Journal of
Theoretical and Applied Electronic Commerce Research

MDPI

Article

What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values

Yuan Meng [1,*], Nianhua Yang [1], Zhilin Qian [1] and Gaoyu Zhang [2]

[1] School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201620, China; yangnianhua@suibe.edu.cn (N.Y.); qian_zhilin@163.com (Z.Q.)
[2] School of Information Management, Shanghai Lixin University of Accounting and Finance, Shanghai 201209, China; 20089840@lixin.edu.cn
* Correspondence: nancymeng@suibe.edu.cn

Received: 18 January 2020; Accepted: 20 October 2020; Published: 20 November 2020

Abstract: Online product reviews play important roles in the word-of-mouth marketing of e-commerce enterprises, but only helpful reviews actually influence customers' purchase decisions. Current research focuses on how to predict the helpfulness of a review but lacks a thorough analysis of why it is helpful. In this paper, feature sets covering review text and context cues are firstly proposed to represent review helpfulness. Then, a set of gradient boosted trees (GBT) models is introduced, and the optimal one, which as implemented in eXtreme Gradient Boosting (XGBoost), is chosen to predict and explain review helpfulness. Specially, by including the SHAP (Shapley) values method to quantify feature contribution, this paper presents an integrated framework to better interpret why a review is helpful at both the macro and micro levels. Based on real data from Amazon.cn, this paper reveals that the number of words contributes the most to the helpfulness of reviews on headsets and is interactively influenced by features like the number of sentences or feature frequency, while feature frequency contributes the most to the helpfulness of facial cleanser reviews and is interactively influenced by the number of adjectives used in the review or the review's entropy. Both datasets show that individual feature contributions vary from review to review, and individual joint contributions gradually decrease with the increase of feature values.

Keywords: online review; review helpfulness; SHAP values; XGBoost; feature contribution; joint feature contribution; individual feature contribution

1. Introduction

In recent years, the rapid development of e-commerce has brought about an explosive growth in online product reviews. When consumers shop online, they pay special attention to the product evaluations from other consumers. In this context, online product reviews have gradually become an online reputation system for e-commerce enterprise. However, as the number of online reviews increases exponentially, the quality of online reviews has become increasingly uneven. It is increasingly difficult and time-consuming for consumers to find helpful review information and for e-commerce enterprises to manage this increasingly massive number of reviews. It is well-known that helpful online reviews form an important element for e-commerce communities, which are embedded with valuable information influencing consumer purchases [1,2]. Moreover, helpful product reviews are relevant to all stakeholders in the online review community, such as consumers, suppliers, retailers, and community platforms. Therefore, the ability to find helpful reviews and conduct causal analysis are common priorities both in industrial and academic fields. In order to more effectively extract valuable information from the vast number of online reviews, both industry and academia have made

great efforts to improve the detection methods for helpful reviews. Although the existing statistical learning methods have a high detection rate for the detection of helpful reviews, they are still difficult to use to obtain reasonable explanations for the main causes and mechanisms underlying the occurrence of helpful reviews. Based on this context, we propose a feature contribution-driven analysis framework for review helpfulness. Our research thoroughly analyzes the formation mechanism of helpful reviews and proposes a corresponding detection method. On the one hand, the proposed method could alleviate the overload pressure caused by massive numbers of online reviews for consumers and enterprises. On the other hand, it provides a useful reference for the e-commerce online community to carry out the task of improving online review governance and optimizing the recommendation mechanism of online review information.

Toward the question of how to determine review helpfulness, a commonly used method for most e-commerce websites is to set up the 'helpful votes' function under each product review. Then, readers' votes are accumulated and utilized to distinguish between helpful and unhelpful reviews. Since such a manual method requires not only a certain amount of time to accumulate 'helpful votes', but also to overcome the evaluation bias during the collection of 'helpful votes', a modeling method is often a preferred solution for this problem [3–5].

Considerable relevant research regards "whether a review is helpful" as a classification problem, which requires extracting possible features from 'labeled review' as an input, and building a prediction model for an 'unlabeled review' combined with a classification algorithm. In the absence of a unified feature extraction standard that can be used to represent a review, researchers commonly extract as many different features as possible to improve the prediction performance to the maximum extent.

In this process, features that improve the performance of a prediction model are chosen as valuable features, which serve as the basis for interpreting which features affect review helpfulness. It is not hard to see such a 'model-based interpretation' is inadequate to some extent, since most commonly used models are 'black-box' models, which leaves some unsolved problems to be tackled.

First, when determining key features based on changes in model performance, there is a range of performance indicators to choose from, such as accuracy, recall, F1-measure, auc, and so on. Key features can also be determined using feature selection methods that come with some tree-based models, such as a Gini impurity decrease or counts of splits of features in trees. Obviously, if the key features incorporated into a model are inconsistent under different evaluation indicators, then the results of a review helpfulness interpretation are not convincing. However, this not-ideal outcome has occurred in most relevant and practical cases thus far, as the features selected by different evaluation indicators are inconsistent. Therefore, we need a more stable and consistent approach to discover the value of features in online product reviews.

Second, it is not clear how various input features should be combined to obtain the prediction result, so it is difficult to quantify the contribution of each feature in a model. Although some tree-based models automatically attribute results to feature importance, it is important to note that 'feature importance' is not the same as 'feature contribution'. The former highlights which features affect model performance, while the latter not only highlights the affecting features, but also directly quantifies the contribution of each feature to the prediction result. In contrast to 'feature importance', 'feature contribution' provides a more intuitive explanation of why a review is helpful.

Third, in practice, various review features, such as many kinds of text or reviewer features, need to be extracted and put into a 'black-box model' to ensure an accurate prediction result is obtained. Inevitably, features interact with each other to influence prediction results. The existing traditional models require predefined feature interaction items and lack the ability to automatically capture any interaction between different features. Therefore, in the related research on review helpfulness prediction, only a few studies involve analyzing the influences of feature interaction items on prediction results.

In this paper, we turned to the recently proposed SHAP (Shapley) values method [6,7] and gradient boosting trees (GBT) models to fill in the research gaps outlined above. The SHAP values

method is a feature attribution method which assigns to each feature a value for a particular prediction, which is helpful for interpreting the prediction result. The method notably provides a strict theoretical improvement from the classic Shapley Value estimation method [8] by ensuring that there is feature consistency and model stability. On the other hand, GBT models such as gradient boosting decision trees (GBDT) [9], eXtreme Gradient Boosting(XGBoost) [10], or Light Gradient Boosting Machine(LightGBM) [11] have been widely applied in various fields such as credit scoring [12] and transportation modes identification [13] in recent years. In addition to having a prediction accuracy advantage, GBT models also possess the benefits of capturing interactions among features without explicitly defining them. Therefore, by combining the SHAP values method with GBT models, it is possible to provide a detailed explanation of why a review is helpful.

We chose two different types of product reviews from Amazon.cn—on headsets and facial cleansers—as our experiment datasets. Based on the information quality theory, we first used a variety of text analysis techniques to extract three dimensions of text features, namely, readability, reliability, and relevancy. In addition, we added important features from reviewers and metadata to ensure feature diversity. Then, we constructed a set of GBT models and multiple sets of baseline ensemble models on the extracted features. Through multiple inter-group and intra-group comparative experiments, we chose the optimal model, XGBoost, as our experimental analysis model. Moreover, we verified the validity of the extracted features through detailed comparative experiments. Based on such comparative experiments, we presented the global contribution and joint feature contribution for a feature on these two kinds of datasets from a global and an individual view, respectively. The experiment results not only explained review helpfulness in detail on both the macro and micro levels, but also helped to comparatively analyze the differences between different product types for understanding review helpfulness.

The remainder of this research is organized as follows. Related work is presented in Section 2. Section 3 describes the research methodology in detail. Then, Section 4 presents the experiment set-up and is followed by Section 5, which presents the results of the experiments. Section 6 discusses the results and findings of the research. Last, Section 7 provides the conclusions, implications, and limitations of this research.

2. Literature Review

2.1. Review Feature Extraction

Heterogeneous features have often been examined by prior research, such as text-related features, reviewer-related features, or metadata features.

Text-related features are commonly extracted from different dimensions of review text with various text analysis techniques [14], such as subjectivity [15,16], linguistics [16,17], readability [16,18], relevancy [19], sentiment [20,21], as well as explained actions and reactions [22]. Chen et al. [15] consider implicit information hidden in the text and take several features into consideration, including a mixture of subjectivity and objectivity. Krishnamoorthy [16] describes a novel method used to automatically extract linguistic features from review texts, and their research results show that linguistic features are better predictors of review helpfulness compared to review metadata, subjectivity, and readability for experiential goods. According to Hu et al. [17], the number of words is a key predictor of helpfulness across three user-controllable filters. Akbarabadi et al. [18] examine the effect of review title features on predicting the helpfulness of online reviews, but they imply that the title characteristics cannot be powerful determinants of online review helpfulness. Chen et al. [19] treat a review as an information item and adopt an IQ framework for feature extraction, which adds evidence to the results that text relevancy facilitates decision-making. Many studies investigate the impacts of sentiment factors on the helpfulness of reviews. Both the findings of [20,21] show that sentiment or expressed emotional arousal in the text affects readers' perceptions of review helpfulness. Different from the above studies which

analyze review text directly, Moore [22] focuses on what individuals explain in reviews, and reveals that actions explanations are more helpful than reactions explanations for utilitarian products.

Reviewer-related features often include the number of reviews posted by specific reviewers [23], or social and reputation features [24]. Zhang et al. [23] consider both reviewer data and metadata are necessary supplements in building helpfulness prediction model, and thus extract the number of reviews posted by specific reviewers in the past and the grade of reviewers to represent reviewer features. Aghakhani et al. [24] identify source credibility as theoretically important variables that affect electronic Word of Mouth(eWOM) adoption on Facebook.

