Management Analytics and Social Insights

www.masi.reapress.com

Manag. Anal. Soc. Insights Vol. 1, No. 1 (2024) 129–144.

Paper Type: Original Article

Examining Discrepancies between Online Product Ratings and Sentiments Expressed in Review Contents

Hung-Li Chang [1], Yu-Lun Liu [2*], Ching-Jui Keng [3], Han-Ling Jiang [4]

[1] College of Management, National Taipei University of Technology; t108749014@ntut.edu.tw.
[2] Kent Business School, University of Kent; y.l.liu@kent.ac.uk.
[3] College of Management, National Taipei University of Technology; cjkeng@ntut.edu.tw.
[4] College of Management, National Taipei University of Technology; t111749016@ntut.edu.tw.

Citation:

Received: 8 August 2023
Revised: 11 November 2023
Accepted:4 March 2024

Chang, H., Liu, Y., Keng, Ch., & Jiang, H. (2024). Examining Discrepancies between Online Product Ratings and Sentiments Expressed in Review Contents. *Management Analytics and Social Insights, 1* (1), 129-144.

Abstract

The existing and most commonly used information systems and consumer studies on the functional aspect of 'reputation systems have two streams: the 'rating systems and the 'review content systems'. Both systems are offered by most of the popular e-commerce retailers to enhance customer communication experiences. However, there is limited research on the relationship between customers' rating scores and the sentiments expressed in their review contents, which significantly affects the reliability of the overall review in the reputation systems. A computational linguistics approach (Semantic orientation approach) using Amazon's (UK) product review data (34,621 reviews) is employed to unveil the gap between the numerical ratings and the sentiments of review contents from the two reputation systems. Results show that although customers give the same high rating for products that have a lower/higher overall rating, the customers' expressions of sentiment in the review content are actually less/more positive compared to products that have a higher/lower overall rating.

Keywords: Reputation system, Rating, review contents, sentiment, Semantic orientation approach (SOA).

1 | Introduction

With the development of the Internet and other technological applications (e.g., mobile apps), retail e-commerce revenues worldwide are projected to reach $4.88 trillion in the next four years [1]. Before and after purchases, more and more Internet consumers are reading and providing, respectively, product 'review comments' and 'rating scores' as ways to evaluate potential purchases and communicate with online retailers and other consumers [2, 3]. Consumers' reviews of the products they have purchased are not just for relating

Corresponding Author: y.l.liu@kent.ac.uk ea

their satisfying/unsatisfying experience to online retailers but also for informing and advising other consumers about their possible purchases [4, 5].

The existing information systems and online consumer studies on the functional aspect of 'reputation systems have two main streams. One stream focuses on the 'rating systems' [6, 7], in which customers provide a (structured) rating score (e.g., from 1 star for very disappointed to 5 stars for very satisfied) regarding their purchase experiences; this score is averaged with other customers' rating scores, and the overall rating (score) is usually displayed on the product's webpage and used as an indicator of product popularity. Most of the existing studies on rating systems reveal that customers tend to give higher scores when they actually feel satisfied by and happy with the product which they have consumed [8, 9], although, in some cases, a peer-to-peer rating reputation system (buyers rate sellers and sellers can also rate buyers, e.g., as they do on eBay) can lead to most of the rating scores being too high and overly positive [10].

Another stream focuses on 'review content (un-structured/open-ended comments) systems' [11]. Consumers write and leave their comments regarding their buying experience on the retailer's website. The content of the reviews can be positive, negative, or a mixture of positive and negative [12, 13]. The word/tone can be strong or soft when reflecting the commenter's sentiments [14]. Unsurprisingly, existing studies show that positive sentiment expressed in review content is positively related to consumers' positive product experiences. When consumers are satisfied/unsatisfied with the products for which they paid, they are likely to express and post positive/negative sentiments in their reviews [15]. Researchers have suggested that these two main online reputation systems are useful forms with which to communicate with customers. In practice, popular e-commerce retailers, such as Amazon, offer both rating and review content reputation systems on their websites. Their customers are invited to rate and comment on products after their purchases. The customers' ID is displayed along with their individual ratings and review content. Potential customers should theoretically be able to expect consistency between the high/low ratings (scores) and positive/negative sentiments expressed in the review contents and reasonably get a chance to have a fair view of the products through reading this information.

Nevertheless, the social comparisons and anchoring effect theories suggest that this might not be so straightforward. Before making judgements, people tend to evaluate their opinions and abilities by comparing themselves to others and then finding objects as standards (anchors). These anchors then affect their subsequent judgements [16]. In the case of reputation systems, before providing their own ratings and review content, customers would review and use other customers' review comments and ratings of the product as their anchors. This customer know-how exchange impacts customers' perceptions of product value and their likelihood of recommending the product [17, 18, 19].

Additionally, recent affective influence studies suggest that people are sensitive to different forms of information and that, in particular, visuals are more powerful than words. In fact, an individual's subsequent intentions and behaviours can be affected differently by the different forms of information that they receive [20]. That is when customers were asked to rate their product experience with a number (e.g., from 1 star to 5 stars) and provide review comments, their ratings and sentiments could have been affected differently by different forms of anchors (information). This difference could lead to a customer's ratings and sentiments exhibiting a gap when the two types of online reputation systems are both offered through one online shopping platform.

The gap between the ratings and sentiments expressed in the review contents may confuse and decrease trust from potential consumers, particularly in the case of products which have an overall positive rating. If the ratings provided by customers are actually biased by the anchor and do not reflect objectively the actual quality and the customer's true experiences with the product, potential customers may have false expectations. False expectations can lead to unhappy customers [21] and damage the reputation of the retailer.

To the best of our knowledge, no study thus far has examined the relationship between online ratings-based and review content-based reputation systems and explored how previous ratings and sentiments may affect customers' ratings and sentiments. Furthermore, most of the research on online consumer and reputation systems only applied experimental methods or questionnaires/surveys. Although these methodologies provide interesting and useful insights, a drawback is the artificiality of the setting. For this study, a computational linguistics approach (semantic orientation approach, SOA) using Amazon's (UK) 'Best Sellers' section product review data for the end of February 2023 (34,621 product reviews) is employed to explore and demonstrate the relationship between consumers' sentiments in their online review contents and their product ratings. Understanding how customers' review contents and ratings might be affected by previous reviews not only informs the design of reputation systems but also empowers online retailers to manage their customers' product experiences.

2 | Literature Review

2.1 | Product ratings and the sentiments of review contents

Before we turn to theoretical considerations, we first define the terms in the given context explicitly, as existing product review studies have used their own terms within different platforms and contexts. As mentioned earlier, many popular e-commerce platforms provide both types of reputation systems. Product review for these platforms is composed of two components: review content and ratings. A rating in this study indicates the structured score (in Amazon's case, the number of stars) given by reviewers (customers). The overall rating of a product is calculated automatically by the reputation system and is the average of all the customers' ratings. As the overall rating provides a comparison criterion for similar products in the same product category, the impact of the overall product rating on customer purchase decisions and product sales is critical [22]. Review content is the body of the review written by reviewers (customers). Review contents can range from one word to paragraphs (unstructured, with no specific word limit). The sentiment expressed in the review content indicates the strength of the review content, be it positive or negative. This includes the use of words (e.g., excellent, good, bad, and terrible) which reflect reviewers' experiences. The interface for Amazon's reputation system, which is a representative e-commerce platform, is used as an example to illustrate these terms (See Figure 1).

