Bias correction methodology for online product ratings using affective text mining and item response theory

Metodología de corrección de sesgo para valoraciones de productos en línea utilizando minería de texto afectiva y teoría de respuesta a ítem

Eduardo Jorquera[1] https://orcid.org/0009-0004-0968-7205
Daniel Cabrera-Paniagua[2*] https://orcid.org/0000-0001-5826-287X
Camilo Gómez-Narváez[3] https://orcid.org/0000-0002-4480-8509
Harvey Rosas[3] https://orcid.org/0000-0002-1751-9701

Recibido 12 de octubre de 2024, aceptado 02 de diciembre de 2024
Received: October 12, 2024 Accepted: December 02, 2024

ABSTRACT

This work presents a methodology for bias correction of product ratings, specifically by considering affective text mining and item response theory (IRT). The novelties of this work are designing a methodology for bias correction in online product ratings, defining an experimental scenario using official data from Amazon Fine Food Reviews, and analyzing promising results obtained from applying the proposed methodology in the experimental scenario. Our experiments reveal that it is possible to conceive an automated method for bias correction in online product ratings using IRT and affective text mining, all the above within a unified methodology.

Keywords: Bias correction methodology, item response theory, affective text mining, amazon fine food reviews.

RESUMEN

Este trabajo presenta una metodología para la corrección de sesgos en las valoraciones de productos, considerando específicamente la minería de texto afectiva y la teoría de respuesta al ítem (TRI). Las novedades de este trabajo son: diseñar una metodología para la corrección de sesgos en las valoraciones de productos online, definir un escenario experimental utilizando datos oficiales de Amazon Fine Food Reviews y analizar los prometedores resultados obtenidos de la aplicación de la metodología propuesta en el escenario experimental. Los experimentos revelan que es posible concebir un método automatizado para la corrección de sesgos en las valoraciones de productos online utilizando la TRI y la minería de texto afectiva, todo lo anterior dentro de una metodología unificada.

***Palabras clave**: Metodología de corrección de sesgos, teoría de respuesta al ítem, minería de texto afectiva, amazon fine food reviews.*

[1] SMU SA. Santiago de Chile. E-mail: eduardoejorqueras@gmail.com
[2] Universidad de Valparaíso. Escuela de Ingeniería Informática. Valparaíso, Chile. E-mail: daniel.cabrera@uv.cl
[3] Consultor independiente, Viña del Mar, Chile. E-mail: camiloegomeznarvaez@gmail.com; hjrosasq@gmail.com
* Autor de correspondencia: daniel.cabrera@uv.cl

INTRODUCTION

Online transactions play a crucial role in modern economies. In the last decade, the use of intelligent systems has been widely extended in online transactions, e.g., e-Commerce [1], transportation and logistics systems [2], stock markets [3], and social environments [4], among others. It is worth noting in all these examples that business processes that were executed exclusively and preferentially in person in the past (i.e., buying in a physical store, travel reservation in a tourism agency, registration in a customer service waiting room, requesting a trip service), can now be performed using online platforms without any issue.

Specifically, in the e-commerce domain, several studies have explored ways to improve service quality in the context of mobile commerce [5], the face of obstacles to adopting mobile commerce by SMEs in the UK [6]. Recognizing the significant factors when consumers buy using online systems is interesting. The factors can range from more objective (such as the price of a product) [7] to subjective ones (such as the idea of value, social status, and direct recommendation) [8]. For its part, [9] explored the role of emotional content from online customer reviews. Meanwhile, [10] studied how product features, other consumer ratings, and single positive-negative consumer reviews influence online purchasing decisions.

One of the most common mechanisms to capture subjective evaluations of third parties about a product or service is a rating score accompanied by an opinion (in the form of a comment), where the rating scores are not necessarily an unbiased representation of the comments. In this regard, some authors have studied the effect of different writing styles in textual comments [11], the presence of emotional content within the text [9], or even the under-reporting evaluation phenomena due to the rating scores from more experienced users [12]. In this sense, it may be possible to argue that rating scores suffer from a degree of bias due to a lack of consistency between rating scores and comments. Therefore, some authors explored alternatives to avoid such problems. For instance, [13] presents a meta-analysis perspective on the impact and influence of reviews on decision-making processes.

On the contrary, [14] tested a non-automated mechanism to avoid bias presented in the opinion texts during new evaluation opportunities. However, to the best of our knowledge, proposals that suggest a methodology for bias correction of product ratings, specifically by considering affective text mining and item response theory (IRT), are impossible to observe. Particularly, IRT has yet to be applied in purchasing systems to explore the relationship between users' comments and reviews about scores and ratings.

Thus, the novelties of the present research work are: 1) designing a methodology for bias correction in online product ratings; 2) defining an experimental scenario by using official data from Amazon Fine Food Reviews; 3) analyzing and discussing promising results obtained from the application of the proposed methodology in the experimental scenario.

The remainder of this paper is organized as follows: The first section contains relevant background; The second section describes the proposed bias correction methodology. An experimental outline is provided in section three to describe a test scenario, including the description of the Amazon Fine Food Reviews dataset and the obtained results. The final section contains a discussion and some conclusions of the work done and possible future work.

LITERATURE REVIEW

Electronic commerce and consumer decisions

Consumer behavior has been studied in different phases of the purchase process in mobile systems [15], the factors that affect trust in mobile commerce [16], or the impatience of online consumers during medical e-commerce shopping [17]. Meanwhile, [11] analyzed the causal relationship between several factors of e-commerce. This study found that information about products or services is mutually influenced by factors such as "convenience," "ease of use system," and "web reputation". Conversely, regarding the use of general recommendation systems for supporting e-commerce processes, there are several proposals for travel [18], movies and web browsing [19], multi-objective recommendation generation [20], and exploring the extension of trust-based recommendation systems using public anonymous information [21], to name a few.