Metadata features are the descriptions of a review itself, such as review rating [17], or review published date [16]. The findings of [17] verify that review rating is a key predictor of review helpfulness. Krishnamoorthy [16] also includes review metadata features in their model for helpfulness prediction.

2.2. Review Helpfulness Prediction Models

Prior studies mainly take advantage of machine learning methods to build review helpfulness prediction models, which treat review helpfulness prediction as a binary classification or a multivariate classification problem. On the basis of feature extraction and feature representation on an annotated review dataset, a specific feature selection method is used to identify the optimal feature set and obtain the optimal classification model. Traditional machine learning methods are widely used in building helpfulness prediction models, such as support vector machine(SVM) [16,19], support vector regression(SVR) [23,25], logistic regression(LR) [26,27], decision tree(DT) [23,28], and ensemble learning models such as random forest(RF) [16,29], bagging classifier [28], or GBDT [20]. Among them, tree ensemble models like random forest or ExtraTrees are considered to be more effective when compared to SVM or LR [16,29]. Recently, popular deep learning models have also been used to predict review helpfulness [3].

In addition to machine learning methods, econometric regression methods are also employed by researchers. Relevant research mostly uses helpful voting information as dependent variable and text or reviewer features as independent variables to build an econometric model for review helpfulness assessment. By analyzing the statistically significant relationships between independent and dependent variables, the influences of review features on review helpfulness are obtained. Commonly used econometric models include multiple regression [15,30,31], Tobit regression [32], and negative binomial regression [33].

Comparatively, the goal of machine learning methods is to extract as many feature items into a model as possible to improve the performance of prediction models. Therefore, machine learning methods bring more accurate prediction results. However, since traditional machine learning models are mostly black-box models, there are still some deficiencies for the identification and interpretation of key features in the models. As for econometric regression methods, the emphasis is to investigate the degree of consistency and the statistical hypothesis between the independent variables and the dependent variable; thus, the number of independent variables is very limited. Although econometric regression methods are more effective for explaining the helpfulness result and finding the key independent variables, the prediction accuracy for review helpfulness is not high due to the strict test hypotheses.

2.3. Interpretation of Review Helpfulness

The purpose of interpreting review helpfulness is to identify key features and to measure the extent to which these key features affect review helpfulness. Therefore, the key features need to be selected by ranking the importance of each extracted feature according to a specific feature engineering method. Existing studies have adopted many different methods to analyze and understand the importance of all kinds of review features on review helpfulness, such as the recursive feature elimination (RFE) method based on the performance comparison on helpfulness prediction models, the feature interpretation

method based on the model itself like GBDT or RandForest, mutual information method, or principal component analysis (PCA).

Malik et al. [34] make use of MSE, RMSE, and RRSE-based error metrics with 10-fold cross-validation to compare model performances and conduct feature selection, in which MSE metric is finally used to measure feature importance. Singh et al. [20] use the GBDT model to predict review helpfulness, and they employ the feature importance metric from GBDT itself to identify key features. Zhang et al. [23] apply recursive elimination of features to infer the most predictable features on review helpfulness based on model performance metrics such as MAE or RMSE, and ten features covering review text and metadata are finally chosen as predictors of review helpfulness. Liu et al. [28] choose mutual information and principal component analysis to explore the utilization of all the features, and they present two different feature sets as the informative features on review helpfulness. In addition, Ghose et al. [29] employ a random forest classifier and three broad categories of features named as reviewer-related features, review subjectivity features, and review readability features for review helpfulness estimation, revealing that using any of such three categories of features can result in a statistically equivalent performance as in the case of using all the available features.

The above studies indicate that the opinions on how to interpret review helpfulness are inconsistent. There are still great uncertainties in identifying affecting features in review helpfulness whether based on the model performance indicators or combined with specific feature engineering methods.

2.4. Research Gap

Although review helpfulness prediction is a hot research issue at present, at least one of the following areas still needs to be improved.

First, most studies extract features from different factors such as text, reviewer, or metadata to represent review helpfulness, but they lack supporting theory basis for explaining why these features are needed to be extracted, especially from the consumers' perspective. In particular, since review text is important unstructured information, a more detailed analysis on how to determine the unified feature forms of the review text is needed. Based on the relevant research, information quality theory provides a good theoretical basis for feature extraction of online reviews. Therefore, it is necessary to reanalyze and redesign the feature forms of on reviews under the information quality theory basis.

Second, although traditional machine learning methods can obtain better performance for review helpfulness prediction, their ability of feature selection and feature interpretation are still weak. Common feature selection methods, such as Gini impurity, PCA, RFE, or MSE/RMS/RRSE based error metrics in different application fields, are not robust enough in relevant research. In particular, these methods fail to provide a consistent feature interpretation result, which leads to inconsistent conclusions about key features affecting review helpfulness. Therefore, improvements in predictive performance and interpretation ability of helpfulness prediction models need to be solved simultaneously.

Third, review helpfulness has not been well understood yet. Existing research mainly focuses on the identification of key features that influence review helpfulness, but seldom introduces the discussion of quantifying the contributions of the key features. In fact, feature contribution is more beneficial for us to identify key affecting features and understand how the key features influence review helpfulness from both the whole and individual view. Meanwhile, the interactions between key features have not been considered in detail in the helpfulness prediction models on reviews. Therefore, it is of great importance to develop an integrated feature interpretation framework on review helpfulness to solve such important questions.

3. Methodology

3.1. Basic Procedure of the Proposed Framework

To conduct an integrated interpretation framework for review helpfulness, we subdivided the procedure into several major parts as shown in Figure 1: (1) data preprocessing; (2) feature extraction; (3) review helpfulness modeling, evaluation, and interpretation.

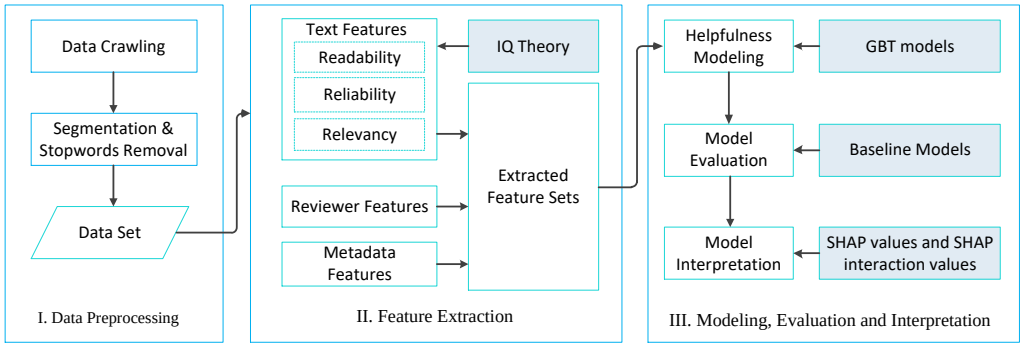

Figure 1. Basic procedure for review helpfulness interpretation.

3.2. Data Preprocessing

Product reviews and relevant data from Amazon.cn are crawled to construct the experimental data. At the data preprocessing stage, the main work includes sentence segmentation, word segmentation, and stopwords removal. It should be noted that Chinese product reviews include both Chinese and English punctuation marks. Thus, based on the sampling statistics of 15,370 Chinese review, the punctuation marks set of sub-sentences is determined, in which the Chinese punctuation marks include '，。！ ？ ；… … () 『』 《》 【】' and English punctuation marks include '.!?;()~{}<>[]'. At the same time, we identified other meaningless punctuation marks in the reviews, such as '["#$%&\'*+/:=@\\^_'|、:‘’“”·''=ノ▽ └┐ ┌]', which can be removed using the regular expression rules. Then, reviews are broken down into sub-sentences for word segmentation and stopwords removal.

3.3. Feature Extraction

This section begins by discussing the core dimensions of review text-related features. Then, it introduces the features extraction of context cues, including reviewer and metadata.

3.3.1. Text Features Based on Information Quality Theory

Based on the information quality (IQ) literature, information quality is user-oriented and multi-dimensional. In [35], the hierarchy of information quality dimensions, namely accessibility, understandability, usefulness and believability, is established accordingly. This framework is based on a longitudinal process by which users acquire, understand, identify, and rely on reliable information to ensure that it is helpful for their decision-making.

Information quality is an abstract concept in nature, so it is particularly important to choose the core dimensions that are generally considered to have a decisive influence on the quality of online reviews. Researchers [36] point that the information quality framework can be used to objectively assess the quality of online reviews if metrics have been developed and the quality dimensions are operationalized relevant to the dataset and task. Thus, in accordance with the longitudinal process of the IQ framework, online consumers can perceive the quality of the review text in following three phases: starting to read reviews, judging the believability of content, and taking advantage of the reviews to make a purchase decision. According to each phase, we proposed a core measure dimension, namely readability, reliability, and relevancy, respectively, as shown in Figure 2. Since accessibility

has nothing to do with the online review community, we include only three core dimensions in our analysis. Derived from the IQ theory, readability measures the expression quality of review text, reliability measures the intrinsic quality of review text, and relevancy measures the utility quality of review text. Next, we detailed the feature extraction methods of each dimension.

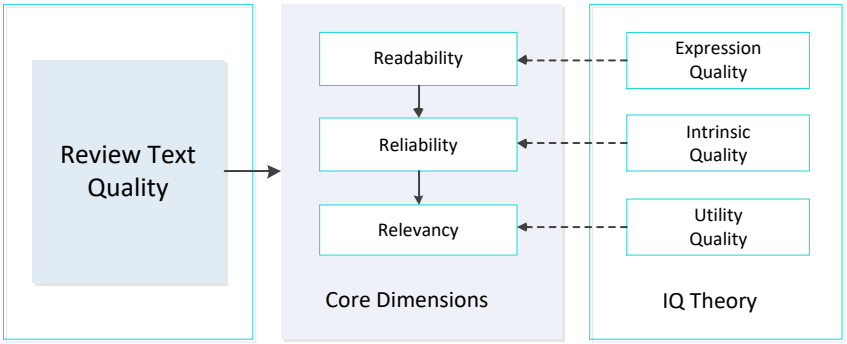

Figure 2. A framework for review text quality measurement.