In addition, as shown in Figure 2, Amazon's reputation system displays ten recently published review contents and ratings on the first page along with the product. Customers can provide comments by using the comment function at the bottom of existing reviews. Customers are invited to provide their review contents and ratings after their purchase, and these new comments will show up on the product page along with the other recently published reviews.

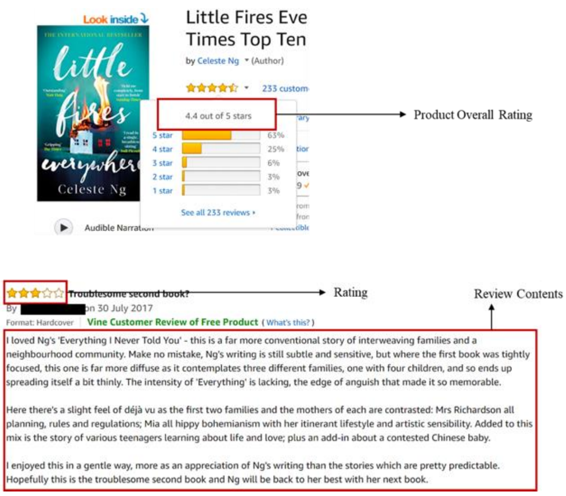

Fig. 1. Components of the Amazon Review interface.

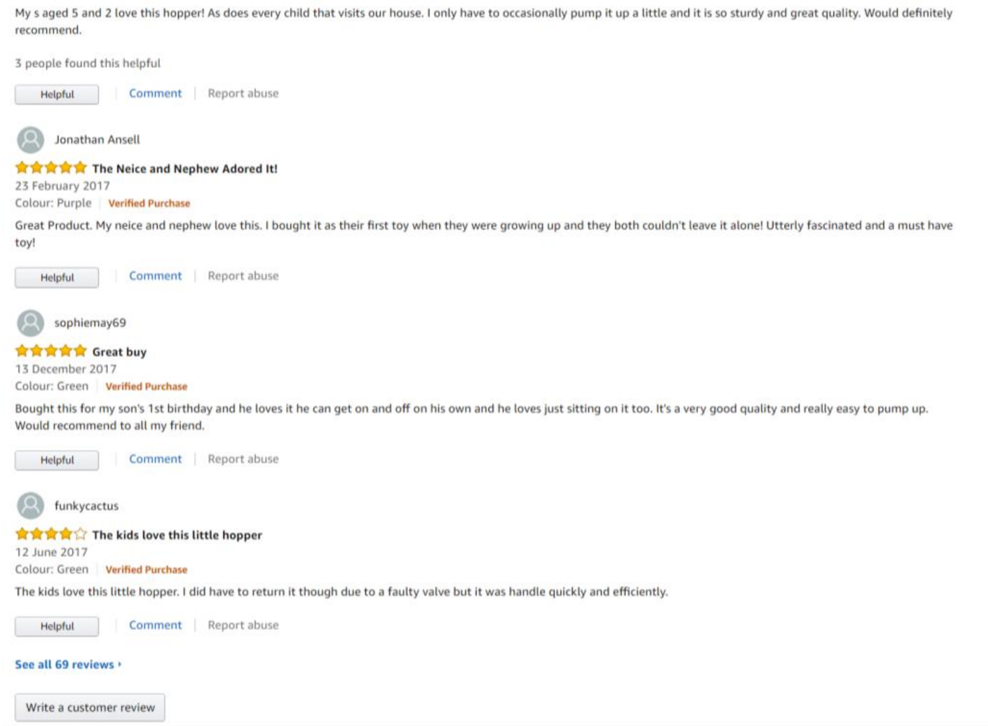

Fig. 2. User Interface (Amazon) of the product review platform (reputation systems)

2.2| Personal image, anchoring effect, and politeness theory

Psychologists indicate that within a community, maintaining a positive personal image and earning self-worth is important to its members [23]. Although interactions and communications are not physically face-to-face, online community members still show significant intentions of protecting their online personal image [2]. E-commerce platforms create communities for sellers and buyers; the profile and activities which a member has and partakes in, respectively, in an online community illustrate and promote the member's online personal image [24].

The social image and identity theories propose that in order to maintain a positive personal image, individuals tend to compare their personal experience and perceptions towards an objective with other community members before making judgements/comments [25]. The social comparison theory indicates that there is a drive within individuals to make accurate self-evaluations [26]. The theory illustrates that individuals evaluate their own opinions by comparing themselves to others in order to reduce uncertainty in these domains. Through reputation systems, customers can use other customers' comments as standards, particularly those comments that were published close to the date when they bought the product and formed their own judgement [27]. Customers do this to avoid submitting a biased or incorrect review, which would negatively affect their image. By selecting standards, a customer can ensure that their experience is not too inconsistent with those of others. An extremely inconsistent experience could be due to personal subjective bias, such as having too high of expectations (compared to others) or lacking relevant product knowledge. For instance, a customer could break a newly purchased product due to not knowing how to use it. Although breaking a product could lead to a negative product experience, rating a product with an extremely low score in the online community and describing this experience in a very negative way could lead to other community members' (other customers') criticisms, including suggestions to read the product guide more carefully before using the product. The commenter also runs the risk of being seen or labelled as unprofessional, negatively influencing their personal image.

Nevertheless, the anchoring effect theory indicates that the tendency for an individual to rely heavily on the initial standard information (aka an 'anchor') when making decisions can lead to a cognitive bias and that those objects near the anchor tend to be assimilated with it [28]. In other words, customers who create their own product reviews by relying on other customers' review contents/ratings (anchors) are expected to be consistent and in line with these existing previous reviews. Even though this could be of help in protecting personal images in the online community, the influence from the anchors could also lead to customers leaving out the facts and their real personal judgements.

In addition, the politeness theory accounts for the redressing of affronts to a person's 'face' by face-threatening acts [29]. Politeness is the expression of the information provider's intention to mitigate threats to face posed by certain face-threatening acts on the part of the information recipients. This situation arises, in particular, when people do not have enough information to refer to when making judgements [30]. In order to maintain a positive personal image, judgements tend to be gentle, conservative, and generous [29] with regard to other community members. Not giving an extremely negative review and rating for the seller's products can not only maintain harmonious relations between the buyers and the sellers, but it can also, more importantly, reduce passively commenters' uncertainties regarding the comments which are made due to personal subjective bias (especially in the case of personally perceived experiences, such as product usage experiences and not for significant product flaws, such as packaging and delivery damages). Therefore, particularly in the early stage of forming product reviews and ratings (anchors), the initial ratings entered into a reputation system will start high, and customer sentiment will be relatively neutral, especially in the case of unsatisfied customers.

