

# Comprehensive helpfulness of online reviews: A dynamic strategy for ranking reviews by intrinsic and extrinsic helpfulness

Jindong Qin<sup>a,b,c,\*</sup>, Pan Zheng<sup>a</sup>, Xiaojun Wang<sup>d</sup>

<sup>a</sup> School of Management, Wuhan University of Technology, Wuhan 430070, Hubei, China

<sup>b</sup> Research Center for Data Science and Intelligent Decision Making, Wuhan University of Technology, Wuhan 430070, Hubei, China

<sup>c</sup> Research Institute of Digital Governance and Management Decision Innovation, Wuhan University of Technology, Wuhan 430070, Hubei, China

<sup>d</sup> School of Management, University of Bristol, Tyndall Avenue, Bristol BS8 1TH, United Kingdom



## ARTICLE INFO

### Keywords:

Online reviews  
Intrinsic helpfulness (IH)  
Extrinsic helpfulness (EH)  
Ensemble neural network model (ENNM)  
Dynamic ranking strategy

## ABSTRACT

Information overload often makes it difficult for consumers to identify valuable online reviews through the traditional “helpful votes” button in the big data era, so it is essential to locate helpful reviews. Unlike the existing efforts that often measure online reviews’ helpfulness one-sidedly, this study takes the intrinsic helpfulness (IH) and extrinsic helpfulness (EH) into account, and the intrinsic-extrinsic comprehensive helpfulness (ICH-ECH) plot can be constructed by ensemble neural network model (ENNM) and time-weighted standard deviation accordingly. Furthermore, this study proposes a measure of EH ignored by previous studies, that is, the percentage of negative replies, which contain useful information that can measure online reviews helpfulness. We corrected it with a time sliding window by an improved iterative Bayesian probability approach (IBPA). In addition, this study further proposes a dynamic time-aware helpfulness ranking (DTAHR) model to dynamically rank reviews and identify beneficial reviews in a short time. We used real data sets from JD.com to conduct all experiments. The experimental results show that the performance of the DTAHR model is significantly better than other strategies. Our findings offer guidelines to evaluate the helpfulness of online reviews from multiple perspectives and rank them dynamically.

## 1. Introduction

Word-of-mouth (WOM) plays a significant role in influencing customers’ attitudes towards products, and purchase decisions [1]. Today, most e-commerce platforms, such as Amazon.com, Yelp.com, JD.com, provide a place for customers to post online reviews. New customers will often refer to these previously posted reviews when making their purchase decisions. These reviews usually contain useful purchase information that customers care about, such as product quality, variety, price. Online reviews do not only help customers get an overview about the product, but also enable the interactive processes such as sharing product/service experience, advocating certain values, and socializing among customers that further develop an online community.

One recent study reported that about 92% of customers read online reviews before making a purchase decision, 89% said that online reviews heavily influenced their purchase decision(favorably or unfavorably), and 82% believed it was better to read online reviews before purchasing than to check with sales staff in-store [2]. However, with the

rapid development of e-commerce, the amount of online reviews available to customers have exploded. Nowadays, a product usually has at least several hundred reviews, and it is almost impossible for customers to read them in detail, which makes it difficult for customers to choose useful ones from the vast number of online reviews, especially when the reviews are not ranked in the right order [3,4]. Therefore, it is essential for a platform to accurately evaluate the helpfulness of online reviews and develop a ranking strategy accordingly [5]. To address these problems, this paper will propose a new analytical method for evaluating the comprehensive helpfulness of online reviews and develop a novel dynamic ranking strategy accordingly.

The “helpfulness” of online reviews is a multi-dimensional concept measured by several indicators and influenced by many variables with different mechanisms. Many current studies on the helpfulness of online reviews are based on the incentive mechanism set up by e-commerce platforms, which often provide a system to record how much people agree with the reviews [6]. Take JD.com for example, it provides customers with a vote button to express their review approval. When

\* Corresponding author at: School of Management, Wuhan University of Technology, Wuhan 430070, Hubei, China.

E-mail address: [qinjindongseu@126.com](mailto:qinjindongseu@126.com) (J. Qin).

customers read a review, they can pay attention to the rating and review text and see how many votes it has attracted from other customers, which allows them to identify the helpfulness of the online review more accurately. There is no consensus in the academic community on how to measure the helpfulness of reviews correctly [7], and many studies measure the helpfulness of reviews in terms of convenience [8–10]. For example, the number of customer votes or the ratio of helpful votes measures the helpfulness of reviews and then ranked accordingly [11]. However, the main disadvantage is that it is difficult to identify the helpfulness of a review when it has few or no votes. Reviews without votes are rarely presented to customers, even though they may be considered helpful. Therefore, only a few helpful reviews are presented to customers, and most of them are eliminated, which undoubtedly increases the burden on customers.

Currently, research on the helpfulness of reviews mainly focuses on two aspects. One is to analyze the factors that affect the helpfulness of reviews and the influencing mechanism, such as the product type [12], the semantics of reviews [13], review types [14], and predict the helpfulness of reviews based on previous analyses [15,16]. In these studies, helpful reviews were measured mainly by the number of votes and the ratio of helpful votes. Another aspect is to explore how to measure the helpfulness of reviews and rank reviews accordingly, such as the semantics of reviews [17], the number and density of attribute words [18], the consistency of mutual information entropy [19], the consistency of a review subset with overall information [5], or an integration of them.

Regarding these previous studies, the helpfulness of reviews is usually regarded as either the usefulness of the reviews themselves (semantics, the number, and density of attribute words, sentiment orientation, etc.) [17–19,5,20,21] or the recognition of customers (the number of votes and the ratio of helpful votes) [12–16]. These two kinds of helpfulness of reviews are not considered comprehensively. Meanwhile, most studies on reviews ranking are also based on one of them. These ranking strategies have the following three problems: (1) relying solely on the usefulness of the reviews themselves, without sufficient reference to the opinions of the wide customer group, may overlook fake reviews and click farming behaviors [22]; (2) relying solely on the recognition of customers may lead to ratchet effect and Matthew effect [23,15]; and (3) few previous studies have considered the effect of the release time of reviews on helpfulness. Both the product and people's perceptions change rapidly, and we should incorporate these changes into the model in time and rank the reviews dynamically. Therefore, the motivation of this study is to measure the helpfulness of reviews more comprehensively from a new perspective and find a strategy to rank reviews reasonably.

This paper proposes an intrinsic-extrinsic comprehensive helpfulness (ICH-ECH) model to measure the review helpfulness from both intrinsic and extrinsic aspects and a dynamic time-aware helpfulness ranking (DTAHR) model to rank reviews accordingly to address these research gaps. Intrinsic helpfulness (IH) is defined as helpfulness derived from review text and characteristics (e.g., the length and sentiment of reviews), which are determined at the review release and are not affected by external conditions like customers and platforms. Extrinsic helpfulness (EH) is defined as the recognition of customers, which is not inherent when the review is released but an external recognition generated through interaction with customers and adjustments from the platform (such as the number of votes and the ratio of helpful votes). Intrinsic-extrinsic review helpfulness should include both elements. Then, IH and EH can calculate the intrinsic comprehensive helpfulness (ICH) and extrinsic comprehensive helpfulness (ECH), and the ICH-ECH plot can be constructed. In addition, we further propose a measure of EH that contains the recognition of other customers: the percentage of negative replies. Then, the reviews can be ranked by DTAHR accordingly.

This paper makes the following key contributions:

(1) This research contributes to the existing literature on review helpfulness [12–16,20,21] by proposing an ICH-ECH model that measures the helpfulness of reviews from both intrinsic and extrinsic aspects. The ICH-ECH plot can be constructed to understand the helpfulness distribution of all reviews on each product considering both the internal review features and the external review context.

(2) This research complements to the existing literature on the extrinsic review helpfulness [12–16] by proposing a new measure of EH (i.e., the percentage of negative replies). By making full use of the text information contained in the replies, we also improve some measures of IH complementing to the existing studies on intrinsic review helpfulness [17–19,5,20,21].

(3) Based on the obtained ICH and ECH of reviews, we propose a DTAHR model to rank review dynamically according to their comprehensive helpfulness and time-aware weight, which can alleviate the ratchet effect and Matthew effect [23,15].

The following is the organization of the paper. In Section 2, some recent work on analyzing the factors and mechanisms that influence the helpfulness of reviews and exploring how to measure the helpfulness of reviews are reviewed. In Section 3, we first introduce the framework of the model and then explain the process of constructing an ICH-ECH plot for each product and ranking online reviews by the DTAHR model. In Section 4, a case study concerning three search products (i.e., cell phone, laptop, camera) and three experience products (i.e., lipstick, running shoes, rice cooker) from JD.com illustrates the use of the proposed methodology and analyze the results. In Section 5, we carry out some discussions and emphasize this study's theoretical and practical implications. Finally, we provide summary reviews and outline the direction of future research in Section 6.

## 2. Related work

This section explains the three lines of research to which this study is related: features and mechanisms affecting the helpfulness of reviews, reviews helpfulness prediction, and measure and rank the helpfulness of reviews.

### 2.1. Features and mechanisms affecting the helpfulness of reviews

The helpfulness of online reviews includes the usefulness of the reviews themselves and the recognition of customers. Extant studies indicate that many features affect the helpfulness of online reviews with different mechanisms. These studies are mainly conducted considering three aspects: reviews (length, depth, readability, semantic information, rating, etc.) [9,24–26,20,21], reviewers (membership level, experience level, social status, number of fans, etc.) [10,27,28], and products (product type, product characteristics, etc.) [29,18]. Rating and review length, as the most intuitive features, were first investigated [30]. The results show that these two factors have a positive impact on the helpfulness of reviews, and the impact is further influenced by product type [31]. It is reported that text readability has a greater impact on review helpfulness than review length [9], and the subjectivity level, language correctness and reviewer reputation of reviews also have a significant impact [10]. Other studies have also pointed out that the helpfulness of the review is not only related to the characteristics of the review, but also depends on the context, such as the psychological distance from the purchase event and the social distance between the review reader and the reviewer [4]. Furthermore, sentiments have different effects on review helpfulness [10].

Except for these standard features, some unique features affect the helpfulness, such as adjacent neighbors [32] and order of reviews [33]. Prior studies suggested that [32], in practice, it is difficult for customers to deal with reviews independently because reviews are presented in a sequence and more likely to be affected by their adjacent neighbors. Another study suggested that [33] the order of reviews was negatively

correlated with the helpfulness of the review and that the negative effect diminished when the reviewer was more professional and experienced, or the reviews were more negative and recently posted. Therefore, the research on features and their influencing mechanism is still important. To better understand the research progress of this area, Table 1 shows the review information used by some critical works of literature in studying the helpfulness of reviews, of which text review, rating and non text review belong to IH and helpful votes and reply belong to EH.

Although existing studies have conducted detailed research on the mechanism of these features affecting review helpfulness, most of them take the number of votes and the ratio of helpful votes as dependent variables, without considering the ratchet effect and Matthew effect leads to the bias of results.

## 2.2. Reviews helpfulness prediction

From the predictive perspective, a larger portion of research work has focused on predicting the helpfulness of each review [37,34]. Previous studies have developed various classification and regression models to predict the helpfulness of reviews, in which the helpfulness is usually divided into several categories, such as Liu et al. [38] defined five classes (i.e., high quality, medium quality, low quality, duplicate, and spam) and train their classification models using manually annotated labels. Zheng et al. [35] and Ghose et al. [10] defined two types (i.e., helpful and unhelpful) by specifying a threshold for the percentage of positive votes. The models used for helpfulness prediction include Neural Network Regression [32], Tobit-regression [39], Random Forest Regression [15], M5P algorithm [40], Support Vector Machine [41], among others. In addition, from the perspective of big data, Zhao et al. [26] use the technical attributes of online text reviews and customers' participation in the review community to predict overall customer satisfaction and explore the influence mechanism of various factors. The independent variables of prediction include review structure (e.g., length and depth), semantic (e.g., product features), syntactic (e.g., number of verbs, number of nouns), lexical features (e.g., n-gram), environmental features (e.g., above and below, ranking), digital features (e.g., rating, release schedule), among others [42]. In addition, the social characteristics of reviewers, such as reputation and historical performance, are also included in the prediction model [43] and classified using the information quality framework [35].