- Readability

Readability refers to the extent to which the review text is easy to understand. It can be operationalized as linguistic features or sentence structural features [16,20,23]. Linguistic features are text features regarding words and sentences in reviews, which may influence readers' reading time. After preprocessing, the total number of words (Nwords), sub-sentences (Nsents), adjectives (Nadj), adverbs (Nadv), verbs (Nverb) and average sub-sentences length (Averlen) are extracted in a review as linguistic features.

As simple linguistic features cannot reveal the relationships between words, we introduced the statistical language model to quantitatively analyze the collocation information of adjacent words in a sentence, so as to calculate a word's probability of occurrence in a sentence. The calculated value can be used to estimate the uniqueness of a sentence's expression. A bigram sentence language model is built as follows:

$$P(X) = \prod_{X_i \in X} p(X_i | X_{i-1}) \tag{1}$$

$P(X)$ represents the occurrence probability of a given sentence X, and X_i represents the word in the sentence X. If the corpus is large enough, Equation (1) can be estimated by the relative frequency of words according to the maximum likelihood estimation and Bernoulli's Large number theorem.

In general, information entropy and perplexity are two metrics used to evaluate language models. The larger the entropy and perplexity, the more unique the sentence structure. For any given sentence X, its entropy and perplexity values are calculated as follows:

$$Entropy(X) = -\sum_{X_i \in X} p(X_i | X_{i-1}) \log_2 p(X_i | X_{i-1}) \tag{2}$$

$$Perplexity(X) = 2^{Entropy(X)} \tag{3}$$

Thus, we extracted the entropy and perplexity of all the sub-sentences in a review and respectively took their average values to evaluate the review's words structure relationships.

- Reliability

Reliability refers to the extent to which the review text is to be trusted. Prior research has pointed that sentiment orientation (positive or negative) and writing style of the review text (subjective or objective) play important roles in determining the degree of review's believability [30,37].

We followed a similar method as [38] to judge the sentiment orientation of each sub-sentence, choosing NB classifiers and feature representations (unigram, bigram or trigram) at the word and

phrase level. Three thousand reviews with either 1-star or 5-star ratings for each product category were selected to build the corpus. The trained model was then used to predict the sentiment orientation of each sub-sentence.

Suppose r^+ denotes the positive sub-sentences in a given review R, and total(r) denotes the sum of positive and negative sub-sentences, we obtained the overall sentiment orientation of R, denoted as *PosSenti*, by the proportion of the positive sub-sentences in all the sentimental sub-sentences, which can be expressed as the following:

$$PosSenti(R) = \frac{count(r^+)}{total(r)} \tag{4}$$

The greater the PosSenti value of R, the more positive R is. Thus, *PosSenti* represents the overall sentiment orientation of R. Meanwhile, in order to measure the degree of mixing sentiment in R, we introduced a variable, denoted as *DevPos*. It is calculated as the deviation between the review R's *PosSenti* and the average *PosSenti* of all the reviews of the same product, which is expressed as the following:

$$DevPos(R) = \left| PosSenti(R) - Avg(\sum PosSenti(R)) \right| \tag{5}$$

Since the average *PosSenti* reflects the equilibrium value of the mixed distribution of positive and negative sentiments for all reviews of a product, the smaller the deviation from the average, the more balanced the mix of two kinds of sentiment orientations. Conversely, it means that the sentiment orientations of most sub-sentences in R are consistent.

We followed the same paradigm of studies in [29,39] to measure the objective degree of each sub-sentence. Three thousand product description sentences and 3000 product reviews were randomly selected to build the corpus. Then, we extracted n-grams (n = 8) features to train the classifier using a dynamic language model classifier, which was used to predict the objective probability of each sub-sentence. Assuming that the objective degree of a sub-sentence r is $obj(r)$, and the total number of sub-sentences in R is $count(r)$, we denoted the overall objective degree of review R as $ObjDegree(R)$, which can be expressed as the following:

$$ObjDegree(R) = \frac{\sum obj(r)}{count(r)} \tag{6}$$

Similar to a mixture of sentiments, a review is often a mixture of styles. Therefore, we took a similar approach to measure the mixed degree of a review's style, denoted by *DevObj*. It is calculated as the deviation between the review R's *ObjDegree* and the average *ObjDegree* of all the reviews belong to the same product, which is expressed as the following:

$$DevObj(R) = \left| ObjDegree(R) - Avg(\sum ObjDegree(R)) \right| \tag{7}$$

- Relevancy

Relevancy refers to the extent to which a review is relevant to the product itself, and it is often operationalized as the quantity of consumer opinions toward a product's attributes or features [19,30].

Online reviews usually include evaluations of multiple attributes of a product (such as appearance, price, effect, logistics, etc.) or multiple evaluations of one or two attributes (such as effect is good and satisfactory, etc.). The former represents the diversity of a review's opinion, while the latter reflects the details of a review's opinion. They both reflect the relevance of the review to the product and thus enable the measurement of the review's relevancy. Therefore, we used the two indicators, namely frequency of attribute-opinion-pair (AttriFreq), and frequency of feature-opinion-pair (FeatureFreq), as a review's relevancy features.

Motivated by the prior study on product feature and opinion extraction from online reviews [40], we first identified the correct feature words and opinion words set of a product. Meanwhile, as some

inexplicit feature words in a review cannot be extracted together, we followed the method of [41] to identify and count them. After identifying the feature–opinion pair in a review, we calculated the semantic similarity of the feature words in a review's feature–opinion pair using the Hownet (http://www.keenage.com), which was used to determine whether any feature words belong to the same attribute, so as to count the total number of attribute–opinion pairs.

3.3.2. Reviewer Features

Studies have shown that reviewer-related features may influence the helpfulness of his or her reviews, such as the number of total reviews the reviewer has published [29] or the reviewer's expertise [33]. Thus, based on the available data on Amazon.cn, we extracted the following three indicators to represent a reviewer's reliability features: ranking (ReviewerRank), total helpful votes (TotalVotes), and average helpful rate of all reviews (AverHelpRate).

3.3.3. Metadata Features

Some metadata features, such as a review's valence and timeliness, serve as context cues to infer a review's quality or helpfulness. Thus, we extracted the review's rating (Rating) and elapsed days from the date of review released to the date of our experiment (Timeliness) as the metadata features.

In summary, 19 features respectively belonging to text, reviewer, and metadata are finally extracted. The specific meaning and abbreviation for each feature are summarized in Table 1.

Table 1. Extracted features and abbreviation.

Factor	Feature Set	Feature Implication	Abbreviation
Text_Readability (TRD)	Linguistics	number of words in a review	Nwords
		number of sub-sentences in a review	Nsents
		Nwords divided by Nsents	Averlen
		number of adjectives in a review	Nadj
		number of adverbs in a review	Nadv
		number of verbs in a review	Nverb
	Structure	average entropy of sub-sentences in a review	Entropy
		average perplexity of sub-sentences in a review	Perplexity
Text_Reliability (TRL)	Sentiment	positive sentiment orientation of a review	PosSenti
		the degree of mixing sentiment of a review	DevPos
	Writing Style	objective degree of a review	ObjSenti
		the degree of mixing style of a review	DevObj
Text_Relevancy (TRE)	Opinion Depth	number of feature-opinion pair in a review	FeatureFreq
	Opinion Diversity	number of attribute-opinion pair in a review	AttriFreq
Reviewer	Identity and Expertise	reviewer' ranking in the community	ReviewerRank
		the total number of helpful votes a reviewer obtained	TotalVotes
		the average helpful rate of all reviews a reviewer obtained	AverHelpRate
Metadata	Review Valence	review rating	Rating
	Timeliness	elapsed days from review published date to the experiment date (the logarithm value)	Timeliness

3.4. Modeling, Evaluation, and Interpretation

Different from common ensemble techniques, such as random forest which relies on simple averaging of models in the ensemble, the core idea of gradient boosting techniques is to add new base-learners to the ensemble sequentially. In doing so, prediction performance of the ensemble model

is improved through such additive base-learners by putting emphasis on the training data that are difficult to estimate. Despite the recent popularity of deep learning, boosting algorithms are more useful in the regime of limited training data, training time, and expertise for parameter tuning when compared to deep learning models. Thus, we employed the state-of-the-art gradient boosting trees models, namely GBDT, XGBoost, and LightGBM, for our modeling job.

3.4.1. Helpfulness Modeling

- GBDT

The original gradient boosting algorithm is proposed by Friedman [9]. Given training data X, m iteration steps, a base-learner function as $g(X)$, and a specific loss function $L(y, f_m(X))$, the model updating equation $f_m(X)$ and gradient descent step size ρ_m are formulated as follows:

$$f_m(X) = f_{m-1}(X) + \rho_m g_m(X) \tag{8}$$

$$\rho_m = \text{argmin}_\rho \sum_{i=1}^{n} L(y_i, f_{m-1}(X_i) + \rho g_m(X_i)) \tag{9}$$

A particular gradient boosting model can be designed with a diverse set of base-learners, which can be classified into three different distinct categories: linear models, smooth models, and decision tree-based models [42]. Of which, decision tree-based ensembles are most frequently used in practice, such as the initial format, gradient boosting decision tree (GBDT), which demonstrates excellent performance in fitting the relationship between multiple heterogeneous input features and target variables. Moreover, tree-based ensembles can capture the influences of variables and their interactions without explicitly defining them in a computationally feasible way. The idea behind it is attributed to the structure of the DT, where each node is split at the most informative feature, and the space of input variables is partitioned into homogenous areas with an if-then rule. Such property makes the tree-based ensemble suitable for helpfulness prediction, thus enabling us to better understand our prediction model.