2.3| Affective influence and forms of information

Psychological studies on affective influence show evidence that individuals are affected by different forms of information. Different forms of information provided by individuals can also affect their personal images differently [20]. Iconology studies [31] demonstrate that people are more sensitive to symbols (e.g., rankings/numbers) and images (e.g., photos) than other forms of information. These forms of information are attractive, forming more directly the first image of an object and building information recipients' mental models more easily [32]. Therefore, when reading and providing this type of information, individuals tend to be more cautious, as it more directly reflects on the individual's self-presentation, identity, and online visual presence [20]. Although purely text-based messages can contain richer information than other forms of information, people are found to be relatively less affected when reading them, especially when a lot of information is provided. People usually do not consider every detail, and some are even influenced/biased by the effect size (i.e., length of the message/word count) of the message when processing text-based information [33]. Therefore, when providing text-based messages, people tend to be relatively more direct in expressing their sentiments.

That is, when customers use previous customers' ratings and review content as anchors, they are more likely to approach doing their own ratings cautiously, particularly when downgrading the product. For instance, when previous customers rated the product as 4-star (out of 5 stars), later customers were likely to give a rating close to 4-star. The later customers would be restrained from giving a low rating (e.g., 1-star or 2-star). As discussed earlier, following this pattern helps customers maintain their personal images and reduces uncertainties, as numbers would more directly show their judgements to other customers in the community than text-based review content. Relatively speaking, customers' expressions of their sentiments in the pure text-based review contents are less affected than numerical ratings by anchoring bias, as text-based information recipients are less likely to look into every detail and identify specifically the different sentiments in all of the text comments. Text-based reviews affect the commenter's image relatively less than numerical scores do (and some systems, such as Amazon's, use gold stars to display the ratings, which makes the ratings even more prominent). Therefore, although customers refer to recently published review content before making their own comments, they are expected to detail their experiences and sentiments relatively directly.

To sum up, the effects of maintaining a personal image, anchoring, and politeness could lead to a gap between online product ratings and the customers' sentiments in text-based review content. In the early stage of forming the overall product ratings and review contents, with little information that could be used as an anchor, customers tend to be relatively polite, particularly when providing ratings. Compared to ratings, the sentiments that customers express in the reviews should more directly reflect their experience but also tend to be relatively neutral at this stage. After more ratings and review content are posted, those ratings and review content become anchors that could influence customers' ratings and review content later. In terms of maintaining a positive personal image, the anchoring effect should remain significant in rating trends, and the variation in ratings between consumers should stay small. Also, due to the politeness effect in the early stage of anchor formation, the ratings (and, therefore, the overall ratings) should remain comparatively high. On the contrary, text-based comments affect personal image less. Hence, sentiments in review contents should remain relatively direct and diverse throughout.

Therefore, it is expected that overall (especially for those products for which the review trend has been formed), those customers who are not satisfied, due to the stronger affective influence and personal image protection evinced by ratings, would tend to be more in line with previous customers' ratings and give the product a high rating but express their negative sentiment in the review content. This phenomenon is expected to be observable, particularly when the overall rating of the products is in the medium range (e.g., 3 - 4 out of 5). It should be observable because these overall medium-ranged rating products might be less satisfying than those products that have a high overall rating (e.g., 4.5 - 5 out of 5). However, as discussed, in order to maintain their image, some unhappy customers would still give a high rating to a product but express their negative sentiment in the review content. Thus, due to the degree of the anchoring effect on different information forms (number vs. text), a gap is expected to exist between the product rating and sentiments in the review content, leading to the following hypothesis:

Hypothesis: Even though customers give the same high rating (e.g., 4 or 5stars out of 5 stars) for products which have a *lower/higher overall rating*, the customers' sentiments in the review content will tend to be relatively *less/more positive* compared to products which have a *higher/lower overall rating*.

3|Method

3.1| Data collection

We collected the product review data from Amazon UK (https://www.amazon.co.uk/) using Python script with the HTML parsing library BeautifulSoup. We used all the Amazon-recommended 23 books to represent experience goods and 13 smart mobile devices to represent search goods from Amazon's 'Best Sellers' section at the end of February 2023 and crawled all the product reviews. It is worth noting that as all the products were from the 'Best Sellers' section and these products were recommended by the Amazon recommendation system, all the products had a relatively positive overall rating. In addition, only the review contents which received 4 or 5 stars from customers were used; the lower-rated review contents (below 3 stars) were excluded, as the data volumes were too small to derive meaningful statistics. In total, 34,261 reviews were collected and used for the analysis.

A computational linguistics approach, known as the semantic orientation approach (SOA), one of the main branches of sentiment analysis, was employed to calculate the sentiment of the review contents. The SOA is based on identifying and selecting sentiment words in test documents. The main concept of this approach is to classify the sentiment of words in a document to infer its semantic orientation, i.e. whether the document has a positive or negative opinion. To classify the sentiment of words, this approach usually uses external data sources, such as a corpus, which has a massive amount of textual data containing sentiment expressions, and a dictionary, which illustrates the polarity of a large number of words. The corpus and dictionary can provide a polarity score between -1 and 1 (-1 means extremely negative, and +1 means extremely positive). Studies based on SOA also use learning algorithms to construct a dictionary and a semantic network between words from a corpus; therefore, an approach using SOA can be considered to be a learning approach, as well.

Compared with a machine learning-based approach, which is another branch of sentiment analysis, SOA is more intuitive in terms of methodology when calculating sentiment. Also, SOA shows consistent accuracy, while the performance of a machine learning approach depends strongly on the quality of the training dataset [34].

In this study, we use SentiWordnet [34] as the predefined sentiment lexicon resource for sentiment classification. It is the most popular open sentiment lexicon resource, which is used for automatic sentiment classification. The synonyms, antonyms, and hierarchies in SentiWordnet, along with sentiment, can be used to determine the polarity of documents [35]. Many sentiment classification tasks extract sentimental words directly from SentiWordNet to avoid a manual sentiment lexicon or build a new lexicon from a massive dataset using the additional learning approach [36]. To classify the sentiment of a given document using a lexicon resource, such as SentiWordNet, the elements of a vectorized document need to be tagged grammatically, i.e., with part-of-speech tags. In corpus and computational linguistics, part-of-speech tagging (POS tagging) is the process which is used to make up a word in a text (corpus) corresponding to a particular part of speech, such as noun, verb, and adjective in this study. Automatic POS tagging has a long history in computational linguistic studies, and its tagging accuracy has reached over 97% [37, 38]. POS-tagged document vectors can be scored by comparing adjectives/adverbs/verbs in documents with their polarity scores in SentiWordNet (for example, the adjective 'bad' has a polarity score of -0.625 and 'worst' has a score of -0.75 in SentiWordNet). The pseudo-code representing the overall procedure for SOA used in this study is presented in Figure 3.

```
FOR every document in the TestDataSet:
    FOR each sentence in the document:
        TaggedSentence = POS(sentence)
        FOR SentiCandidate (adverb, adjective, and verb) in TaggedSentence:
            PolarityScore += LookupSentiWordNet(SentiCandidate)
            TotalCandidateCount++
    AveragePolarity = PolarityScore/TotalCandidateCount
```

Fig. 1. Pseudo code for SOA approach in this study

3.2| Analysis design and method

As hypothesized, the research question for this study comes from the expected gap between the ratings and sentiments of reviews, and the main factor leading to this gap could be hinted at by the product's overall rating. If the gap does not exist, the overall ratings should reflect the average sentiment of customers' review content. But, if the gap does exist, then there should be inconsistencies in sentiments across products which have the same rating but different overall ratings.

