Are Review Helpfulness Score and Review Unhelpfulness Score Two Sides of The Same Coin or Different Coins?

Warut Khern-am-nuai
McGill University
warut.khern-am-nuai@mcgill.ca

Yinan Yu
University of Memphis
yyu4@memphis.edu

Abstract

Online review platforms have increasingly incorporated the review evaluating system (i.e., a system that allows users to evaluate whether reviews are helpful/unhelpful) to assist review readers and encourage review contributors. However, although we have extensive knowledge about the review helpfulness score, our insights regarding its counterpart, the review unhelpfulness score, are lacking. Addressing this limitation is important because many researchers have adopted the review unhelpfulness score assuming that it is driven by intrinsic review characteristics while practitioners also implicitly assume that the unhelpfulness score can identify low-quality reviews. The primary objective of this work is to verify whether the review unhelpfulness score is influenced by intrinsic review characteristics that drive review helpfulness score. We find that unlike review helpfulness score, unhelpfulness score is not driven by intrinsic review characteristics, and that helpfulness voters behave significantly different than unhelpfulness voters. Further implications and future directions are also discussed.

1. Introduction

Over the years, the online review system has become an important source of information for consumers [1]. As such, it has significant impacts on consumers' decision-making process [2] and product sales [3]. However, an issue that has increasingly become apparent in recent years, especially among large online review platforms, is an issue of information overload where review readers are unable to process the sheer amount of reviews available [4, 5]. Although several review platforms have explored advance technologies such as review recommendation systems [e.g., 6, 7] to alleviate such a problem, a common approach that most review platforms employ is to leverage peer evaluations in the form of review evaluation system [8]. In this system, review readers can vote for reviews that they deem helpful (or unhelpful). Then, the platform uses this helpfulness/unhelpfulness scores to prioritize or filter out certain set of reviews to reduce the effort and time required for users to find good reviews to read. In this regard, prior academic literature has extensively studied the review helpfulness score, ranging from the source of the review helpfulness [9] to its implications [10]. However, the review unhelpfulness score, which is also widely employed together with the review helpfulness score on most online review platforms (e.g., Target.com, Macys.com), receives much less attention in the academic literature. More importantly, even with limited understandings of how the review unhelpfulness score actually works, many studies on online reviews [e.g., 10, 11] consistently used it as a measure or construct based on an underlying assumption that the review unhelpfulness score and the review helpfulness score are two sides of the same coin (i.e., the existing knowledge on the review helpfulness score can be directly applied to the review unhelpfulness score on an opposite direction). This short paper aims to formally examine such a conventional wisdom. Particularly, we propose the following research questions: *1) Is the review unhelpfulness score driven by reviews' intrinsic characteristics? 2) How do unhelpfulness voters and helpfulness voters behave differently?*

We aim to investigate the proposed research question in two directions. The first direction is related to the relationship between the review helpfulness/unhelpfulness score and the intrinsic characteristics of the reviews. In that regard, Mudambi and Schuff [9] have empirically demonstrated that intrinsic review characteristics are the primary factor that influences the review helpfulness score. Particularly, the length of the reviews positively affects the review helpfulness score. Meanwhile, reviews with extreme ratings (i.e., 1-star or 5-star) are more helpful than reviews with moderate ratings (i.e., 2-, 3-, and 4-star) for experience goods. Several follow-up studies have also shown consistent results in this regard. For example, Pan and Zhang [12] have shown that the length of the reviews is positively correlated with the review helpfulness score.

URI: https://hdl.handle.net/10125/71403
978-0-9981331-4-0

Meanwhile, there is a scant set of evidence which suggests that the unhelpfulness score may also be driven by intrinsic review characteristics. For instance, one of the most common reasons for users to rate reviews as unhelpful is "lack of information" [13]. Hence, it is important to investigate if the review unhelpfulness score is driven by intrinsic review characteristics as in the case of the review helpfulness score. Specifically, we examine two types of intrinsic review characteristics. The first type of characteristics is quantitative review measurements, such as review rating, length, and the number of photos attached, etc. The second type of characteristics is textual features, such as topics distribution, sentiment, and readability, which are obtained using text mining techniques.

The second direction is regarding the characteristics of the voters. Although several prior works have empirically examine the voting behavior in the case of the review helpfulness score [e.g., 8], there is virtually no evidence of the profile of those who cast the review unhelpfulness votes, even though their behavior could profoundly impact the trustworthiness of the review unhelpfulness score. As such, it is important to study whether the voters behave similarly between those casting review helpfulness score and those casting review unhelpfulness score. Particularly, we examine whether helpfulness voters and unhelpfulness voters are similar in terms of their involvement in the platform, and the diversity of their votes.

To operationalize our research agenda, we collaborate with a large restaurant review platform in Asia to obtain a rich dataset. Interestingly, our analyses demonstrate that review helpfulness scores are significantly different than review unhelpfulness scores in multiple aspects. The helpfulness score appears to be driven by both intrinsic quantitative review characteristics and review textual features, while it is not the case for the unhelpfulness score. In addition, helpfulness voters are much more involved with the platform than unhelpfulness voters are, although the helpfulness scores are much less diverse than the unhelpfulness scores are. Lastly, unhelpfulness votes are more evenly submitted by voters, while helpfulness votes are more concentrated (i.e., most are casted by a small group of voters).