- XGBoost

XGBoost was recently proposed by Chen and Guestrinis [10]. Based on the original framework of gradient boosting [9], it uses K additive trees to approximate the output $\hat{y}_i$ as the following:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), f_k \in F \tag{10}$$

Here, f_k is an independent Classification and Regression Tree (CART) at each of the k steps which maps the input variables X_i to y_i, and F is the space of functions containing all CARTs. Different from the original gradient boosting algorithm, XGBoost aims at minimizing the regularized objective function defined as below:

$$Obj = \sum_{i=1}^{n} \iota(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), f_k \in F \tag{11}$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$. The regularized objective function contains two parts: the training loss function ι and the regularization term Ω. The training loss ι measures the difference between the predicted value $\hat{y}_i$ and the true value y_i. The regularization term Ω measures the complexity of model, which helps to smooth the final learnt weight to avoid overfitting.

XGBoost also introduces two important techniques, namely shrinkage and column subsampling. Shrinkage technique scales the newly added weights at each step of boosting, thus helping to reduce the influence of each tree and overfitting as well. Column subsampling chooses only a random subset of input features in building a given tree, for speeding up the training process [43].

- LightGBM

Although a few effective optimizations have been adopted in XGBoost, the efficiency and scalability are still unsatisfactory in the case of high feature dimensions and large data size. To alleviate the problem, Ke et al. [11] proposed a novel algorithm based on gradient boosting trees, namely LightGBM.

Conventional tree-based gradient boosting models need to scan all the data instances and then determine the split points by estimating the information gain, leading to computational complexities proportionally increasing to both the number of features and the number of instances. LightGBM utilizes two novel techniques, named as GOSS (Gradient-based One-side Sampling) and EFB (Exclusive Feature Bundling) methods, to reduce the number of data instances and the number of features, which help speed up the training process of boosting over 20 times as original GBDT algorithm while achieving almost the same accuracy. Notably, GOSS keeps all the instances with large gradients and performs random sampling on the instances with small gradients, since small gradients imply their training errors are small and are already well trained. To avoid changing the original data distribution by much, GOSS also introduces a constant multiplier $\frac{1-a}{b}$ to amplify the sampled data with small gradients when calculating the information gain. Additionally, EFB bundles those mutually exclusive features in the sparse feature space into a single feature by a greedy algorithm to reduce the number of features without hurting the accuracy of split point determination by much.

3.4.2. Model Evaluation

- Baseline Models

Two families of ensemble techniques, Bagging [44,45] and Adaboost [46], are often combined with a given learning algorithm to improve their performance and robustness in applications. For performance comparisons, we introduced popular ensemble models as baseline ensemble models, including Bagging, Adaboost techniques with DT, LR, and SVM as base learners, respectively, namely Bagging-DT, Bagging-LR, Adaboost-DT, Adaboost-LR, and Bagging-SVM. Excellent ensembles such as RF [47] and ExtraTrees [48] are also included.

- Evaluation Metrics

Three commonly used evaluation metrics are adopted in this research to measure the performance of the GBT models and baseline models, including Accuracy (ACC), F1-measure (F1) and AUC. Both datasets adopt a five-fold cross validation to calculate the average of such three metrics for model performance comparison.

After a classification task is completed, samples are divided into four parts: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). The four parts of samples can be presented in the confusion matrix as shown in Table 2.

Table 2. Confusion matrix.

True Condition	**Predicted Condition**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

ACC is the ratio of the number of the corrected samples to the total number of samples, which is defined as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

F1-measure, also known as F-score, is the weighted harmonic average of Precision and Recall and is often used to evaluate the quality of classification models. Precision, Recall, and F1-measure are defined as the following:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - measure = \frac{\text{Recall} * \text{Precision} * 2}{\text{Recall} + \text{Precision}} \tag{15}$$

AUC (area under the curve) is a metric which computes the area under the receiver operating characteristic (ROC) curve. A ROC curve is a graphical plot which is created by plotting the TPR (TPR = $\frac{TP}{TP+FN}$) vs. the FPR (FPR = 1 − TPR). It shows the performance of a binary classifier at various threshold settings. By computing the area under the roc curve, the curve information is summarized in one value. The larger the value, the better the classifier's performance is.

3.4.3. Model Interpretation

- SHAP Values

Due to the specific structure of DT as base learner in the gradient boosting tree models, it is straightforward to obtain the valuable features through the trained model. Specifically, every node in a DT is a condition on a single feature designed to split the dataset. The measure based on which the locally optimal condition is chosen is either Gini impurity or information gain/entropy for classification task. Thus, according to this measure, feature importance can be ranked by the averaged impurity decrease from each feature over all the trees in the ensemble.

However, the ranking of feature importance found by the model is not enough to explain an individual prediction. For example, we have no idea about why the model makes a 'helpful' prediction for a review and how each feature contributes to the final outcome.

To gain insight into how individuals evaluate review helpfulness, we turned to the feature attribution methods, in which the explanatory model g is a linear function of the feature attribution values:

$$g(z' = \phi_0 + \sum\nolimits_{i=1}^{M} \phi_i z_i') \tag{16}$$

where M is the number of features, ϕ_i is the feature attribution value of feature i, and z_i' represents feature i being observed ($z_i' = 1$) or not ($z_i' = 0$). We regarded the feature attribution value as 'feature contribution'.

Intuitively, the model g can be used to interpret both a single prediction and the entire model based on the average feature attribution across all the observations. Thus, it is suitable to interpret our review helpfulness prediction model.

A thing to note is how to calculate the value of ϕ_i in Equation (16). Lundberg et al. [7] recently proposed SHAP values as a measure of feature attribution value based on a unification of ideas from game theory. Given a model f, and a set S containing non-zero indexes in z', the classic Shapley values attribute ϕ_i to each feature can be formulated as follows:

$$\phi_i = \sum\nolimits_{S \in N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f(S \cup \{i\} - f(S))] \tag{17}$$

where N is the set of all input features.

Due to the challenge of estimating ϕ_i in the above equation with traditional Shapley values, Lundberg et al. [7] introduced a tree SHAP value estimation algorithm (SHAP values) for GBT models. More details of tree SHAP algorithm can be referred to [7]. As shown in [7], the SHAP value method is the only consistent and locally accurate individualized feature attribution method when compared to

Saabas method [49]. It also shows consistent results in the global feature attribution across all data instances when compared to Gain and Split. Therefore, we used the Tree SHAP algorithm to explain the feature contribution to the individual review helpfulness and the helpfulness prediction model.

- SHAP Interaction Values

As some specific features are predictive in conjunction with the other features, measuring the interactions between features is a problem that cannot be ignored. Different from the contribution of an individual feature described in the previous section, the contribution of feature interactions is called joint feature contributions. Based on the Shapley interaction index [7], which follows from similar axioms as SHAP values, the joint feature contribution value $\phi_{i,j}$ between feature i and j can be obtained as follows:

$$\phi_{i,j} = \sum_{S \in N\{i,j\}} \frac{|S|!(M-|S|-Z)!}{Z(M-1)!} \nabla_{i,j}(S), \tag{18}$$

when $i \neq j$, and

$$\nabla_{i,j}(S) = f(S \cup \{i,j\}) - f(S \cup \{i\}) - f(S \cup \{j\}) + f(S) \tag{19}$$

Since it is relatively easy to capture the pairwise relationship between joint features in GBT models, we further quantified the feature interactions by Equations (18) and (19), thus enabling the estimation of the joint contribution of interactive features on review helpfulness model.

4. Experiment Set-up

4.1. Dataset

We collected products reviews and related data of headsets and facial cleanser from Amazon.cn to build the experiment dataset. There were more than 50 reviews for each product correspondingly. The data collection period was from January 2016 to April 2016. Each review of the two datasets was then labelled as helpful or unhelpful based on the total number of helpful votes it received. To be specific, the reviews with five or more useful votes were labelled as helpful, while those that did not receive useful votes were labelled as unhelpful. To check the validity of the 'helpfulness' label, we randomly picked about 2500 reviews respectively from the helpful and unhelpful (to ensure balance) to build the annotation datasets for both categories of products. Two project team members were invited to annotate each review of the two annotation datasets without labels. Each member read the content of each review independently with visible information of the reviewer and metadata, and then they answered the question "whether the review is helpful for your purchase?". By calculating the Kappa statistic, the annotation result of a member was compared with the result of its helpfulness label. The calculation results were 82.9% and 84.2%, respectively, indicating the reviews with 'helpful vote' selected in our annotation datasets were ideal for our experiment.

After preprocessing, the preliminary statistical results of some intuitive features among the two datasets are presented in Table 3.

In total, there were 2406 helpful reviews and 2594 unhelpful reviews in the headset dataset, and 2515 helpful reviews and 2485 unhelpful reviews in the facial cleanser dataset. On average, facial cleanser reviews were more detailed than headset reviews with longer review words (43.5 vs. 31.2), more product features embedded in the reviews (3.6 vs. 2.8), and more information entropy (8.14 vs. 7.96). As for the perspective of review sentiment, headset reviews were more positive and objective than facial cleanser reviews, showing a higher degree of positive sentiments (0.64 vs. 0.55), a higher degree of objective sentiments (0.42 vs. 0.38), and a higher rating (4.25 vs. 4.13). It can also be observed that there were no significant differences between the two kinds of product reviews in terms of average sentence length, published date, and reviewer ranking.

Table 3. Description of headset and facial cleanser datasets.

Feature Item	Headset Reviews	Facial Cleanser Reviews
helpful/helpless	2406/2594	2515/2485
average review length	31.2	43.5
average number of sentences	6.7	9.3
average sentence length	7.2	7.1
average positive degree	0.64	0.55
average objective degree	0.42	0.38
average entropy	7.96	8.14
average feature frequency	2.8	3.6
average elapsed days	318	320
average reviewer rank	794,000	786,000
average rating	4.25	4.13

(Source of data: https://www.amazon.cn).

