Using XGBoost and Skip-Gram Model to Predict Online Review Popularity

Lien Thi Kim Nguyen1, Hao-Hsuan Chung2, Kristine Velasquez Tuliao1, and Tom M. Y. Lin1

Abstract
Review popularity is similar to awareness and information accessibility components: Both have a profound effect on customer purchase decisions. Therefore, this study proposes a new method for predicting online review popularity that combines the extreme gradient boosting tree algorithm (XGBoost), to extract key features on the bases of ranking scores and the skip-gram model, which can subsequently identify semantic words according to key textual terms. Findings revealed that written reviews had higher review popularity than non-textual reviews (reviewer and product factors). Moreover, the proposed method achieved higher prediction accuracy than the traditional ridge regression technique of Root Mean Squared Logarithmic Error (RMSLE). The main factors affecting review popularity and key reviewers for specific textual terms were also identified. Findings could help vendors identify key influencers for their product promotion and then support the design of word-suggestion systems for online reviews.

Keywords
online word of mouth, review popularity, extreme gradient boosting tree algorithm, skip-gram model, predictive models

Introduction
In recent years, together with the rapid growth of the internet and web 2.0 applications, numerous online products related to word-of-mouth (WOM) communities have emerged that provide user-generated content (e.g., online product review) conveniently and at low cost (Ghose & Ipeirotis, 2010). Hence, customer reviews have become today’s electronic WOM (eWOM) for the current generation of customers and business managers. Reviewers on these online platforms devote time and energy to writing reviews and sharing useful information with prospective consumers to help them evaluate products and make purchase decisions (Zhao et al., 2013; Zhu & Zhang, 2010). Online reviews are therefore particularly important in the purchase of products and services, as a source of advice for consumers to narrow the information gap and reduce uncertainty in their decisions (Dimoka et al., 2012). Service providers and manufacturers value these online reviews because they may reveal genuine concerns of online customers and provide useful market intelligence (Forman et al., 2008; A. H. Huang et al., 2015).

Online consumer reviews are accessible to many readers online; readers often perceive online reviews as more credible and more trustworthy than traditional advertising (L. Huang et al., 2007). Based on these advantages, prior studies on online reviews have focused on the effect of reading reviews on readers’ attitudes (Doh & Hwang, 2009; M. Lee et al., 2009). These studies serve as conceptual antecedents of some decision and behavioral outcomes of interest (e.g., purchasing a laptop computer, movie, or camera).

The number of readers who have been attracted to reading an online review is called review popularity; a review with high review popularity is considered attractive, similar to awareness components (awareness and knowledge on the product, favorable attitudes or feelings toward the product, and action on the acquisition of the product) and information accessibility components (ambiguous information and nonambiguous information; J. Wu, 2017). Therefore, obtaining high review popularity is critical for driving sales; a review with many readers means potential for more reviewers who will create further online reviews—an effective strategy for promoting products (R. Huang & Sarigollu, 2012; Luan et al., 2019). Moreover, to inspire other people to read a review, the review content written by a reviewer is a crucial factor. However, optimal selection of key reviewers from many reviewers must still be answered (Y. M. Li et al., 2011). Therefore, investigating what factors influence

1National Taiwan University of Science and Technology—Taiwan Tech, Taipei
2Taiwan Marketing Research Co., Limited, Taipei

Corresponding Author:
Lien Thi Kim Nguyen, Department of Business Administration, National Taiwan University of Science and Technology—Taiwan Tech, No. 43, Sec. 4, Keelung Rd., Taipei 106. Email: d10508809@mail.ntust.edu.tw

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
review popularity and how to select key reviewers are relevant questions for research scholars and online marketers.

A few studies in the existing literature on eWOM analyze factors of online review popularity, mostly focused on helpfulness prediction based on helpfulness votes (Chua & Banerjee, 2016; Ghose & Ipeirotis, 2010; Hu & Chen, 2016; S. Lee & Choeh, 2016; Z. Liu & Park, 2015; Mudambi & Schuff, 2010; Siering et al., 2018; Yang et al., 2020). To evaluate the helpfulness of reviews, these studies selected famous websites, such as Amazon and Trip Advisor, where online readers can submit review helpful votes. However, review helpfulness cannot be determined if the voting mechanism is ineffective, or without a helpful vote function; furthermore, there may be cases where new reviews are yet to receive votes, which would mean that their level of helpfulness could not be determined (Singh et al., 2017). In these cases, the review popularity can be used to estimate the helpfulness evaluation, as the two are closely related favorable attitudes toward product (J. Wu, 2017). In addition, previous studies analyze the effectiveness of online reviews using many aspects, such as sentiment factors, review extremity, review length, review depth, product type, review title, and review content for helpfulness prediction. However, a combination of reviewer characteristics, product description, and the nature of the words used in the reviews have not been extensively investigated in these studies.

