Ratio Mathematica Volume 54, 2025

Comparison of aggregation methods used in online reviews: a critical analysis

Maria Grazia Olivieri[*]
Elisa Fiorenza[†]

Abstract

With digitalization, online reviews have become a crucial component of consumer decisions and are an important source of feedback on products and services. Users can make informed choices using online review systems such as e-commerce platforms, social media, and others. However, the collection and analysis of these reviews raises a number of questions regarding reliability, representativeness and accuracy. Since they directly influence users' choices, the methods used to aggregate review scores are extremely important. In this article, we explore the most common aggregation methods for rating online reviews on the Amazon platform, with a particular focus on the arithmetic mean. Although these methods are easy to implement, they do not always accurately reflect the overall quality of a product. We then critically analyze how more sophisticated approaches can minimize some of the limitations of the mean, such as the so-called recency bias; propose more robust solutions and alternative approaches that can provide a more accurate representation of the overall rating, improving the quality of aggregated ratings.

Keywords: aggregation methods; arithmetic mean; online reviews; alternative methods; recency bias; weighting.

2010 AMS subject classification: 62P20, 91C20, 91B82[‡]

---

[*] Ecampus University, Novedrate, Italy; mariagrazia.olivieri@uniecampus.it.
[†] Universitas Mercatorum, Rome, Italy; elisa.fiorenza@unimercatorum.it
[‡] Received on February 26, 2025. Accepted on June 1, 2025. Published on June 30, 2025. DOI:10.23755/rm.v54i0.1674. ISSN: 1592-7415. eISSN: 2282-8214.

1. Introduction

In the e-commerce and service industries, online reviews have become an important factor influencing customers' purchasing behavior and business strategies, and thus their purchasing decisions [9]. With millions of reviews shared every day, review sites offer a seemingly comprehensive overview of the products and services they offer. Users rely on these reviews to make purchasing decisions, while recommendation systems use review aggregations to recommend products. The aggregation methods that are used to calculate the overall rating of a service or product online do not always guarantee the reliability and usefulness of the aggregated reviews. Platforms, such as Amazon and the like, use simple statistical methods as a review aggregation system, which they then integrate with more specific algorithms. Fundamental to user trust is the ability to correctly combine reviews into a single score. In order to aggregate individual ratings into a complete one, which reflects the quality of the service or product, online platforms collect thousands of ratings; this requires more effective and fair tools and methods [10]. Among them, the arithmetic mean is most often used because it is considered easy to implement and because it is believed to provide a fair representation of users' opinions; however, it has several critical problems. Among the main problems are its sensitivity to extreme values, the arithmetic mean can be significantly affected by extremely positive or negative reviews, which do not necessarily reflect the average consumer experience [30]; bias in skewed distributions, in cases of highly skewed review distributions (e.g., if most reviews are positive, but there are some very negative reviews), the arithmetic mean may not accurately represent the quality perceived by the average consumer; furthermore, it ignores variability, that is, the arithmetic mean ignores the dispersion of ratings, which can be an important indicator of product or service quality [19]. This paper provides a critical analysis of the arithmetic mean, and proposes alternative methods that can provide more robust and representative approaches to address these limitations. It is very important that review platforms adopt more sophisticated aggregation techniques to improve the quality of information provided to consumers and ensure more informed purchasing decisions. User credibility, quality of review content, sentiment expressed, and consistency between reviews are some of the qualitative factors that could influence the final rating that these conventional methods tend to ignore. Recency bias is also a major problem with these systems [16]. When more recent reviews, which often represent temporary changes, updates, or trends, have an excessive impact on the overall review aggregation, recency bias occurs. This phenomenon can lead to a distortion of a product's rating, giving an inaccurate image of its real value in the long run. Review aggregation is a complex process that uses different statistical and machine learning techniques to summarize the opinions expressed by users. In this paper, we will examine how recency bias affects these methodologies, discuss its implications and propose solutions to reduce its effect [11]. The aim is to provide a comprehensive and informed picture of how online review platforms could evolve to better respond to users' needs, contributing to improve the quality of information available in the decision-making process.

The paper is structured in 4 Sections: in Section 1 the Introduction, in Section 2 we present a brief literature review on the different aggregation methods that online platforms use for review evaluation; in Section 3 we analyze the basic arithmetic mean method in

Amazon online reviews, its critical issues and proposing alternative aggregation methods. Furthermore, recency bias and its manifestation in online reviews are described, then its influence on the different aggregation methods proposed; in Section 4 we present the conclusions.