With this notion in mind, we designed the data analysis to verify the impact of the overall rating on the gap between the rating and customers' sentiments in the review comments. After calculating the sentiment score of all the collected reviews (comments) using the computational linguistics approach (sentiment calculator), we sorted and summarised the review data by 1) the rating for each product and 2) the calculation of the average sentiment of the review comments for differently rated reviews as a form of 'sentiment table' (the processes are depicted in Figure 2). We repeated this process for all of the products (23 books and 13 smart mobile devices) and compared all of the sentiment tables.

Presumably, all the review comments which have the same rating should have a similar level of sentiment, regardless of whether or not the products follow the rating scheme proposed by the review interface. Suppose significantly different levels of sentiment for reviews with the same rating can be found across different products with different overall ratings. In that case, we can indirectly affirm the misalignment between ratings and review sentiments. It is worth noting the point we are arguing is that the approach being utilized relies on indirect, instead of direct, affirmation, i.e., we are not able to define the 'absolute' sentiment of a 5-star or 4-star rated review. We can only investigate the gap between ratings and the sentiments of review contents by identifying the variation in sentiments within the review contents with the same rating across different products.

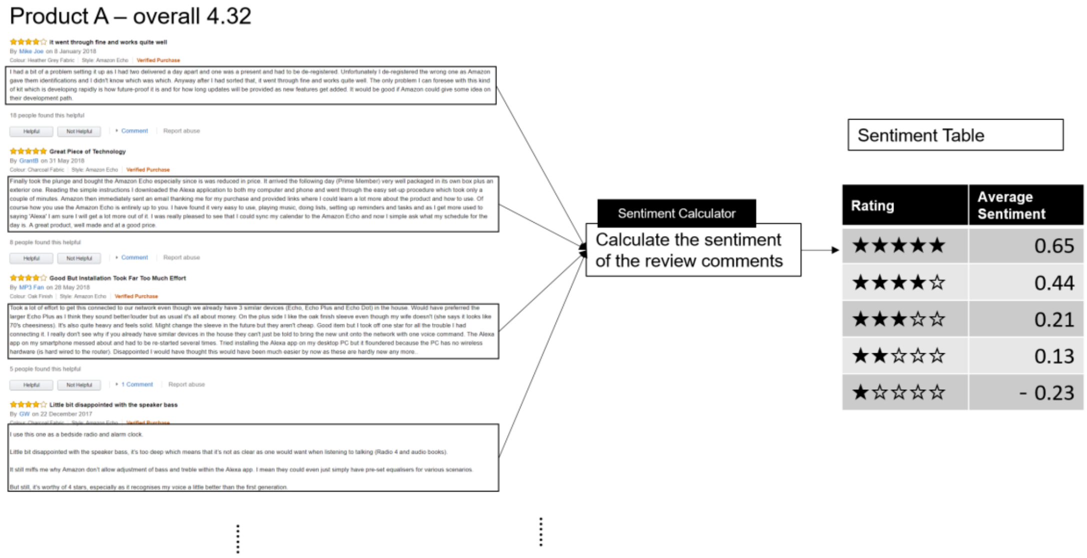

Fig. 2. Example of creating a sentiment table from product reviews

4| Results

4.1| Testing of book category

As indicated earlier, we investigated two product categories – best-selling books and smart mobile devices. This is to ensure that any gap found between the sentiments of review contents and ratings is a generalized phenomenon and not moderated by product type (experience goods or search goods). Table 1 represents the aggregate of the sentiment tables for all 23 books and the average sentiments of the review contents when the books received a 4-star or 5-star rating from customers.

Table 1. Aggregated sentiment table for the book category

Book ID	Overall Rating	Average Sentiment of Review Contents	
		Customers who rated the product 4-star	Customers who rated the product 5-star
Book 1	4.7649	0.3325	0.3840
Book 2	4.1667	0.2065	0.2605
Book 3	4.6541	0.3210	0.3667
Book 4	4.1488	0.1751	0.1957
Book 5	4.6096	0.2853	0.3097
Book 6	4.5423	0.3498	0.3396
Book 7	4.0743	0.2492	0.2609
Book 8	4.2356	0.3069	0.3543
Book 9	4.3295	0.3789	0.3798
Book 10	3.7316	0.2497	0.2482
Book 11	4.4460	0.3427	0.3601
Book 12	4.2536	0.2183	0.2663
Book 13	4.6265	0.4531	0.4527
Book 14	3.5021	0.2162	0.2079
Book 15	4.1380	0.3419	0.3277
Book 16	4.4647	0.3159	0.3026

Book 17	3.8829	0.1817	0.2552
Book 18	4.1098	0.2355	0.2823
Book 19	4.4716	0.3162	0.3471
Book 20	3.6372	0.1300	0.1139
Book 21	4.3667	0.2625	0.3470
Book 22	4.2492	0.2541	0.2433
Book 23	4.0872	0.2505	0.2495

To investigate the relationship between product overall rating and average sentiment expressed in these 5-/4-star rated review contents, we conducted a linear regression analysis to statistically affirm the impact of overall ratings on the sentiment of review contents. As depicted in Figure 3, the average sentiments for the 5-star rated review contents were positively impacted by the overall product reviews. The 5-star rated review contents in higher overall rated products (4-star to 5-star) were more positive than the 5-star rated review contents in lower overall rated products (3-star to 4-star). That is to say, even though the product received the same 5-star rating from these customers, the level of sentiment varied within the overall product rating. This showed that a gap between the ratings and sentiments expressed in the review contents existed and that this was more observable in products which had a lower overall rating; some of the 5-star reviews in lower-rated products were considered to be not positive enough (not that positive) to be rated 5-star. The significant positive relationship [$t = 5.984$, $p < .01$] between the average sentiments of the 5-star rated review contents and overall rating supported this finding. In addition, more than 61% of the variability in the average sentiments expressed in the 5-star-rated review contents can be explained by the product's overall ratings. This indicated that product overall rating is an important factor in the inconsistency between sentiments in review contents and ratings.