Previous review helpfulness prediction models generally perform well on training sets and test sets but deficiently on future predictions, especially later reviews [30]. This is probably because these models lack the training to account for changes in the environment (such as the progress of products and the evolution of people's taste) [44]. Both the product and people's perceptions change rapidly. These prediction models must capture these changes in real-time and modify the model to

**Table 1**

Review information used by some key literatures in studying the helpfulness of reviews.

Study	IH			EH	
	Text review	Rating	Non text review	Helpful votes	Reply
Korfiatis et al. [9]	✓	✓	✗	✓	✗
Park and Nicolau [27]	✗	✓	✗	✓	✗
Zhou and Guo [33]	✓	✗	✗	✓	✗
Lee and Choeh [34]	✓	✓	✗	✓	✗
Zheng et al. [35]	✓	✗	✓	✓	✗
Zhang et al. [19]	✓	✗	✗	✗	✗
Guo et al. [30]	✗	✗	✗	✓	✗
Bi et al. [36]	✓	✓	✗	✗	✗
Sun et al. [18]	✓	✗	✗	✓	✗
Our study	✓	✓	✓	✓	✓

make more accurate predictions, such as time reference to market and consumer feedback.

## 2.3. Measure and rank the helpfulness of reviews

Unlike predicting review helpfulness with the number and ratio as a dependent variable, some studies [30,11,19,5,45,26] have measured helpfulness by some specific criteria and ranked the reviews accordingly. When reviews get fewer votes, considering the number and ratio as the helpfulness of reviews will become inaccurate or even deviate seriously. Therefore, Guo et al. [30] proposed an iterative Bayesian probability approach (IBPA) to evaluate the helpfulness automatically and experimentally demonstrated that the method significantly improved the accuracy of the training model. Using a new system based on Bayesian statistics, Wang et al. [11] proposed two simple methods to improve existing ranking methods and demonstrate the scientificity of the ranking method through simulation experiments. These methods effectively alleviate the problem of undervoting, but these studies did not consider the features of intrinsic helpfulness, leading to a one-sided result.

Zhang et al. [5] and Zhang et al. [19] introduced the consistency of general information and mutual information entropy to measure the helpfulness of reviews, and solved them by formulating the consistency issue as an optimization problem and proposing a heuristic algorithm accordingly. On this basis, Wang et al. [45] measured the helpfulness of reviews from a combined perspective of consistency and time-awareness in light of product features and sentiment orientations. These studies all treat the reviews as a set and rank them by the consistency of a review subset with overall information. In addition, the sentiment and textual characteristics of reviews can also measure the helpfulness [26]. These methods are based on the intrinsic features of the reviews, ignoring the extrinsic features (number of votes and the ratio of useful votes) and the subjective opinions of the reviewer, which may lead to bias in the results.

## 2.4. Summary

This section provides an overview of some important literature regarding features and mechanisms affecting the helpfulness of reviews, reviews helpfulness prediction, and measure and rank the helpfulness of reviews. The features in Section 2.1 and independent variables in Sections 2.1 and 2.2 are considered as IH and the dependent variable in Section 2.2 and measures in Section 2.3 are considered as EH in this paper. Overall, while previous studies [9,24–26,10,27–29,18] have demonstrated the influence of intrinsic and extrinsic features on reviews helpfulness, few studies have attempted to measure the helpfulness of reviews from both intrinsic and extrinsic aspects. Therefore, this research bridges the literature gap by proposing an ICH-ECH model that measures the helpfulness of reviews from both intrinsic and extrinsic aspects and proposing a DTAHR model to rank review dynamically according to their comprehensive helpfulness and time-aware weight.

## 3. Methodology

This section introduces constructing an ICH-ECH plot for each product and presents a DTAHR model to rank online reviews. Fig. 2 shows the framework of the methodology, which is composed of three phases as follows:

- Phase 1. Extracting attribute words and calculating IH and EH.
- Phase 2. Determining the weights of IH and EH and calculating ICH and ECH.
- Phase 3. Constructing ICH-ECH plot and ranking reviews by DTAHR.

Let  $IH^i$  ( $i = 1, 2, \dots, 14$ ) and  $EH^i$  ( $i = 1, 2$ ) denote the measurement indicators of IH and EH, and the specific information is shown in

**Table 2**  
Variable interpretation.

Variable	Explain	Type
IH <sup>1</sup>	Consistency between the review and overall information	Continuous
IH <sup>2</sup>	Number of attribute words in the review	Continuous
IH <sup>3</sup>	Density of attribute words in the review	Continuous
IH <sup>4</sup>	Number of negative sentences in the review	Discrete
IH <sup>5</sup>	Number of neutral sentences in the review	Discrete
IH <sup>6</sup>	Number of positive sentences in the review	Discrete
IH <sup>7</sup>	Proportion of negative sentences in the review	Continuous
IH <sup>8</sup>	Proportion of neutral sentences in the review	Continuous
IH <sup>9</sup>	Proportion of positive sentences in the review	Continuous
IH <sup>10</sup>	Number of image in the review	Discrete
IH <sup>11</sup>	Rating of the review	Discrete
IH <sup>12</sup>	Whether the reviewer a JD plus member	Binary
IH <sup>13</sup>	Whether the review have a video	Binary
IH <sup>14</sup>	Whether the reviewer have a image	Binary
EH <sup>1</sup>	The number of votes in the review	Discrete
EH <sup>2</sup>	The percentage of negative replies in the review	Continuous

**Table 2.** In the first phase, online reviews are divided into reply & votes, text reviews, and non-text reviews. Specifically, attribute words of the product are extracted from reply and text reviews using latent Dirichlet allocation (LDA). Then we can obtain IH<sup>1</sup>, IH<sup>2</sup>, ..., IH<sup>9</sup> by the attribute words and text reviews. IH<sup>10</sup>, IH<sup>11</sup>, ..., IH<sup>14</sup> can be obtained by non-text reviews and EH<sup>1</sup>, EH<sup>2</sup> can be obtained by reply & votes. In the second phase, we can calculate the weights of these IH by ensemble neural network model (ENNM) and EH by the degree of dispersion. Under the value of IH and EH, the ICH and ECH can be obtained. In the final phase, based on the obtained ICH and ECH, the ICH-ECH plot of each product can be constructed. The detailed descriptions of Phase 1 are illustrated in Sections 3.1, 3.2, 3.3, Phase 2 are presented in Section 3.4, and Phase 3 are illustrated in Section 3.5.

### 3.1. Processing of replies

In this subsection, we propose a new EH, ignored in previous studies: the percentage of negative replies. Take JD.com as an example; after posting a review, other customers can not only vote on the review but also publish some replies under the review, as shown in Fig. 1 (has been translated into English). To the best of our knowledge, many responses are questioning and negative ones, which carry a large amount of information that customers are concerned about. Reply 1 and 2 in Fig. 1 show customers' concern about battery and fingerprint lock. At the same time, the percentage of negative replies can also be used to measure customers' recognition of a review. Although a review looks helpful, it

may not be recognized by customers because there can be fake and click farming behaviors [22], as Reply 3 and 4 shown in Fig. 1.

In the following, we process reviews from two aspects: attribute word extraction and the identification of sentiment orientations.

#### 3.1.1. Attribute words extraction based on LDA

LDA is an unsupervised generative probabilistic model that can extract topics from a large number of online reviews, each consisting of a set of related words of similar meanings [46]. Studies [47,48] have shown that LDA is an effective approach to extract product attribute words from online reviews. In this study, we determine the number of topics by log-likelihood [49]. We rank these obtained related words by word frequency, then manually remove those unrelated, and finally get the important attribute words. The overall process consists of two steps: (1) preprocessing of replies and (2) attribute words extraction.

##### (1) Preprocessing of replies.

The replies contain some noise and irrelevant data. To improve the effectiveness of attribute word extraction, we designed some metrics, such as the number of data replicates, data length, and then eliminated the data according to specific rules. After excluding the irrelevant data, we split the text and filtered the stop, negative, affective, and degree words to obtain the preprocessed replies.

##### (2) Attribute words extraction.

Let  $\mathbb{R}_{\text{review}} = \{R_1, R_2, \dots, R_N\}$  denote the set of online reviews and listed in descending order of release time, where  $R_n$  is the  $n$ th review in  $\mathbb{R}_{\text{review}}$ . Let  $\mathbb{R}_{\text{reply}} = \{r_1, r_2, \dots, r_N\}$  denote the reply sets of all reviews, where  $r_n$  is the reply set of  $R_n$ . Let  $r_n = \{r_n^1, r_n^2, \dots, r_n^M\}$  denote the all replies of  $R_n$ , where  $r_n^m$  is the  $m$ th reply of the  $n$ th review. The pre-processed replies are input into the LDA model to extract each topic's subject and related words. We adjusted the results by manually filtering the noise since some words and synonyms may be in the extracted topics. Then the attribute words of replies can be denoted as  $f_{\text{reply}} = \{f'_1, f'_2, \dots, f'_P\}$ , where  $f'_p$  is the  $p$ th frequent attribute word of replies. Similarly, the attribute words of reviews can be obtained and denoted as  $f_{\text{review}} = \{f^R_1, f^R_2, \dots, f^R_Q\}$ , where  $f^R_q$  is the  $q$ th frequent attribute word of reviews.

#### 3.1.2. Identification of sentiment orientation

In this subsection, we identify the sentiment orientation of text by the API of Baidu AI Cloud. We choose Baidu AI Cloud as our sentiment analysis tool because the corpus data of this research is Chinese, and Baidu AI Cloud is one of the largest AI cloud service markets in China, ranking first in the NLP (natural language processing) market share in China. Refer to the sentiment classification threshold of Baidu AI Cloud, the sentiment category  $E^*(r_n^m)$  can be obtained by Eq. (1).

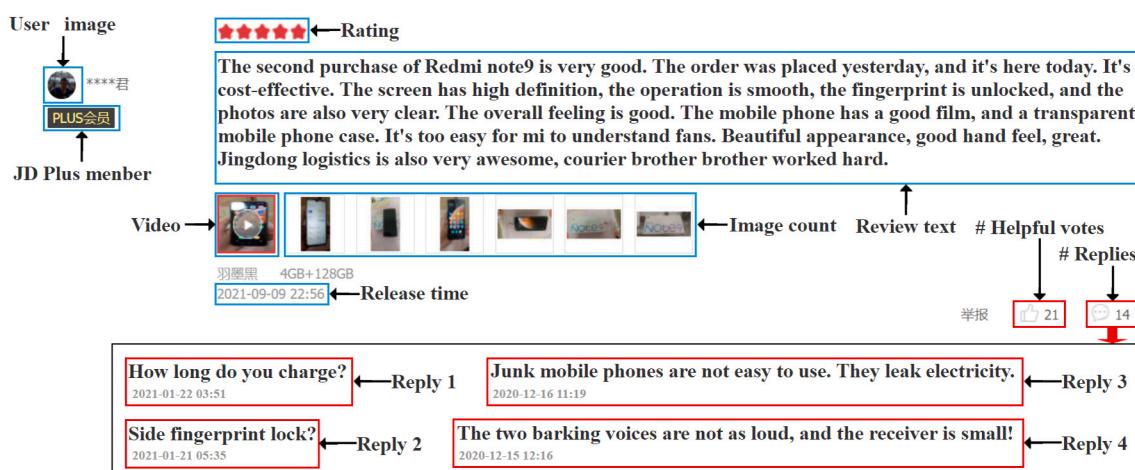


Fig. 1. An example of product reviews and replies on JD.com.

