cases (i.e., the correlation coefficients of sales rankings and the correlation coefficients of the sentiment scores). The results show that the effect of filtering out unreliable reviewers (U) is not negligible, since deciding whether or not to filter out unreliable reviewers is the only difference between the adjusted ratings, where α is 0, and the original ratings (see Eq. (6) and Table 2).

According to the results in Tables 7 and 8 and Figs. 7 and 8, the correlation coefficients slightly increase as the value of α increases in both cases. This means the adjusted star ratings are more accurate for estimating true product quality when the higher adjusting effects are applied. The average correlation coefficients of sales rankings (i.e., 0.7422) are substantially greater than the average correlation coefficients of sentiment scores (i.e., 0.3556). It is therefore assumed that product sales rankings can be used for more accurate approximated validations of the adjusted customer ratings than reviewers' sentiment scores. However, further analysis is necessary in the future. For example, the *p*-values of reviewers' sentiment scores are significantly less than the *p*-values of product sales rankings.

Table 9 shows that the most reviewers are classified as not decided (about 84%) since they only have a single review under their account. A research expansion to classify reviewers only having one review and adjust their ratings will be considered in the future. Table 9 indicates that the number of realistic reviewers is the highest among reviewers who has multiple reviews (about 49%). It is postulated that the proportion of unbiased reviewers (i.e., realistic reviewers) and the proportion of biased (i.e., optimistic or pessimistic) or unreliable reviewers are almost the same on *Amazon.com*, but the further analysis is necessary. Table 9 also indicates that the number of optimistic reviewers is considerably larger than the number of pessimistic reviewers. This result shows that customers tend to write reviews when they are substantially satisfied with their purchased products.

Figure 9 shows the distributions of the original star ratings and the adjusted star ratings for 739 products that are listed in *Amazon Best Sellers* in *Electronics*. The original ratings in Fig. 9 follow a *J*-shaped distribution and correspond with Hu et al.'s findings (i.e., the *J*-shaped distribution in Fig. 3) [15]. The distributions of the adjusted ratings, where α=0.5 and α=1, are slightly shifted from a *J*-shaped distribution to a normal distribution. The shift is not significant, because the most reviewers are classified as not decided in this case study. However, the distributions in a range of 4 stars to 5 stars are relatively significantly changed, because the number of optimistic reviewers is greater than the number of pessimistic reviewers.

Table 7 The correlation coefficients and the *p*-values of product sales rankings

		Correlation coefficient of product sales rankings	*p*-value
Original ratings (baseline)		0.6867	–
Adjusted ratings with α=	0	0.7385	0.0217
	0.1	0.7393	0.0197
	0.2	0.7401	0.0183
	0.3	0.7408	0.0170
	0.4	0.7416	0.0154
	0.5	0.7423	0.0146
	0.6	0.7430	0.0136
	0.7	0.7436	0.0125
	0.8	0.7443	0.0116
	0.9	0.7449	0.0107
	1	0.7455	0.0102
	Average	0.7422	–

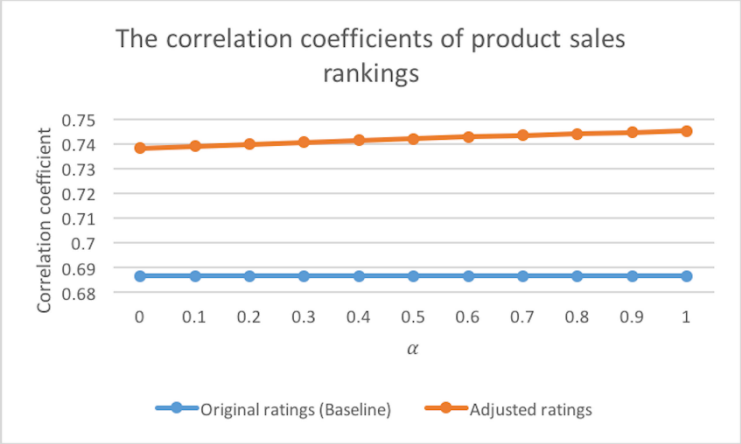

Fig. 7 The correlation coefficients of product sales rankings

Table 8 The correlation coefficients and the *p*-values of reviewers' sentiment scores

	Correlation coefficient of reviewers' sentiment scores	*p*-value
Original ratings (baseline)	0.3268	–

Adjusted ratings with α=	0	0.3536	0.0002
	0.1	0.3541	0.0002
	0.2	0.3545	0.0001
	0.3	0.3548	0.0001
	0.4	0.3552	0.0001
	0.5	0.3556	0.0001
	0.6	0.3559	0.0001
	0.7	0.3563	≃ 0.0000
	0.8	0.3567	≃ 0.0000
	0.9	0.3570	≃ 0.0000
	1	0.3574	≃ 0.0000
	Average	0.3556	–