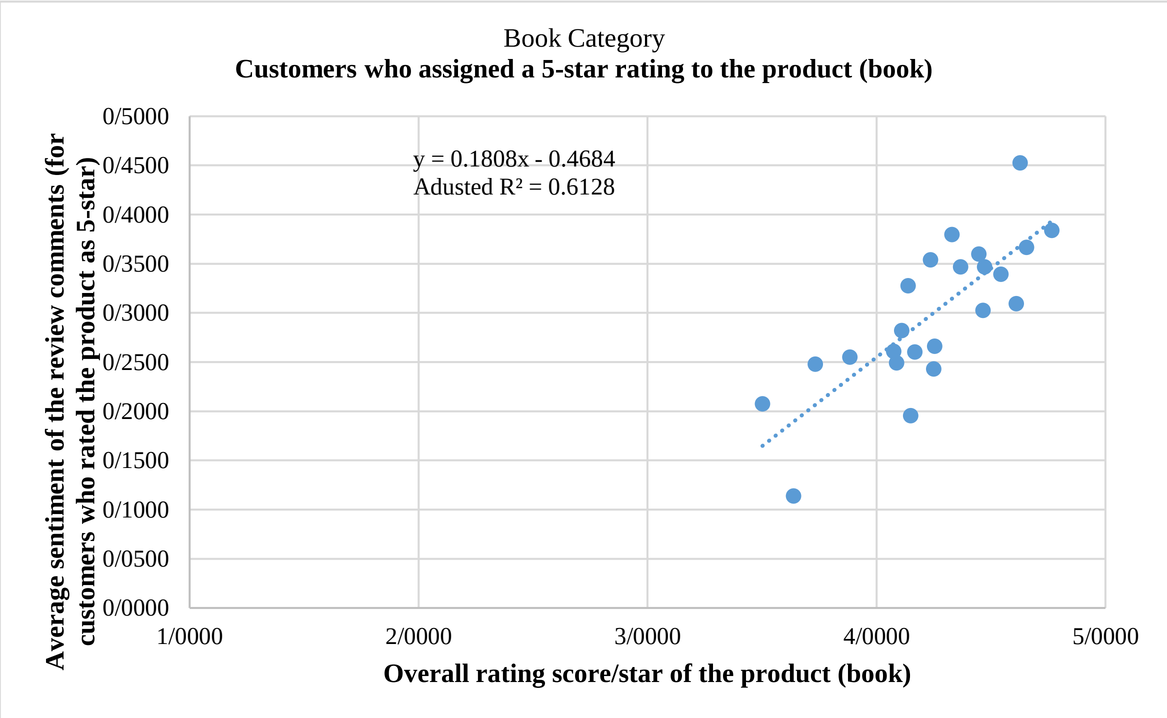

Fig. 3. Overall rating score vs. the average sentiment of the 5-star rated review contents (book category)

Table 2. Regression analysis results for 5-star rated reviews (book category)

Model	Beta	SE	T statistic	Sig.
(Constant)	-0.468	0.128	-3.647	0.002
Overall Rating	0.181	0.030	5.985	< 0.01
R-Square	0.630			

Adjusted R-Square	0.613
Sig of model (p)	< 0.01

A similar result was obtained for the book reviews, which had been rated with a 4-star rating by customers (see Figure 6). The Adjusted R-square [= .476] was slightly less than that of the 5-star rating case, but the relationship was also found to be significant [t = 4.581, p < .01] (See Table 3).

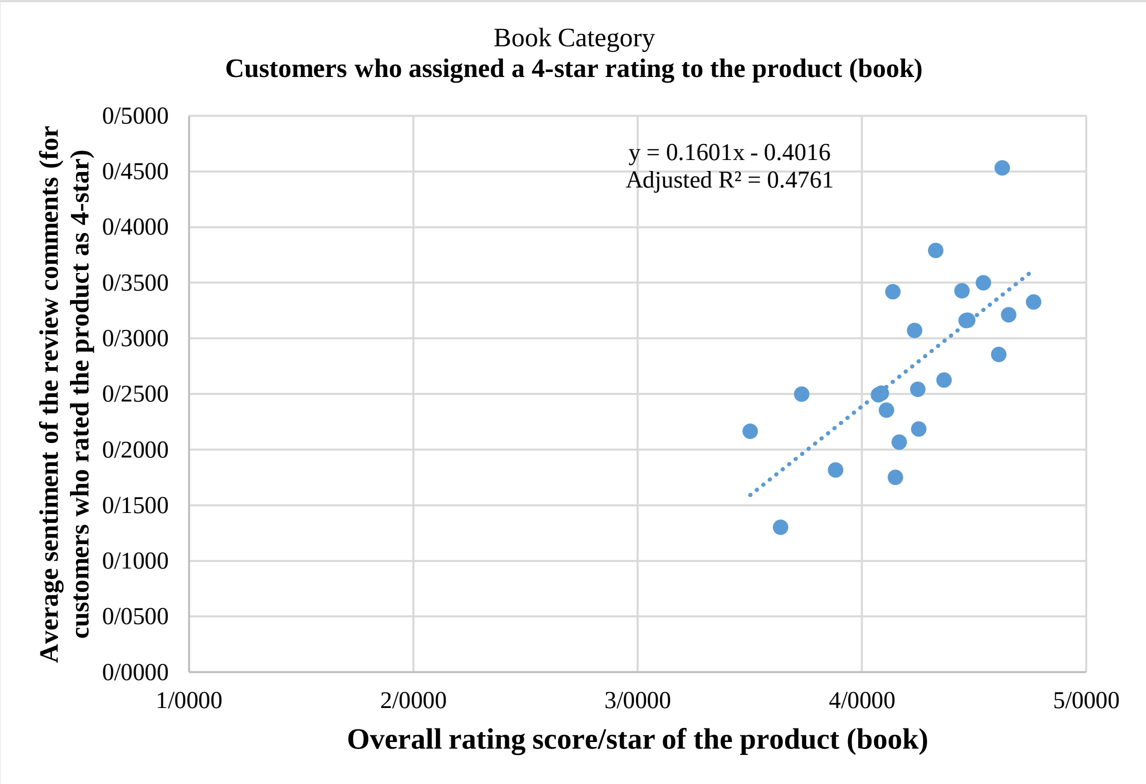

Fig. 4. Overall rating score vs. the average sentiment of the 4-star rated review contents (book category)

Table 3. Regression analysis results for 4-star rated reviews (book category)

Model	Beta	SE	T statistic	Sig.
(Constant)	-0.402	0.148	-2.703	0.013
Overall Rating	0.160	0.034	4.581	< 0.01
R-Square	0.499			
Adjusted R-Square	0.476			
Sig of model (p)	< 0.01			

4.1| Testing of smart mobile device category

Table 4 presents the average sentiment of the 5-star and 4-star rated review contents of the 13 smart mobile products. To investigate the relationship between product overall rating and the average sentiment expressed in 5-/4-star rated review contents, the same analyses as those used for the book category products were conducted.

Table 4. Aggregated sentiment table for the smart mobile device category

Book ID	Overall Rating	Average Sentiment of Review Contents

		Customers who rated the product 4-star	Customers who rated the product 5-star
Smart Mobile device 1	4.1877	0.2985	0.3970
Smart Mobile device 2	4.1923	0.3762	0.4199
Smart Mobile device 3	3.2895	0.2943	0.3346
Smart Mobile device 4	4.4803	0.3805	0.4976
Smart Mobile device 5	3.8816	0.3252	0.4318
Smart Mobile device 6	3.8059	0.3128	0.3704
Smart Mobile device 7	4.2000	0.3927	0.4519
Smart Mobile device 8	3.3361	0.2519	0.4004
Smart Mobile device 9	4.3398	0.3682	0.4988
Smart Mobile device 10	4.0876	0.3374	0.4125
Smart Mobile device 11	4.2596	0.3138	0.4846
Smart Mobile device 12	4.1454	0.3962	0.4761
Smart Mobile device 13	4.2583	0.3244	0.4802

As presented in Figure 5 and Table 5, the product's overall rating showed a significant relationship [$t = 4.383$, $p < .01$] with the average sentiments of the 5-star rated review contents. It should be noted that 60.2% of the average sentiment's variability can be explained by overall product rating in this case.