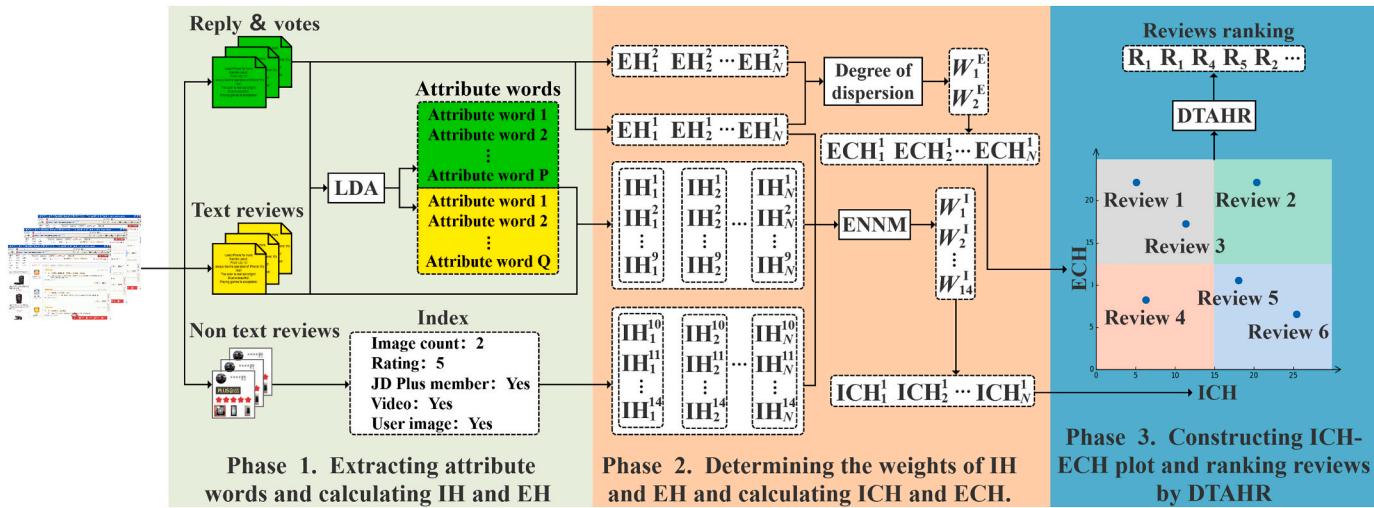


Fig. 2. The framework of constructing ICH-ECH plot and ranking reviews by DTAHR.

$$E^* \begin{pmatrix} r_n^m \\ r_n^m \end{pmatrix} = \begin{cases} \text{Negative} & 0 \leq E(r_n^m) < 0.4 \\ \text{Neutral} & 0.4 \leq E(r_n^m) < 0.6 \\ \text{Positive} & 0.6 \leq E(r_n^m) \leq 1 \end{cases} \quad (1)$$

where  $E(r_n^{mk})$  is the sentiment orientation score of  $r_n^m$  obtained by Baidu AI Cloud ( $E(r_n^m) \in [0, 1]$ ).

### 3.2. Calculating the IH of reviews

IH is the inherent attribute representing the usefulness of the reviews themselves. These attributes have been determined at the review release and are not affected by external conditions like customers and platforms.

This study selects 14 most representative IH measures:  $IH^1, IH^2, IH^3, IH^4, IH^5, \dots, IH^9$  are the subdivisions of reviews sentiment, which is one of the most essential features of reviews [36,18,19,22,23,9,24–26,20,21];  $IH^{10}, IH^{11}, \dots, IH^{14}$  is introduced by previous study [18] and our investigation on JD.com, which is in line with intuition. Previous studies [30,19,18,5,45,36,22,23,9,24–26,20,21] have shown that these IH measures have a representative impact on the helpfulness of reviews. We use replies to improve 3 of them: the consistency of a review with comprehensive information ( $IH^1$ ), the number of attribute words ( $IH^2$ ), and the density of attribute words ( $IH^3$ ). Let  $IH_n^i$  denote the value of  $IH^i$  of review  $R_n$ .

### 3.3. Calculating the EH of review

The EH is not inherent when the review is released but an external recognition generated through interaction with customers and adjustments from the platform. In addition to the above number of votes and the percentage of helpful votes, EH also includes ranking reviews and others.

In this study, we propose a new measure of EH, that is, the percentage of negative replies. Negative replies can reflect the level of customer distrust, and cooperating with the number of votes can more comprehensively measure the helpfulness of a review. For example, the IH of a review is high. Still, the number and percentage of negative replies are also high, because of the customer dissatisfaction with fake information about the review. Thus, click farming behavior is identifiable through the percentage of negative replies. We chose the number of votes and the percentage of negative replies as the measures of EH. Next, we will introduce the calculation process of EH<sup>2</sup>.

Since most online reviews suffer from an inadequate reply, using the

percentage of negative replies to represent customer distrust of reviews can sometimes be inaccurate. For example, 50 out of 100 replies are negative would be more accurate than 1 out of 2 negative replies, even though both percentages of negative replies are 50%. Guo et al. [30] proposed an IBPA model to solve this kind of problem. Let  $E$  denote the prior probability of percentage of negative replies. Because the number of negative replies for a review follows a binomial distribution (negative and not negative), the prior distribution of  $E$  can be generally configured as beta distribution, that is,  $E \sim Beta(\alpha, \beta)$  [30].

However, we consider that the prior distribution of the percentage of negative replies is not invariant but varies over time. We refine the original model to allow the fitted prior distributions to capture this variation. The effect of the release time of each review on the prior distribution was not considered when fitting the previous distribution of the invariant by Guo et al. [30]. The closer the reviews are released, the more likely they will have the same prior distribution. Hence, we respectively use the release time of each review as a reference to choose reviews with similar release times to fit the prior probability of the percentage. We draw on the concept of the time sliding window for reference. In each time sliding window, we fit the prior probability of the percentage and then correct the data in the center of the window.

Let  $E_n$  denote the percentage of negative replies of review  $R_n$ . We arrange  $\mathbb{R}_{\text{review}}$  in descending order of release time and let  $t_{R_n}$  denote the release time of  $R_n$  (i.e.,  $t_{R_1} \leq t_{R_2} \dots \leq t_{R_N}$ ). Therefore, taking the  $E_n$  as a reference, the time sliding window  $T_n^E$  can be denoted by Eq. (2).

$$T_n^E = \{E_i | t_{R_n} - \Delta \leq t_{R_i} \leq t_{R_n} + \Delta, k_i > 0\} \quad (2)$$

where  $\Delta$  is the radius of the time sliding window, that is, the size of the time sliding window is  $2\Delta$ , and the size of the time sliding window varies when near the start and end points.

Let  $\mu_{E_n}$  and  $\sigma_{E_n}$  denote the mean and variance of  $T_n^E$ . Let  $x_n$  and  $k_n$  denote the number of negative replies and the total number of replis of  $R_n$ . The framework of IBPA with time sliding window is shown in Fig. 3 and the specific operation steps for prior distribution fitting are shown in Algorithm 1.<sup>1</sup> Algorithm 1 is mainly divided into three steps: (1) slide the time window and search the data in it; (2) fit the value of  $\alpha$  and  $\beta$  in the time window by function IBPA; and (3) correct the data in the center of

<sup>1</sup> After obtaining  $\mu_{E_n}$  and  $\sigma_{E_n}^2$ , the parameters  $\alpha$  and  $\beta$  can be fitted through the method of moments as follows:  $\alpha = \mu_{E_n} \left( \frac{\mu_{E_n}(1-\mu_{E_n})}{\sigma_{E_n}^2} - 1 \right)$ ,  $\beta = (1-\mu_{E_n}) \left( \frac{\mu_{E_n}(1-\mu_{E_n})}{\sigma_{E_n}^2} - 1 \right)$ .

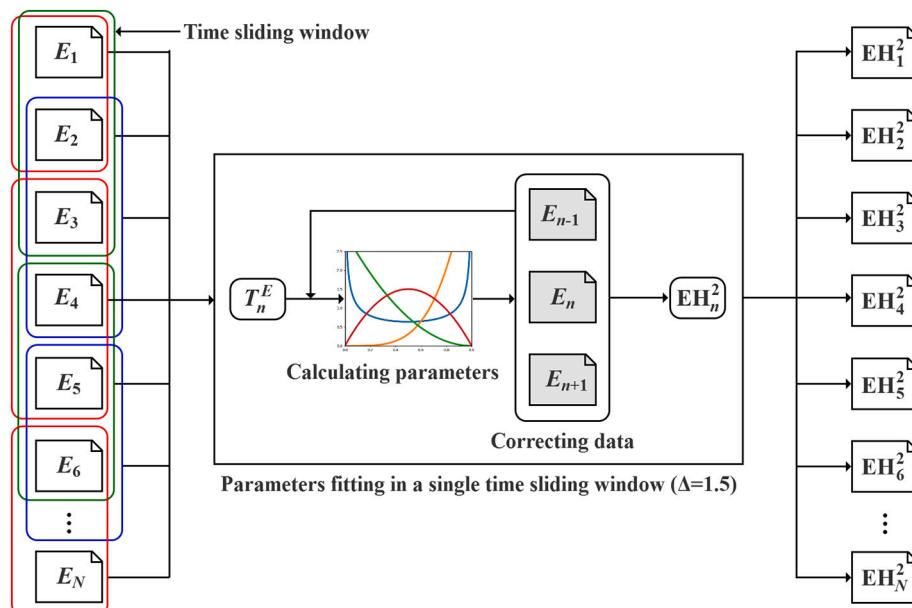


Fig. 3. The framework of IBPA with time sliding window.

the time window by the fitted  $\alpha$  and  $\beta$ . It is worth noting that not all data in the time sliding window will converge due to data sparsity. We only iterate once for those non-convergent data to correct the prior probability. Let  $EH_n^i$  denote the value of  $EH^i$  of review  $R_n$ .