2. Literature review

In this section, we discuss a review of prior literature that is closely related to this paper. Particularly, we survey prior works in multiple discipline including information systems, marketing, and computer science that specifically studied review evaluation, review helpfulness scores, and review unhelpfulness scores.

2.1. Review evaluation

Prior literature in review evaluation has studied this process from several perspectives. Most of the papers in this area study the intrinsic characteristics that contribute for review helpfulness votes and find that several review characteristics indeed have a significant impact. For example, Mudambi and Schuff [9] demonstrate that review length and review valence have significant impact on the helpfulness scores that the review attains while the product type (search goods vs. experience goods) moderates such an impact. Relatedly, Wu, et al. [14] show that review valence, review length, and review readability are the characteristics that drive review helpfulness scores. Furthermore, Eslami, et al. [15] utilize multiple research methodologies, including sentiment analysis, PLS-SEM, ANOVA, and Artificial Neural Networks, to show that review length is the review characteristic that has the most influence on review helpfulness votes.

This stream of literature is also closely connected to another stream where the focus is to predict the helpfulness of the reviews based on review characteristics (rather than establishing correlation or causal inferences between review characteristics and review helpfulness scores). For instance, Kim, et al. [16] develop a model based on the Support Vector Machine (SVM) regression to predict the helpfulness scores of reviews on Amazon based on variety of features. Their model shows promising results with the rank correlations of up to 0.66. In the same way, Liu, et al. [17] develop a non-linear regression model based on several factors to predict the review helpfulness votes. They test the model with the data from IMDB movie reviews and shows that their approach is highly effective in predicting the helpfulness votes. In the meantime, Xiong and Litman [18] attempt to extend the predictive models developed to predict the helpfulness votes of product reviews to the context of peer reviews. They develop a model using the Support Vector Machine (SVM) regression with the radial basis function (RBF) kernel. They conduct an experiment to show that their model performs well when predicting the helpfulness votes in a peer-review system, using the peer-review corpus which was collected from web-based peer-review system in an introductory history class at the undergraduate level.

2.2. Review helpfulness score

Apart from the stream of literature that studies the connection between intrinsic review characteristics and review helpfulness scores, there is another related stream that studies how the helpfulness votes of the reviews are driven by external factors. Connors, et al. [13] conduct a controlled experiment and show that reviews written by a self-identified expert are voted as more helpful than those that are not even though the content is the same. Relatedly, Ngo-Ye and Sinha [19] propose a hybrid text regression model based on reviews' characteristics, particularly on the RFM (Recency, Frequency, Monetary Value) dimensions to predict the review helpfulness votes. Also, Baek, et al. [20] use the data from Amazon to show that reviewers' credibility is an important factor that leads to reviews obtaining helpfulness votes. In addition to the reviewer characteristics, there is also a substream of prior works that focuses on identifying other external factors that may impact review helpfulness scores. For example, Yu, et al. [8] utilize the data from a large restaurant review platform and multiple econometric methods to empirically demonstrate that review helpfulness scores are driven by the level of social interactions of review writers prior to the review submission.

Different from the previous two streams of research mentioned above, another perspective that prior studies have investigated regarding review helpfulness scores is on the usage side of the scores. For instance, Ghose and Ipeirotis [21] show a connection between review helpfulness scores and the impact of online reviews on product sales. Similarly, Chen, et al. [10] show an evidence that reviews with more helpfulness votes are more likely to impact consumers' purchase decision, and eventually impact product sales. In addition, the review helpfulness score is also widely used in the literature as a proxy of review quality. For example, Khern-am-nuai, et al. [22] use review helpfulness scores as a proxy to measure review quality, in addition to review length and readability. In the same way, Wang, et al. [23] also use review helpfulness scores, along with review length, to measure review quality. Evidently, this use of review helpfulness scores is widely-adopted in the literature [e.g., 24, 25].

2.3. Review unhelpfulness score

Although the topics related to review evaluation and review helpfulness scores are heavily discussed in the literature. Another score that appears alongside the review helpfulness score, the review *unhelpfulness* score, has attracted significantly less attention from researchers. This score is usually used as another factor that represents review characteristics to perform several operations such as detecting review spams [26] and ranking reviews [27]. In the meantime, the most popular usage of the review unhelpfulness score in the literature is to use it as a discount factor when measuring review quality through the review helpfulness score. In other words, many studies have measured review quality by subtracting review unhelpfulness scores from review helpfulness scores [e.g., 10]. This practice relies upon an assumption that review helpfulness scores and review unhelpfulness scores are voted in a similar manner and that they represent intrinsic review characteristics (i.e., high quality reviews get helpfulness votes while low quality reviews get unhelpfulness votes). Unfortunately, although there exists evidence that the assumption regarding review helpfulness scores may hold, there is virtually no prior evidence with regard to whether review unhelpfulness scores follow that assumption. Furthermore, Connors, et al. [13] use a survey to extract factors that induce review readers to vote reviews as unhelpful and many of them are not directly related to intrinsic review quality but related more to personal perception (e.g., "Overly Emotional," "Irrelevant Comments").