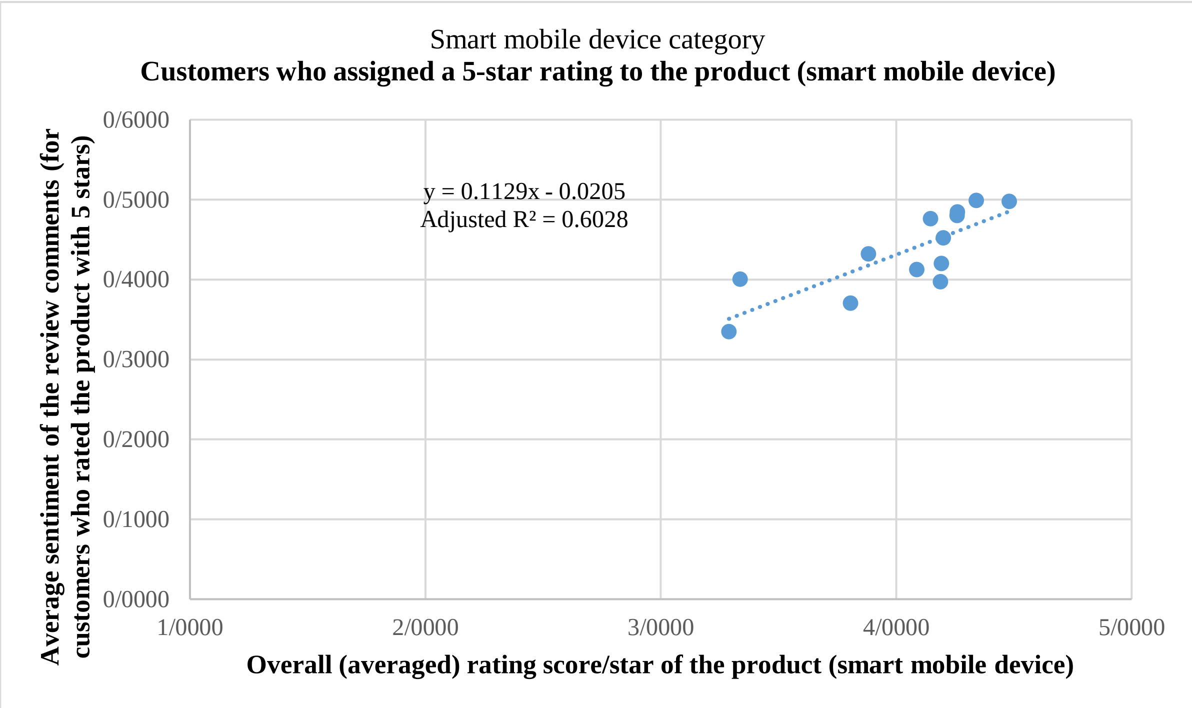

Fig. 4. Overall rating score vs. the average sentiment of the 5-star rated review contents (smart mobile device category)

Table 3. Regression analysis results for 5-star rated reviews (smart mobile device category)

Model	Beta	SE	T statistic	Sig.
(Constant)	-0.020	0.104	-0.196	0.848
Overall Rating	0.113	0.026	4.383	0.001
R-Square	0.635			
Adjusted R-Square	0.603			
Sig of model (p)	0.002			

The results for the average sentiments of the 4-star rated review contents in the smart mobile device category also showed a pattern similar to the previous cases above [$t = 3.312$, $p < .01$] (Figure 8, Table 6). In this case, 45.4% of the average sentiment's variability could be explained by the overall product rating, which also supported reasonably the importance of the influence of the product's overall rating on the gap between sentiments expressed in review contents and ratings.

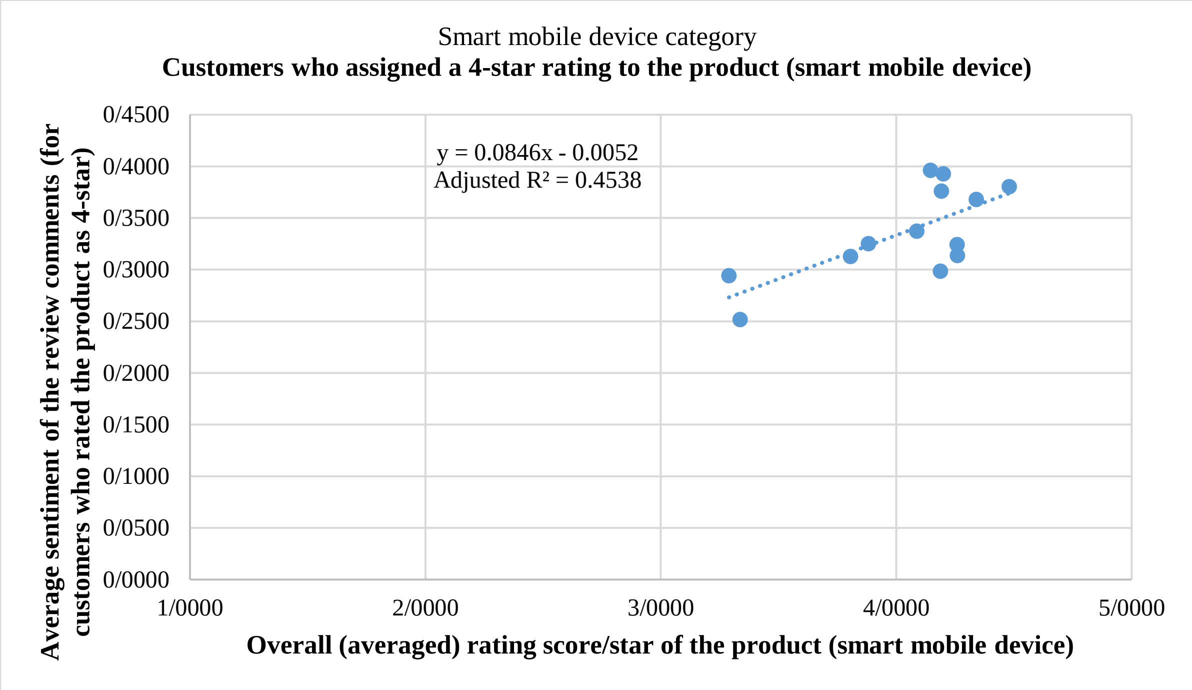

Fig. 5. Overall rating score vs. the average sentiments of the 4-star rated review contents (smart mobile device category)

Table 3. Regression analysis results for 4-star rated reviews (smart mobile device category)

Model	Beta	SE	T statistic	Sig.
(Constant)	-0.005	0.103	-0.049	0.961
Overall Rating	0.085	0.026	3.312	0.007
R-Square	0.499			
Adjusted R-Square	0.454			
Sig of model (*p*)	0.006			

There is a slight variation in the Adjusted R-square results for the four regression models, but all of the results are consistent in terms of the relationship between the product's overall rating and the average sentiment expressed in the reviews. Therefore, these results support the hypothesis by demonstrating that even though customers give the same high rating for products which have a lower/higher overall rating, the customers' expressions of sentiment in the review content tend to be relatively less/more positive compared to products which have a higher/lower overall rating.