**Algorithm 1** IBPA with time sliding window distribution fitting algorithm.

```

Input: Training review set  $\mathbb{R}_{\text{review}} = \{R_1, R_2, \dots, R_N\}$ , voting set  $VR = \{(x_1, k_1), (x_2, k_2), \dots, (x_N, k_N)\}$ , review release time set  $TR = \{t_{R_1}, t_{R_2}, \dots, t_{R_N}\}$ , time sliding window radius  $\Delta$ 
Output: Corrected percentage of negative replies  $EH^2 = \{EH_1^2, EH_2^2, \dots, EH_N^2\}$ 
1: function IBPA( $\mathbb{R}_{\text{review}}$ ,  $VR'$ ,  $x_n, k_n$ )
2:    $\alpha, \beta \leftarrow 0$   $\triangleright$  initialize distribution parameters
3:   repeat
4:      $\alpha' \leftarrow \alpha, \beta' \leftarrow \beta$   $\triangleright$  record current parameters values
5:      $EH^2' \leftarrow \emptyset$ 
6:     for  $R_i \in \mathbb{R}'_{\text{review}}$  do  $\triangleright$  calculate percentage of negative replies based on fitted distribution
7:        $EH^2' \leftarrow (x'_i + \alpha)/(k'_i + \alpha + \beta)$ 
8:        $EH^2' \leftarrow EH^2' \cup \{EH^2'_i\}$ 
9:     end for
10:     $\mu \leftarrow \text{mean}(EH^2')$   $\triangleright$  calculate mean ( $\mu$ ) and variance ( $\delta^2$ ) of percentage of negative replies
11:     $\delta^2 \leftarrow \text{var}(EH^2')$ 
12:     $\alpha \leftarrow \mu \times (\mu \times (1 - \mu)/\delta^2 - 1)$   $\triangleright$  update distribution parameters
13:     $\beta \leftarrow (1 - \mu) \times (\mu \times (1 - \mu)/\delta^2 - 1)$ 
14:    until  $\alpha' = \alpha, \beta' = \beta$ 
15:     $EH_n^2 \leftarrow (x_n + \alpha)/(k_n + \alpha + \beta)$   $\triangleright$  correct percentage of negative replies
16:  return  $EH_n^2$ 
17: end function
18: function CorrectProportion( $\mathbb{R}_{\text{review}}$ ,  $VR, TR, \Delta$ )
19:    $EH^2 \leftarrow \emptyset$ 
20:   for  $R_n \in \mathbb{R}_{\text{review}}$  do  $\triangleright$  choose center of time sliding window
21:      $\mathbb{R}'_{\text{review}} \leftarrow \emptyset, VR' \leftarrow \emptyset$ 
22:     for  $R_j \in \mathbb{R}_{\text{review}}$  do  $\triangleright$  choose reviews of time sliding window based on center
23:       if  $t_{R_j} > t_{R_n} - \Delta \text{ and } t_{R_j} \leq t_{R_n} + \Delta \text{ and } k_j > 0$  then
24:          $\mathbb{R}'_{\text{review}} \leftarrow \mathbb{R}'_{\text{review}} \cup \{R_j\}$ 
25:          $VR \leftarrow VR \cup \{(x_j, k_j)\}$ 
26:       end if
27:     end for
28:      $EH_n^2 \leftarrow \text{IBPA}(\mathbb{R}'_{\text{review}}, VR', x_n, k_n)$ 
29:   end for
30:    $EH^2 \leftarrow EH^2 \cup \{EH_n^2\}$ 
31:  return  $EH^2$ 
32: end function

```

### 3.4. Calculating the ICH and ECH

The IH and EH are two different aspects of the helpfulness of reviews. Dynamic measurement and comparison of IH and EH helpfulness will help understand reviews' overall helpfulness and formulate appropriate helpfulness ranking. We can calculate the ICH and ECH and construct the ICH-ECH plot according to the obtained IH and EH.

ICH and ECH represent the comprehensive value of IH and EH, respectively. Let  $ICH_n$  and  $ECH_n$  denote the ICH and ECH of review  $R_n$  and they can be calculated by Eq. (3)–(4).

$$ICH_n = \sum_{i=1}^{14} W_i^I \overline{IH}_n^i \quad (3)$$

$$ECH_n = W_1^E \overline{EH}_n^1 + W_2^E (1 - \overline{EH}_n^2) \quad (4)$$

where  $\overline{IH}_n^i$  is the normalized  $IH^i$  of review  $R_n$ ;  $\overline{EH}_n^i$  is the normalized  $EH^i$  of review  $R_n$ ;  $W_i^I$  and  $W_i^E$  is the weights of  $\overline{IH}_n^i$  and  $\overline{EH}_n^i$ .

To avoid the subjectivity of manually assigning weights, we introduce a machine learning approach to calculate  $W_i^I$  ( $i = 1, 2, \dots, 14$ ). We select the reviews posted long enough on the first page of similar products as the sample set, take the number of votes ( $\overline{IH}_n^i$ ) in these samples as the dependent variable, and each measure ( $\overline{IH}_n^i$ ,  $i = 1, 2, \dots, 14$ ) as the independent variable. As the reviews on the first page are posted long enough to be considered thoroughly browsed, their votes can better reflect the customer's attitude.

However, the results are full of randomness due to the large sample size of training and random initial parameters. Moreover, since each review was posted at a different time, different weights should be assigned to them when used as training samples. To address these limitations, we introduce an ensemble neural network model (ENNM) [50] to calculate  $W_i^I$ . Neural network (NN) is an efficient prediction method [51,50], especially in the case of nonlinear, multicollinearity, and other complex relationships. Although the neural network's main task is to make predictions, it can also be used to determine the weight of various measures [51,50]. ENNM integrates multiple back propagation neural networks (BPNN) and comprehensively considers their training results. The structure of the improved ENNM is shown in Fig. 4.

To better solve our problems, we have made some improvements to

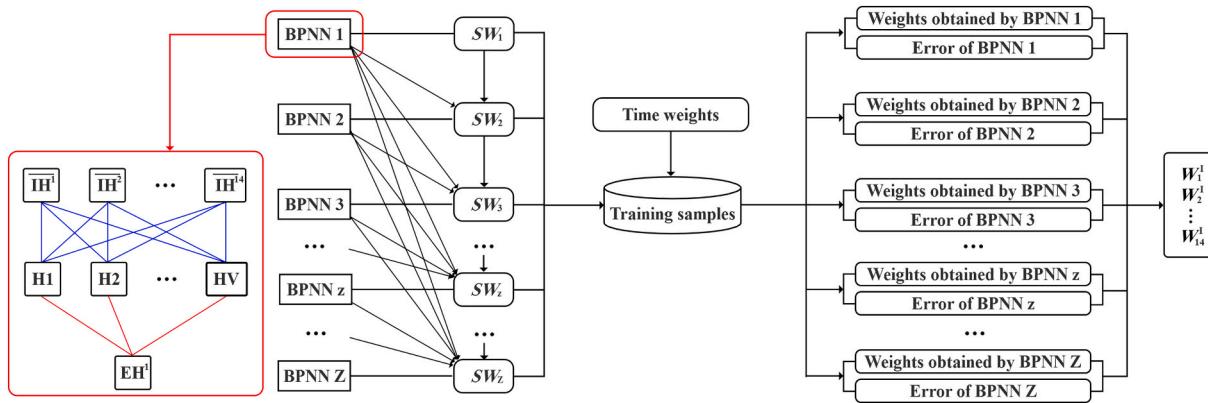


Fig. 4. The structure of the improved ENNM.

ENNM. The following three points reflect the improvements:

- (1)The method of an exponential moving average is used to update the weight of samples to make the weight update smoother.
- (2)When calculating the absolute error of each sample training result, time weight is introduced to realize the perception of time of training result.
- (3)When determining the parameter updating range, we refer to the absolute error of each sample training result to make the parameter updating method more flexible (the original method only sets a threshold value, and the weight of samples whose error exceeds the threshold value is updated according to fixed parameters).

Let  $w_v^z$  ( $v = 1, 2, \dots, V, z = 1, 2, \dots, Z$ ) denote the weight between output unit and the  $v$ th hidden unit in the  $z$ th BPNN. Let  $w_{hv}^z$  denote the weight between  $\text{IH}^i$  and the  $v$ th hidden unit in the  $z$ th BPNN. The weight of  $\text{IH}^i$  in  $z$ th BPNN can be calculated by Eq. (5).

$$MW_i^z = \sum_{v=1}^V w_{hv}^z w_v^z \quad (5)$$

Let  $NW_z$  denote the weight of  $z$  neural network and it can be calculated by Eq. (6). Let  $\overline{NW}_z$  denote the normalized value of  $NW_z$ . Then the weight of  $\text{IH}^i$  can be calculated by Eq. (7).

$$NW_z = e^{-\Xi_z} \quad (6)$$

$$W_i^1 = \sum_{z=1}^Z \overline{NW}_z \cdot MW_i^z \quad (7)$$

where  $\Xi_z$  is the absolute error in  $z$ th BPNN.

As the foregoing method uses the number of votes of fully browsed reviews as the dependent variable, we can only get the weight of each  $\text{IH}_n^i$ . For those reviews not fully viewed, it is difficult to measure the impact of the number of votes and the percentage of negative replies on the comprehensive helpfulness of the reviews. Therefore, we choose the time-weighted standard deviation to determine  $W_i^E$ . To make the total weights of ICH and ECH equal, we make  $\sum_{i=1}^{14} W_i^E = \sum_{i=1}^2 W_i^E$ . Based on the weights of IH and EH, the ICH and ECH can be obtained, and the ICH-ECH plot of each product can be constructed.

### 3.5. Ranking reviews by DTAHR model

Most research on review ranking is based on one aspect of IH and EH. These ranking strategies have the following three problems: (1) relying solely on IH, without sufficient reference to the group's opinions, may overlook fake and click farming behaviors; (2) relying solely on EH ranking may lead to ratchet or Matthew effect; and (3) few previous

studies have considered the effect of the release time of reviews on helpfulness. Both the product and people's perceptions change rapidly, and we should incorporate these changes into the model in time to dynamically rank the helpfulness of reviews.

Inspired by Wang et al. [45] considering the time dynamics of online reviews, we propose a DTAHR model in this study. The proposed model mitigated the ratchet and Matthew effects by recently rewarding reviews. Newly released reviews have more chances to rank in the front and thus have a chance to get a greater EH. To assign different weights to reviews posted at different times, we define a family of functions  $\mathcal{H} = \{g : N \rightarrow R | 0 \leq g(t) \leq 1, g'(t) \leq 0, t \geq 0\}$ , which represents the decay of information caused by the passage of time. To the best of our knowledge,  $g(t) = e^{-\beta t}$  is the most common function in family of functions  $\mathcal{H}$  [45]. The framework of DTAHR model is shown in Fig. 5. Fig. 5 shows that the model includes two phases, that is,

- Phase 1. Calculating the time-aware weight of reviews.
- Phase 2. Ranking reviews.

The detailed descriptions of the DTAHR model are illustrated as follows.

Let  $H_n(t)$  denote the comprehensive helpfulness of  $R_n$ , where  $t$  is the time since the review was posted.  $H_n(t)$  consists of two parts: ICH $_n$  and ECH $_n$ , and it can be obtained by Eq. (8).

$$H_n(t) = (1 + f_n(t)) \cdot \text{ICH}_n + \text{ECH}_n(t) \quad (8)$$

where  $f_n(t)$  is the reward function of  $R_n$ , which belongs to family  $\mathcal{H}$ , that is, when review  $R_n$  is first posted,  $f_n(t)$  takes the maximum value and then decreases over time. ECH $_n(t)$  denote that the value of ECH $_n$  would change over time (ECH $_n(0) = 0$ ), because after  $R_n$  is posted, it will constantly interact with customers, getting votes and replies.

We make the reward function  $f_n(t) = e^{-\lambda_n t}$ , where  $\lambda_n$  denote the intensity of reward of  $R_n$ . So the next step is to determine  $\lambda_n$ .

Let  $\mathcal{T}^R = \{T_1^R, T_2^R, \dots, T_N^R\}$  denote the time sliding window set of  $\mathbb{R}_{\text{review}}$ , where  $T_n^R$  denote the  $n$ th time sliding window centered on  $R_n$  and it can obtained by Eq. (9).

$$T_n^R = \{R_i | t_{R_n} - \Delta \leq t_{R_i} \leq t_{R_n} + \Delta\} \quad (9)$$

where  $\Delta$  is the radius of the time sliding window.  $T_n^R$  should be divided equally along the width of review posting time (i.e., fix the value of  $\Delta$ ) to ensure that each  $T_n^R$  contains the same time length. Therefore,  $T_n^R$  include all reviews posted within a fixed time length.