Another body of knowledge corresponds to social commerce. This concept is associated with a business model where buyers and sellers carry out business using online social networks [22]. Social commerce is characterized by user-generated content as the primary information source for consumers; this mobile-intelligent terminal is increasingly significant for social commerce and social relationships, which undoubtedly play a critical role in consumer purchasing decisions. Several authors have explored consumers' perceptions of social commerce benefits [23], integrating social commerce quality and elements for co-creating value [24].

Conversely, another body of knowledge corresponds to the analysis of online reviews [25], [26]. In [27], the effects of culture on consumer consumption and the generation of online reviews were explored. It was analyzed whether information from culturally personalized reviews (based on nationality) influences consumers' intentions to recommend an agency or company. In addition, it was examined whether cultural differences impact reviews in terms of the valence and dispersion of ratings and content of reviews. The analysis used a regression approach for review ratings and the Linguistic Inquiry and Word Count (LIWC) method for the textual content of reviews. Meanwhile, in [28], the content factors of a review (quality of the argument) were studied: comprehensiveness, clarity, readability and relevance, and valence (the positive or negative meaning of the message).

Machine learning applied to e-commerce

Considering the massive data availability based on large scales of data production, data analytic methodologies have played a key role. They aimed to extract and analyze relevant information, which, in most cases, needs to be more evident. Due to the complex data structures, relations, and patterns, machine learning (ML) algorithms are still common for extracting knowledge from data.

In [29], a comprehensive review of the use of ML was presented, including techniques such as clustering, classification, prediction, or text analysis (in the case of sentimental analysis and product online reviews), which are often used in this context. In [30], a hierarchical deep neural network was used to deliver a product ranking, extracting latent features in online reviews for Amazon products. In [31], sentiment analysis and fuzzy set theory were used for the same purpose but incorporated consumer preferences.

Bias correction using IRT

Applications of IRT can be found in different ambits, such as education (teacher and class assessment) [32] and streaming platforms [33]. In [34], a multi-facet item response theory (MFIRT) approach was proposed to examine the effects of product attributes, reviewer criticality, consumption situation, product type, and time in assessing latent customer satisfaction. The MFIRT approach predicts product performance more accurately than alternative methods and provides novel insights to inform marketing strategies. In psychology, particularly considering the analysis of subjective ratings, [35] proposed a multitasking method based on IRT for automatic response style removal of the valence and arousal level of artificial faces, specifically in the content-independence context. In [36], eye movement tracking was analyzed using IRT towards a possible application in consumer decisions on the shopping process. In this work, the authors analyzed the relationship between the visual process and decision-making, finding key parameters that appear to impact decision-making behavior.

To the best of our knowledge, IRT has yet to be applied to study the relationship between user comments and reviews, and scores and ratings. This opens the possibility of exploring alternative approaches for its application in e-commerce, education, and management, among other fields.

METHODOLOGY

The proposed bias correction methodology consists of an ensemble method that takes advantage of the simultaneous modeling of text features and rating scores. This method provides an alternative to capturing latent information from reviews, considering affective feedback from the buyer. Finally, this knowledge is combined as the input of an IRT model, aiming to correct a possible bias. The entire methodology is summarized in Figure 1.

Text preprocessing

Text mining is a collection of techniques for processing non-structured data in the form of text so that information can be derived [37]. There are

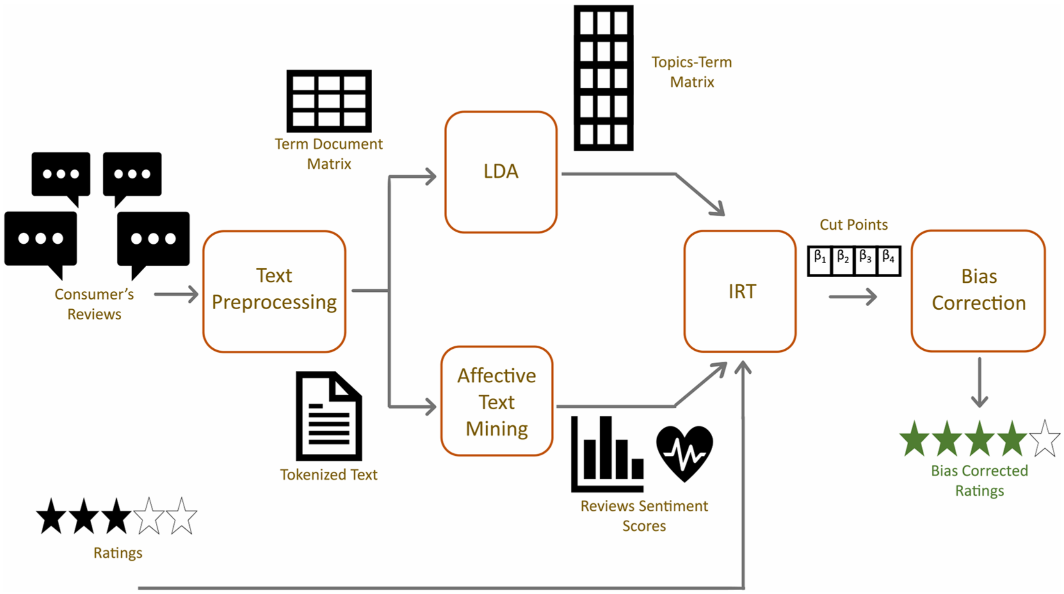

Figure 1. General diagram of the proposed bias correction methodology.

multiple approaches in the interest of extracting valuable information from text, but most commonly, a first step consists of data cleaning (also known as data wrangling in a preprocessing stage), where the text is prepared for further analysis. This stage includes eliminating non-informative words (stop words), removing prefixes and suffixes (stemming), and splitting each review into unique words (tokenization); the first two are used as input for topic modeling, while the remaining are used for affective content extraction.