In summary, prior literature has extensively studied review evaluation and review helpfulness scores. It has demonstrated that the review helpfulness score tends to be directly related to intrinsic characteristics of reviews and hence it could be a consistent estimator of review quality. On the other hand, the review unhelpfulness score is understudied but it has been used in prior works to indicate reviews with lower quality. The primary objective of this paper is to identify whether this assumption holds by using an empirical analysis on data from a restaurant review platform.

3. Research Context, Data, and Method

We collaborate with a large review platform in Asia to investigate our research question. Although the platform allows user reviews of several types of venue (e.g., theaters, public attractions, etc.) to be posted on the website, most of the reviews on the platform are for restaurants that the reviewers visit in the past. Apart from textual content of the reviews, contributors are also asked to issue a star rating (ranged from 1 to 5) for their overall evaluation of the venue. Similar to several third-party review platforms, this website adopts the review helpfulness voting system as the way to measure how review readers evaluate the quality of a review and to assist readers to decide which reviews to read. Specifically, the users can vote either "helpful" or "not helpful" for each review.

In our collaboration, the platform provides us with three datasets. The first dataset consists of consumer reviews information (e.g., review id, reviewer id, the date of review, textual content of the reviews, the associated star rating, the number of photos attached etc.) in June 2016. The second dataset includes the timestamps of helpfulness and unhelpfulness votes for each review in the first dataset. This dataset allows us to track which review reader casts helpfulness or unhelpfulness vote for which review at which time. The platform also provides us with an access to the corresponding review reading logs (e.g., the review, the reviewer reader, and the date and time of reading). Using the review reading logs, we can control for the exposure to the reviews (since reviews with higher number of reads would naturally obtain higher (un)helpfulness votes). Lastly, the platform provides us with one month of review generating activities information (May 2016) of voters who vote at least once for reviews generated in June 2016.

We first conduct review level analysis to examine how quantitative review characteristics affect the helpfulness and unhelpfulness votes. Our dependent variables are *Helpfulness* and *Unhelpfulness*, which are measured by the number of helpfulness or unhelpfulness votes received by a review within 30 days after the reviews are posted on the platform, respectively. In our first dataset, there were 14,515 reviews generated on the platform in June 2016 that received at least one vote (either helpfulness or unhelpfulness) in 30 days. Among them, 14,503 reviews received helpfulness votes and 462 reviews received unhelpfulness votes. Conceptually, the explanatory variables that we will be using in our analysis are review extremity, review depth, and review richness. We measure review extremity by the star rating associated with the review. In the meantime, review depth is measured by the number of words of a review. Note that this variable captures the amount of information contained in the review and it is usually correlated with the level of effort the reviewer put in writing this review [22]. Review richness is measured by the number of photos attached, which captures the information richness of a review. We also control for the number of times that the review is accessed by review readers (*View*) and the chronological order of the review for a restaurant (*Rank*), to capture the potential impacts caused by exposure and review rank.

We report summary statistics for review quantitative characteristics in Table 1 below. It demonstrates that reviews in our sample obtain a 3.87 star rating on average. There are roughly 493 characters in each review, while the longest review contains 10,205 characters. Reviewers on average submit 6 photos for each of their reviews. Reviews have been read for 50 times and the rank in order is about 33 on average. Lastly, the average helpfulness votes and unhelpfulness votes are about 7.51 and 0.03, with large standard deviation of 11.83 and 0.19, respectively, which suggests that our sample contains reviews with a sufficient variation in the number of votes received.

Table 1. Summary statistics (quantitative characteristics)

Variable	Mean	Std. Dev.	Min	Max
Rating	3.87	0.86	1	5
Length	493.05	570.65	0	10,205
Photo	5.57	5.82	1	85
View	49.80	129.46	0	7,026
Rank	32.62	54.77	1	464
Helpfulness	7.51	11.83	0	107
Unhelpfulness	0.03	0.19	0	3

Apart from analyzing the quantitative review measurements mentioned above, we explore further into the content of review text to better understand how textual features affect the helpfulness and unhelpfulness votes received. Specifically, we conduct topic modeling, sentiment analysis, and readability analysis to examine the impact of review content.

We adopt topic modeling approach, based on the highly cited Latent Dirichlet Allocation (LDA) model [28], to discover semantic structures hidden in the textual content of reviews in our sample. LDA is an unsupervised clustering model for discovering abstract topics of a collection documents and generating a predefined number of topics. In LDA, each review is modeled as a mixture of various topics, which, in turn, are modeled as term distributions. We use the *scikit-learn* package for Python [29] to analyze review text. We run the model with three, four, five, and six topics, respectively, and inspect the distribution of terms. We find that the choice of four topics delivers the lowest perplexity, which is a commonly used measure for the evaluation of topic models [28], and the most meaningful distribution. According to top five terms of each of the four topics listed in Table 2, we label these topics food and meal, service, restaurant atmosphere, and drink and dessert, respectively.