2. Literature review

In the literature, several articles have discussed the use of different online review aggregation techniques, from the traditional arithmetic mean to the use of the median, probabilistic models and more sophisticated advanced approaches such as Bayesian models; in order to improve the accuracy of aggregated reviews, reduce bias and provide a more representative summary of users' experiences. Several researchers then provide an in-depth analysis of various review aggregation methods, discussing both the limitations of traditional approaches and the advantages of alternative methods to improve the reliability and representativeness of ratings in online review sites. According to [26], the authors present a comprehensive overview of opinion analysis and sentiment analysis, considering different online review aggregation methods and analyzing the challenges in combining feedback from different sources. The paper describes the use of arithmetic mean as a traditional aggregation technique, and then analyzes more advanced techniques such as review weighting. [12] address the problem of rating score prediction in online review sites by examining the importance of the textual content of reviews. The authors propose an approach that integrates textual information with numerical ratings of reviews to improve rating prediction, suggesting that prediction errors can be reduced compared to using the arithmetic mean, or purely numerical tools, by using well-integrated textual information. [28] considers different online review aggregation techniques, among which he offers a comprehensive overview of sentiment analysis and opinion mining. He analyzes the tone and emotion behind user reviews, combining these techniques with numerical aggregation methods in order to obtain more accurate and representative ratings of products or services. [5] study the impact that online reviews have on product prices and competition among major online retailers. The authors, in order to reduce bias, analyze the use of aggregation systems, proposing more reliable methods, capable of influencing customers' opinions and actions. [22] in their study address the problem of incorrect reviews, demonstrating, through a method that automatically detects "opinion spam", how aggregation techniques can be used to find and delete biased reviews; influencing the reliability of review aggregation systems. [27] present a method based on sentiment analysis technique and intuitionistic fuzzy set theory to rank alternative products through online reviews. In the method, the online reviews of alternative products regarding the features are crawled using crawler software. Again, [14] propose two mechanisms for classifying online product reviews: one classifies reviews based on their expected usefulness, consumer orientation, and the other, which involves a producer orientation, classifies reviews based on their expected effect on sales. These tools have the ability to combine econometric analysis with text mining techniques in general, with subjectivity analysis in particular. [8] study different aggregation techniques in e-commerce contexts and the economic role of online reviews. The authors highlight how the distribution of scores and the reputation of users can influence the aggregation system. The paper highlights in particular the fact that trust in reviews depends both on the average quality of scores and on the historical behavior of users who contribute to reviews. The overall reliability of ratings can be improved by using more advanced aggregation techniques, such as weighting reviews by user reputation. In fact, traditional methods, such as simple arithmetic mean, are not sufficient to address the complexities caused by recency bias, low-quality reviews, and differences in opinion. In [23], the author focuses on the representativeness of online review distributions to study how extreme bias and conformity affect it; and to determine whether the representativeness is altered by online review requests. The success of the Internet economy depends on representative consumer reviews online. [29] in his study he considers different scenarios, describing their properties. Specifically, the author considers a type of bias that occurs in online product reviews in the sequence in which these reviews are written. [13] in their study demonstrate that dimensional review bias, which arises from consumers' heterogeneity and their multidimensional product preferences and experiences, weakens information transfer among consumers. The authors use a novel text mining approach to identify and quantify two types of dimensional distortion from text reviews: dimensional rating bias and dimensional preference bias. Furthermore, they add a quantitative approach to reduce dimensional rating bias. They examined the effectiveness and application of bias and de-bias techniques in unidimensional and multidimensional rating systems.

3. Critical analysis of the aggregation methods

Online review aggregation is a critical activity for online platforms such as Amazon, TripAdvisor, Yelp, and other e-commerce and travel sites, which collect millions of user opinions. The ability to correctly aggregate reviews into a single score (e.g., the average rating of a product or service) is essential to user trust in the platforms themselves [17]; Amazon is one of the largest and most influential online review platforms in the world. Users create a large database of reviews by writing reviews of the products they purchase; these reviews help other customers in their purchasing decisions, making them more informed, because they influence the ranking of the products themselves and their visibility on the platform. However, to be reliable and transparent, reviews depend largely on how individual scores are combined into an overall rating. Amazon currently uses the arithmetic mean as a basic method to aggregate review scores; it is a simple tool to determine the average score of star ratings left by users. It has several limitations that risk compromising the accuracy and representativeness of the final rating. The aim of this article is to analyze the use of the arithmetic mean as an aggregation method for evaluating online reviews, examine its limitations; and then compare it with other alternative aggregation methods that could improve the accuracy and reliability of ratings. Median, weighted mean, sentiment analysis, machine learning offer promising solutions to overcome the limitations of traditional arithmetic method, improve the representativeness of ratings, and reduce the bias effects of extreme ratings [4]. The implementation of more sophisticated and transparent approaches could not only improve consumer trust, but also promote a fairer and more reliable review platform.