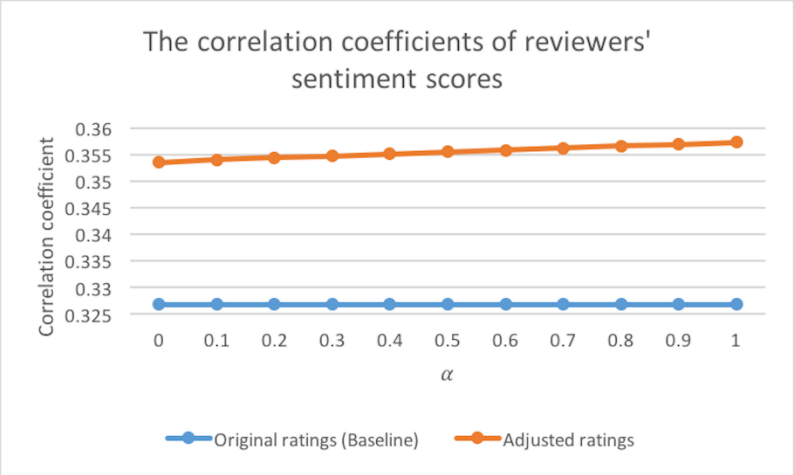

Fig. 8 The correlation coefficients of reviewers' sentiment scores

Table 9 Reviewer classification results

Reviewer type	Number
Optimistic (O)	1,824
Pessimistic (P)	46
Realistic (R)	2,205
Unreliable (U)	394
Not decided (N)	22,977
Total	27,446

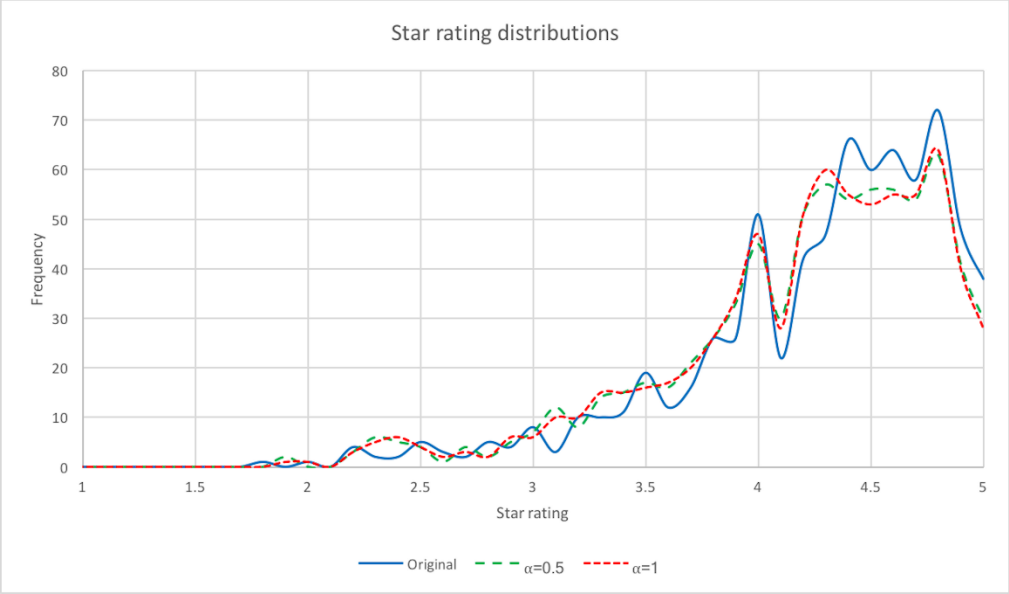

Fig. 9 The distributions of the original star ratings and the adjusted star ratings for 739 products

6 CONCLUSION

The objective of the proposed research is to reduce the bias of customer ratings caused by not only unreliable reviewers but also by optimistic/pessimistic reviewers on online commerce websites. An unsupervised numerical data-driven model is proposed to provide adjusted star ratings based on customers' rating histories and tendencies, instead of using human-labeled training data.

The proposed work is comprised of five main steps. First, customers' numerical ratings and their past rating histories are extracted from online commerce company websites. Extracted customer ratings are then normalized to a 1-5 scale if they use a different scale. Filtering customer ratings written by unverified users is exploited to improve performance and reduce the bias caused by spammers and fake reviewers. A minimum distance classifier then classifies reviewers as realistic, optimistic, pessimistic, or unreliable based on their rating histories and product sales rankings. Finally, the original customer ratings are converted to the adjusted ratings.