5| Discussions

5.1| Implicatons

The data analysis results show evidence of the existence of a gap between ratings and sentiments expressed in the review contents, and the gap is also affected significantly by the overall product ratings. These findings contribute to online customer research and electronic word-of-mouth (eWOM) studies by demonstrating the affective influence of forms of information on customers' ratings and sentiments in online review comments. In order to maintain a positive image in the online community, customers conduct social comparisons with other customers via referencing their review contents and rating scores. However, customers evaluate and perceive information given in different forms (numerical vs. text-based indicators) differently. Customers are more sensitive to numbers and are affected more effectively by the anchoring effect when reading and making comparisons between forms of numerical information. When giving product ratings, they tend to follow the previously published reviewers' ratings, and due to the impact of politeness, particularly in the early stage of forming the overall product rating, the ratings remain relatively high. However, customers are less sensitive to text-based content; when providing review comments, the sentiments tend to reflect relatively directly their experiences. These contradictions lead to the gap between the sentiments expressed in review comments and the rating scores, which is especially observable when the product has a relatively low overall rating. Customers who are not very happy give high ratings but express their negative sentiments in the review comments.

The results provide useful insights into reputation systems for managers and designers. The product's overall rating has been demonstrated to be a critical factor affecting customer purchase decisions, and most of the existing reputation systems only generate the overall product ratings by simply averaging individual customers' ratings. The analysis results in this study raise very generic doubt about how exactly the product's overall rating reflects the customers' opinions. When it comes to the comparisons of two products with high vs. low overall ratings, the difference in the overall ratings between the two products can be smaller than their actual difference in terms of sentiments due to the inflated ratings for the product with the lower overall rating (i.e., the differences between overall rating 4.5 and 4.0 might be more than that expressed by the numerical gap of 0.5). Due to the structures of the rating schemes and interfaces provided by the reputation systems, the overall product ratings can distort the rating-based product comparisons. In order to provide a more objective review that gives consumers a way to match their expectations and product performance, the system algorithm should also consider the sentiments of the review comments. Another possible system design suggestion for resolving this issue is to improve the interface of ratings-based reputation systems by providing explicit criteria for rating. Also, automatic rating recommendations for review commenters based on a machine-learning approach could present an alternative option. Using such a system would allow commenters to compare their intended rating with the proposed rating generated by the comparison of other reviewers' reviews. Most state-of-the-art reputation systems based on product reviews still lack a systematic solution for avoiding reviewers' incorrect anchorings and over/under ratings. Hence, interface and system design practitioners need a solution to overcome these problems. For product management, retailers should not only put efforts into satisfying customers who literally gave the product a low rating (e.g., 1-star/2-star), but they should also look into customers' sentiments, as expressed in the review contents and deal with customers' negative sentiments carefully, even though these unhappy customers may have given the product a high rating score.

This study also has methodological contributions. Apart from the traditional behavioural approach, we adopted a data analytics approach based on computational linguistics to investigate the phenomena and revisit the theory regarding human behaviours in reputation systems. By combining a computational linguistic approach using SOA and statistical analysis, we quantified and visualized the gap between customers' ratings and sentiments expressed in review contents.

5.2| Conclusion and suggestions for future works

This study is the first to examine and demonstrate the gap between the two major reputation systems. This phenomenon has been found across product types, i.e., not only for those who experience goods for which customers usually refer to other customers' review comments while making their judgements due to the lack of objective indicators, but also for search goods, which can be compared relatively objectively using structured indicators. That is, the effect of anchors and the affective influence of the form of information could override the influence of the objective indicators and affect the final customer product reviews.

The directions for future works fall into two categories: expanding datasets and adopting new methodologies. It is suggested that future studies adopt the big data analytics approach to collecting and analyzing massive product review datasets from various e-commerce sources and product categories. This will enable more generalized arguments regarding rating behaviour on online platforms and how the gap between rating and sentiment expressed in a review comment can be affected by product categories. In addition, as highlighted earlier, this study took items for the two categories analyzed from Amazon UK's 'Best Sellers' recommend product list for data analysis; therefore, all of the selected products had an overall rating score above 3-star (neutral). Future studies are encouraged to examine products with a low overall rating score (e.g., below 3-star) and test to see if the product's overall rating score could be a moderator in the gap between the two major reputation systems. This study can be expanded upon and strengthened by combining the data analytics approach with lab experiments. An experimental setting may provide the theoretical backbone behind customers' rating behaviours. In particular, anchoring behaviour could be investigated thoroughly in lab experiments, and the results could be combined with the data analytics results to maximize the synergy of the two different approaches.

There is some work which should be done to take the next step; however, we believe that this topic has real potential in terms of theoretical and practical contributions by examining the gap between ratings and sentiments in review content, something which has not been considered and examined previously. Apart from the existing discourses on review helpfulness and the rating-revenue relationship, which have been covered extensively, we hope this paper opens the door to the exploration of the generic issues of human behaviour in product reviews and attracts further attention from researchers and practitioners alike.

Author Contribution

Conceptualization, Yu-Lun Liu, Hung-Li Chang and Han-Ling Jiang.; Methodology, Yu-Lun Liu; Software, Yu-Lun Liu; Validation, Hung-Li Chang and Ching-Jui Keng; formal analysis, Han-Ling Jiang; investigation, Han-Ling Jiang; resources, Yu-Lun Liu; data maintenance, Yu-Lun Liu; writing-creating the initial design, Yu-Lun Liu; writing-reviewing and editing, Han-Ling Jiang and Hung-Li Chang; visualization, Han-Ling Jiang; monitoring, Ching-Jui Keng; project management, Yu-Lun Liu. All authors have read and agreed to the published version of the manuscript.

Funding

It is declared that this research involved no external funding and was undertaken solely with the researchers' own means.