The traditional methods of aggregation

One of the most common methods of aggregating reviews on platforms like Amazon is the arithmetic mean. This method uses a star rating system (for example, 1 to 5 stars) to determine the average score of each review a product has received. A numerical summary of the overall quality of a product can be obtained using the arithmetic mean, which helps users make better choices. A vacuum cleaner, sold on Amazon, has received numerous reviews that include the following star ratings (Table 1):

	Stars				
	Stars (st_i)				
	5*	4*	3*	2*	1*
Reviews (r_i)	50	30	10	5	5

Table 1. 100 reviews of an Amazon product

The arithmetic mean of the review score, which we indicate with $\bar{x}$, can be calculated with the following formula where st_i indicates the stars and r_i the reviews obtained:

$$\bar{x} = \frac{\sum_{i=1}^{n}(st_i \cdot r_i)}{\sum_{i=1}^{n}(r_i)} \tag{1}$$

$$\bar{x} = \frac{(5 \cdot 50) + (4 \cdot 30) + (3 \cdot 10) + (2 \cdot 5) + (1 \cdot 5)}{50 + 30 + 10 + 5 + 5} = \frac{415}{100} = 4.15$$

In this case, the average product rating $\bar{x}$ is 4.15 stars.

The arithmetic mean has obvious advantages: it is a simple method to calculate and implement, which provides an immediate summary of reviews; the average rating is easily understandable for users and gives them a quick indication of the overall quality of a product [25]. If the reviews are far apart and appear to be conflicting, using the average as an aggregation method may not be appropriate to express the reality of user opinion. If a product has many extremely positive reviews and only a few 1-star reviews, the average may not accurately represent the quality of the product, creating a misleading impression. If a product has 92 5-stars reviews and 6 1-star reviews, the average will be 4.75 (approximate), which may seem very positive, but may not reflect the fact that some people have very negative reviews. Again, the mean does not take into account the content of the reviews, only the numerical results. For example, a 3-stars review may be very thorough and point out small flaws in the product, while a 5-stars review may be very superficial. The mean cannot distinguish between a rating that describes a product as "good but with room for improvement" and a rating that describes a product as "excellent." Finally, it does not capture sentiment, which means that the arithmetic mean cannot distinguish between positive and negative reviews that talk about different aspects of the product, such as quality or customer service. It does not indicate which aspects are praised or criticized. When using the average as an aggregation tool for online reviews, all reviews, regardless of when they were recorded, contribute equally to the final score. Recency bias can distort the perception of product quality, especially for customers who only look at the most recent reviews. A product on Amazon has the following reviews: 100 past reviews, many of which were very positive, with an average rating of 4.8 stars; ten additional reviews were written for the product in the last few weeks, with an average rating of 2-stars, indicating emerging quality issues. If someone looks only at the most recent reviews, they might think the product has a very low rating, even though most of the reviews have been positive. However, if the arithmetic mean does not take these recent changes into account, the average score may still be high, representing an overall rating that does not accurately reflect recent issues.

$$before\ recent\ reviews = \frac{(4.8 \cdot 100)}{100} = 4.8$$

$$after\ recent\ reviews = \frac{(4.8 \cdot 100) + (2 \cdot 10)}{110} = 4.55$$

Users typically rely more on recent reviews in their purchasing decisions because they believe they are more reliable than older reviews. It goes without saying that if a product has been reviewed negatively in the recent period, the user may have a negative impression, even if the rating derived from the arithmetic mean is high, due to older positive reviews. Therefore, recency bias significantly influences perceptions of online reviews when scores are aggregated using the arithmetic mean. To address this issue, platforms such as Amazon are working to implement systems that balance older and more recent reviews, giving customers a more accurate impression of the quality of the product. Despite its limitations, the arithmetic mean remains a simple and useful tool for having a general overview of product reviews; combining it with other aggregation methods would lead to a more accurate and complete analysis.