The time-aware intensity measurement could be derived through a comprehension of both macro and micro perspective. At the macro and micro level, the unit of analysis is the time window and reviews in it. Let  $t_{n_{\max}}$  denote the maximum time in time sliding window  $T_n^R$ ,  $t_{\text{now}}$  denote

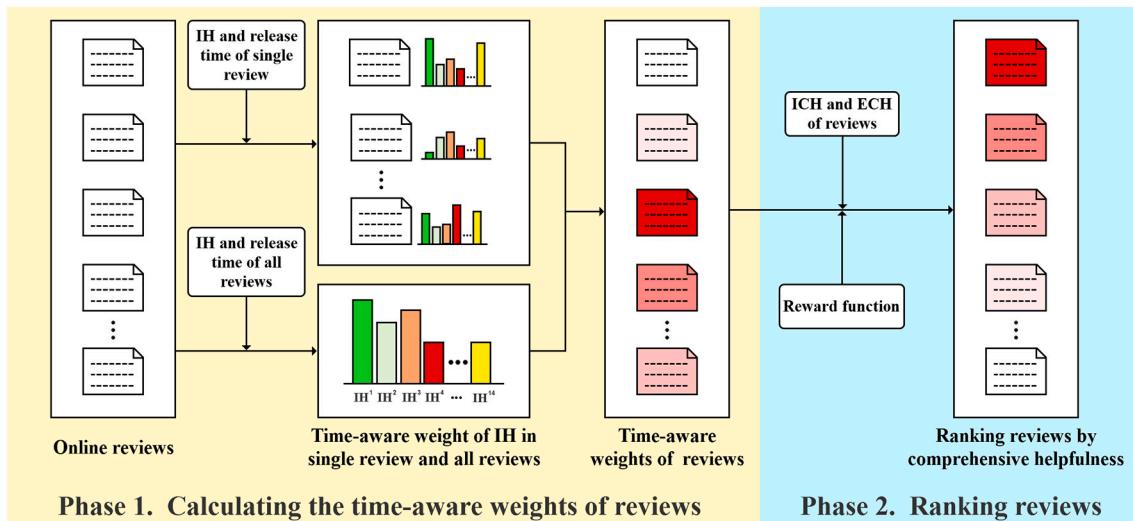


Fig. 5. The framework of DTAHR model.

the present time. The time-aware intensity of  $IH^i$  can be calculated as the sum product of the weight of and within the sliding time window  $T_n^R$ , as illustrated in Eq. (10).

$$TW^i = \sum_{T_n^R \in \mathcal{T}} \sum_{R_i \in T_n^R} h_2 \left( t_{\text{now}} - t_{n_{\max}} \right) \cdot h_3(t_{n_{\max}} - t_{R_i}) \cdot \overline{IH^i_n} \quad (10)$$

where function  $h_2$  indicates the decay of the whole time sliding window at the macro level and function  $h_3$  indicates the decay inside a time sliding window at the micro level ( $h_2$  and  $h_3$  belong to family  $\mathcal{H}$ ).

Let  $TW_n^i$  denote the time-aware intensity of  $R_n$  on  $IH^i$  and it can be calculated by Eq. (11).

$$TW_n^i = h_1(t_{\text{now}} - t_{R_i}) \cdot \overline{IH^i_n} \quad (11)$$

where function  $h_1$  indicates the decay of time weight of  $R_n$ .

The time dimension plays a vital role in providing consumers with consistent information, and the remote reviews might be out of date. Therefore,  $TW^i$  tend to reflect the recency of  $IH^i$  in all reviews and  $TW_n^i$  tend to reflect the recency of  $IH^i$  in  $R_n$ . According to  $TW^i$  and  $TW_n^i$ , the time-aware intensity of  $R_n$  can be obtained by Eq. (12).

$$TW_{R_n} = \sum_{i=1}^{14} TW^i TW_n^i \quad (12)$$

To normalize  $TW_{R_n}$ , we set  $TW = 14 \sum_{T_n^R \in \mathcal{T}} \sum_{R_i \in T_n^R} h_2(t_{\text{now}} - t_{n_{\max}}) \cdot h_3(t_{n_{\max}} - t_{R_i}) \cdot h_1(t_{\text{now}} - t_{R_i})$  as the upper bound of  $W_{R_n}$ . Therefore, the  $\lambda_n$  can be calculated by Eq. (13).

$$\lambda_n = \gamma \left( 1 - \frac{TW_{R_n}}{TW} \right) \quad (13)$$

where  $\gamma$  is the regulation factor used to adjust the range of  $\lambda_n$ .  $TW_{R_n}$  denotes the time-aware intensity of  $R_n$  and it represents the total score of all measures that people pay attention to in the recent period contained in  $R_n$ . Hence, if  $R_n$  has larger  $W_{R_n}$ , it should decay more slowly, that is,  $\lambda_n$  is smaller.

According to Eq. (8)–(13), the comprehensive helpfulness  $H_n(t)$  of review  $R_n$  at time  $t$  can be obtained. Thus, at time  $t$ , we can rank all reviews according to  $H_n(t)$ , and the ranking will change with time.

To verify the performance of the DTAHR model, we design a simulation experiment and develop some indicators to measure our ranking effect.

#### 4. Case study

This section selected several specific products from JD.com to illustrate how to construct an ICH-ECH plot and rank them by the DTAHR model. These selected products are divided into **search products** and **experience products**. The relevant data comes from JD.com (<https://www.jd.com>), one of China's largest e-commerce platforms. Next, we introduce the data used in the case study. It is then followed with experimental results.

##### 4.1. Data collection

In this case study, six products are chosen, including three search products: cell phone, laptop, and camera; three experience products: lipstick, running shoes, and rice cooker. A Python program is developed to collect online reviews posted on JD.com after 2018 automatically. The example of collected datasets is shown in Fig. 1, which includes the text content of reviews, as well as non-text content such as ratings, user image, and the number of pictures. By December 2021, we collected 25168 reviews and 24856 replies. After excluding some invalid data, 22908 reviews and 21456 replies were used for analysis. The relevant information of the collected reviews is shown in Table 3.

##### 4.2. Attribute words extraction

According to Section 3.1, we extracted the attribute words of reviews and replies of each product. As the extracted topics contain some noise words or have the same meaning, we need to filter the noise and manually adjust the synonyms. Table 4 shows the 10 words with the highest frequency in the reviews and replies of each product, in which bold indicates that the attribute words are from the reply, bracketed bold indicates that the attribute words are from the reviews, and replies and the remaining attribute words are from the reviews.

##### 4.3. Calculating the IH and EH of reviews

According to the attribute words of the reviews and replies, we can calculate the value of  $IH_n^i$  ( $i = 1, 2, \dots, 14, n = 1, 2, \dots, N$ ). We set the reward parameter to 0.5, that is,  $\rho = 0.5$ . The EH is not inherent when the review is released but an external recognition from customers and platforms. There are two types of EH: the number of votes ( $EH_n^1$ ) and the percentage of negative replies ( $EH_n^2$ ). When calculating  $EH_n^2$ , we set the time sliding window to 20 days, that is,  $\Delta = 10$ . According to the

**Table 3**

The relevant information of the collected reviews.

	Search product			Experience product			Total
	Cell phone	Laptop	Camera	Lipstick	Running shoes	Rice cooker	
Total number of reviews	4827	3947	3915	2767	4563	2889	22908
Total number of replies	7208	6473	4342	1064	1077	1292	21456

**Table 4**

Attribute words extraction.

Search product			Experience product		
Cell phone	Laptop	Camera	Lipstick	Running shoes	Rice cooker
Speed <b>(Screen)</b>	Appearance <b>(Keyboard)</b>	Appearance <b>(Camera lens)</b>	Colour	<b>(Sole)</b>	<b>(Function)</b>
Effect	Speed	Effect	<b>Fake</b>	<b>(Quality)</b>	Appearance <b>(Inner liner)</b>
Appearance <b>(Screen)</b>	Performance	Performance	Effect <b>(Quality goods)</b>	Comfort	Capacity
Sound effect <b>(Price)</b>	Effect	<b>Price</b>	Moist	Appearance	Texture
Standby time	Performance	Speed	Texture	Size	Time
<b>Charger</b>	Game	Service	<b>Special counter</b>	Air permeability	Heat preservation
Feel	After sale	<b>Machine</b>	Skin colour	<b>Customer service</b>	<b>(Rice)</b>
Signal	Customer service	Quality goods	<b>Customer service</b>	Appearance	Quality
	Power on	Photo	Official	Details	Guoba

aforementioned process, we can obtain the IH and EH of each review.

#### 4.4. Constructing the ICH-ECH plot

To construct the ICH-ECH plot, we could first obtain the weights of  $\overline{IH}^i, i = 1, 2, \dots, 14$ . We select the reviews posted long enough on the first page of products as the sample set, take the number of votes (EH<sup>1</sup>) in these samples as the dependent variable and each measure ( $\overline{IH}^i, i = 1, 2, \dots, 14$ ) as the independent variable. As the reviews posted long enough on the first page can be considered fully browsed, their votes can reflect customers' attitudes better.

According to Section 3.4, we can construct the ENNM to calculate the weights of  $\overline{IH}_n^i$ . Let  $z$  ( $z = 1, 2, \dots, Z$ ) denote the number of neural networks. In this study, we set the total number of neural networks  $Z = 500$  and the number of hidden layer neurons  $V = 128$ . The obtain value of  $W_i^j$  ( $i = 1, 2, \dots, 14$ ) change with  $z$  are shown in Fig. 6. Fig. 6 shows that with the increase of  $z$ , the value of  $W_i^j$  tends to be stable, which means that the randomness of initial parameters and the bias (such as noise and outliers) of a single NN are are eliminated partly. Then we can obtain

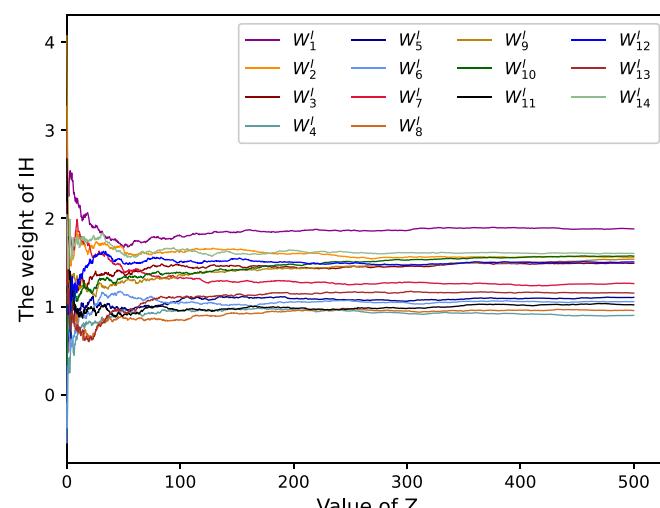


Fig. 6. Weights of  $\overline{IH}^i, i = 1, 2, \dots, 14$  change with the number of neural network.

that  $W_1^E = 11.59$  and  $W_2^E = 7.0023$ .

According to the weights obtained, we can use Eq. (3)–(4) to calculate the ICH and ECH of each review and then construct the ICH-ECH plot of each product. Fig. 7 shows the ICH-ECH plot of six products. As can be seen from Fig. 7, most of the points are concentrated in the lower part, that is, the ECH of most reviews is low. Our finding focuses mainly on the point in the lower right corner, that is, the reviews with high ICH and low ECH. As these reviews are considered to have great potential, especially when they have just been released. Due to the ratchet effect and Matthew effect, early reviews will get more votes and replies, making it difficult for customers to see late reviews. Therefore, greater weights should be added to those released reviews with high ICH. This study proposes a DTAHR model to calculate the comprehensive helpfulness of reviews and rank them dynamically.

#### 4.5. Ranking reviews by DTAHR model

We can calculate its ICH immediately for the review just released. Even so, the ECH is not available because the review's ECH is obtained by its interaction with customers, which takes time. If the reviews are only ranked according to ICH, ECH, or the sum of ICH and ECH, it will be unfavorable to the reviews just released and aggravate the ratchet effect and Matthew effect. The main idea of the DTAHR model is to give a weight reward to the review just released and determine the decay rate of the weight reward according to the time-aware weight of the review.

To verify the performance of our model, we design a simulation experiment. Refer to Wang et al. [45], when calculating the comprehensive helpfulness, we set the time sliding window radius  $\Delta = 10, h_1(t) = h_2(t) = e^{-0.05t}, h_3(t) = e^{-0.1t}, \gamma = 0.1$ . Refer to Wang et al. [11] and actual data of JD.com, we set  $\Delta T = 60s, \lambda_{\Delta T}^C = 0.05$  (i.e., an average of 72 customers arrive every day),  $m = 20, \varphi = 10, p = 0.1$  and  $k = 1$ .