Next, we conduct sentiment analysis to determine whether the polarity of review text (i.e., whether a review is positive, negative, or neutral) impact the helpfulness and unhelpfulness votes a review receives. We use a lexicon based sentiment analysis tool, TextBlob, in Python [30] to calculate a sentiment score for each review based on the review content. The score, which is an average polarity score of each word

in that review, is in the range of -1 to 1, where 1 means positive statement and -1 means a negative statement.

Lastly, we analysis if the text readability affects the number of helpfulness and unhelpfulness a review obtains. We use a Python package NLTK [31] to calculate Gunning–Fog (GF) index, which estimates the years of formal education a person needs to understand the text on the first reading and is commonly adopted in IS literature [e.g., 22, 32]. A higher readability score indicates that the review is harder to understand.

Table 2. Top terms of topics in reviews

Topic	Top terms
1 Food and meal	pork, rice, eat, chicken, delicious
2 Service	order, time, price, staff, menu
3 Restaurant atmospheres	shop, good, restaurant, atmosphere, come
4 Drink and dessert	ice, sweet, cream, tea, coffee

We report summary statistics of textual features in Table 3. It shows that on average, 30.93% of reviews' content focuses on discussing foods of the restaurants, 26.24% of the content talks about the drink and dessert, 22.03% of the content is about the service and the other 20.80% of the content is about the restaurants' atmospheres. The mean sentiment of review text is 0.23, which is slightly positive. The mean value of review readability score is 7.77.

Table 3. Summary statistics (textual features)

Variable	Mean	Std. Dev.	Min	Max
% of Topic 1	0.3093	0.3271	0.0003	0.9964
% of Topic 2	0.2203	0.2737	0.0005	0.9936
% of Topic 3	0.2080	0.2967	0.0003	0.9970
% of Topic 4	0.2624	0.2798	0.0003	0.9934
Sentiment	0.23	0.18	-0.78	1
Readability	7.77	3.64	0.40	50.77

4. Empirical Analysis

Our empirical analysis consists of two parts. We first conduct review level analysis to examine how intrinsic review characteristics affect (un)helpfulness votes. Second, we conduct user level analysis to examine how helpfulness voters behave differently from unhelpfulness voters.

4.1. Review helpfulness score, review unhelpfulness score, and intrinsic review characteristics

We first examine the relationships between review helpfulness/unhelpfulness scores and intrinsic quantitative review characteristics. For this analysis, we rely on the following regression specification:

$$DV = \alpha_0 + \alpha_1 Rating + \alpha_2 Rating^2 + \alpha_3 Log(Length) + \alpha_4 Log(Photo) + \alpha_5 Log(View) + \alpha_6 Log(Rank) + \varepsilon.$$

Note that our dependent variables (*Helpfulness* and *Unhelpfulness*) are count data. Therefore, we utilize the Negative Binomial Regression for our analysis. It is also worth noting that we add the quadratic term for *Rating* because previous studies have suggested that the relationship between review extremity and helpfulness is non-linear [9]. Additionally, since review length, the number of photos attached, the number of views received, and review rank are skewed in nature, we apply the natural log transformation on them (1 + x, where x is the variable of interest) to provide a better model fit.

Table 4. Regression results on quantitative review characteristics

	Helpfulness	*Unhelpfulness*
$Rating$	0.550***	-0.542
	(0.107)	(0.440)
$Rating^2$	-0.100***	0.092
	(0.014)	(0.059)
$Log(Length)$	0.163***	0.099
	(0.019)	(0.077)
$Log(Photo)$	0.394***	0.030
	(0.044)	(0.127)
$Log(View)$	0.236***	0.308***
	(0.031)	(0.061)
$Log(Rank)$	-0.095***	-0.031
	(0.014)	(0.041)
Constant	-0.825**	-4.332***
	(0.214)	(0.875)
Observations	14,515	14,515
Note: 1) Robust standard errors clustered by reviewer are in parentheses; 2) *** p < 0.001, ** p < 0.01, * p < 0.05		

Table 4 reports the results of the regression analysis. The first column presents the regression results where the main dependent variable is *Helpfulness*. The results are largely consistent with previous studies. Specifically, the relationship between both *Rating* and *Rating*^{2} and *Helpfulness* are statistically significant. The positive coefficient of *Rating* and the negative coefficient of *Rating*^{2} indicate an inverted-U relationship between rating and helpfulness votes, which is similar to the results of Mudambi and Schuff [9]. In other words, reviews that have either too high or too low ratings are associated

with lower levels of helpfulness votes than reviews with moderate ratings. In addition, longer reviews and reviews with more photos attached are more likely to obtain more helpfulness votes, since such reviews tend to be perceived as more informative and of higher quality, which is also consistent with previous works [9]. Intuitively, reviews that are read more by the users obtain more helpfulness votes. The coefficient of *Log*(*Rank)* is negatively significant, indicating that reviews obtain more helpfulness votes when they stay longer on the platform.