The weighted mean μ is a method certainly more complex than the arithmetic mean and is used when you want to give greater weight to certain reviews or categories of reviews. Unlike the arithmetic mean, where each value has equal weight, the weighted average multiplies each value (such as a review score) by a "weight" that changes its importance. In online reviews on Amazon or other e-commerce platforms, the weighted average is used to assign different weights to reviews based on factors such as:

- date of publication (recency bias);
- user reliability (for example, whether a review comes from a verified user or a user who actually purchased the product);
- quality of the review (detailed reviews, reviews with photos, reviews that receive approval from other users as "helpful");
- rating by feature categories (e.g. quality, durability, ease of use).

To calculate it, given n ratings x_i, which represent the stars received, they are multiplied by a specific weight w_i (which can be assigned based on different factors) and then the results are added. The total is then divided by the sum of the weights w_i:

$$\mu = \frac{\sum_{i=1}^{n}(x_i \cdot w_i)}{\sum_{i=1}^{n}(w_i)} \tag{2}$$

It is specified that Amazon, in assigning weights to the reviews useful for calculating the weighted average, uses several combined factors to decide how much a review "counts". The main criteria with which Amazon assigns weights: verified purchase; review date; helpfulness voted by other users; reviewer reliability; review content.

Let us now consider a weighted approach in which more importance is given to the most recent reviews; assigning them an increasing weight and using a weighting system based on the publication date. Specifically, the most recent review (review 10) is given the greatest weight, and decreasing weights for reviews that are more distant in time (Table 2).

Reviews	*Time*	*Stars*	*Weights*
Review 1	1 december	2 stars	0.5
Review 2	3 december	3 stars	0.9
Review 3	3 december	1 star	1.1
Review 4	6 december	2 stars	1.4
Review 5	8 december	3 stars	1.8
Review 6	9 december	2 stars	2
Review 7	10 december	3 stars	2.4
Review 8	11 december	5 stars	2.7
Review 9	13 december	5 stars	3
Review 10	14 december	5 stars	3.4

Table 2. Amazon reviews in two weeks considering relative weights

Let's calculate the weighted average μ:

$$\mu = \frac{(2 \cdot 0.5) + (3 \cdot 0.9) + (1 \cdot 1.1) + (2 \cdot 1.4) + (3 \cdot 1.8) + (2 \cdot 2) + (3 \cdot 2.4) + (5 \cdot 2.7) + (5 \cdot 3) + (5 \cdot 3.4)}{0.5 + 0.9 + 1.1 + 1.4 + 1.8 + 2 + 2.4 + 2.7 + 3 + 3.4} = 3.63$$

The weighted average turns out to be about 3.63, while the arithmetic mean is 3.1. As for the benefits of weighted average, this aggregation method allows you to give more weight to higher quality reviews, increasing the accuracy of the product rating; it can help give more weight to the latest reviews, which may indicate improvements or changes to the product; it allows for the application of custom weights for a variety of factors, such as user verification, review quality, date, etc., improving the accuracy of the aggregation.

But it also has some critical issues: determining the right weights can be difficult and requires subjective evaluation; there may be a risk of manipulating the results, for example, by giving too much weight to positive reviews and reducing the importance of negative reviews if the weights are not determined correctly. Finally, the quality of the final result strongly depends on the quality of the assigned weights; if these are not chosen carefully, the weighted average may not accurately reflect the overall experiences of users [33].
If the weight assigned to recent reviews is greater than that assigned to older ones, the final score will suffer from recency bias, that is, the product may seem better or worse than it actually is based on past reviews.

A product received 10 initial very positive reviews (average of 5-stars) and then, in the last few days, received 5 very negative reviews (average of 1-star). If you give more weight to recent reviews in the weighted average, the final rating of the product will be influenced mainly by the more negative reviews, even though most of the past reviews were positive.

- Initial reviews (average 5-stars rating): 10 reviews, weight 1
- Recent reviews (average 1-star rating): 5 reviews, weight 2 (given the greater weight for recent reviews)

We will have:

$$\mu = \frac{(5 \cdot 10 \cdot 1) + (1 \cdot 5 \cdot 2)}{(10 \cdot 1) + (5 \cdot 2)} = 3$$

Although most of the reviews are positive, the final score of the product is heavily influenced by the most recent reviews, which are all negative. Recency bias, when more recent reviews are given very high weights, can have a significant impact on the weighted average of reviews; in this case, it leads to a lowering of the final score, giving a negative impression of the product. Platforms like Amazon, in order to provide a more accurate representation of the quality of the product, try to balance the weight of reviews based on a number of factors, but there is always the possibility that more recent reviews influence the final rating too much.
Let's always consider the same Amazon product, let's see what happens if we use the median as an aggregation method. The median is another method of aggregating reviews that can be used as an alternative to the arithmetic mean or weighted mean. Unlike the mean, which treats all review scores equally, the median focuses on the central score of a data set, that is, the value that separates the lower half from the upper half of the sample. In the context of online reviews, the median is calculated by sorting all reviews in ascending or descending order and taking the value that is in the center of the distribution [1]. If there is an even number of data points, the median is the average of the two middle values:

1. Sort review scores in ascending order

$$[1, 2, 2, 2, 3, 3, 3, 5, 5, 5]$$

2. The median is the average of the two central values being the number of even reviews (10 reviews),

$$Med = 3$$

In this case, recency bias does not significantly impact the median; because the median depends on the relative position of the scores and not on the magnitude of the most recent scores. The most recent reviews are all positive (5 points), but the median remains the same. If there were many more high reviews, the median could still remain stable unless the total number of reviews changes the center position. In other words, the median is less sensitive to changes in review value than the arithmetic mean, which is directly affected by extreme scores (such as very high recent reviews). When you want to avoid biases due to particularly positive or negative recent reviews, the median can be a useful aggregation method. It has several advantages over the arithmetic mean. It is less sensitive to extreme values, or outliers, in fact very high or very low ratings do not influence the median. Additionally, the median provides a more realistic view of what "most" users think of the product. If a product has many negative reviews but also many extremely positive reviews, the mean may not accurately reflect the average user experience. In such cases, the median may be a more representative measure of the "typical" user experience. However, it also has some issues; it doesn't reflect the overall distribution, meaning it doesn't provide information about the number of reviews in each rating range (e.g., 1-star or 5-stars). The median is unable to show that there are a sizable number of unhappy consumers when a product has a large number of 1-star ratings but few 5-stars evaluations. It also disregards the "strength" of the reviews; for example, two 5-stars reviews and one 1-star review may provide a 5-stars median, but it won't make the disparity in the scores obvious. The median is less prone to recency bias than the mean since it concentrates on the reviews' core value. To avoid bias, it is important to ensure that all reviews are adequately considered and that any weights are carefully calibrated.

Sophisticated approaches

Sentiment analysis

Sentiment analysis is a Natural Language Processing (NLP) technique that aims to determine whether a text expresses a positive, negative, or neutral sentiment. When applied to online reviews, sentiment analysis can provide a general indication of how users perceive a product or service [20]. However, recency bias (the tendency to give more weight to recent reviews than older ones) can affect the results of sentiment analysis.

Let's consider the same Amazon product, this time we take into account the sentiment expressed by consumers (Table 3). Recent reviews are much more positive than early reviews, and so recency bias could impact the overall sentiment analysis of the product.

Review	*Time*	*Content*	*Sentiment*
Review 1	1 december	*"I wouldn't buy it again!"*	Negative
Review 2	3 december	*"The design is nice, but the display quality isn't the best."*	Neutral
Review 3	3 december	*"I've been having problems with the*	Negative

		camera, I can't take decent pictures."	
Review 4	6 december	*"The phone is very slow, the battery drains quickly."*	Negative
Review 5	8 december	*"Audio is great, but battery life could be better."*	Neutral
Review 6	9 december	*"I don't like the operating system, too complicated."*	Negative
Review 7	10 december	*"Overall satisfactory product, but there are better!"*	Neutral
Review 8	11 december	*"This phone is awesome! All day battery life, beautiful screen!"*	Positive
Review 9	13 december	*"Very satisfied, great display and long lasting battery."*	Positive
Review 10	14 december	*"Awesome phone, amazing speed, highly recommended!"*	Positive

Table 3. Amazon reviews considering the sentiment

The most recent reviews (last 3) are all very positive, while the oldest reviews (first 7) are mostly negative or neutral. Now let's see how the sentiment analysis might change depending on whether we consider only the most recent reviews or all reviews.

<u>Without considering recency bias</u>

Suppose we apply a simple sentiment analysis to all reviews (without weighting them by date) using a sentiment analysis model that classifies reviews as positive, negative, or neutral.

	Sentiment		
	Negative	Neutral	Positive
Review	4	3	3

Since there are more negative reviews (4) than positive reviews (3), the overall sentiment will be negative, even though there are many recent positive reviews. However, the influence of older reviews may diminish the impact of recent positive reviews. Recency bias has not been taken into account, and so the final result is still affected by many initial reviews that are negative. The overall rating may not accurately reflect current reality, especially if the product or service has recently received a series of very positive reviews; the simple sentiment average is also influenced by older negative reviews.