After preprocessing, a term-document matrix is constructed by calculating the inverse term document frequency for each word in each review (or document in text mining terminology). This matrix has as many rows as different reviews are in the dataset (corpus) and as many columns as unique words; this would be the principal input for the topic modeling in the following stage [38].

Affective text mining

A technique used to extract patterns and information directly from text data is known as affective text mining [37]. This technique is based on frequency counts of predefined words associated with an affective grade or score (obtained from fixed affective dictionaries), words, phrases, or documents. One of the most common choices is to assign negative scores (between –1 and 0) to words that contain negative sentiments and positive scores (between 0 and 1) to words that contain positive ones, considering a 0 score for neutral words.

In decision-making processes, affective text mining has gained some relevance and impact, especially by summarizing reviews and comments of some products or services, focusing on highlighting the most positive and negative aspects of the evaluated items. For instance, [39] analyzed the emotional content using ML and affective text mining in online hotel reviews, finding a negative correlation between negative sentiment content and helpfulness scores. Meanwhile, [40] applied this technique to extract and summarize product features and their corresponding affective responses based on online product descriptions and consumer reviews.

A term-document matrix was necessary for the LDA step to calculate the affective scores related to each review A tokenization step must also be added to the same matrix. The outline for associating an affective score to a review requires first matching the tokenized words of a review with a predefined affective dictionary, linking them to an affective score to for each matched word. Second, the average of the affective scores is calculated, and finally, this procedure is repeated in each review, giving a unique affective score, as there are many reviews in the corpus.

Topic modeling using LDA

Due to its versatility, Latent Dirichlet Allocation (LDA) has been used in many areas, e.g., for identifying co-occurrence relationships for different terms [41]. LDA can also be used with web scraping to summarize textual web content [42].

This method allows the modeling of a corpus of documents as a mixture of components starting from a single document that is considered a realization of multinomial random variables that, on their own, can be viewed as representations of underlying topics. Thus, each word is thought to be generated from a single topic, and different words in a document may have come from different topics. Each document is represented as a list of mixed proportions for this mixture of components, which is therefore reduced to a probability distribution on a fixed set of topics. This distribution is a reduced description associated with the document [43]. In a more practical text mining context, LDA requires a term-document matrix as input and, in return, provides the distribution of topics among the whole set of documents. The topic distribution for each review is an additional useful output from LDA for this work.

Item response theory

Since rating scores are a prevalent practice in many fields, some analytical approaches have been applied to summarize and extract relevant information about evaluations of products, services, and abilities of knowledge, among others. Depending on the context of the analysis, some examples of these analytical tools are mean averages of scores and the correlation between scores and other variables of interest, such as gender, sex, and age, among others. Due to the complex nature of measurement and the poor capacity of the aforementioned analytical tools to incorporate or extract information, such as different abilities that respondents may present, IRT emerges as a statistical model to perform several significant tasks to merge latent information of respondents and the obtained scores [32].

IRT allows measuring the performance of each item through latent variables as part of a logistic-based model. An important assumption for this is that these variables are hidden and, as such, unknown, and they can only be measured indirectly, given a particular response. These latent variables usually represent abilities, skills, or other particular characteristics associated with a respondent [32]. Moreover, these models can distinguish between the respondent's severity (represented by γ) and the unbiased performance of an item (represented by θ); this is done through a Bayesian assembly of the model [44]. According to [32], the relation between an obtained score of an item and the respondent's latent characteristics can be presented as follows:

$$\begin{aligned} Prob\left[Y_{is} = k \middle| \theta_i, \gamma_{is}, \beta\right] = \\ logit^{-1}\left(\theta_i - \gamma_{is} - \beta_{j(k-1)}\right) - \\ logit^{-1}\left(\theta_i - \gamma_{is} - \beta_{jk}\right) \end{aligned} \tag{1}$$

where $Prob[Y_{is} = k|\theta_i, \gamma_{is}, \beta]$ is the conditional probability of obtaining a specified score (Y) equal to k given an i-th item and a s-th respondent, and $logit^{-1}(x) = (1 + e^{-x})^{-1}$ is an inverse logistic function. The calculation of all the parameters can be performed by computational Bayesian likelihood-based methods, as is detailed in [32], using RStan [45]. This work focuses extracting of the respondent's bias associated with an item. Thus, given equation (1), a linear structure (denoted by η) is added to merge the textual information extracted in the previous steps of this methodology and the latent parameters that IRT models (performance and severity). The additive η structure is incorporated as follows:

$$\begin{aligned} Prob\left[\Upsilon_{ijs} = k \middle| \theta_i, \gamma_{is}, \eta_{is}, \beta\right] = \\ logit^{-1}\left(\theta_i - \gamma_{is} - \eta_{is} - \beta_{(k-1)}\right) - \\ logit^{-1}\left(\theta_i - \gamma_{is} - \eta_{is} - \beta_{k}\right) \end{aligned} \tag{2}$$

where $\eta_{is} = AF_{is} * (T_{is} * \delta)$ is a proposed linear structure (as a weighted regression) to add information from the textual reviews to calculate of possible bias in the scores due to some respondent's evaluation and the respective topic and affective scores. Specifically, AF_{is} is a calculated affective score for the i-th product and s-th respondent review and $T_{is} = (T_{1is}, T_{2is}, ..., T_{tis})$ is the topic distribution of the s-th respondent review about the i-th product. Finally, $\delta = (\delta_1, \delta_2, ..., \delta_t)$ is a vector of regression parameters representing a fixed effect of the t-th topic in the s-th respondent review of the i-th product.