The results related to the primary interest of this work is reported in the second column. In other words, these results present the regression analysis where the main dependent variable is the *Unhelpfulness Votes.* Intuitively, if review readers cast the unhelpfulness votes based on intrinsic review characteristics, we should expect to see the relationship between the independent variables and dependent variable that is on the opposite side of the results observed in Column (1). For example, shorter reviews and reviews without photos attached should be more likely to receive unhelpfulness votes. Interestingly, it turns out that the coefficients of all variables reflecting reviews' intrinsic characteristics (e.g., rating, length, the number of photos attached, etc.) are statistically insignificant at p-value < 0.05. The only coefficient that is statistically significant is that of *Log(View)*. The positive coefficient there indicates that reviews that are read more by the readers have higher chance to receive more review unhelpfulness votes, which is consistent with the intuition based on the effect of exposure. This result provides us with an evidence that review unhelpfulness votes may not be driven by intrinsic review characteristics, at least not the ones that drive review helpfulness votes. It also cautions academic researchers who utilize the review unhelpfulness votes as a proxy to identify reviews with lower quality that such an assumption might not be valid.

In addition to the results based on quantitative review measurements, we next explore the relationship between review helpfulness scores, review unhelpfulness scores, and intrinsic review characteristics that are based on textual features. Here, we continue to use the same regression specification, but the independent variables change. We now focus on the impact of review textual features (i.e., topics distribution, sentiment, and readability) on the number of helpfulness and unhelpfulness votes received by the review. Again, we utilize the Negative Binomial Regression for our analysis. Since each review is represented by the probability that it belongs to one of the four topics, Topic 4 is omitted due to perfect multicollinearity.

Table 5 reports the result of this analysis. The first and second column present the regression results corresponding to dependent variable *Helpfulness* and *Unhelpfulness*, respectively. According to the first column, compared to reviews heavily focusing on Topic 4 that discusses drink and dessert, reviews involving other topics that discuss foods, services, and atmosphere of the restaurant are more likely to obtain helpfulness votes. In particular, it appears that reviews discussing restaurant atmosphere (Topic 3) is able to attract the most helpfulness votes, followed by those discussing foods (Topic 1), and then services (Topic 2). The coefficient of text sentiment is negatively significant at 0.1% level, indicating that reviews with higher sentiment tend to be perceived as less helpful. Text's readability does not seem to have a significant impact on helpfulness scores. For the other dependent variable, *Unhelpfulness*, we find that coefficients of all these variables reflecting reviews' textual features shown in the second column are statistically insignificant. This result shows that review unhelpfulness votes are not driven by review textual features either.

Table 5. Regression results on Review textual features

	Helpfulness	*Unhelpfulness*
Topic 1	0.733***	0.202
	(0.148)	(0.368)
Topic 2	0.660***	0.419
	(0.125)	(0.284)
Topic 3	1.047***	0.584
	(0.132)	(0.356)
Sentiment	-0.967***	-0.443
	(0.150)	(0.278)
Readability	0.013	0.015
	(0.008)	(0.014)
Constant	1.505***	-3.716***
	(0.130)	(0.245)
Observations	14,515	14,515
Note: 1) Robust standard errors clustered by reviewer are in parentheses; 2) *** p < 0.001, ** p < 0.01, * p < 0.05		

Lastly, we combine quantitative review characteristics and textual features together as a set of independent variables and report the results in Table 6. The impacts of review quantitative characteristics on helpfulness votes shown in the first column are statistically the same as those reported in Table 4. For review textual features, the coefficient of Topic 2 becomes insignificant. It suggests that compared to reviews focusing on drink and dessert, reviews talking more about foods (Topic 1) and restaurant atmosphere (Topic 3) are more likely to receive helpfulness votes.

Additionally, the coefficients of *Readability* is negatively significant. It indicates that reviews with more esoteric text are more likely to be perceived as helpful. Consistently, we observe no significant impact of either quantitative review characteristics or review textual features on unhelpfulness votes. *Log(View)* is the only significant independent variable, which captures the effect of exposure.

Table 6. Regression results on quantitative review characteristics and textual features

	Helpfulness	*Unhelpfulness*
$Rating$	0.422***	-0.616
	(0.095)	(0.409)
$Rating^2$	-0.079***	0.104
	(0.013)	(0.055)
$Log(Length)$	0.171***	0.090
	(0.020)	(0.070)
$Log(Photo)$	0.379***	0.015
	(0.043)	(0.132)
$Log(View)$	0.234***	0.312***
	(0.030)	(0.064)
$Log(Rank)$	-0.083***	-0.022
	(0.014)	(0.040)
Topic1	0.375***	0.254
	(0.106)	(0.375)
Topic2	0.155	0.292
	(0.099)	(0.296)
Topic3	0.663***	0.611
	(0.085)	(0.335)
Sentiment	-0.274***	-0.471
	(0.090)	(0.321)
Readability	-0.013***	0.006
	(0.004)	(0.017)
Constant	-0.847***	-4.426***
	(0.221)	(0.885)
Observations	14,515	14,515
Note: 1) Robust standard errors clustered by reviewer are in parentheses; 2) *** p < 0.001, ** p < 0.01, * p < 0.05		

In summary, the results discussed earlier empirically demonstrate that reviews' quantitative characteristics and textual features have asymmetrical impacts on review helpfulness and unhelpfulness scores. While these intrinsic review characteristics significantly affect the helpfulness votes a review received, their connection to the review unhelpfulness votes is particularly weak (i.e., there is no statistically significant connection between intrinsic review characteristics and review unhelpfulness scores). Hence, the common practice of deducting unhelpfulness votes from total votes to represent the review quality that is commonly used in the literature may not reflect the true nature of these scores.