4.2. Hyper-Parameter Tuning

Normally, machine learning algorithms have a few dozen hyper-parameters needing to be configured prior to model training. Hyper-parameter configurations have a significant impact on model performance, especially for GBDT, XGBoost, or LightGBM, which have a substantial number of hyper-parameters. Recently, the strategy of sequential model-based optimization (SMBO) [50] has shown to be a better alternative to grid search for optimizing the hyper-parameters of machine learning algorithms [12,51]. The basis for hyper-parameters tuning is to optimize a mapping function over a configuration space, which specifies the hyper-parameter values to be explored for each hyper-parameter. We employed a SMBO method to tune the hyper-parameters for the baseline models and GBT models. A five-fold cross-validation accuracy was used to find the optimal hyper-parameter setting and the corresponding loss, and the total number of evaluations 'n_calls' or the number of iterations in bagging or Adaboost were all set to 100.

5. Experimental Results

5.1. Comparison of Performance for Review Helpfulness Prediction

On the basis of the five types of feature sets extracted from the two categories of datasets, the performances of the baseline models and GBT models were then compared by a five-fold cross-validation on the training set. All models were optimized with their corresponding optimal parameters. The results are summarized in Tables 4 and 5.

Table 4. Model comparison results on headset dataset.

Type	Model	ACC (%)	F1 (%)	AUC
Ordinary Models	DT	74.24	73.17	0.8248
	LR	72.37	68.77	0.7720
	SVM	54.60	61.85	0.5969
Baseline Ensemble Models	Bagging-DT	75.80	75.19	0.8423
	AdaBoost-DT	73.17	71.74	0.8047
	Bagging-LR	72.47	68.74	0.7720
	AdaBoost-LR	74.72	71.25	0.8152
	Bagging-SVM	62.90	50.00	0.5528
	RF	75.77	75.83	0.8406
	ExtraRF	75.77	74.99	0.8376
GBT Models	GBDT	76.77	75.94	0.8495
	XGBoost	77.25	76.50	0.8498
	LightGBM	76.45	75.76	0.8416

(Source of data: https://www.amazon.cn).

Table 5. Model comparison results on Facial cleanser dataset.

Type	Model	ACC (%)	F1 (%)	AUC
Ordinary Models	DT	68.18	69.50	0.7443
	LR	63.74	63.41	0.6798
	SVM	51.41	55.79	0.5438
Baseline Ensemble models	Bagging-DT	69.49	70.31	0.7618
	AdaBoost-DT	66.50	67.21	0.7215
	Bagging-LR	67.56	64.94	0.7307
	AdaBoost-LR	68.38	66.13	0.7571
	Bagging-SVM	54.17	50.30	0.5421
	RF	69.93	70.52	0.7674
	ExtraRF	68.99	70.13	0.7615
GBT Models	GBDT	70.10	70.28	0.7742
	XGBoost	70.17	70.70	0.7762
	LightGBM	68.99	69.74	0.7657

(Source of data: https://www.amazon.cn).

From Tables 4 and 5, the following consistent conclusions can be found.

First, ordinary models such as DT or LR with some ensemble techniques performed better than their corresponding ordinary models. For example, Bagging-DT using the bagging technique on decision tree (DT) outperformed DT in terms of performance, with ACC, F1, and AUC increased by 1.56%, 2.02%, 0.0175, and 1.31%, 0.81%, and 0.0175, respectively, for the two types of datasets. Similarly, the performance of ordinary logistic regression (LR) can be improved greatly when ensembled as AdaBoost-LR with the ACC, F1, and AUC increased by 2.35%, 2.48%, 0.0432 and 4.64%, 2.72%, and 0.073, respectively, on the two types of datasets. The results imply the effectiveness of ensemble techniques. The results also indicate that the performances of SVM in the ordinary model and ensembles were worse than those of DT or LR.

Second, RF model performed excellently, and its result was the most consistent among all the baseline ensemble models on both types of datasets. Notably, for the headset dataset, RF performance was comparable to Bagging-DT, and for the facial cleanser dataset, RF was the best performing model of all the baseline models with 69.93% ACC, 70.62% F1, and 0.7674 AUC.

Third, by comparing the GBT models with the best benchmark ensemble model RF, it can be inferred that XGBoost model performed best in the GBT model group. In particular for the headset dataset, XGBoost had a strong advantage than RF with ACC and F1 increased by 1.48% and 0.67%, and for the facial cleanser dataset, XGBoost also demonstrated a weak advantage in the evaluation indicators, with ACC, F1, and AUC increased by 0.01%, 0.07%, and 0.01. As for GBDT, it was slightly ahead of RF on the headset dataset with ACC, F1, and AUC increased by 1%, 0.11%, and 0.009. Meanwhile, GBDT was comparable to RF on the facial cleanser dataset with ACC and AUC increased by 0.17% and 0.0068, but F1 decreased by −0.24%. Therefore, in terms of overall performance, GBDT still slightly outperformed RF in our experiment datasets. Compared to the results of LightGBM and RF, it can be found that LightGBM did not demonstrate significant performance advantages over RF. For the headset dataset, LightGBM only had the ACC increased by 0.68%, while for the facial cleanser dataset, LightGBM was slightly weaker than RF with ACC, F1, and AUC decreased by 0.94%, 0.78%, and 0.0017, respectively.

The performance comparison results reveal that GBT models achieved better prediction results for the review helpfulness prediction problem compared to ordinary and baseline ensemble models, especially as implemented in XGBoost. Therefore, XGBoost was adopted for the feature analysis and model interpretation of review helpfulness prediction in the following sections.

5.2. Analysis of Feature Validity

This section examines the validity of the feature sets generated in our research. As can be seen from Table 1, five broad feature sets were extracted to represent reviews' helpfulness: (1) text-readability features (TRD), (2) text-reliability features (TRL), (3) text-relevancy features (TRV), (4) reviewer features, and (5) metadata features. The first three are text-related features. To examine their validities, we first used each subset of the text-related features to build prediction models, and then we used all the text-related features (Text-ALL) to build models again. Moreover, we built the subsequent models by incrementally adding other sets of features. Meanwhile, we adopted a baseline feature set used in [29] for comparison (Baseline). We evaluated each model in the same way as above, using five-fold cross validation and reporting the metrics of ACC, F1, and ROC. The evaluations of feature validity are presented in Table 6.

Table 6. Evaluations of the feature validity.

Dataset	Feature Set	ACC (%)	F1 (%)	ROC
Headset	Baseline	74.28	74.27	0.8298
	Text-TRD	75.37	75.17	0.8197
	Text-TRL	71.32	69.47	0.7683
	Text-TRV	71.12	66.12	0.7740
	Text-ALL	75.44	75.33	0.8329
	+Reviewer	75.92	75.63	0.8374
	+Meta-data	77.25	76.50	0.8498
Facial Cleanser	Baseline	67.84	68.15	0.7319
	Text-TRD	67.04	68.08	0.7270
	Text-TRL	65.22	66.48	0.7111
	Text-TRV	67.32	68.28	0.7394
	Text-ALL	69.68	70.20	0.7651
	+Reviewer	69.79	70.22	0.7691
	+Meta-data	70.17	70.70	0.7762

(Source of data: https://www.amazon.cn).

The results reveal that our total feature sets were superior to the baseline feature set with ACC, F1, and ROC increased by 2.97%, 2.23%, 0.02 and 2.33%, 2.55%, and 0.04, respectively, across two kinds of datasets. Since the main differences between our feature sets and the baseline feature sets were text structural features and relevancy features, the results indicate these features were beneficial to improve the helpfulness of the model's performance.

Another finding is that using any subset of text-related features resulted in a subtle difference in model performance, while using all available subsets of text-related features obtained optimal results. This seems to imply two possible explanations. One explanation is that the three feature sets representing readability, reliability, and relevancy contain some collinear features; thus, some features are interchangeable. This finding is also supported in [29]. Another explanation is that there may be interactions between certain text-related features, which result in a slight improvement in model performance when adopting the total feature sets. Therefore, the next section discusses feature redundancy and the interactions between features.

5.3. Results of Feature Contribution on Helpfulness

5.3.1. Global Feature Contribution

To better interpret our optimal helpfulness model implemented in XGBoost, we applied the Tree SHAP method [7], which has proven to be a powerful tool for confidently interpreting GBT models. Tree SHAP first measures the contribution each feature has on the model output (Tree SHAP values) for individuals in the training dataset. Then, the global feature contributions are ranked according to

the mean (|Tree SHAP|) across all samples. The global feature contributions derived from the XGBoost helpfulness model are shown in Figure 3.

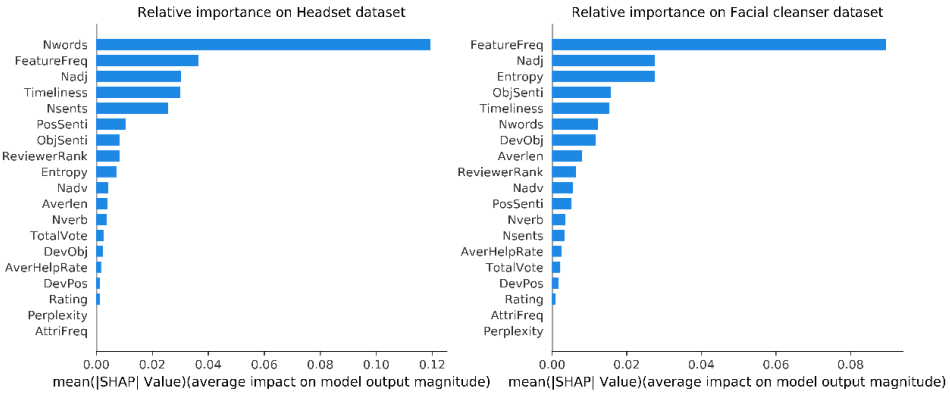

Figure 3. Global feature attribution across two datasets. (Source of data: https://www.amazon.cn).