Bias correction

IRT works with biased measures to reduce them; if such bias evaluation is overlooked, each statistical estimation produced would be biased [46], and for this reason, it is common for testing systems to use it [47]. Furthermore, IRT models that use social-emotional learning [48] but do not include a sentimental factor, which is crucial to reducing possible bias. According to [32], a group of additional parameters (equation (1) and equation (2)) could be calculated in an IRT model. These values are denoted by $\beta = (\beta_1, \beta_2, ..., \beta_{k-1})$ and correspond to $k - 1$ parameters depending on k as the possible rating scores in equation (1).

The $k-1$ β-parameters are obtained so that it is possible to redefine the original rating scores after calculating the unbiased performances (θ). These parameters are the cutting points that will provide the final output of the proposed bias correction methodology (Figure 1). In summary, IRT requires evaluation data (where the user's severity is reflected) to subtract it from the product performance; this allows to make a simulation of unbiased performance estimation.

EXPERIMENTAL OUTLINE AND RESULTS

Data description

This study used the Amazon Fine Food Reviews [49] dataset was used. This data collection contains 568,454 reviews of 74,258 unique food products from 1999 to 2012. Each review contains vital information about specific features evaluated by users. Table 1 summarizes the description of each available variable.

For this dataset, every review is considered once the reviewer (user ID) purchased the item involved. In this case, the items were fine food products identified with a unique identification (product ID), and the user could make an evaluation with rating stars (score) and write a review. In addition, during item purchasing processes, other buyers can also evaluate whether other comments were helpful (helpfulness index).

The proposed bias correction methodology's effectiveness was explored using a subset of Amazon Fine Food Reviews, keeping only the rating scores and reviews in the list of variables (Table 1). In addition, it was only considered for products having more than 100 reviews, corresponding to 5,157 users and 122 unique products, for a total of 12,694 reviews.

Text preprocessing step

As detailed in the previous section, the Amazon Fine Food Reviews dataset was subject to data wrangling, removing all the stop words and considering only the term stem for each word. In this process, a term-document matrix was obtained, consisting of 12,694 rows and 9,596 columns. Each column represents the entire set of reviews but in a slightly lower dimension set, keeping only the words that contain non-redundant information. For each component of the matrix, a term-frequency inverse-document-frequency was calculated.

Conversely, all the reviews were divided word by word to extract their individual affective content. A simple visual summarization can be seen in a word cloud in Figure 2, where words in the center such as "like," "tea," "flavor," and "food" appear more often in the corpus, while other words in the perimeter appear far less often.

Table 1. Amazon Fine Food Reviews variable description.

Variable	Description
Product ID	Unique code to identify a product.
User ID	Identification of each Amazon user who evaluated and/or purchased a product.
Profile name	Identifying the name of Amazon users.
Helpfulness numerator	Number of Amazon users who evaluate a comment as helpful.
Helpfulness denominator	Total time which a comment was evaluated.
Score	Evaluation rating star (from 1 to 5).
Time	The date a comment was issued on.
Summary	An abbreviated version of a product's comment.
Text	Phrases written by consumers to evaluate specific features of a purchased product.

Figure 2. Word cloud of considered amazon fine food reviews.

LDA step

As beneficial as it might appear, the visualization in (Figure 2) lacks the depth necessary to uncover fundamental patterns in the corpus, primarily related to the potential existence of bias in the reviews. Nor is there a possible relationship with the corresponding rating. LDA was used to explore the latent information within the reviews. Table 2 shows 3 of the 5 topics obtained using LDA. Each column contains the words related to a particular topic. The order of the words follows the probability of belonging to the given topic; that is, the first word is the most likely to belong to the topic. Although this analysis uncovers latent information, it fails to achieve the goal of detecting possible bias in the scores related to the reviews to some degree.

Affective text mining step

Alongside the LDA analysis, affective text mining was conducted over the tokenized reviews. The sentiment extraction from the text of each review was carried out term by term and averaged afterward. A simple comparison of the performance using different sentiment dictionaries (AFINN, Bing, and NRC) resulted in the selection of Bing [50] due to its capacity to characterize negative content more effectively. In this way, a sentiment value was obtained for each review. In Figure 3, a color-coded word cloud is shown, where the words used in a mostly positive context, such as “love,” “sweet,” or “recommend,” are in blue. In red are words used with a negative connotation, such as “jerky,” “bad,” or “bitter.” The topics document matrix and the affective scores are obtained, where the inputs of an IRT model are according to equation (2); this makes it possible to project on a continuum scale.

Table 2. Example of 10 words belonging to 3 of the 5 obtained topics using LDA.

Topic 3	Topic 4	Topic 5
cat	taste	tea
food	like	flavor
eat	good	taste
treat	flavor	green
dog	snack	organ
can	sweet	drink
one	bar	good
chicken	sugar	enjoy
love	just	great

IRT step

Once the topic document matrix and the affective scores are calculated, these values can be incorporated as the corresponding input and the corresponding rating scores of each product. The proposed IRT model (equation (2)) requires calculating all parameters for a posterior bias correction of the original rating scores distribution (Table 3).

After computing all the parameters in equation (2), 122 performance parameters (θ) and 5,157 severity

Figure 3. Word cloud considering affective content, blue possible, and red negative.

Table 3. Rating scores distribution of Amazon Fine Food Reviews used sample.

Rating Score	Frequency
1	1.255
2	761
3	1.237
4	2.531
5	6.910

parameters (γ), as well as the topic regression parameters (δ), were obtained. The different contributions for each topic in equation (2) are interesting to note, given the sign of the parameters. In the case of the first and fifth topics, there is an inverse association concerning the original rating scores (Table 4).

As depicted in the methodology, one of the main contributions of using IRT models is the capacity to recalculate a distribution of the original scores as a continuous set of scores and then obtain the performance parameters seen in Figure 4.