4.2. Helpfulness voters vs. unhelpfulness voters

Our next set of analyses focus at the voter level to inspect how unhelpfulness voters behave differently from helpfulness voters. There are 3,722 voters who vote for reviews generated in June 2016. We define helpfulness voters as voters who only vote up for other reviews (i.e. casting helpfulness votes), and unhelpfulness voters as those who only vote down for other reviews (i.e. casting unhelpfulness votes). We perform the student t-test on the review volume, the average of content length, the average number of photos per review, the number of comments submitted by the voters in one month before focal reviews (only 3,599 voters generate reviews in the May 2016) as proxies to measure their engagement to the review platform. Formally, our comparison specification is:

$H_0: \mu(O_{helpfulness\ voters}) = \mu(O_{unhelpfulness\ voters})$

$H_a: \mu(O_{helpfulness\ voters}) \neq \mu(O_{unhelpfulness\ voters})$,

where the variable of interest O consists of four variables. First, we measure *ReviewCount*, which is the total number of reviews written by each reviewer before the data collection date. This variable essentially measures the engagement level that the reviewers have with the platform. The second variable is *Length*, which is the average length of reviews written by each reviewer before the data collection date. It is generally used to measure the amount of effort exerted by the reviewers, which is also another proxy for review platform engagement. The third variable is *Photos*, which measures the average number of photos reviewers attaches in each of their reviews. Since taking photos require additional and premeditate effort, this variable is also a good proxy to measure review platform engagement. Lastly, we include *Comment* as the fourth variable of interest. This variable counts the total number of comments that each reviewer made on other reviews before the data collection date. Commenting is one of the social networking features that the platform offers to stimulate user-to-user interactions. In that regard, this variable also measures the platform engagement (and user engagement as well).

The results of our exploratory analysis at the voter-level are presented in Table 7 below. Perhaps not so surprisingly, we find that helpfulness voters and unhelpfulness voters are vastly different. Specifically, unhelpfulness voters are significantly less engaged with the review platform than the helpfulness voters. For example, in terms of the review-contributing behavior, unhelpfulness voters write reviews almost four times less often than helpfulness voters (0.52 vs.

2.05). In the same way, the average length of the review is also significantly different. The mean of the average length of reviews written by helpfulness voters is 116.04 while that of unhelpfulness voters is 39.55. Again, the difference in review length, which generally captures the effort in review writing [22] is almost 4 times. As for the photos attached in the reviews written, helpfulness voters have 1.20 photos attached in their reviews on average while unhelpfulness voters attach only 0.30 photos in their reviews on average. Lastly, for the engagement with other uses, which we measure using the comment sent by the voters, although the magnitude of the comments sent by helpfulness voters is much higher than that of unhelpfulness voters, the variance of this variable is also significantly high. As a result, the difference between the two groups turns out to be statistically insignificant.

Table 7. Engagement of helpfulness versus unhelpfulness voters

Variable	Helpfulness Voters (Mean)	Unhelpfulness Voters (Mean)	T-value (p-value)
ReviewCount	2.05	0.52	3.17 (0.002)
Length	116.04	39.55	3.55 (0.000)
Photo	1.20	0.30	4.76 (0.000)
Comment	1.12	0.11	1.07 (0.284)

In addition to the differences between helpfulness voters and unhelpfulness voters in terms of their involvement in the platform, we are also interested in how the diversity of each vote type differs. In that regard, we calculate the Gini coefficient of the votes, based on the vote volume of each voter. Gini coefficient is widely used in the literature to measure the diversity of user behavior [e.g., 22, 33]. The Gini coefficient ranges from zero (highest diversity) to one (highest concentration). In our context, the Gini coefficient of zero represents the case where each voter contributes only one vote, so the votes are distributed equally among all voters. Meanwhile, the Gini coefficient of one represents the scenario where all votes are casted by a single voter. Following the literature, we calculate the Gini coefficient by constructing a curve that is similar to the Lorenz curve [34] and dividing the area between the Lorenz curve and the 45-degree line by the total area under the 45-degree line.