<u>With recency bias</u>

Now consider the impact of recency bias. Suppose we decide to weight the most recent reviews so that they have a higher weight than the older reviews (Table 4). A simple technique would be to assign increasing weight to the most recent reviews (e.g., the highest weight goes to the latest review, the second highest weight goes to the second-to-last review, and so on).

Review	*Weights (w_i)*
Review 1	0.5
Review 2	0.9
Review 3	1.1
Review 4	1.4
Review 5	1.8
Review 6	2
Review 7	2.4
Review 8	2.7
Review 9	3
Review 10	3.4

Table 4. Amazon reviews by assigning a weight

Each positive, negative or neutral review is then weighted according to the associated weight. The weighted average of sentiment *s* is calculated, denoted by μ_s:

$$\mu_s = \frac{\sum_{i=1}^{n}(s \cdot w_i)}{\sum_{i=1}^{n}(w_i)} \tag{3}$$

s	*Negative*	*Neutral*	*Positive*
	-1	0	1

$$\mu_s = \frac{(-1 \cdot 0.5) + (0 \cdot 0.9) + (-1 \cdot 1.1) + (-1 \cdot 1.4) + (0 \cdot 1.8) + (-1 \cdot 2) + (0 \cdot 2.4) + (1 \cdot 2.7) + (1 \cdot 3) + (1 \cdot 3.4)}{0.5 + 0.9 + 1.1 + 1.4 + 1.8 + 2 + 2.4 + 2.7 + 3 + 3.4}$$

$$= 0.21$$

The weighted sentiment μ_s is around 0.21, which is a positive value, indicating that overall, considering the recency bias, the overall sentiment of the reviews is positive. This is a significant change from the unweighted average (which would have given a negative or neutral value), thanks to the influence of very positive recent reviews. By applying recency bias (weighting the most recent reviews), sentiment analysis can provide a more accurate assessment of current user perceptions, as more recent reviews are emphasized [2]. This approach takes into account changes in sentiment over time.

Cluster Evaluation System

Clustering is a sophisticated aggregation method that allows you to group reviews into homogeneous classes based on similar properties, such as score, review content, sentiment, or themes [15]. Clusters are formed, and an aggregate score is calculated for each group [31]; the impact of recency bias can strongly influence clustering results, especially if recent reviews are significantly more positive than older ones.

A product, such as a smartwatch, receives 10 reviews over the course of a month. The reviews are distributed such that the first 7 reviews are negative or neutral, while the last 3 reviews are very positive. The impact of recency bias can lead to a bias in clusters, causing newer (positively oriented) reviews to be treated as more relevant, while older (often negative) reviews are excluded or undervalued in the clustering process.

Reviews	*Time*	*Stars*	*Content*
Review 1	1 december	2 stars	*"Battery life is too short."*
Review 2	3 december	3 stars	*"Interface is complicated to use."*
Review 3	3 december	1 star	*"Not satisfied, connection with phone is unstable."*
Review 4	6 december	2 stars	*"I don't like the operating system, too complicated."*
Review 5	8 december	3 stars	*"Watch is ok, but not what I expected."*
Review 6	9 december	2 stars	*"Display quality is poor."*
Review 7	10 december	3 stars	*"Design is nice, but I don't find many useful features."*
Review 8	11 december	5 stars	*"Great smartwatch, easy to use and very accurate."*
Review 9	13 december	5 stars	*"I love this smartwatch! Excellent features and beautiful design."*

Review 10	14 december	5 stars	*"I love it, perfect for sports and daily life!"*

Table 5. Amazon reviews for clustering

Older reviews (December 1st to December 10th) are generally negative or neutral, while newer reviews are all very positive.

Review clustering without considering recency bias

Reviews can be classified into two clusters with a clustering method such as K-means: one for positive ratings and one for negative ratings. The clustering is based only on the ratings and content of the reviews, without considering the period in which they were written. Clustering process:

1. Rating features: Each review is visualized as a point in the feature space, where the rating and sentiments (positive, negative, or neutral) are the main features.

2. Application of K-means: it is an unsupervised clustering algorithm, which means that it does not require a labeled dataset to train it. The goal of K-means is to divide a dataset into K number of clusters, where each cluster contains data points that are similar to each other [18]. The method works iteratively and tries to minimize the variability within each cluster while maximizing the distance between the clusters. K-means follows these steps [24]:

- Selection of two initial clusters (one positive and one negative).
- Assignment of reviews to clusters based on their proximity to centroids (mean values of clusters).
- Recomputation of centroids based on reviews assigned to each cluster.
- Iteration until the final clusters are established.