Bias correction step

As shown in Figure 4, the continuous estimated distribution suggests the possible existence of four groups of score ratings, separated into three values: approximately −3, 0, and 3. This suggests that the initial 5 rating scores can be reduced to only 4, considering the cutting points also calculated (β); this, however, could change. Considering the cutting points (Table 5), the rating scores' first suggestion is also modified. It can be seen in Figure 5 that the continuous performance scale is separated only into two groups with the last calculated cutting point ($\beta4 = -1.454$). Finally, the bias correction suggests

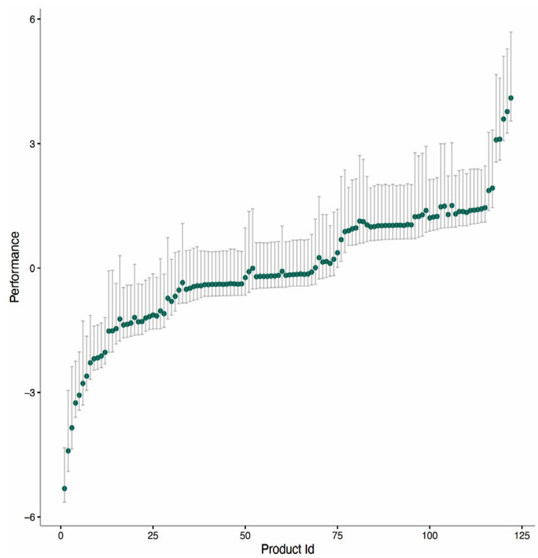

Figure 4. Product performance distribution obtained using the proposed IRT model with their corresponded estimated confidence intervals.

Table 4. Calculated topic regression parameter of the proposed IRT model 2.

Parameter	Mean	Standard Error	Q_{25}	Q_{75}
δ_1	–3.286	0.003	–4.239	–2.309
δ_2	4.176	0.003	3.277	5.084
δ_3	3.644	0.002	2.907	4.378
δ_4	1.566	0.003	0.753	2.392
δ_5	–4.296	0.002	–5.131	–3.439

Table 5. Calculated topic regression parameter of the proposed IRT model 2.

Parameter	Mean	Standard Error	Q_{25}	Q_{75}
β_1	–15.478	0.002	–15.740	–14.689
β_2	–11.258	0.002	–11.485	–10.596
β_3	–7.165	0.001	–7.366	–6.579
β_4	–1.454	0.001	–1.641	–0.912

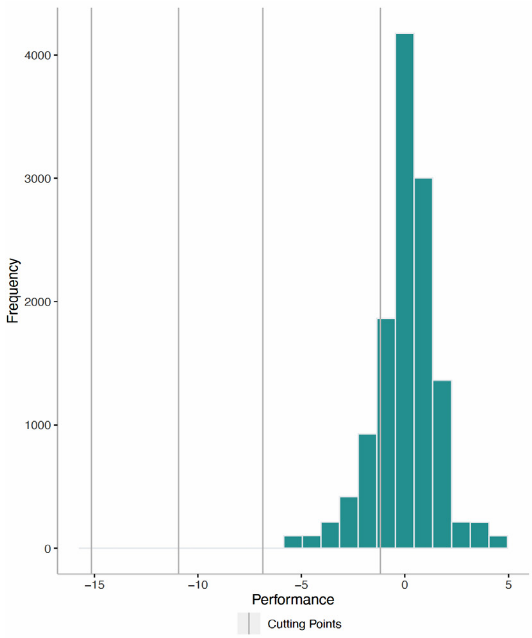

Figure 5. Product performances distribution incorporating the cutting points.

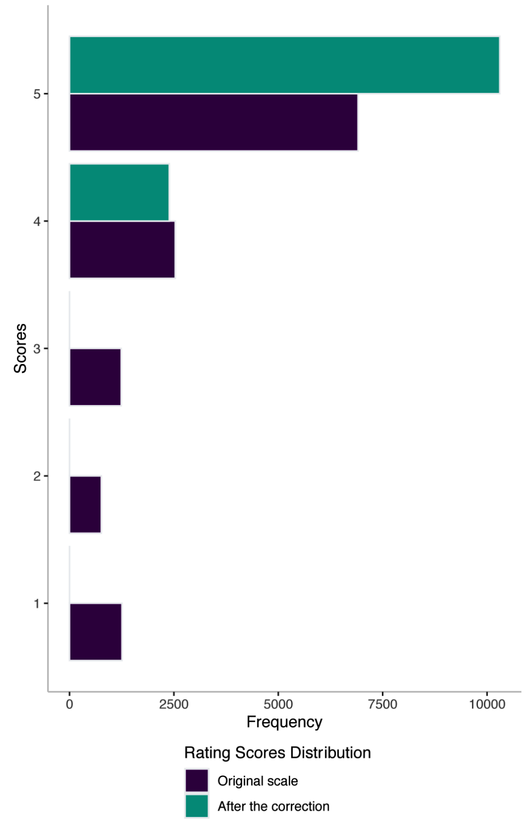

Figure 6. Comparison between the original scores and the unbiased corrected ones obtained using IRT models.

that the new scores are 4 and 5 stars, giving no product scores of 3 or less as the original score distribution. Figure 6 shows a comparison between the original and corrected score ratings.

DISCUSSION

When the buying processes take place, it is almost impossible for a potential buyer to read and take advantage of all the available comments. Given this difficulty, exploring a possible relationship between the reviews and the averages scores evaluations are also an extremely difficult task. Despite some degree of discrepancies between the comments and the rating scores, accommodating buyers' personal preferences is a tremendous challenge.