Let  $NR_1$  and  $NR_2$  denote the number of initial reviews and added reviews in the simulation experiment. We selected 200 reviews from the mobile phone dataset (evenly distributed posting time). Then, we sorted these reviews according to the posting time and selected first 100 reviews as the initial reviews ( $NR_1 = 100$ ). According to the posting time, we added the remaining 100 reviews one by one ( $NR_2 = 100$ ). After adding each review, we will calculate the score and ranking according to different ranking strategies. At present, common review ranking strategies include ranking by votes (replaced by ECH in this paper), ranking by review content (replaced by ICH in this paper), and ranking by review release time. In addition, we also compared two other review ranking

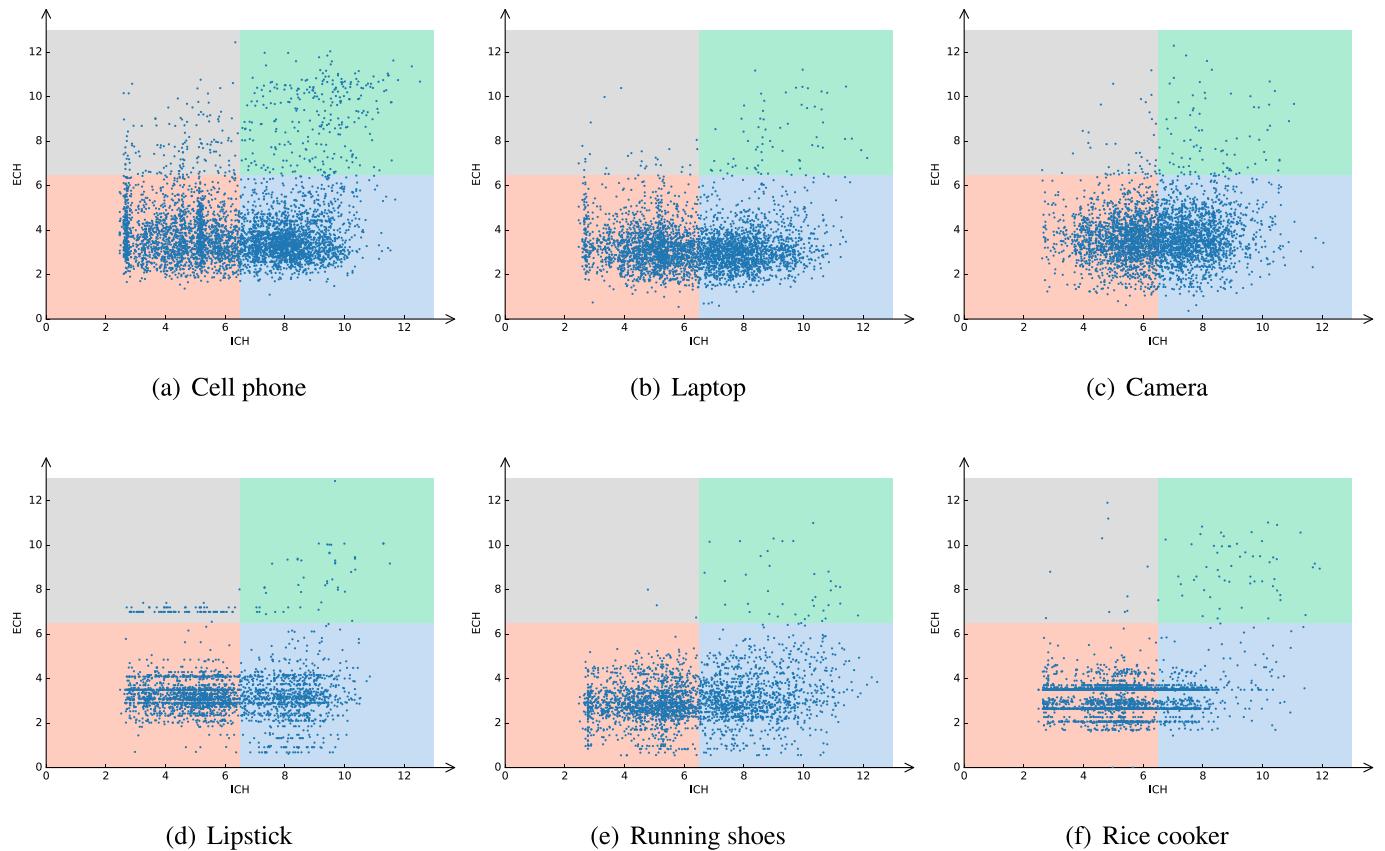


Fig. 7. ICH-ECH plot of each product.

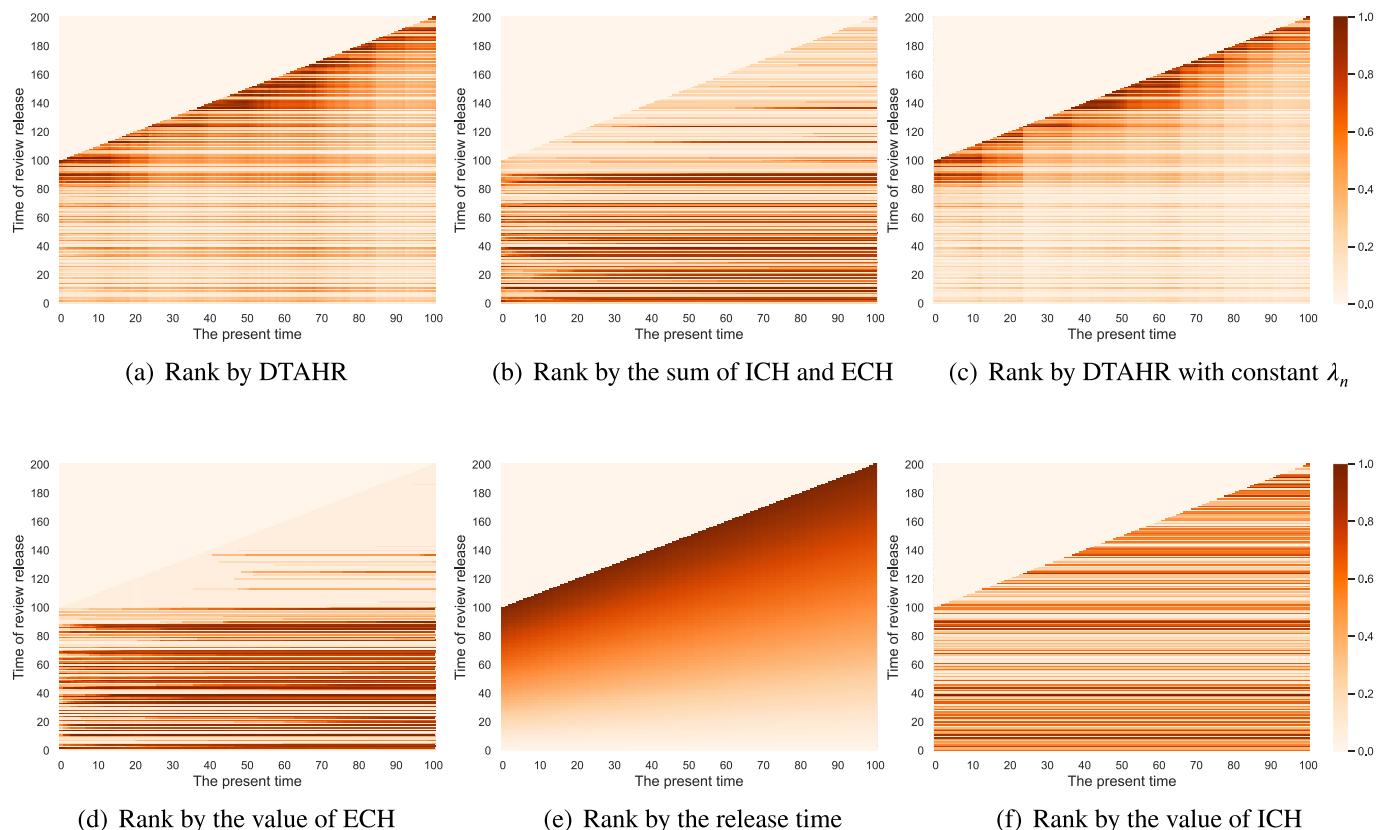


Fig. 8. The change of reviews ranking over time under different ranking strategies.

strategies: ranking by the sum of ICH and ECH and ranking by DTAHR with constant  $\lambda_n$ .

**Fig. 8** shows the dynamic change of comprehensive helpfulness  $H$  of each review under different ranking strategies (to compare the effects of different strategies, we normalized  $H$ ). In **Fig. 8**, the x-axis represents the current time, and the y-axis represents the release time of each review. Each column of data in the figure can be regarded as the distribution of  $H$  of all reviews at the current time.

It can be seen from **Fig. 8(b)**, when ranking reviews according to the sum of ICH and ECH, it is difficult for newly released reviews to rank at the top. In contrast, early posted reviews can always occupy the home page because of the ratchet effect and Matthew effect. **Fig. 8(d)** show that when reviews are ranked according to the value of ECH, it is more difficult for newly released reviews to rank at the top. Moreover, compared with **Fig. 8(b)**, the ratchet effect and Matthew effect is more serious. In **Fig. 8(f)**, when reviews are ranked according to the value of ICH, although the newly released reviews have the opportunity to rank at the top, it not only ignores the customer's opinions (which are reflected in ECH), but also leads to ratchet effect. In **Fig. 8(e)**, when reviews are ranked according to the release time, although the ratchet effect and Matthew effect is effectively solved and the newly released reviews are ranked at the top, those beneficial reviews cannot be found. In **Fig. 8(c)**, when reviews are ranked according to DTAHR with constant  $\lambda_n$ , the decay rate of the initial reward of each review is the same. Although this can also alleviate the ratchet effect and Matthew effect, it cannot perceive the time weight of the review. Moreover, as shown in **Fig. 8(a)**, when DTAHR ranks reviews, it can not only effectively alleviate the ratchet effect and Matthew effect (give the newly released reviews a chance to rank at top), but also sense the time weight of reviews. Most darker horizontal lines in **Fig. 8(a)** and (c) can correspond to **Fig. 8(b)** and (f), indicating that DTAHR can find useful reviews.

To more specifically illustrate the effects of these six ranking strategies, we selected six evaluation criteria to measure their ranking results: the percentage of reviews with 0 votes ( $EC_1$ ), the percentage of replies with 0 replies ( $EC_2$ ), the average time for new reviews to obtain one vote ( $EC_3$ ), the average time for new reviews to obtain 10 votes ( $EC_4$ ), the number of changes of reviews in the top 20 ( $EC_5$ ), and the variance of voting distribution ( $EC_6$ ). **Table 5** shows the evaluation criteria values of each ranking strategy. There are four kinds of basic information, the number of reviews ( $BI_1$ ), the number of votes ( $BI_2$ ), the number of replies ( $BI_3$ ), and the number of negative replies ( $BI_4$ ).

Consistent with our expectations, the strategy ranking by release time performs better on  $EC_1, EC_2, EC_3, EC_4$ , but performs poorly on  $EC_5, EC_6$  (The more moderate the value of  $EC_5, EC_6$ , the better the performance of the strategy). This means that strategy ranking by release time can only solve the ratchet effect and Matthew effect but cannot find helpful reviews. In the remaining five strategies, the effects of ranking by DTAHR and ranking by DTAHR with constant  $\lambda_n$  are significantly better

than other strategies, but DTAHR with constant  $\lambda_n$  cannot perceive the time weight of the review.