This study predicts online review popularity for cosmetics using a new method that combines the extreme gradient boosting tree algorithm (XGBoost) and the skip-gram model to analyze the effect of non-textual reviews (reviewer and product-related factors) and textual reviews. It has been observed in previous investigations that the type of reviewer, product features, and review content contribute the most to the prediction of reviewer popularity. Taiwan’s cosmetics industry accounted for 0.7% of the global cosmetics market, valued at approximately US$4 billion. Skincare products accounted for 52.4% of the value (Taiwan Cosmetic Market, 2018). The beauty market’s heaviest consumers are also the most likely to engage in online reviews, with more than 70% saying that comment posts motivate them to buy products (Stargage, 2017). By proposing a new method to investigate the factors that affect online review popularity, this study addresses four research questions:

**Research Question 1:** Which type of variables (non-textual review or textual review) effectively contributes to review popularity?

**Research Question 2:** Which textual term and its semantic words contribute to create the highest popularity of a review?

**Research Question 3:** Who are the key reviewers based on these important terms, who create high review popularity?

**Research Question 4:** Which machine learning methods provide a high predictive accuracy in the popularity prediction model?

The study primarily aims to predict online review popularity, where the popularity level is defined based on the total number of users who read the review. Two types of popular machine learning methods were used for the experiment: ridge regression and XGBoost. An effective prediction model for review popularity is built using the XGBoost method. Theoretically, the results of this study contribute to prior research in eWOM by highlighting the influential role of textual reviews and their contribution to review popularity. More specifically, the research sheds light on the strong relationship between important textual review terms and their semantic words in the comments and review popularity. The practical contribution is based on the study’s proposed model: Online vendors can analyze customers’ online reviews to identify important terms that are then recommended to reviewers for use to achieve high review popularity, which will increase company sales. The findings could also help vendors to identify key influencers for their product promotion. The major contributions are summarized as follows:

- The study highlights the importance of certain textual terms and semantic words for review popularity.
- The study builds an effective model to predict the popularity of online reviews using a combination of the XGBoost and skip-gram models.
- It employs two different machine learning methods to study prediction models.
- It extracts the important values of non-textual and textual review features and determines the key factors that increase review popularity.
- It helps online retailers to develop online marketing strategies and word-suggestion systems to support reviewers in achieving high review popularity. In addition, the study identifies key reviewers for product promotion.

The rest of the article is structured as follows: The “Literature Review” section reviews the related literature and is followed by the methods of data collection and analysis in the “Research Methodology” section. The results are presented in the “Results” section. Next, the “Discussion” and “Implication” sections presents the discussion and implications. The “Conclusion” section provides the conclusion. Finally, the “Limitations and Future Research” section outlines the study’s limitations and directions for future research.

**Literature Review**

**eWOM**

WOM is a consumer-dominated channel of marketing communication where the sender is independent of the market (Brown et al., 2007). The development of network technology and the ubiquitous distribution of the internet have transformed traditional face-to-face WOM communication into eWOM.
communication. Consumers now use the internet to share experience and opinions by writing their product experience and reading peer consumers’ product evaluation on different platforms, such as retailers’ websites, brand communities, independent websites, customer blogs, and so on (M. Lee & Youn, 2009). Online customer reviews therefore affect consumers’ purchase decisions (Zhao et al., 2013), company product sales (Zhu & Zhang, 2010), and online helpfulness evaluations (Malik & Hussain, 2018; Singh et al., 2017). Online customer reviews are a type of eWOM: Customer reviews are more reliable than the information provided by sellers because they offer noncommercial, detailed, experiment-based, and up-to-date information (Hennig-Thurau et al., 2004; Yoo & Gretzel, 2011). Moreover, online reviews influence consumers and businesses considerably because they transcend spatio-temporal limitations by enabling simultaneous communication among numerous people (Bickart & Schindler, 2001).

In an online WOM environment, while some consumers provide experience-based product information (reviewers), others read product information provided by peer consumers (readers). Researchers have examined the effect of online reviews from different perspectives. For instance, online reviews influence e-commerce sales and people’s awareness of products (Álvarez et al., 2007; Chevalier & Mayzlin, 2006; Godes & Mayzlin, 2004); a study by Cheng and Ho (2015) indicated that readers considered reviews more practical and useful when written by reviewers with many followers, a high level of expertise in review writing, and high image and word counts. Other studies have investigated the effect of online reviews on consumers’ purchase decisions and evaluations (Hung, 2017; Kumar & Benbasat, 2006; D. Yin et al., 2014; Zhao et al., 2013). Another stream of research has explored the moderating role of product and consumer-related characteristics in the effect of online reviews, including product involvement and reviewer self-disclosure (Zhu & Zhang, 2010). In the advertising field, most studies have focused on exploring the effect of online reviews and their influence on product attitudes (Zablocki et al., 2019) and brand awareness (Zhao et al., 2015). Other research has investigated factors that influence online review readers’ evaluation of review helpfulness, for instance, the rating of a product based on the sentiment score of a review (F. Wang & Karimi, 2019) or on the linguistic category features of the text (Krishnamoorthy, 2015). These studies have concluded that online product reviews increase interaction among customers, customer ties, and the frequency and amount of time customers spend on a website (Mudambi & Schuff, 2010).