	Clusters	
	1 (Negative)	*2 (Positive)*
Number of Reviews	7	3

Cluster 1 (Negative): Reviews with a score of 2 and 3, or Reviews: 1, 2, 3, 4, 5, 6, 7.
Cluster 2 (Positive): Reviews with a score of 5, or Reviews: 8, 9, 10.

In this case, recency bias has not been taken into account. The older reviews, which are predominantly negative, are all grouped into the negative cluster, while the more recent reviews are all in the positive cluster. While the clustering appears to accurately reflect overall sentiment, it does not take into account the effect that more recent reviews may have in balancing the overall sentiment. If a product has significantly improved over time, the clusters will not adequately reflect this without considering the importance of the date.

Clustering weighted for recency bias

Now let's introduce a recency bias into the clustering process. Consider the Table 2. Performing clustering with weights:

- Review Weighting: Each review is given a higher weight if it is more recent.
- Weighted Clustering: We use a clustering algorithm that takes into account the weights assigned to each review. The review with the highest weight will have a greater influence on the final cluster result.

With recency bias applied, the clustering algorithm will be more influenced by recent reviews with high scores. Therefore, even though the number of negative reviews may be higher, recency bias will cause recent reviews with high ratings to influence the clustering result more. Clustering result with recency bias:

- Cluster 1 (Negative): Clustering may likely assign some of the older reviews (those with low scores) to the negative cluster, but the influence of recent reviews with high scores will cause negative reviews to be less prominent.
- Cluster 2 (Positive): The cluster of positive reviews will likely be much larger, due to the influence of the higher weights of recent reviews.

In this case, the recency bias will have a significant impact on the clustering result. Recent reviews (very positive) will shift the center of the positive cluster, while older reviews (often negative) will be less influential in determining the location of the center of the negative cluster. Clustering is more influenced by recent reviews, which are much more positive [21]. However, this approach could lead to a distorted representation of the overall sentiment of the product, as the clusters focus on the most recent reviews, ignoring older reviews that may be negative.

4. Conclusions

Online reviews are a critical tool for consumers, but traditional aggregation methodologies, such as arithmetic mean, fail to provide reliable results when reviews have skewed distributions, recency bias, or heterogeneous content; arithmetic mean struggles to capture the nuanced and dynamic nature of user feedback. Quantitatively, the analysis highlighted, for example, in a case where 90% of the reviews were positive but only 10% were recent and negative, the arithmetic mean remained high (4.75), masking potential product issues. More sophisticated and accurate solutions are offered by the implementation of alternative methods such as clustering, sentiment analysis, machine learning and MCDM [32]. For example, applying weights to the review dates shifted the sentiment score from slightly negative to positive, clearly showing the influence of recent improvements in product quality. As for multi-criteria methods, as expected from future research, the integration of these techniques appears promising, allowing aggregation based on a broader range of variables, and they also improve the reliability of overall ratings. These sophisticated methods can help Amazon shoppers receive more accurate and representative product ratings. Continued innovation in aggregation techniques is essential to improve the online shopping experience and ensure that reviews accurately represent the quality of products sold. In the future, we plan to try to use multi-criteria decision-making methods, which consider multiple aspects of reviews (quality, usefulness, truthfulness, user involvement) and which are gaining attention for their ability to address the problems of traditional methods.

References

[1] Abdel-Hafez, A., Xu, Y., & Jøsang, A. (2014, September). A rating aggregation method for generating product reputations. In Proceedings of the 25th ACM conference on Hypertext and social media (pp. 291-293).

[2] Agarwal, B., Mittal, N., Bansal, P., & Garg, S. (2015). Sentiment analysis using common-sense and context information. Computational intelligence and neuroscience, 2015(1), 715730.

[3] Bączkiewicz, A., Kizielewicz, B., Shekhovtsov, A., Wątróbski, J., & Sałabun, W. (2021). Methodical aspects of MCDM based e-commerce recommender system. Journal of Theoretical and Applied Electronic Commerce Research, 16(6), 2192-2229.

[4] Benlahbib, A. (2020). Aggregating customer review attributes for online reputation generation. IEEE Access, 8, 96550-96564.

[5] Chevalier, J., & Goolsbee, A. (2003). Measuring prices and price competition online: Amazon. com and BarnesandNoble. com. Quantitative marketing and Economics, 1, 203-222.

[6] Dahooie, J. H., Raafat, R., Qorbani, A. R., & Daim, T. (2021). An intuitionistic fuzzy data-driven product ranking model using sentiment analysis and multi-criteria decision-making. Technological Forecasting and Social Change, 173, 121158.