Let  $n$  and  $m$  denote the total numbers of reviews and the size of time window, respectively. The time complexity of DTAHR model is  $O(14mn + n + n\log(n))$ , where  $O(14mn)$  and  $O(n)$  can be obtained by **Eqs. (10)** and **(11)**,  $O(n\log(n))$  is the time complexity of ranking  $H$  of each review. The remaining five review ranking strategies only involve ranking operations, so their time complexity is  $O(n\log(n))$ . Therefore, the time complexity of DTAHR model is slightly higher than other strategies. But if  $m \ll n$  (i.e., there are enough reviews), the disadvantage of DTAHR model on time complexity can be ignored.

#### 4.6. Robustness of DTAHR model

The proposed DTAHR model, which ranks reviews based on comprehensive helpfulness within the same product category, is significantly better than other strategies in the simulation experiment. Therefore, to examine the robustness of the proposed model, it is worthwhile to discuss further the impacts of  $NR_1$  and  $NR_2$  on calculating the comprehensive helpfulness of reviews and ranking reviews.

In the previously presented simulation experiments, the number of initial reviews and added reviews are set to  $NR_1 = 100$  and  $NR_2 = 100$ . To investigate the impacts of  $NR_1$  and  $NR_2$ , we further study the cases where  $NR_1$  is set as 0, 100, 200 and  $NR_2$  is set as 100, 200, 300, respectively, and examine the performance of the DTAHR model, in comparison with other ranking strategies. To compare the effects of different ranking strategies, experiments were also conducted to compare the evaluation criteria of each ranking strategy in various combinations of  $NR_1$  and  $NR_2$ . The results are shown in **Fig. 9**, where the x-axis represents the combinations of  $NR_1$  and  $NR_2$ , the y-axis shows the value of evaluation criteria. According to the result, the performance of each ranking strategy is consistent with the conclusion of **Section 4.5**: the effectiveness of the DTAHR model is significantly better than other strategies.

## 5. Discussion and implications

To further analyze the advantages and disadvantages of the proposed model, further discussions are carried out according to two aspects: (1) implications for research and (2) implications for practice.

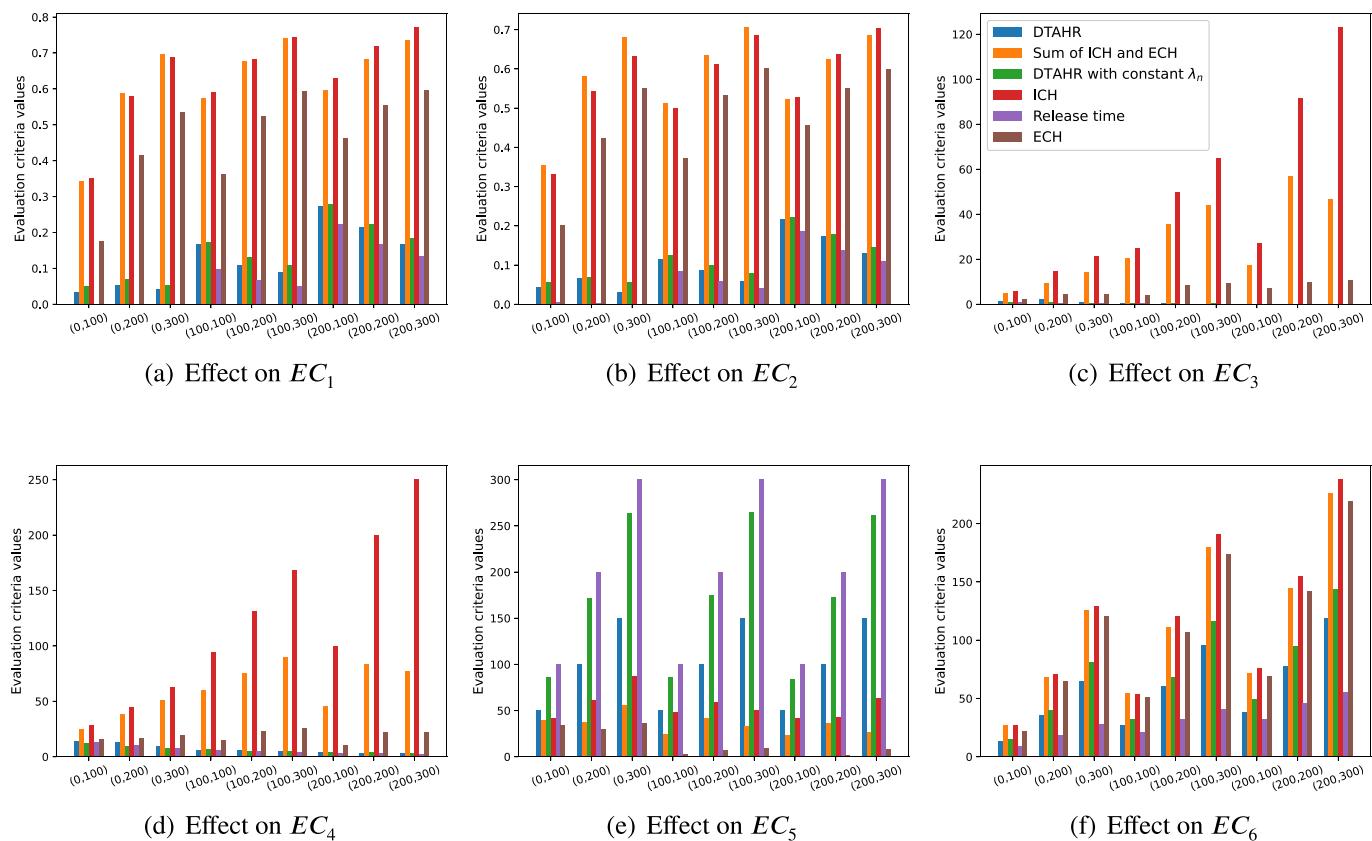
### 5.1. Implications for research

Nowadays, there is a growing literature on studying online reviews helpfulness from different perspectives, such as text mining [9,24–26], the number of votes [18], the ratio of helpful votes [30]. This study explores the helpfulness composition of online reviews from a new perspective. We divide online reviews helpfulness into IH and EH,

**Table 5**  
Evaluation criteria values of each ranking strategy.

Ranking strategy	Basic information				Evaluation criteria					
	$BI_1$	$BI_2$	$BI_3$	$BI_4$	$EC_1$	$EC_2$	$EC_3$	$EC_4$	$EC_5$	$EC_6$
Sum of ICH and ECH	200	3018.54 (15.9)	12354.05 (121.7)	6694.26 (79.7)	0.58 (0.0116)	0.52 (0.0125)	16.14 (4.62)	46.42 (11.77)	18.54 (4.41)	33.19 (0.38)
Release time	200	3020.46 (15.7)	12321.83 (120.1)	7823.66 (89.5)	0.11 (0.0086)	0.09 (0.0109)	0.61 (0.11)	10.15 (0.54)	100.00 (0)	12.75 (0.21)
ICH	200	3020.82 (16.2)	10975.47 (123.4)	5143.56 (70.3)	0.39 (0.0132)	0.41 (0.0124)	5.34 (1.17)	17.79 (1.95)	5.32 (0.89)	30.45 (0.37)
ECH	200	3020.50 (17.3)	10966.25 (141.0)	5411.21 (95.2)	0.61 (0.0117)	0.52 (0.0139)	25.63 (7.79)	62.71 (10.67)	36.15 (8.80)	35.15 (0.48)
DTAHR with constant $\lambda_n$	200	3017.15 (16.1)	10966.35 (150.3)	5595.66 (93.8)	0.30 (0.0137)	0.23 (0.0139)	1.24 (0.27)	10.29 (0.5918)	71.95 (2.57)	17.63 (0.2998)
DTAHR	200	3016.96 (15.2)	12316.26 (131.1)	5035.56 (91.2)	0.29 (0.0126)	0.23 (0.0131)	1.36 (0.29)	10.86 (0.76)	53.71 (3.40)	20.58 (0.38)

**Note:** The data in brackets are standard deviation.



**Fig. 9.** The effects of ranking strategies in different combinations of  $NR_1$  and  $NR_2$ .

representing two important streams of online reviews research. We analyze the measures of IH and EH, quantified them using ensemble neural network and Bayesian inference, and finally obtain an ICH-ECH plot. Through the ICH-ECH plot, we can obtain a general understanding of the IH and EH distribution of a product in a short time to assist decision-making. Our research findings provide a novel approach for the research of online review helpfulness.

In addition, we propose a DTAHR model that ranks reviews dynamically. The ranking of online reviews is drawing an increasing attention in recent studies [5,19,11,45]. Due to the sharp increase in the number of online reviews, a good ranking mechanism is important to alleviate customers' pressure of reading abundant reviews. Previous studies have found that time has a significant impact on the helpfulness of online reviews [52,45]. By introducing the concept of time-aware weight, this paper rewards the helpfulness of reviews posted recently. With time, the reward weight decreases to alleviate the ratchet effect and Matthew effect. The DTAHR model proposed in this study can effectively identify real-time and helpful reviews and is applicable to e-commerce websites with fast update and large number of reviews (such as JD.com and Amazon.com). Furthermore, the DTAHR model may also screen out those fake reviews, even if the fake reviews are challenging to be identified by e-commerce platforms. On account of these fake reviews, with a large ICH, customers can identify the fake information contained therein to reduce the misleading helpfulness of fake reviews. Therefore, the proposed DTAHR model can also provide a possible direction for detecting fake reviews.

## 5.2. Implications for practice

Our research findings also offer plenty of implications for business. Through the ICH-ECH plot, managers can understand the helpfulness distribution of all reviews on a product and make corresponding decisions. In addition, when using the DTAHR model to rank reviews,

those reviews posted for a long time and have high ICH and low ECH can be regarded as potential fake reviews because they carry a lot of information that customers do not recognize. Managers can collect and extract features for more accurate fake review filtering for these reviews.

Furthermore, in this study, ENNM is used to estimate the weight of each IH. Compared with a single neural network, the advantage is that the results obtained are more stable and accurate. At the same time, the weight of each IH can also give managers some enlightenment to assist them in making decisions and improving WOM. For example, as shown in Fig. 6, the three IH with the highest weight are:  $IH^1$ ,  $IH^{14}$  and  $IH^{10}$ . The three IH with the lowest weight is:  $IH^5$ ,  $IH^8$  and  $IH^{11}$ . Platforms and e-tailers can get the following information: (1) the information consistency of review, user image, and the number of pictures greatly impact the helpfulness of reviews. In the future, customers should be encouraged to post more information and pictures and set user images; and (2) the number, percentage, and score of neutral sentences have little effect on the helpfulness of reviews.

## 6. Conclusions, limitations and future research

Customer replies to online reviews reflect customers' concerns about the product, which should be considered when calculating the helpfulness of reviews. Most studies ignore the role of replies, which contain helpful information, such as sentiment orientations, product quality, price. Based on the prior work, this study introduces the indicator of the percentage of negative replies to measure the helpfulness of reviews. IH and EH are two different aspects of review helpfulness. Most studies have considered only one aspect rather than a comprehensive perspective that contains both aspects. Thus, after adding the information of reply, this study proposes a method to measure the review helpfulness from both intrinsic and extrinsic aspects and calculates the ICH and ECH by using the various measures of EH and IH, thus constructing the ICH-

ECH plot. Finally, we propose a new ranking strategy that changes dynamically over time.

This study can be extended in several ways. For instance, our data crawls from a single website (JD.com), and whether these conclusions apply to other e-commerce platforms needs to be further verified. In addition, this study only expands a special form of online review: reply. More review forms, such as merchant reply and follow-up review, can be considered in the future. Furthermore, because the weight of EH is challenging to determine, we use the dispersion of data as their weight. Other weighting mechanisms can be explored in the future. Finally, although our analysis uses the review and reply data posted on JD.com, the evaluation of the proposed DTAHR model is based on numerical experiments. One future research avenue is to implement the proposed approach and evaluate its performance. Of course, the implementation will require cooperation with the merchants and e-commerce platform, but will certainly generate valuable insights.