In Figure 3, the *x*-axis is essentially the average magnitude change in model output when a feature is integrated out of the model. The features are ordered by the absolute sum value of their effect magnitudes on the model. It can be first inferred that feature contributions varied across different product categories, with some specific features contributing far more than other features. For example, for the headset dataset, Nwords dominated the other features, while for the facial cleanser dataset FeatureFreq stood out as the most import predictor. Both Nwords and FeatureFreq contributed significantly to the model outputs. The results reflect that, on the whole, headset consumers were more concerned with the content elaborateness, while facial cleanser consumers were more interested in the evaluation of some attributes of the product. Another finding is that in our extracted features, some contributed little or nothing to the model outputs, such as AttriFreq, Perplexity, Rating, and DevPos. Thus, it is necessary to re-evaluate the model performance to decide whether to exclude such low contribution or no contribution features. This issue will be discussed in the following section.

5.3.2. Individual Feature Contribution

As Tree SHAP values are derived from an individualized model interpretation approach, an individualized interpretation for each sample can be obtained from the model. Figure 4 presents some insights into how the contribution of an individual feature on the model output is affected by its value. The x position of the dot is the impact of the feature on the review helpfulness, and the color of the dot represents the value of that feature for the review.

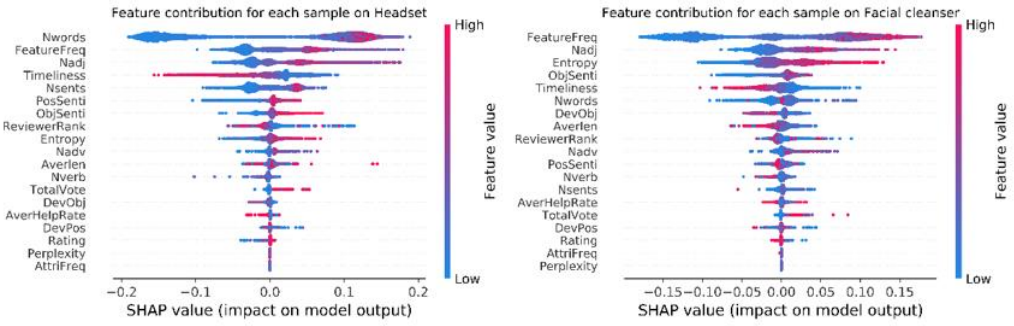

Figure 4. Comparison of individual feature contributions. (Source of data: https://www.amazon.cn).

Figure 4 reveals that there is a certain degree of linear relationship between the contributions of some features and their values. We can take the features that contributed the most to the two datasets respectively as examples. For the headset reviews, the more positive (PosSenti), the more

objective (ObjSenti), and the more unique expression (Entropy) were easier to be understood as useful. This trend was only obvious in the Entropy feature of the facial cleanser dataset. In the opposite, the older (Timeliness) of the review, the less likely it was to be helpful for both kinds of reviews. For most of the remaining features, the relationships between their contributions and their value were non-linear, such as the number of words (Nwords) and frequency of features (FeatureFreq) in the headset dataset, or FeatureFreq and the number of adjectives (Nadj) in the facial cleanser dataset.

Another finding is that individual feature contributions varied across reviews, even with the same feature values, reflecting as a broad range of impacts on the model output in Figure 4, such as some high-contribution features like Nwords, FeatureFreq, and Nadj in the headset dataset, or FeatureFreq, Nadj, and Entropy in the facial cleanser dataset. This may imply that other features may influence the impacts of these high-contribution features; thus, it is necessary to capture the joint contributions of these features from the model.

5.4. Feature Reduction Based on Global Contribution

The unimportant review features were determined according to the reverse order of the global feature contribution obtained in Section 5.3.1. They were removed one by one from the XGBoost model, and at the same time the 5-fold cross validation(cv) ACC, F1, and AUC results were recalculated accordingly. Figure 5 shows the performances of the model on the remaining features after removing features in turn on both types of datasets.

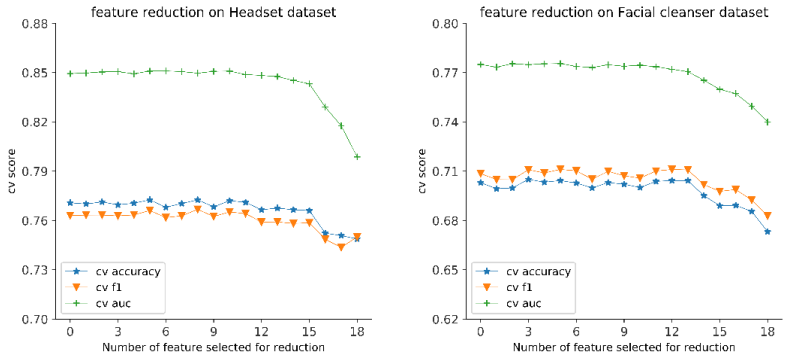

Figure 5. Feature reduction comparison on headset and facial cleanser datasets. (Source of data: https://www.amazon.cn).

The first finding across two datasets is that eliminating a certain number of low-contribution features does not bring about obvious changes in model performances. For the headset dataset, it can be observed that until the top fifteen low-contribution features were selected to be removed, the model showed a downward trend on the indicators ACC, F1, and AUC. For the facial cleanser dataset, the corresponding number of features to be removed was thirteen. The findings imply that effective feature selection can be performed according to the global contribution of features. Another finding is that only a few features played decisive roles in assessing the helpfulness of reviews. Overall, the key features that affected the helpfulness of headset reviews were the following four features in turn: the number of words (Nwords), frequency of product features (FeatureFreq), the number of adjectives (Nadj), and elapsed days (Timeliness); the key features that influenced the helpfulness of facial cleanser reviews corresponded to the frequency of product features (FeatureFreq), the number of adjectives (Nadj), expression uniqueness (Entropy), objective sentiment degree (ObjSenti), the number of words (Nwords), and elapsed days (Timeliness).

5.5. Results of Joint Feature Contribution on Review Helpfulness

5.5.1. Global Joint Feature Contribution

The SHAP interaction values method was used to automatically capture the joint feature contributions embedded in the features, and then the contribution of each feature on the model output was decomposed into two-part effects: the main contribution of a feature itself and the joint contributions between the feature and other features. Figure 6 shows the ranking results of the main feature contribution and joint feature contribution with the proportion of feature contribution to the total contribution.

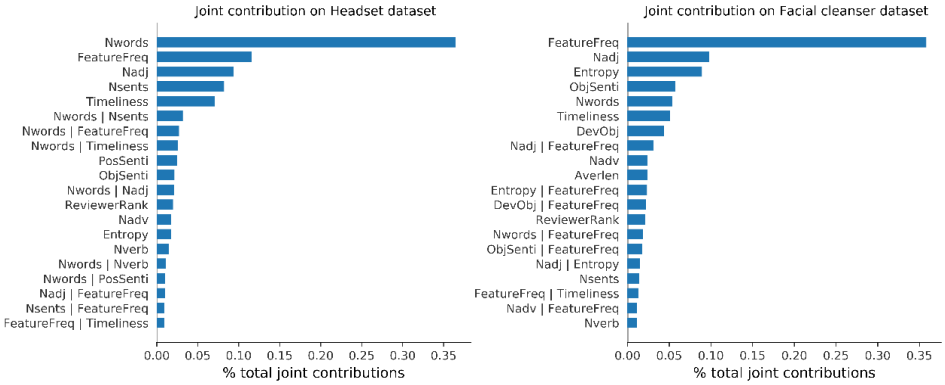

Figure 6. Joint feature contributions across two datasets. (Source of data: https://www.amazon.cn).

After stripping off the main feature contribution, more joint contributions measured from the headset dataset occurred between Nwords and Nsents, Nwords and FeatureFreq, Nwords and Timeliness, as well as Nwords and Nadj. It means the impact of the number of words in a review on review helpfulness was interactively affected by features such as the number of sentences, the feature frequency, elapsed days, and the number of adjectives. Therefore, it is concluded that for the headset buyers, the more acceptable reviews are usually those with a certain length, many sub-sentences, some opinions on the product features, and also published recently.

As for the facial cleanser dataset, more joint contributions occurred between FeatureFreq and Nadj, FeatureFreq and Entropy, FeatureFreq and Devobj, as well as FeatureFreq and Nwords. It means the impact of prominent features such as the number of product features in a review on review helpfulness is affected by features like the number of adjectives, expression uniqueness, objective sentiment deviation, and the number of words. It can also be inferred that the reviews that are more acceptable to facial cleanser buyers usually contain product feature opinions, unique expression structures, objective sentiment orientations different from the average level, and review text with a certain length. Moreover, according to the ranking results of main feature contribution, it can be found that the more prominent the main contribution of a feature is, the more likely it is to jointly contribute to other important features.

5.5.2. Individual Joint Feature Contribution

To observe how a prominent feature interacts with other features, we further mapped the value of the prominent feature against its SHAP value in the samples of the whole dataset and colored the value of several other features with strong interactions on the prominent feature.

Figure 7 demonstrates that for the headset dataset, even if most reviews' lengths (Nwords) were less than 70, the extent to which length impacts the prediction differed, as shown by the obvious vertical dispersion of dots at a length less than 70. This means other features affect the contribution of Nwords. When the word length exceeded 70, the interaction effect was significantly reduced because the vertical dispersion of the dots was distinctly reduced. Based on the distribution of the red sample

dots above the Y-axis 0.0 and the value color of Nsents, FeatureFreq, Nadj, and Nverb on the sample dots, it can be concluded that the greater the number of sub-sentences, product feature frequency, adjectives, and adverbs there are, the more positively Nwords will contribute. Meanwhile, the dots color of Timeliness shows that Timeliness lowered the contribution of Nwords with a length above 70. This means that the earlier the review is published, the more likely it is to reduce the effect of Nwords on review helpfulness. By the dots color of PosSenti, it shows that PosSenti increased the contribution of Nwords with a length above 70. It means that the more positive the review, the more likely it is to increase the effect of Nwords on review helpfulness with a review length above 70.