Figure 1. Lorenz curve (helpfulness vs. unhelpfulness voters)

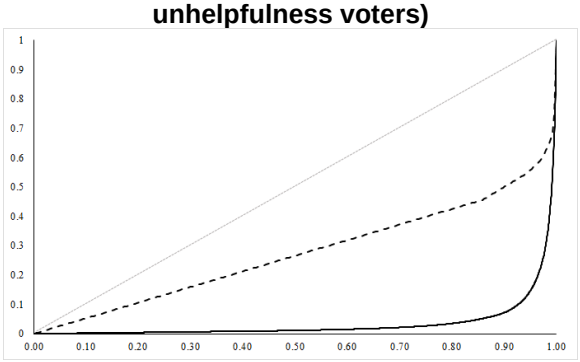

Figure 1 shows the diversity of the review helpfulness score (solid black line) and the diversity of the review unhelpfulness score (dotted black line), compare to the line of perfect equality (the 45-degree grey line). Interestingly, the review unhelpfulness score is much more diverse than the review helpfulness score. The Gini coefficient of the review helpfulness score is 0.930, while the Gini coefficient of the review unhelpfulness score is only 0.453. The permutation test [35] confirms that such a difference is statistically significant at p-value < 0.05.

The results from our empirical analyses are particularly interesting from the platform's perspective. Unlike the case of review helpfulness score, most intrinsic review characteristics do no influence the unhelpfulness votes. In addition, unhelpfulness voters are significant different than helpfulness voters in terms of their engagement to the review platform. Each unhelpfulness voter also casts the vote to only a few reviews, making the diversity of the unhelpfulness vote much higher.

5. Conclusions

The review platforms have been increasingly incorporating the review evaluation system to assist consumers' decision making, improve their satisfaction, and enhance the content quality on the platform. While many platforms adopt the review evaluation system that allows users to vote for both helpful and unhelpful reviews, most existing studies that involve review evaluations only focus on examining review helpfulness scores. Meanwhile, our knowledge on the nature of negative review evaluation (e.g., review unhelpfulness score) is very limited. Even then, many prior works that utilize review helpfulness and unhelpfulness score implicitly assume that both types of scores are related to reviews' intrinsic characteristics. The primary purpose of this study is to validate this assumption by paying specific attention to the factors that influence review

unhelpfulness votes. Particularly, we are interested in identifying whether review unhelpfulness votes are driven by intrinsic review characteristics or not, and whether the helpfulness voters and unhelpfulness voters are systematically different. We operationalize our research agenda by using a unique dataset obtained through a collaboration with a large restaurant review platform in Asia. Our analysis demonstrates that, unlike the review helpfulness scores, the review unhelpfulness scores do not appear to be driven by intrinsic review characteristics, including both review quantitative measures and textual features. We also find that helpfulness voters are significantly different than unhelpfulness voters in terms of their engagement with the platform and the concentration level of their voting behavior. As the first step of our effort to investigate the unhelpfulness votes, this result cautions against the commonly adopted assumption in the literature that review unhelpfulness score is an opposite side of the review helpfulness score.

Moving forward, we are interested to expand this research project in several directions, including exploring further into the mechanism of why people cast unhelpfulness votes and how these votes might affect platform users' behavior and the platform itself.

First, although our results demonstrate that common review quantitative measurements (e.g., ratings and length) and textual features (e.g., sentiment and readability) do not appear to influence review unhelpfulness votes, it is possible that the scores are influenced by some other factors, such as certain concrete emotion and text controversy. In particular, we would focus on analyzing reviews that received both helpfulness and unhelpfulness votes to investigate this question. Second, we plan to expand beyond the realm of the driving factors that influence review unhelpfulness scores. For instance, investigating the diversity of unhelpfulness votes based on restaurants may provide some interesting insights that yield a better view of the underlying mechanism behind the negative review evaluation. Third, it would also be interesting to empirically examine whether the label "review unhelpfulness" is misleading. For example, if we can establish that review unhelpfulness scores are not influenced by intrinsic review characteristics, but are driven by review readers' personal perception (e.g., readers' disagreement with the review content), then an alternative label such as "dislike" might better reflect the true meaning of this score. In this regard, we plan to investigate, conditional on our findings regarding the driving factors of review unhelpfulness scores, whether the change in the label improves review readers' satisfaction and platform's welfare. This aspect of the study could also improve platform design to better capture quality versus relevance dimensions of online reviews. In that regard, the insights would be useful for both academic researchers and platform managers. Lastly, we attempt to investigate the potential influence caused by unhelpfulness votes on both vote receivers' behavior and platform's prosperity. If users cast unhelpfulness votes based on their personal perception or preference rather than the quality of a review, it might lead to vote receivers' negative reactions such as casting unhelpfulness votes for others as revenge, reduce their engagement with the platform or even quit the platform. If that is the case, users might not trust the evaluation system and other content generated on the platform anymore, which may hinder the platform's long-term development.

6. References

[1] Q. Ye, R. Law, and B. Gu, "The impact of online user reviews on hotel room sales," *International Journal of Hospitality Management*, vol. 28, no. 1, pp. 180-182, 2009.

[2] I. E. Vermeulen and D. Seegers, "Tried and tested: The impact of online hotel reviews on consumer consideration," *Tourism management*, vol. 30, no. 1, pp. 123-127, 2009.