A case study involving real-world customer ratings from an electronic commerce company (i.e., *Amazon.com*) validates this work. The results indicate that the adjusted customer ratings have statistically higher correlation coefficients of the estimated product qualities and customer sentiments than the original customer ratings. It is concluded that the adjusted customer ratings can be helpful for both customers' future purchasing decisions and predicting future market sales instead of the original customer ratings.

Future work will present attributes (other than an average of customer ratings, a standard deviation of customer ratings, and a correlation coefficient, which are used in this research) for more accurate reviewer classification. More accurately approximated validations of the adjusted ratings, other than using product sales rankings and reviewers' sentiment scores, will be considered in the

future. The authors will also propose an advanced model to adjust not only overall product ratings but also sub-ratings, if they are available (e.g., product sub-ratings provided by *J. C. Penney*), for better performance and product feature discovery.

ACKNOWLEDGMENT

The authors would like to acknowledge the NSF I/UCRC Center for Healthcare Organization Transformation (CHOT), NSF I/UCRC grant #1624727 for funding this work. Any opinions, findings, or conclusions found in this paper are those of the authors and do not necessarily reflect the views of the sponsors. The authors acknowledge Lawrence Lee for the programming aspects of this research.

REFERENCES

[1] Y. Jiang, J. Shang, and Y. Liu, "Maximizing customer satisfaction through an online recommendation system: A novel associative classification model," *Decision Support Systems*, vol. 48, no. 3, pp. 470–479, Feb. 2010.

[2] N. Hu, I. Bose, N. S. Koh, and L. Liu, "Manipulation of online reviews: An analysis of ratings, readability, and sentiments," *Decision Support Systems*, vol. 52, no. 3, pp. 674–684, Feb. 2012.

[3] S. Tuarob and C. S. Tucker, "Automated discovery of lead users and latent product features by mining large scale social media networks," *Journal of Mechanical Design*, vol. 137, no. 7, 071402, 2015.

[4] A. S. Singh and C. S. Tucker, "Investigating the Heterogeneity of Product Feature Preferences Mined Using Online Product Data Streams," in *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2015, V02BT03A020–V02BT03A020.

[5] S. woo Kang and C. S. Tucker, "Automated Mapping of Product Features Mined From Online Customer Reviews to Engineering Product Characteristics," in *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2016, V01BT02A023–V01BT02A023.

[6] S. Lim and C. S. Tucker, "A Bayesian Sampling Method for Product Feature Extraction From Large-Scale Textual Data," *Journal of Mechanical Design*, vol. 138, no. 6, 061403, 2016.

[7] A. S. Singh and C. S. Tucker, "A machine learning approach to product review disambiguation based on function, form and behavior classification," *Decision Support Systems*, vol. 97, pp. 81–91, 2017.

[8] S. Rose, N. Hair, and M. Clark, "Online Customer Experience: A Review of the Business-to-Consumer Online Purchase Context: Online Customer Experience," *International Journal of Management Reviews*, vol. 13, no. 1, pp. 24–39, Mar. 2011.

[9] S. S. Srinivasan, R. Anderson, and K. Ponnavolu, "Customer loyalty in e-commerce: an exploration of its antecedents and consequences," *Journal of retailing*, vol. 78, no. 1, pp. 41–50, 2002.

[10] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research*, vol. 43, no. 3, pp. 345–354, 2006.

[11] P.-Y. Chen, S. Dhanasobhon, and M. D. Smith, "All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon.com," *Available at SSRN: https://ssrn.com/abstract=918083*, May 2008.

[12] S. Aral, "The problem with online ratings," *MIT Sloan Management Review*, vol. 55, no. 2, p. 47, 2014.

[13] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 519–528.

[14] S. M. Mudambi and D. Schuff, "What makes a helpful review? A study of customer reviews on Amazon. com," *MIS quarterly*, vol. 34, no. 1, pp. 185–200, 2010.

[15] N. Hu, P. A. Pavlou, and J. J. Zhang, "Why do online product reviews have a J-shaped distribution? Overcoming biases in online word-of-mouth communication," *Communications of the ACM (2009)*, vol. 52, no. 10, pp. 144–147, Mar. 2007.

[16] S. Lim, C. S. Tucker, and S. Kumara, "An unsupervised machine learning model for discovering latent infectious diseases using social media data," *Journal of Biomedical Informatics*, vol. 66, pp. 82–94, Feb. 2017.

[17] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, 2010, vol. 1, pp. 492–499.