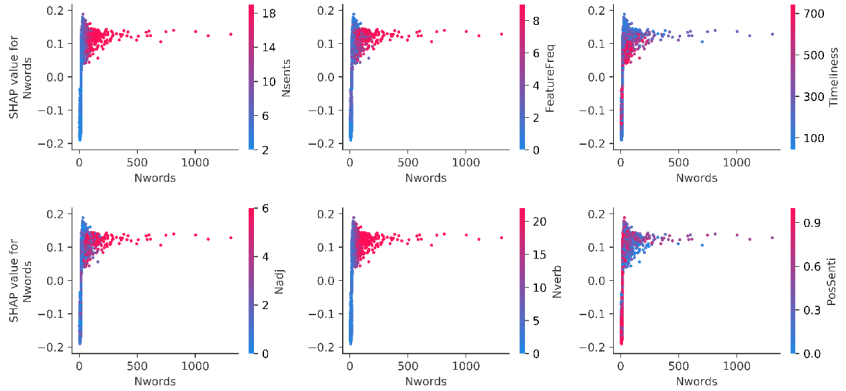

Figure 7. Individual joint feature contribution on the headset dataset. (Source of data: https://www.amazon.cn).

As for the facial cleanser dataset, similar findings can be found from the obvious dispersion change trend of dots on the FeatureFreq in Figure 8. As the frequency of product feature contained in a review increased, the interaction effect acts on FeatureFreq decreased significantly. According to the dispersion of the red sample dots above the Y-axis 0.0 and the value color of Nadj, Nwords, Entropy, and ObjSenti on the sample dots, it can be inferred that the greater the number of adjectives (Nadj) and words (Nwords), the more unique expression (Entropy), and the more objective (ObjSenti) in sentence expression, the more positive impacts on FeatureFreq's contribution will produce. By the dots color of DevObj, it moderately shows that a greater objective sentiment deviation increased FeatureFreq's contribution. It means that the greater the difference between the objective sentiment of the review and the average level, the more likely it is to increase the effect of FeatureFreq on review helpfulness. By the dots color of Timeliness, it shows most dots with positive contributions (greater than 0.0 on the y-axis) had a small Timeliness value. This finding is consistent with headset dataset, which implies that the more recent the review date, the more likely FeatureFreq is to increase review helpfulness.

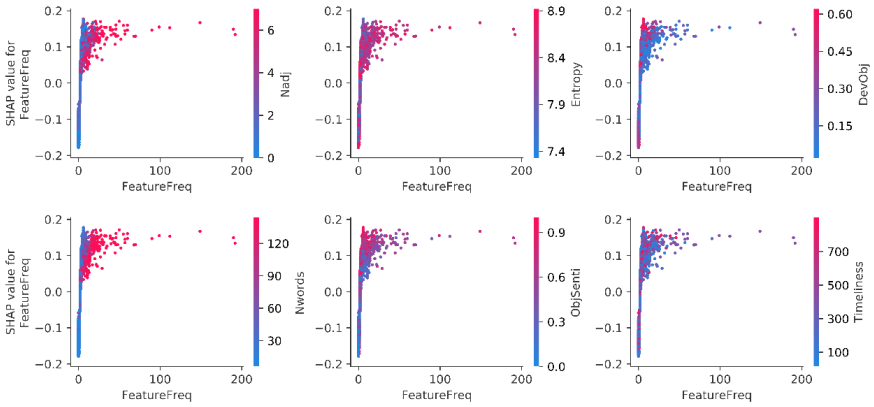

Figure 8. Individual joint feature contribution on the facial cleanser dataset. (Source of data: https://www.amazon.cn).

6. Further Discussion

To begin with, the above research results indicate that text-related review features and published time mainly affect review helpfulness. However, for different product types, the specific text-related features that play influential roles are different. From the experimental results, the features that contribute the most to the helpfulness of headset reviews are the following four features in order: the number of words (Nwords), product feature frequency (FeatureFreq), the number of adjectives (Nadj), and the elapsed days from review published date to experiment date (Timeliness). It can be inferred that when consumers purchase headset products, the reviews they want to refer to generally include more detailed product descriptions, especially the feature opinions on certain aspects of the headset, and are published recently. In contrast, the features that contribute the most to the helpfulness of facial cleanser reviews are the following features in order: product feature frequency (FeatureFreq), the number of adjectives (Nadj), information entropy(Entropy), objective sentiment degree (ObjSenti), the number of words (Nwords), and the elapsed days from review published date to experiment date (Timeliness). This indicates when consumers purchase facial cleanser, they not only pay attention to the opinions, product details, or date of the reviews, but they also pay attention to the objectivity and expression uniqueness of the reviews, and they may prefer more objective and personalized reviews.

Additionally, the experimental results also show that both the interactions between text-related features and between text-related feature and review date indeed affect review helpfulness. The more the features have impacts on the review helpfulness, the greater the interactions with other features. The results of this study show that the number of words is the most important feature affecting the helpfulness of headset reviews, and it has the most obvious interactions with features like the number of sentences, the number of product features, the number of adjectives, and review published date. A similar finding is also found in the facial cleanser dataset, where the feature frequency has the greatest impact on review helpfulness, and also has the greatest interactions with the number of adjectives, information entropy of the review, objectivity of the review, and the review length.

Finally, from the experiment results, it can be found that the influences of most features on review helpfulness show non-linear and dynamic trends, which can be seen from the feature contributions and interactive feature contributions of individual reviews. Notably, with the increase in feature value, the effect of interactions between features on review helpfulness decreases gradually.

Similar to the research related to online review helpfulness mentioned in the literature review, this paper also verifies the main influence of text-related features and timeliness of reviews on review helpfulness, such as review length, product features frequency, sentiment degree, and writing style. However, the experiment results contain more differences. First of all, this study extracts the syntactic structure (expression uniqueness) feature of text, namely entropy, and verifies that expression uniqueness plays an important role, which expands the extraction scope and form of review text features. This suggests that it is not only necessary to examine the character or word features of review text, but it also necessary to consider the structure and collocation between characters or words. Secondly, this study not only identifies the important features impacting review helpfulness, but it also further quantifies the feature importance, that is, the contribution of each feature, and analyzes the impact of each feature on review helpfulness in detail. Finally, based on the investigation of important features, this study adds to the investigation of interactions between features. By analyzing the contributions of global joint features and individual joint features, the degree and trend of feature interactions are analyzed in detail.

7. Conclusions

In this paper, we have answered the question of why a review is helpful from both the macro and micro levels by measuring the feature contribution with SHAP values and SHAP interaction values. Combined with the optimal GBT model implemented in XGBoost to help modeling review helpfulness, we identified features that influence review helpfulness, quantified those feature contributions, and automatically captured the interaction effects between them. Through experimental analysis on

two types of datasets, this paper reveals the main feature contribution and joint feature contribution of headset reviews and facial cleanser reviews. The results of the comparative analysis and visual analysis on multiple groups of experiments explain the formation mechanism of helpful reviews for the two kinds of products. Some meaningful conclusions can be drawn from our experiment. (1) Both datasets indicate that a few features contribute the most, while most features contribute less or have no contribution. Specifically, for the headset dataset, Nwords (review length) dominated the other features; while for facial cleanser dataset, FeatureFreq (feature opinion frequency in a review) stood out as the most import predictor. (2) There are indeed interactions between features. We found prominent features are easier to interact with other features. (3) By visualizing the relationships between feature values and feature contributions, we found there are linear relationships between features' contributions and their values in a few features, and most of them are nonlinear relationships. Meanwhile, by visualizing the relationship between feature values and joint feature contributions, we found the interactions between features gradually decreased with the increase of feature values.

Different from prior research which mainly focuses on explaining feature importance based on helpfulness of the model performance, this study provides a more stable, comprehensive, and detailed analysis of review helpfulness on the basis of feature contribution. Combined with the SHAP value method and XGBoost model, this study first introduces a stable calculation method for quantifying feature contributions and avoids the inconsistency problem of feature analysis caused by multiple evaluation indicators in the previous studies. Moreover, unlike the previous studies mostly focusing on examining the direct impacts of multiple features on review helpfulness, this study supplements the interactions between features, captures, and quantifies the degree of joint contribution through SHAP value method and XGBoost model. In doing so, this study reveals the influence of feature value on review helpfulness from the micro level for the first time by visualizing the relationships among feature value, feature contribution, and feature joint contribution for an individual review. Notably, this kind of influence is nonlinear in most cases, and the interaction relationship is dynamic and complex.

In the e-commerce context, enterprises are faced with many uncertainties in managing user-generated content. Based on a feature contribution-driven analysis framework on review helpfulness, this study provides some specific implications to eliminate these uncertainties, especially for IT managers, professionals, as well as academics. First, for the IT managers, the establishment of an effective and accurate online reputation management system is the most concerned issue. The review evaluation method based on feature contribution analysis proposed in this study is conducive to more accurate identification of helpful reviews, so as to establish an effective management way of online reviews based on reviews' helpfulness evaluation results. This will further help promote the value of online reviews and enabling the role of IT technology in economic performance. Second, for the professionals, they are not only concerned with the helpfulness results of online reviews, they are more concerned with what factors contribute to the evaluation results. This study analyzes and measures feature contributions and joint feature contributions on review helpfulness by analyzing readability, reliability, relevance, and metadata. In doing so, professionals could analyze the reasons for the formation of review values from the factors of multiple dimensions of reviews, so as to develop corresponding review management strategies. It will be of benefit to guide reviewers to publish useful reviews and identify reviews that are helpful to consumers. Meanwhile, professionals can further explore the valuable information in the helpful reviews to guide relevant business work, such as to conduct innovative design of new products or to implement marketing communication decisions based on helpful reviews. Third, it is always an area of focus for academics to interpret the helpfulness model. This study introduces Shapley Additive exPlanations method into the XGBoost model, so as to decompose the contribution of each feature from the helpfulness evaluation results. To our knowledge, this is the first time that an external explanatory tool has been used to explain the influence of review features on review helpfulness in relevant studies. This has a special implication for researchers to combine external interpretation tools with evaluation model to uncover the black-box nature of the traditional helpful evaluation model and improve its explanatory ability.