Besides showing the effectiveness of the proposed methodology, the experiments conducted with the Amazon Fine Food Reviews, suggest that all products tend to have good evaluations. Almost every product tends to be evaluated with a rating score of 5. An interesting result is the presence of negative emotional content and how the proposed IRT model can balance these negative scores to reduce the bias and give a high score once the correction of scores was performed.

The observed shift in ratings, where high scores (4 and 5) were adjusted more significantly than lower scores, can be attributed to the interplay of sentiment analysis and the IRT model. Even highly rated products often contain reviews with negative sentiment, potentially indicating inflated scores due to various biases. By incorporating sentiment information, the IRT model can identify and adjust for this discrepancy, leading to a more balanced representation of product quality. Moreover, lower-rated products might already have a more consistent relationship between sentiment and scores, resulting in fewer adjustments after bias correction. This phenomenon underscores the importance of considering both sentiment and rating scores in understanding true product evaluations.

Score ratings and comments are prevalent mechanisms when evaluating products and services, especially in e-commerce. Due to the large number of available reviews on online platforms, methodologies can be tested to capture latent content within comments and use this information to remove any possible bias in the rating scores. In this work, text mining techniques (LDA and affective text mining) were used to extract information from the comments. Additionally, IRT models were used to further remove bias in the related rating score evaluations of the products.

In support of decision-making processes by potential buyers, there are interesting research opportunities as possible extensions of this work. The question of how to provide valuable and comprehensive visualizations using scores and text summarization still needs to be answered. A possible input would be to consider viewing topics to characterize the performance of specific features of products in e-commerce, having each topic as a dimension of the overall product score.

While the proposed methodology offers a promising approach to bias correction, it is significant to consider its potential impact on the end-user experience. The adjusted ratings, while aiming for greater objectivity, might deviate from the original scores provided by users. It is crucial to ensure transparency by clearly communicating to users that the displayed ratings have undergone bias correction. Additionally, providing explanations or visualizations highlighting the key factors influencing the adjustments could enhance user understanding and trust in the system. It is worth noting that this study did not explicitly address extreme scenarios such as sarcasm or review bombing, which could pose challenges for sentiment analysis techniques relying on dictionaries.

Other possible extensions would be the study of integrating this methodology into personalized recommendation systems, which would allow users to explore features according to their interests and further set performance ranges to refine searches. Finally, the proposed techniques could be used to analyze the consistency between scores and reviews. Another possible extension would be to consider all the proposed methodology's techniques and encapsulate them within online reviews and rating analysis engines for further incorporation into an autonomous purchase decision-making system. Then, purchasing decisions could be delegated to an intelligent system, which, using a decision model that considers both objective (for instance, the price of a product) and subjective aspects (for reviews and ratings), could autonomously select and purchase a product on behalf of the human customer. In this context, better analyzing and understanding the reviews and ratings generated by others could bring the intelligent system's purchasing decision closer to the customer's decision in the same scenario.

CONCLUSIONS

This research presented the design of a methodology for bias correction in online product ratings based on the combined forces of ML algorithms such as LDA, affective text mining, and IRT. The proposed approach was tested using the Amazon Fine Food Reviews dataset. Our experiments reveal that it is possible to conceive an automated method for bias correction in online product ratings by using IRT and affective text mining, all the above within a unified methodology.

According to the main results, some limitations can be listed. One constraint was the nature of the dataset in relation to the food products domain due to the presence of skewed score ratings accumulated in high scores. A second drawback is the availability of high computing process structures to perform the experiments with the entire dataset or with different domains.

Future work will extend the analysis to other diverse datasets beyond the Amazon Fine Food Reviews to further validate the proposed methodology's robustness and generalizability. This will allow us to assess the effectiveness of the bias correction approach across different product categories and user demographics, ensuring its broader applicability in real-world e-commerce scenarios.

Some extensions of this work open the possibility of exploring alternative approaches to possible applications in e-commerce, education, and management, among other domains. More concretely, further research can be conducted to analyze the possible effects that a pool of keywords (dictionary) might have on bias corrections as an additional feature incorporated into the IRT models. In the same way, writing styles or key phrases can be examined for further analysis. Additionally, it is possible to reduce skewed performance because the data uses merging categories.

REFERENCES

[1] H. Kaur, A. Rai, S.S. Bhatia, and G. Dhiman, "MOEPO: A novel Multi-objective Emperor Penguin Optimizer for global optimization: Special application in ranking of cloud service providers," *Engineering Applications of Artificial Intelligence*, vol. 96, Art. no. 104008, 2020, doi: 10.1016/j.engappai.2020.104008.

[2] D. Cabrera, C. Cubillos, E. Urra, and R. Mellado, "Framework for Incorporating Artificial Somatic Markers in the Decision-Making of Autonomous Agents," *Applied Sciences*, vol. 10, no. 20, Art. no. 7361, 2020, doi: 10.3390/app10207361.

[3] D. Cabrera-Paniagua, T.T. Primo, and C. Cubillos, "Distributed Stock Exchange Scenario Using Artificial Emotional Knowledge," in *Advances in Artificial Intelligence, IBERAMIA 2014. Lecture Notes in Computer Science*, vol. 8864, A.L.C. Bazzan and K. Pichara, Eds., 2014, pp. 649-659, doi: 10.1007/978-3-319-12027-0_52.

[4] J. Yang, X. Zou, W. Zhang, and H. Han, "Microblog sentiment analysis via embedding social contexts into an attentive LSTM," *Engineering Applications of Artificial Intelligence*, vol. 97, Art. no. 104048, 2021, doi: 10.1016/j.engappai.2020.104048.

[5] C. Kaatz, "Retail in my pocket- replicating and extending the construct of service quality into the mobile commerce context," *Journal of Retailing and Consumer Services*, vol. 53, Art. no. 101983, 2020, doi: 10.1016/j.jretconser.2019.101983.