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

[4] L. A. G. Camacho and S. N. Alves-Souza, "Social network data to alleviate cold-start in recommender system: A systematic review," *Information Processing & Management*, vol. 54, no. 4, pp. 529-544, 2018.

[5] S. Zhou and B. Guo, "The order effect on online review helpfulness: A social influence perspective," *Decision Support Systems*, vol. 93, pp. 77-87, 2017.

[6] M. Salehan, M. Mousavizadeh, and M. Koohikamali, "A Recommender System for Online Consumer Reviews," 2015.

[7] S. Zhang, M. Salehan, A. Leung, I. Cabral, and N. Aghakhani, "A recommender system for cultural restaurants based on review factors and review sentiment," 2018.

[8] Y. Yu, W. Khern-am-nuai, A. Pinsonneault, and Z. Wei, "The Role of Social Technologies in Review Evaluation and Online Platform Implications," 2020.

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

[10] P.-Y. Chen, S. Dhanasobhon, and M. D. Smith, "All reviews are not created equal: The disaggregate

impact of reviews and reviewers at amazon. com," SSRN 2008.
[11] J. Zheng, Y. Tan, and G. Yin, "Does Help Help? An Empirical Investigation of Review-In-Review in User-Generated Content System," SSRN 2016.
[12] Y. Pan and J. Q. Zhang, "Born unequal: a study of the helpfulness of user-generated product reviews," *Journal of Retailing*, vol. 87, no. 4, pp. 598-612, 2011.
[13] L. Connors, S. M. Mudambi, and D. Schuff, "Is it the review or the reviewer? A multi-method approach to determine the antecedents of online review helpfulness," in *2011 44th Hawaii International Conference on System Sciences*, 2011: IEEE, pp. 1-10.
[14] P. F. Wu, H. Van Der Heijden, and N. Korfiatis, "The influences of negativity and review quality on the helpfulness of online reviews," in *International conference on information systems*, 2011.
[15] S. P. Eslami, M. Ghasemaghaei, and K. Hassanein, "Which online reviews do consumers find most helpful? A multi-method investigation," *Decision Support Systems*, vol. 113, pp. 32-42, 2018.
[16] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on empirical methods in natural language processing*, 2006: Association for Computational Linguistics, pp. 423-430.
[17] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and predicting the helpfulness of online reviews," in *2008 Eighth IEEE international conference on data mining*, 2008: IEEE, pp. 443-452.
[18] W. Xiong and D. Litman, "Automatically predicting peer-review helpfulness," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 2011: Association for Computational Linguistics, pp. 502-507.
[19] T. L. Ngo-Ye and A. P. Sinha, "The influence of reviewer engagement characteristics on online review helpfulness: A text regression model," *Decision Support Systems*, vol. 61, pp. 47-58, 2014.
[20] H. Baek, J. Ahn, and Y. Choi, "Helpfulness of online consumer reviews: Readers' objectives and review cues," *International Journal of Electronic Commerce*, vol. 17, no. 2, pp. 99-126, 2012.
[21] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *IEEE transactions on knowledge and data engineering*, vol. 23, no. 10, pp. 1498-1512, 2011.
[22] W. Khern-am-nuai, K. Kannan, and H. Ghasemkhani, "Extrinsic versus intrinsic rewards for contributing reviews in an online platform," *Information Systems Research*, vol. 29, no. 4, pp. 871-892, 2018.
[23] S. Wang, P. Pavlou, and J. Gong, "On Monetary Incentives, Online Product Reviews, and Sales," in *International Conference on Information Systems (ICIS), Dublin, Ireland*, 2016.
[24] D. Qiao, S.-Y. Lee, A. Whinston, and Q. Wei, "Incentive Provision and Pro-Social Behaviors," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.
[25] Y. Wang, P. Goes, Z. Wei, and D. Zeng, "Production of Online Word‐of‐Mouth: Peer Effects and the Moderation of User Characteristics," *Production and Operations Management*, 2019.
[26] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010: ACM, pp. 939-948.
[27] S. Kawate and K. Patil, "An Approach for Reviewing and Ranking the Customers' Reviews through Quality of Review (QoR)," *ICTACT Journal on Soft Computing*, vol. 7, no. 2, 2017.
[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993-1022, 2003.
[29] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825-2830, 2011.
[30] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, and E. Dempsey, "Textblob: simplified text processing," *Secondary TextBlob: Simplified Text Processing*, vol. 3, 2014.
[31] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 2002, pp. 63-70.
[32] P. B. Goes, M. Lin, and C.-m. Au Yeung, ""Popularity effect" in user-generated content: Evidence from online product reviews," *Information Systems Research*, vol. 25, no. 2, pp. 222-238, 2014.
[33] D. Lee and K. Hosanagar, "How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment," *Information Systems Research*, vol. 30, no. 1, pp. 239-259, 2019.
[34] J. L. Gastwirth, "A general definition of the Lorenz curve," *Econometrica: Journal of the Econometric Society*, pp. 1037-1039, 1971.
[35] P. I. Good, *Permutation, parametric, and bootstrap tests of hypotheses*. Springer Science & Business Media, 2006.