### CRediT authorship contribution statement

**Jindong Qin:** Conceptualization, Methodology, Writing-original-draft, Writing-review-editing, Funding-acquisition, Supervision. **Pan Zheng:** Conceptualization, Methodology, Data-curation, Writing-original-draft. **Xiaojun Wang:** Writing-review-editing, Supervision.

### Data availability

No data was used for the research described in the article.

### Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Projects 71701158 and 72071151, MOE (Ministry of Education in China) Project of Humanities and Social Sciences (17YJC630114), and the Natural Science Foundation of Hubei Province (2020CFB773).

### References

- [1] X. Yang, G.F. Yang, J.N. Wu, Y.Z. Dang, W.G. Fan, Modeling relationships between retail prices and consumer reviews: A machine discovery approach and comprehensive evaluations, *Decis. Support Syst.* 145 (2021).
- [2] K.K.Y. Kuan, K.L. Hui, P. Prasarnphanich, H.Y. Lai, What makes a review voted? An empirical investigation of review voting in online review systems, *J. Assoc. Inf. Syst.* 16 (1) (2015) 48–71.
- [3] Q. Jones, G. Ravid, S. Rafaeli, Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration, *Inf. Syst. Res.* 15 (2) (2004) 194–210.
- [4] D.Z. Yin, S.D. Bond, H. Zhang, Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews, *MIS Quart.* 38 (2) (2014) 539–560.
- [5] Z.Q. Zhang, G.Q. Chen, J. Zhang, X.H. Guo, Q. Wei, Providing consistent opinions from online reviews: a heuristic stepwise optimization approach, *Informs J. Comput.* 28 (2) (2016) 236–250.
- [6] J.N. Wu, Review popularity and review helpfulness: a model for user review effectiveness, *Decis. Support Syst.* 97 (2017) 92–103.
- [7] D.Z. Yin, S. Mitra, H. Zhang, When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth, *Inf. Syst. Res.* 27 (1) (2016) 131–144.
- [8] H. Baek, J. Ahn, Y. Choi, Helpfulness of online consumer reviews: readers' objectives and review cues, *Int. J. Electron. Commer.* 17 (2) (2012) 99–126.
- [9] N. Korfiatis, E. Garcia-Barriocanal, S. Sanchez-Alonso, Evaluating content quality and helpfulness of online product reviews: the interplay of review helpfulness vs. review content, *Electron. Commer. Res. Appl.* 11 (3) (2012) 205–217.
- [10] A. Ghose, P.G. Ipeirotis, Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics, *IEEE Trans. Knowl. Data Eng.* 23 (10) (2011) 1498–1512.
- [11] J.N. Wang, J.Z. Du, Y.L. Chiu, Can online user reviews be more helpful? Evaluating and improving ranking approaches, *Inf. Manage.* 57 (8) (2020).
- [12] D. Weathers, S.D. Swain, V. Grover, Can online product reviews be more helpful? Examining characteristics of information content by product type, *Decis. Support Syst.* 79 (2015) 12–23.
- [13] Q.D. Li, D.D.J. Zeng, D.J.J. Xu, R.R. Liu, R.H. Yao, Understanding and predicting users' rating behavior: a cognitive perspective, *Informs J. Comput.* 32 (4) (2020) 996–1011.
- [14] A. Qazi, K.B.S. Syed, R.G. Raj, E. Cambria, M. Tahir, D. Alghazzawi, A concept-level approach to the analysis of online review helpfulness, *Comput. Hum. Behav.* 58 (2016) 75–81.
- [15] S. Krishnamoorthy, Linguistic features for review helpfulness prediction, *Expert Syst. Appl.* 42 (7) (2015) 3751–3759.
- [16] M. Salehan, D.J. Kim, Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics, *Decis. Support Syst.* 81 (2016) 30–40.
- [17] J.E. Fresnedo, D. Gefen, A semantic measure of online review helpfulness and the importance of message entropy, *Decis. Support Syst.* 125 (2019).
- [18] X.Y. Sun, M.X. Han, J. Feng, Helpfulness of online reviews: examining review informativeness and classification thresholds by search products and experience products, *Decis. Support Syst.* 124 (2019).
- [19] J. Zhang, C. Wang, G.Q. Chen, A review selection method for finding an informative subset from online reviews, *Informs J. Comput.* 33 (1) (2021) 280–299.
- [20] N.F. Ibrahim, X.J. Wang, A text analytics approach for online retailing service improvement: evidence from twitter, *Decis. Support Syst.* 121 (2019) 37–50.
- [21] N.F. Ibrahim, X.J. Wang, Decoding the sentiment dynamics of online retailing customers: time series analysis of social media, *Comput. Hum. Behav.* 96 (2019) 32–45.
- [22] C.X. Jiang, J. Zhu, Q.F. Xu, Dissecting click farming on the Taobao platform in China via PU learning and weighted logistic regression, *Electron. Commer. Res.* 22 (1) (2022) 157–176.
- [23] A.Y.K. Chua, S. Banerjee, Analyzing review efficacy on amazon.com: does the rich grow richer?, *Comput. Hum. Behav.* 75 (2017) 501–509.
- [24] Y. Pan, J.Q. Zhang, Born unequal: a study of the helpfulness of user-generated product reviews, *J. Retail.* 87 (4) (2011) 598–612.
- [25] H. Hong, D. Xu, G.A. Wang, W.G. Fan, Understanding the determinants of online review helpfulness: a meta-analytic investigation, *Decis. Support Syst.* 102 (2017) 1–11.
- [26] Y.B. Zhao, X. Xu, M.S. Wang, Predicting overall customer satisfaction: big data evidence from hotel online textual reviews, *Int. J. Hosp. Manag.* 76 (2019) 111–121.
- [27] S.W. Park, J.L. Nicolau, Asymmetric effects of online consumer reviews, *Ann. Tour. Res.* 50 (2015) 67–83.
- [28] X. Xu, Examining consumer emotion and behavior in online reviews of hotels when expecting managerial response, *Int. J. Hosp. Manag.* 89 (2020).
- [29] G. Cui, H.K. Lui, X.N. Guo, The effect of online consumer reviews on new product sales, *Int. J. Electron. Commer.* 17 (1) (2012) 39–57.
- [30] X.H. Guo, G.Q. Chen, C. Wang, Q. Wei, Z.Q. Zhang, Calibration of voting-based helpfulness measurement for online reviews: an iterative bayesian probability approach, *Informs J. Comput.* 33 (1) (2021) 246–261.
- [31] D. Schuff, S. Mudambi, What makes a helpful online review? A study of customer reviews on amazon.com, *MIS Quart.* (2010).
- [32] J.H. Du, J. Rong, H. Wang, Y. Zhang, Neighbor-aware review helpfulness prediction, *Decis. Support Syst.* 148 (2021).
- [33] S.S. Zhou, B. Guo, The order effect on online review helpfulness: a social influence perspective, *Decis. Support Syst.* 93 (2017) 77–87.
- [34] S. Lee, J.Y. Choeh, Predicting the helpfulness of online reviews using multilayer perceptron neural networks, *Expert Syst. Appl.* 41 (6) (2014) 3041–3046.
- [35] X.L. Zheng, S. Zhu, Z.X. Lin, Capturing the essence of word-of-mouth for social commerce: assessing the quality of online e-commerce reviews by a semi-supervised approach, *Decis. Support Syst.* 56 (2013) 211–222.
- [36] Jian-Wu Bi, Yang Liu, Zhi-Ping Fan, Jin Zhang, Wisdom of crowds: conducting importance-performance analysis (IPA) through online reviews, *Tour. Manag.* 70 (2019) 460–478.
- [37] Y.J. Park, Predicting the helpfulness of online customer reviews across different product types, *Sustainability* 10 (6) (2018).
- [38] J. Liu, Y. Cao, C.Y. Lin, Y. Huang, Z. Ming, Low-quality product review detection in opinion summarization, in: Proc Joint Conference on Empirical Methods in Natural Language Processing & Computational Natural Language Learning, 2007.
- [39] M. Siering, J. Muntermann, B. Rajagopalan, Explaining and predicting online review helpfulness: the role of content and reviewer-related signals, *Decis. Support Syst.* 108 (2018) 1–12.
- [40] Y.H. Hu, K.C. Chen, Predicting hotel review helpfulness: the impact of review visibility, and interaction between hotel stars and review ratings, *Int. J. Inf. Manag.* 36 (6) (2016) 929–944.
- [41] X.H. Yu, Y. Liu, J.X. Huang, A.J. An, Mining online reviews for predicting sales performance: a case study in the movie domain, *IEEE Trans. Knowl. Data Eng.* 24 (4) (2012) 720–734.
- [42] A. Dash, D.S. Zhang, L.N. Zhou, Personalized ranking of online reviews based on consumer preferences in product features, *Int. J. Electron. Commer.* 25 (1) (2021) 29–50.
- [43] M.P. O'Mahony, B. Smyth, A classification-based review recommender, *Knowl.-Based Syst.* 23 (4) (2010) 323–329.
- [44] G. Shmueli, O.R. Koppius, Predictive analytics in information systems research, *MIS Quart.* 35 (3) (2011) 553–572.
- [45] C. Wang, G.Q. Chen, Q. Wei, A temporal consistency method for online review ranking, *Knowl.-Based Syst.* 143 (2018) 259–270.
- [46] S. Tirunillai, G.J. Tellis, Mining marketing meaning from online chatter: strategic brand analysis of big data using latent dirichlet allocation, *J. Mark. Res.* 51 (4) (2014) 463–479.
- [47] Y. Guo, S.J. Barnes, Q. Jia, Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation, *Tour. Manag.* 59 (2017) 467–483.

- [48] S. Tirunillai, G.J. Tellis, Mining marketing meaning from online chatter: strategic brand analysis of big data using latent dirichlet allocation, *J. Mark. Res.* 51 (4) (2014) 463–479.
- [49] T.L. Griffiths, M. Steyvers, Finding scientific topics, *Proc. Natl. Acad. Sci. USA* 101 (2004) 5228–5235.
- [50] J.W. Bi, Y. Liu, Z.P. Fan, E. Cambria, Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model, *Int. J. Prod. Res.* 57 (22) (2019) 7068–7088.
- [51] P. Phillips, K. Zigan, M.M.S. Silva, R. Schegg, The interactive effects of online reviews on the determinants of swiss hotel performance: a neural network analysis, *Tour. Manag.* 50 (2015) 130–141.
- [52] R. Filieri, C.F. Hofacker, S. Alguezaui, What makes information in online consumer reviews diagnostic over time? The role of review relevancy, factuality, currency, source credibility and ranking score, *Comput. Hum. Behav.* 80 (2018) 122–131.

**Jindong Qin** is currently an Associate Professor with the School of Management, Wuhan University of Technology, where he serves as the Director of the Research Center for Data Science and Intelligent Decision Making. He has published more than 60 papers in journals

indexed by SCI/SSCI, such as IIE Transactions, European Journal of Operational Research, Journal of the Operational Research Society, Annals of Operations Research, Information Sciences, Knowledge-Based Systems, and Applied Soft Computing. His current research interests include machine learning and decision analysis.

**Pan Zheng** is a master candidate of School of Management, Wuhan University of Technology. His research interest includes data mining and machine learning.

**Xiaojun Wang** is a Professor of Operations Management at the School of Management, University of Bristol. His current research predominantly focuses on supply chain risk and resilience, low carbon manufacturing, eco-design, sustainability, and social media research. His research outputs have been published in many international journals, including Production and Operations Management, European Journal of Operational Research, British Journal of Management, Computers in Human Behaviour, Omega, International Journal of Production Economics, International Journal of Production Research, and Journal of the Operational Research Society.