[6] N.P. Rana, D.J. Barnard, A.M.A. Baabdullah, D. Rees, and S. Roderick, "Exploring barriers of m-commerce adoption in SMEs in the UK: Developing a framework using ISM," *International Journal of Information Management*, vol. 44, pp. 141-153, 2019, doi: 10.1016/j.ijinfomgt.2018.10.009.

[7] X. Tang and Z. Song, "Neurological effects of product price and evaluation on online purchases based on event-related potentials," *Neuroscience Letters*, vol. 704, pp. 176-180, 2019, doi: 10.1016/j.neulet.2019.04.019.

[8] D. Cabrera, N. Araya, H. Jaime, C. Cubillos, R. Vicari, and E. Urra, "Defining an Affective Algorithm for Purchasing Decisions in E-Commerce Environments," *in IEEE Latin America Transactions*, vol. 13, no. 7, pp. 2335-2346, 2015, doi: 10.1109/TLA.2015.7273796.

[9] J. Guo, X. Wang, and Y. Wu, "Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions," *Journal of Retailing and Consumer Services*, vol. 52, Art. no. 101891, 2020, doi: 10.1016/j.jretconser.2019.101891.

[10] B. Von Helversen, K. Abramczuk, W. Kopeć, and R. Nielek, "Influence of consumer reviews on online purchasing decisions in older and younger adults," *Decision Support Systems*, vol. 113, pp. 1-10, 2018, doi: 10.1016/j.dss.2018.05.006.

[11] L. Abdullah, R. Ramli, H.O. Bakodah, and M. Othman, "Developing a causal relationship among factors of e-commerce: A decision making approach," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 10, pp. 1194-1201, 2020, doi: 10.1016/j.jksuci.2019.01.002.

[12] B. Bhole and B. Hanna, "The effectiveness of online reviews in the presence of self-selection bias," *Simulation Modelling Practice and Theory*, vol. 77, pp. 108-123, 2017, doi:10.1016/j.simpat.2017.05.005.

[13] K. Li, Y. Chen, and L. Zhang, "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*, vol. 55, Art. no. 102107, 2020, doi: 10.1016/j.jretconser.2020.102107.

[14] G. Askalidis, S.J. Kim, and E.C. Malthouse, "Understanding and overcoming biases in online review systems," *Decision Support Systems*, vol. 97, pp. 23-30, 2017, doi: 10.1016/j.dss.2017.03.002.

[15] C.H. Tseng and L.F. Wei, "The efficiency of mobile media richness across different stages of online consumer behavior," *International Journal of Information Management*, vol. 50, pp. 353-364, 2020, doi: 10.1016/j.ijinfomgt.2019.08.010.

[16] S. Sarkar, S. Chauhan, and A. Khare, "A meta-analysis of antecedents and consequences of trust in mobile commerce," *International Journal of Information Management*, vol. 50, pp. 286-301, 2020, doi: 10.1016/j.ijinfomgt.2019.08.008.

[17] V. Priya, S. Subha, and B. Balamurugan, "Analysis of performance measures to handle medical E-commerce shopping cart abandonment in cloud," *Informatics in Medicine Unlocked*, vol. 8, pp. 32-41, 2017, doi: 10.1016/j.imu.2017.03.003.

[18] S. Renjith, A. Sreekumar, and M. Jathavedan, "An extensive study on the evolution of context-aware personalized travel recommender systems," *Information Processing & Management*, vol. 57, no. 1, Art. no. 102078, 2020, doi: 10.1016/j.ipm.2019.102078.

[19] I. Mazeh and E. Shmueli, "A personal data store approach for recommender systems: enhancing privacy without sacrificing accuracy," *Expert Systems with Applications*, vol. 139, Art. no. 112858, 2020, doi: 10.1016/j.eswa.2019.112858.

[20] A. Jain, P.K. Singh, and J. Dhar, "Multi-objective item evaluation for diverse as well as novel item recommendations," *Expert Systems with Applications*, vol. 139, Art. no. 112857, 2020, doi: 10.1016/j.eswa.2019.112857.

[21] L. Ardissono and N. Mauro, "A compositional model of multi-faceted trust for personalized item recommendation," *Expert Systems with Applications*, vol. 140, Art. no. 112880, 2020, doi: 10.1016/j.eswa.2019.112880.

[22] B. Feng, Q. Ye, and B.J. Collins, "A dynamic model of electric vehicle adoption: The role of social commerce in new transportation," *Information & Management*, vol. 56, no. 2, pp. 196-212, 2019, doi: 10.1016/j.im.2018.05.004.

[23] X. Wang, X. Lin and M.K. Spencer, "Exploring the effects of extrinsic motivation on consumer behaviors in social commerce: Revealing consumers' perceptions of social commerce benefits," *International Journal of Information Management*, vol. 45, pp. 163-175, 2019, doi: 10.1016/j.ijinfomgt.2018.11.010.

[24] T. Hu, H. Dai, and A.F. Salam, "Integrative qualities and dimensions of social commerce: Toward a unified view," *Information & Management*, vol. 56, no. 2, pp. 249-270, 2019, doi: 10.1016/j.im.2018.09.003.

[25] L.T.T. Tran, "Online reviews and purchase intention: A cosmopolitanism perspective," *Tourism Management Perspectives*, vol. 35, Art. no. 100722, 2020, doi: 10.1016/j.tmp.2020.100722.

[26] Y. Lamb, W. Cai, and B. McKenna, "Exploring the complexity of the individualistic culture through social exchange in online reviews," *International Journal of Information Management*, vol. 54, Art. no. 102198, 2020, doi: 10.1016/j.ijinfomgt.2020.102198.