Limited by experimental conditions, this study only includes the reviews from two typical product types. At the same time, due to the uncertainty of online review quality, the quantitative measure model of review helpfulness and the interactions in the model need to be further improved. The feature representations of review helpfulness could also be further explored and designed.

Future research could be conducted from the following aspects. First, collect as much review data as possible, cover as much product data as possible, and verify the robustness of the experiment conclusions. Second, explore as many feature forms of review helpfulness as possible, either from review text or others, to enhance the performance of prediction model. Last, only two levels of review helpfulness (either helpful or unhelpful) are considered in this research; thus, multiple levels of review helpfulness can be further analyzed in the future.

Author Contributions: Y.M. contributed to the conception of the research, performed the experiment, and wrote the manuscript. N.Y. helped perform the data analysis with constructive discussions. Z.Q. performed the data crawling. G.Z. helped perform the data preprocessing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Shanghai Philosophy and Social Sciences Planning Project, grant number 2020BGL009.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ren, G.; Hong, T. Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews. *Inf. Process. Manag.* **2019**, *56*, 1425–1438. [CrossRef]
2. Malik, M.; Hussain, M.A. An analysis of review content and reviewer variables that contribute to review helpfulness. *Inf. Process. Manag.* **2018**, *54*, 88–104. [CrossRef]
3. Eslami, S.P.; Ghasemaghaei, M.; Hassanein, K. Which online reviews do consumers find most helpful? A multi-method investigation. *Decis. Support Syst.* **2018**, *113*, 32–42. [CrossRef]
4. Krestel, R.; Dokoohaki, N. Diversifying customer review rankings. *Neural Netw.* **2015**, *66*, 36–45. [CrossRef] [PubMed]
5. Siering, M.; Muntermann, J.; Rajagopalan, B. Explaining and predicting online review helpfulness: The role of content and reviewer-related signals. *Decis. Support Syst.* **2018**, *108*, 1–12. [CrossRef]
6. Lundberg, S.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 4768–4777.
7. Lundberg, S.M.; Erion, G.G.; Lee, S.-I. Consistent individualized feature attribution for tree ensembles. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1–9.
8. Shapley, L. A value for n-person games. *Ann. Math. Stud.* **1953**, *28*, 307–317.
9. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
10. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
11. Ke, G.L.; Meng, Q.; Finley, T.; Wang, T.F.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 3149–3157.
12. Xia, Y.; Liu, C.; Li, Y.; Liu, N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Syst. Appl.* **2017**, *78*, 225–241. [CrossRef]
13. Xiao, Z.; Wang, Y.; Fu, K.; Wu, F. Identifying Different Transportation Modes from Trajectory Data Using Tree-Based Ensemble Classifiers. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 57. [CrossRef]
14. Humphreys, A.; Wang, R.J.-H. Automated Text Analysis for Consumer Research. *J. Consum. Res.* **2017**, *44*, 1274–1306. [CrossRef]

15. Chen, X.; Sheng, J.; Wang, X.; Deng, J. Exploring Determinants of Attraction and Helpfulness of Online Product Review: A Consumer Behaviour Perspective. *Discret. Dyn. Nat. Soc.* **2016**, *2016*, 1–19. [CrossRef]
16. Krishnamoorthy, S. Linguistic features for review helpfulness prediction. *Expert Syst. Appl.* **2015**, *42*, 3751–3759. [CrossRef]
17. Hu, Y.-H.; Chen, K.; Lee, P.-J. The effect of user-controllable filters on the prediction of online hotel reviews. *Inf. Manag.* **2017**, *54*, 728–744. [CrossRef]
18. Akbarabadi, M.; Hosseini, M. Predicting the helpfulness of online customer reviews: The role of title features. *Int. J. Mark. Res.* **2018**, *62*, 272–287. [CrossRef]
19. Chen, C.C.; Tseng, Y.-D. Quality evaluation of product reviews using an information quality framework. *Decis. Support Syst.* **2011**, *50*, 755–768. [CrossRef]
20. Singh, J.P.; Irani, S.; Rana, N.P.; Dwivedi, Y.K.; Saumya, S.; Roy, P.K. Predicting the "helpfulness" of online consumer reviews. *J. Bus. Res.* **2017**, *70*, 346–355. [CrossRef]
21. Yin, D.; Bond, S.D.; Zhang, H. Keep Your Cool or Let it Out: Nonlinear Effects of Expressed Arousal on Perceptions of Consumer Reviews. *J. Mark. Res.* **2017**, *54*, 447–463. [CrossRef]
22. Moore, S.G. Attitude Predictability and Helpfulness in Online Reviews: The Role of Explained Actions and Reactions. *J. Consum. Res.* **2015**, *42*, 30–44. [CrossRef]
23. Zhang, Z.; Qi, J.; Zhu, G. Mining customer requirement from helpful online reviews. In Proceedings of the 2nd International Conference on Enterprise Systems, Shanghai, China, 2–3 August 2014; pp. 249–254.
24. Aghakhani, N.; Karimi, J.; Salehan, M. A Unified Model for the Adoption of Electronic Word of Mouth on Social Network Sites: Facebook as the Exemplar. *Int. J. Electron. Commer.* **2018**, *22*, 202–231. [CrossRef]
25. Ngo-Ye, T.L.; Sinha, A.P.; Sen, A. Predicting the helpfulness of online reviews using a scripts-enriched text regression model. *Expert Syst. Appl.* **2017**, *71*, 98–110. [CrossRef]
26. Schindler, R.M.; A Bickart, B. Perceived helpfulness of online consumer reviews: The role of message content and style. *J. Consum. Behav.* **2012**, *11*, 234–243. [CrossRef]
27. Weathers, D.; Swain, S.D.; Grover, V. Can online product reviews be more helpful? Examing characteristics of information content by product type. *Decis. Support Syst.* **2015**, *79*, 12–23. [CrossRef]
28. Liu, Y.; Jin, J.; Ji, P.; Harding, J.A.; Fung, R.Y. Identifying helpful online reviews: A product designer's perspective. *Comput. Des.* **2013**, *45*, 180–194. [CrossRef]
29. Ghose, A.; Ipeirotis, P.G. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Trans. Knowl. Data Eng.* **2010**, *23*, 1498–1512. [CrossRef]
30. Chua, A.Y.; Banerjee, S. Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. *Comput. Hum. Behav.* **2016**, *54*, 547–554. [CrossRef]
31. Li, Q.; Cui, J.; Gao, Y. The influence of social capital in an online community on online review quality in China. In Proceedings of the 48th Hawaii International Conference on System Sciences, Kauai, HI, USA, 5–8 January 2015; pp. 562–570.
32. Fang, B.; Ye, Q.; Kucukusta, D.; Law, R. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tour. Manag.* **2016**, *52*, 498–506. [CrossRef]
33. Hong, H.; Xu, D. Research of online review helpfulness based on negative binary regress model. In Proceedings of the 12th International Conference on Service Systems and Service Management (ICSSSM), Guangzhou, China, 22–24 June 2015; pp. 1–5.
34. Malik, M.; Hussain, A. Helpfulness of product reviews as a function of discrete positive and negative emotions. *Comput. Hum. Behav.* **2017**, *73*, 290–302. [CrossRef]
35. Wang, R.Y.; Reddy, M.P.; Kon, H.B. Toward quality data: An attribute-based approach. *Decis. Support Syst.* **1995**, *13*, 349–372. [CrossRef]
36. Otterbacher, J. Helpfulness in online communities: A measure of message quality. In Proceedings of the 27th International Conference on Human Factors in Computing Systems, Boston, MA, USA, 4–9 April 2009; pp. 955–964.
37. Rausser, G.C.; Simon, L.; Zhao, J. Rational exaggeration and counter-exaggeration in information aggregation games. *Econ. Theory* **2015**, *59*, 109–146. [CrossRef]
38. Zhang, Z.; Ye, Q.; Zhang, Z.; Li, Y. Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Syst. Appl.* **2011**, *38*, 7674–7682. [CrossRef]

39. Pang, B.B.; Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004; pp. 271–278.
40. K Liu, K.; Xu, L.; Zhao, J. Extracting opinion targets and opinion words from online reviews with graph co-ranking. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistic (ACL), Baltimore, MD, USA, 23–25 June 2014; pp. 314–324.
41. Hai, Z.; Chang, K.; Cong, G.; Yang, C.C. An Association-Based Unified Framework for Mining Features and Opinion Words. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*, 1–21. [CrossRef]
42. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **2013**, *7*, 21. [CrossRef] [PubMed]
43. Sheridan, R.P.; Wang, W.-M.; Liaw, A.; Ma, J.; Gifford, E.M. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353–2360. [CrossRef] [PubMed]
44. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 23–140. [CrossRef]
45. Breiman, L. Pasting Small Votes for Classification in Large Databases and On-Line. *Mach. Learn.* **1999**, *36*, 85–103. [CrossRef]
46. Freund, Y.; E Schapire, R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
47. Breiman, L. Random forest. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
49. Saabas, A. Interpreting Random Forests. August 2015. Available online: http://blog.datadive.net/random-forest-interpretation-with-scikit-learn/ (accessed on 5 September 2018).
50. Hutter, F.; Hoos, H.H.; Leyton-Brown, K. Sequential model-based optimization for general algorithm configuration. In Proceedings of the 5th International Conference on Learning and Intelligent Optimization, Rome, Italy, 17–21 January 2011; pp. 507–523.
51. Levesque, J.-C.; Gagne, C.; Sabourin, R. Bayesian hyperparameter optimization for ensemble learning. In Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence, New York, NY, USA, 25–29 June 2016; pp. 437–446.