Data Availability

All the data are available in this paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Topic: E-commerce worldwide. (2024, May 16). Statista. https://www.statista.com/topics/871/online-shopping/#topicOverview

[2] Cheung, C. M. K., & Lee, M. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decision Support Systems*, *53*(1), 218–225. https://doi.org/10.1016/j.dss.2012.01.015

[3] Sardar, A., Manzoor, A., Shaikh, K. A., & Ali, L. (2021). An Empirical examination of the impact of EWOM information on young consumers' online purchase intention: Mediating role of eWOM Information Adoption. *SAGE Open*, *11*(4), 215824402110525. https://doi.org/10.1177/21582440211052547

[4] Wahab, O. A., Bentahar, J., Otrok, H., & Mourad, A. (2015). A survey on trust and reputation models for Web services: Single, composite, and communities. *Decision Support Systems*, *74*, 121–134. https://doi.org/10.1016/j.dss.2015.04.009

[5] Changchit, C., & Klaus, T. (2019). Determinants and impact of online reviews on product satisfaction. *Journal of Internet Commerce*, *19*(1), 82–102. https://doi.org/10.1080/15332861.2019.1672135

[6] Chen, B., Guo, J. Q., Tseng, B. L., & Yang, J. (2011). User reputation in a comment rating environment. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2020408.2020439

[7] Karthik, R., & Ganapathy, S. (2021). A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce. *Applied Soft Computing*, *108*, 107396. https://doi.org/10.1016/j.asoc.2021.107396

[8] Thoma, V., & Williams, A. (2013). The devil you know: The effect of brand recognition and product ratings on consumer choice. *Judgment and Decision Making*, *8*(1), 34–44. https://doi.org/10.1017/s1930297500004484

[9] Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, *58*(4), 696–707. https://doi.org/10.1287/mnsc.1110.1458

[10] Resnick, P., & Zeckhauser, R. (2004). Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. In *Emerald (MCB UP) eBooks* (pp. 127–157). https://doi.org/10.1016/s0278-0984(02)11030-3

[11] Mishra, A., & Satish, S. (2016). EWOM: Extant Research Review and Future Research Avenues. *Vikalpa*, *41*(3), 222–233. https://doi.org/10.1177/0256090916650952

[12] Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52. https://doi.org/10.1002/dir.10073

[13] Wu, Y., Liu, T., Teng, L., Zhang, H., & Xie, C. (2021). The impact of online review variance of new products on consumer adoption intentions. *Journal of Business Research*, *136*, 209–218. https://doi.org/10.1016/j.jbusres.2021.07.014

[14] Bornfeld, B., Rafaeli, S., & Raban, D. R. (2014). Electronic Word-of-Mouth Spread in Twitter as a Function of Message Sentiment. *Proceedings of the Fourth International Conference on Social Eco-Informatics*, 6–14. https://www.thinkmind.org/articles/sotics_2014_1_20_60011.pdf

[15] Kim, S., Kandampully, J., & Bilgihan, A. (2018). The influence of eWOM communications: An application of online social network framework. *Computers in Human Behavior*, *80*, 243–254. https://doi.org/10.1016/j.chb.2017.11.015

[16] Liu, Y., & Keng, C. (2014). Cognitive dissonance, social comparison, and disseminating untruthful or negative truthful EWOM messages. *Social Behavior and Personality*, *42*(6), 979–995. https://doi.org/10.2224/sbp.2014.42.6.979

[17] Lee, M., & Youn, S. (2009). Electronic word of mouth (eWOM). *International Journal of Advertising*, *28*(3), 473–499. https://doi.org/10.2501/s0265048709200709

[18] Kang, M. G., Sun, B., Liang, T., & Mao, H. (2022). A study on the influence of online reviews of new products on consumers' purchase decisions: An empirical study on JD.com. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.983060

[19] Li, K., Chen, Y., & Zhang, L. (2020). Exploring the influence of online reviews and motivating factors on sales: A meta-analytic study and the moderating role of product category. *Journal of Retailing and Consumer Services*, *55*, 102107. https://doi.org/10.1016/j.jretconser.2020.102107

[20] Winkielman, P., & Gogolushko, Y. (2018). Influence of suboptimally and optimally presented affective pictures and words on Consumption-Related behavior. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02261

[21] Shan, G., Zhou, L., & Zhang, D. (2021). From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems*, *144*, 113513. https://doi.org/10.1016/j.dss.2021.113513

[22] Babic-Rosario, A., Sotgiu, F., & De Valck, K. (2019). Conceptualizing the electronic word-of-mouth process: What we know and need to know about eWOM creation, exposure, and evaluation. *Journal of the Academy of Marketing Science*, *48*(3), 422–448. https://doi.org/10.1007/s11747-019-00706-1

[23] Calhoun, C., & Benhabib, S. (1994). Situating the self: gender, community and postmodernism in contemporary ethics. *The Philosophical Review*. https://doi.org/10.2307/2185801

[24] Huang, Z. H., & Benyoucef, M. (2013). From e-commerce to social commerce: A close look at design features. *Electronic Commerce Research and Applications*, *12*(4), 246–259. https://doi.org/10.1016/j.elerap.2012.12.003

[25] Ma, X., Khansa, L., Deng, Y., & Kim, S. S. (2013). Impact of prior reviews on the subsequent review process in reputation systems. *Journal of Management Information Systems*, *30*(3), 279–310. https://doi.org/10.2753/mis0742-1222300310

[26] Gerber, J. P., Wheeler, L., & Suls, J. (2018). A social comparison theory meta-analysis 60+ years on. *Psychological Bulletin*, *144*(2), 177–197. https://doi.org/10.1037/bul0000127

[27] Gruen, T. W., Osmonbekov, T. D., & Czaplewski, A. J. (2006). eWOM: The impact of customer-to-customer online know-how exchange on customer value and loyalty. *Journal of Business Research*, *59*(4), 449–456. https://doi.org/10.1016/j.jbusres.2005.10.004

[28] Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*(3), 437–446. https://doi.org/10.1037/0022-3514.73.3.437

[29] Locher, M. A., & Watts, R. J. (2005). Politeness theory and relational work. *Journal of Politeness Research*, *1*(1). https://doi.org/10.1515/jplr.2005.1.1.9

[30] Rees, C. E., & Knight, L. (2008). Thinking 'no' but saying 'yes' to student presence in general practice consultations: politeness theory insights. *Medical Education*, *42*(12), 1152–1154. https://doi.org/10.1111/j.1365-2923.2008.03173.x

[31] Panofsky, E. (2018). Studies in iconology. In *Routledge eBooks*. https://doi.org/10.4324/9780429497063

[32] Babin, L. A., & Burns, A. C. (1997). Effects of print ad pictures and copy containing instructions to imagine on mental imagery that mediates attitudes. *Journal of Advertising*, *26*(3), 33–44. https://doi.org/10.1080/00913367.1997.10673527

[33] Yüce, P., & Highhouse, S. (1998). Effects of attribute set size and pay ambiguity on reactions to 'help wanted' advertisements. *Journal of Organizational Behavior*, *19*(4), 337–352. https://doi.org/10.1002/(sici)1099-1379(199807)19:4

[34] Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *Language Resources and Evaluation*. http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf

[35] Devitt, A., & Ahmad, K. (2007). Sentiment Polarity Identification in Financial News: A cohesion-based approach. *Meeting of the Association for Computational Linguistics*, 984–991. http://www.tara.tcd.ie/bitstream/2262/25729/1/DevittAhmad_acl2007final.pdf

[36] Hung, C., & Lin, H. (2013). Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification. *IEEE Intelligent Systems*, *28*(2), 47–54. https://doi.org/10.1109/mis.2013.1

[37] Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In *Lecture Notes in Computer Science* (pp. 171–189). https://doi.org/10.1007/978-3-642-19400-9_14

[38] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. https://doi.org/10.3115/1073445.1073478