[27] J.M. Kim, M. Jun, and C.K. Kim, "The Effects of culture on consumers' consumption and generation of online reviews," *Journal of Interactive Marketing*, vol. 43, pp. 134-150, 2018, doi: 10.1016/j.intmar.2018.05.002.

[28] V. Srivastava and A.D. Kalro, "Enhancing the Helpfulness of online consumer reviews: the role of latent (content) factors," *Journal of Interactive Marketing*, vol. 48, pp. 33-50, 2019, doi: 10.1016/j.intmar.2018.12.003.

[29] I. Portugal, P. Alencar, and D. Cowan, "The use of machine learning algorithms in recommender systems: A systematic review," *Expert Systems with Applications*, vol. 97, pp. 205-227, 2018, doi: 10.1016/j.eswa.2017.12.020.

[30] H.C. Lee, H.C. Rim, and D.G. Lee, "Learning to rank products based on online product reviews using a hierarchical deep neural network," *Electronic Commerce Research and Applications*, vol. 36, Art. no. 100874, 2019, doi: 10.1016/j.elerap.2019.100874.

[31] Y. Liu, J.W. Bi, and Z.P. Fan, "Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory," *Information Fusion*, vol. 36, pp. 149-161, 2017, doi: 10.1016/j.inffus.2016.11.012.

[32] T. Eckes, *Introduction to many-facet rasch measurement*, Berlin, Germany: Peter Lang GmbH, 2011.

[33] Z. Ye and J. Sun, "Comparing Item Selection Criteria in Multidimensional Computerized Adaptive Testing for Two Item Response Theory Models," in *3rd International Conference on Computational Intelligence and Applications (ICCIA)*, Hong Kong, China, 2018, pp. 1-5, doi: 10.1109/ICCIA.2018.00008.

[34] L. Peng, G. Cui, Y. Chung, and C. Li, "A multi-facet item response theory approach

to improve customer satisfaction using online product ratings," *Journal of the Academy of Marketing Science*, vol. 47, no. 5, pp. 960-976, 2019, doi: 10.1007/s11747-019-00662-w.

[35] S. Kumano and K. Nomura, "Multitask Item Response Models for Response Bias Removal from Affective Ratings," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Cambridge, United Kingdom, 2019, pp. 1-7, doi: 10.1109/ACII.2019.8925539.

[36] L. Chen, Y. Tu, X. Xia, L. Wang, and L. Bao, "Predicting Decision-making Using Item Response Theory," in *2019 3rd International Conference on Circuits, System and Simulation (ICCSS),* Nanjing, China, 2019, pp. 229-234, doi: 10.1109/CIRSYSSIM.2019.8935601.

[37] T. Jo, *Text Mining*, *Studies in Big Data Challenge*, vol. 45, Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-319-91815-0.

[38] E. Jorquera and H. Rosas, "Visual summarization of online news through topics using Latent Dirichlet Allocation," in *7th Latin American Conference on Networked and Electronic Media (LACNEM 2017)*, Valparaiso, Chile, 2017, pp. 1-6. doi: 10.1049/ic.2017.0038.

[39] S. Chatterjee, "Drivers of helpfulness of online hotel reviews: A sentiment and emotion mining approach," *International Journal of Hospitality Management*, vol. 85, Art. no. 102356, 2020, doi: 10.1016/j.ijhm.2019.102356.

[40] W.M. Wang, Z. Li, Z.G. Tian, J.W. Wang, and M.N. Cheng, "Extracting and summarizing affective features and responses from online product descriptions and reviews: A Kansei text mining approach," *Engineering Applications of Artificial Intelligence*, vol. 73, pp. 149-162, 2018, doi: 10.1016/j.engappai.2018.05.005.

[41] C. Wan, Y. Peng, K. Xiao, X. Liu, T. Jiang, and D. Liu, "An association-constrained LDA model for joint extraction of product aspects and opinions," *Information Sciences*, vol. 519, pp. 243-259, 2020, doi: 10.1016/j.ins.2020.01.036.

[42] X. Xie, Y. Fu, H. Jin, Y. Zhao, and W. Cao, "A novel text mining approach for scholar information extraction from web content in Chinese*," Future Generation Computer Systems*, vol. 111, pp. 859-872, 2020, doi: 10.1016/j.future.2019.08.033.

[43] J.C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet Allocation," *The Art and Science of Analyzing Software Data*, New York, NY: Morgan Kaufmann, 2015, pp. 139-159. doi: 10.1016/B978-0-12-411519-4.00006-9.

[44] J.P. Fox, Bayesian Item Response Modeling: Theory and Applications, New York, NY: Springer New York, 2010, doi: 10.1007/978-1-4419-0742-4.

[45] Stan, "RStan: the R interface to Stan," mc-stan.org. Accessed: 1 May, 2024. [Online]. Available: http://mc-stan.org/

[46] A. Grover *et al.,* "Bias Correction of Learned Generative Models using Likelihood-Free Importance Weighting," 2019, arXiv: 1906.09531.

[47] T. Sakumura and H. Hirose, "Bias Reduction of Abilities for Adaptive Online IRT Testing Systems," in *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, Kumamoto, Japan, 2016, pp. 450-455, doi: 10.1109/IIAI-AAI.2016.122.

[48] D. Bolt, Y.C. Wang, R.H. Meyer, and L. Pier, "An IRT Mixture Model for Rating Scale Confusion Associated with Negatively Worded Items in Measures of Social-Emotional Learning," *Applied Measurement in Education*, vol. 33, no. 4, pp. 331-348, 2020, doi: 10.1080/08957347.2020.1789140.

[49] J.J. McAuley and J. Leskovec, "From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews," in *Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro Brazil*: ACM, 2013, pp. 897-908, doi: 10.1145/2488388.2488466.

[50] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA, 2004, pp. 168-177, doi: 10.1145/1014052.1014073.