[18] J. Zhan, H. T. Loh, and Y. Liu, "Gather customer concerns from online product reviews – A text summarization approach," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2107–2115, Mar. 2009.

[19] R. Menon, L. H. Tong, S. Sathiyakeerthi, A. Brombacher, and C. Leong, "The needs and benefits of applying textual data mining within the product development process," *Quality and reliability engineering international*, vol. 20, no. 1, pp. 1–15, 2004.

[20] R. Rai, "Identifying key product attributes and their importance levels from online customer reviews," in *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2012, pp. 533–540.

[21] T.-L. Wong and W. Lam, "Learning to extract and summarize hot item features from multiple auction web sites," *Knowledge and Information Systems*, vol. 14, no. 2, pp. 143–160, Feb. 2008.

[22] T.-L. Wong and W. Lam, "An unsupervised method for joint information extraction and feature mining across different Web sites," *Data & Knowledge Engineering*, vol. 68, no. 1, pp. 107–125, Jan. 2009.

[23] F. Wu and B. A. Huberman, "Opinion formation under costly expression," *ACM Transactions on Intelligent Systems and Technology*, vol. 1, no. 1, pp. 1–13, Oct. 2010.

[24] C. Tucker and H. Kim, "Predicting emerging product design trend by mining publicly available customer review data," in *Proceedings of the 18th International Conference on Engineering Design, Impacting Society through Engineering Design*, Lyngby/Copenhagen, Denmark, 2011, vol. 6.

[25] Y. Liu, J. Jin, P. Ji, J. A. Harding, and R. Y. Fung, "Identifying helpful online reviews: a product designer's perspective," *Computer-Aided Design*, vol. 45, no. 2, pp. 180–194, 2013.

[26] T. Ferguson, M. Greene, F. Repetti, K. Lewis, and S. Behdad, "Combining Anthropometric Data and Consumer Review Content to Inform Design for Human Variability," in *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2015, V02BT03A022–V02BT03A022.

[27] N. Archak, A. Ghose, and P. G. Ipeirotis, "Show me the money!: deriving the pricing power of product features by mining consumer reviews," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 56–65.

[28] M. McGlohon, N. S. Glance, and Z. Reiter, "Star Quality: Aggregating Reviews to Rank Products and Merchants," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010, pp. 114–121.

[29] L. Qiu, J. Pang, and K. H. Lim, "Effects of conflicting aggregated rating on eWOM review credibility and diagnosticity: The moderating role of review valence," *Decision Support Systems*, vol. 54, no. 1, pp. 631–643, Dec. 2012.

[30] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013, pp. 409–418.

[31] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 939–948.

[32] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 191–200.

[33] L. M. Willemsen, P. C. Neijens, F. Bronner, and J. A. de Ridder, "'Highly Recommended!' The Content Characteristics and Perceived Usefulness of Online Consumer Reviews," *Journal of Computer-Mediated Communication*, vol. 17, no. 1, pp. 19–38, Oct. 2011.

[34] N. Hu, P. A. Pavlou, and J. Zhang, "Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication," in *Proceedings of the 7th ACM conference on Electronic commerce*, 2006, pp. 324–330.

[35] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting Burstiness in Reviews for Review Spammer Detection," *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pp. 175–184, 2013.

[36] J. Q. Zhang, G. Craciun, and D. Shin, "When does electronic word-of-mouth matter? A study of consumer product reviews," *Journal of Business Research*, vol. 63, no. 12, pp. 1336–1341, Dec. 2010.

[37] K. Kousha and M. Thelwall, "Can Amazon. com reviews help to assess the wider impacts of books?," *Journal of the Association for Information Science and Technology*, vol. 67, no. 3, pp. 566–581, 2016.

[38] M. Friedman and A. Kandel, *Introduction to pattern recognition: statistical, structural, neural and fuzzy logic approaches*, vol. 32. World Scientific Publishing Co Inc, 1999.

[39] T. Popoviciu, "Sur les équations algébriques ayant toutes leurs racines réelles," *Mathematica*, vol. 9, pp. 129–145, 1935.

[40] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, vol. 6. Cambridge: MIT press, 2001.

[41] R. A. Fisher, "On the probable error of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3–32, 1921.

[42] S. Bird, "NLTK: the natural language toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions*, 2006, pp. 69–72.

[43] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-Based Recommendations on Styles and Substitutes," presented at the Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015, pp. 43–52.

[44] J. McAuley, R. Pandey, and J. Leskovec, "Inferring Networks of Substitutable and Complementary Products," presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 785–794.

LIST OF TABLES

LIST OF FIGURES