[7] Decker, R., & Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. International Journal of Research in Marketing, 27(4), 293-307.

[8] Dellarocas, C., & Wood, C. A. (2008). The sound of silence in online feedback: estimating trading risks in the presence of reporting bias. Management science, 54(3), 460-476.

[9] Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. Decision support systems, 45(4), 1007-1016.

[10] Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001, April). Rank aggregation methods for the web. In Proceedings of the 10th international conference on World Wide Web (pp. 613-622).

[11] Eryarsoy, E., & Piramuthu, S. (2014). Experimental evaluation of sequential bias in online customer reviews. Information & Management, 51(8), 964-971.

[12] Ganu, G., Elhadad, N., & Marian, A. (2009, June). Beyond the stars: improving rating predictions using review text content. In WebDB (Vol. 9, pp. 1-6).

[13] Ge, Y., & Li, J. (2015). Measure and mitigate the dimensional bias in online reviews and ratings. In ICIS.

[14] Ghose, A., & Ipeirotis, P. G. (2006, December). Designing ranking systems for consumer reviews: the impact of review subjectivity on product sales and review quality. In Proceedings of the 16th annual workshop on information technology and systems (pp. 303-310).

[15] Gionis, A., Mannila, H., & Tsaparas, P. (2007). Clustering aggregation. Acm transactions on knowledge discovery from data (tkdd), 1(1), 4-es.

[16] Han, S., & Anderson, C. K. (2020). Customer motivation and response bias in online reviews. Cornell Hospitality Quarterly, 61(2), 142-153.

[17] Hellerstein, J. M., Haas, P. J., & Wang, H. J. (1997, June). Online aggregation. In Proceedings of the 1997 ACM SIGMOD international conference on Management of data (pp. 171-182).

[18] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 178-210.

[19] Ivanova, O., & Scholz, M. (2017). How can online marketplaces reduce rating manipulation? A new approach on dynamic aggregation of online ratings. Decision Support Systems, 104, 64-78.

[20] Jamadi Khiabani, P., Basiri, M. E., & Rastegari, H. (2020). An improved evidence-based aggregation method for sentiment analysis. Journal of Information Science, 46(3), 340-360.

[21] Ji, F., Cao, Q., Li, H., Fujita, H., Liang, C., & Wu, J. (2023). An online reviews-driven large-scale group decision making approach for evaluating user satisfaction of sharing accommodation. Expert Systems with Applications, 213, 118875.

[22] Jindal, N., & Liu, B. (2008, February). Opinion spam and analysis. In Proceedings of the 2008 international conference on web search and data mining (pp. 219-230).

[23] Karaman, H. (2021). Online review solicitations reduce extremity bias in online review distributions and increase their representativeness. Management Science, 67(7), 4420-4445.

[24] Lakshmanaprabu, S. K., Shankar, K., Gupta, D., Khanna, A., Rodrigues, J. J., Pinheiro, P. R., & de Albuquerque, V. H. C. (2018). Ranking analysis for online customer reviews of products using opinion mining with clustering. Complexity, 2018(1), 3569351.

[25] Lin, S. (2010). Rank aggregation methods. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5), 555-570.

[26] Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA.

[27] Liu, Y., Bi, J. W., & Fan, Z. P. (2017). Ranking products through online reviews: a method based on sentiment analysis technique and intuitionistic fuzzy set theory. Information Fusion, 36, 149-161.

[28] Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature.

[29] Mackiewicz, J. (2008). Reviewer motivations, bias, and credibility in online reviews. In Handbook of research on computer mediated communication (pp. 252-266). IGI Global.

[30] Orlikowski, W. J., & Scott, S. V. (2014). What happens when evaluation goes online? Exploring apparatuses of valuation in the travel sector. Organization Science, 25(3), 868-891.

[31] Pourghebleh, B., & Navimipour, N. J. (2017). Data aggregation mechanisms in the Internet of things: a systematic review of the literature and recommendations for future research. Journal of Network and Computer Applications, 97, 23-34.

[32] Preethi, J. S., Abirami, A. M., Askarunisa, A., Priya, G. S., & Sankaragomathy, E. (2015). Applying MCDM techniques for ranking products based on online customer feedback. Int. J. Knowl. Based Comput. Syst, 3, 13-20.

[33] Viviani, M., & Pasi, G. (2017). Quantifier guided aggregation for the veracity assessment of online reviews. International Journal of Intelligent Systems, 32(5), 481-501.