This study analyzes the effect of experience goods (cosmetics) reviews on review popularity; the use of influential review content, reviewer variables, and product factors in the prediction model offer robust results.

**Online Non-Textual Reviews**

**Reviewer-related factors.** Online retailers increasingly seek to understand reviewer factors’ influence on online reader trust, usefulness evaluation, review popularity, and purchase decisions (Y. Liu et al., 2008; D. Wang et al., 2013), and how to strategically utilize these factors to drive business. Previous literature on online reviews has focused on reviewer self-disclosure (e.g., name, photo, location, and gender); the literature describes disclosure in various ways: as communication of personal information, thoughts, and feelings to other people (Srivastava & Kalro, 2019); cultural background, gender, age, and personality (Gao et al., 2017); reviewer information display on the website (Racherla & Friske, 2012); reviewer characteristics, such as nicknames, hobbies, and interests (Ghose & Ipeirotis, 2010); and disclosure of personal identity, expertise, and reputation (Z. Liu & Park, 2015). Hu and Chen (2016) demonstrated that the time a reviewer published a review, the number of days that the review remained visible, and the number of reviews with the same rating at the time of the review all had an impact on the evaluation of the helpfulness of the review. In addition, another study considers user information, venue, and content factors to predict the popularity of micro-reviews (Vasconcelos et al., 2014).

One reason that reviewer characteristics have an impact on online consumers is that they are an indicator of the credibility of the review source (X. Wu et al., 2020). Both online reviews and retailer websites provide a variety of information about reviewers, including rankings, summaries of past contributions, and special badges, such as top 50 or 100. Based on that information, consumers infer credibility, and this affects their evaluation of the reviews (Ghose & Ipeirotis, 2010). In addition, online review system characteristics and informant characteristics have an influence on the perceived informativeness and trustworthiness of review popularity, from the perspective of readers: These factors ultimately affect readers’ intention to use the related service (Johanes et al., 2016). In this study, we extend the reviewer factors of online WOM communication by analyzing reviewer factors’ effect on review popularity; the factors include reviewer gender, age, active index on the website, reviewer star sign (e.g., Leo, Aries, Taurus), reviewer’s related topics, number of articles written by a reviewer, review publishing date, and how a reviewer experiences products.

**Product-related factors.** Given that potential consumers often lack experience with products due to physical distance (Kim, 2014), they usually compare alternative products based on information found in reviewers’ product evaluations in the comments, such as appearance, quality, and practical value of a product (X. Li et al., 2019; D. H. Park et al., 2007). Product uncertainty refers to consumer’s difficulty in evaluating product attributes and predicting how a product will perform in the future, which is a major impediment to online markets (Dimoka et al., 2012). When consumers consider buying products online, they often read online reviews about the products to understand the features of the products. Each product has multiple features, and consumers have different preferences for each feature of each product. Therefore, it is
necessary to solve the problem of product feature extraction based on online reviews (Fan et al., 2020).

Many scholars have pointed out the importance of product feature extraction from different perspectives. For instance, a study conducted by Zhu and Zhang (2010) considered product categories and consumer characteristics in relation to consumers’ evaluation of online review helpfulness. D. Wang et al. (2013) proposed the SumView technique, which is a web-based review summarization system for automatically extracting the most representative expressions and customer opinions on various product aspects. This method can lead to more user satisfaction than the results without summaries. Mudambi and Schuff (2010) adopted a linear regression model to investigate the characteristics of helpful online reviews on Amazon.com. Their findings revealed that review extremity, review depth, and product type were prominent factors with respect to helpfulness.

The current study proposes a new method to analyze the dimension of product factors such as brand, rating, attribute, and pictures in a review. In our proposed method, these factors are ranked based on their relative importance in contributing to increase review popularity.

**Textual Review**

Understanding the factors of a review that attracts popularity is crucial for both vendors and reviewers (Liang, 2016). Vendors invest substantial resources in content engineering and are constantly seeking optimal means of managing and improving the digital content in their portion of cyberspace. Therefore, knowing what makes a review attractive is critical to effective content engineering and the overall success of their internet marketing efforts. Empirical research has indicated that two main components influence the value of reviews: review content of the review itself and sentiment of review (Al-Natour & Turetken, 2020). The first component has been investigated in previous studies to assess the value of a review by analyzing its content, such as review depth (Bjerier et al., 2015; S. Lee & Choeh, 2016), review length (Eslami et al., 2018), writing style and timeliness of the review (S. Park & Nicolau, 2015), and signals expressed in the review content and basic text characteristics (Huffaker et al., 2011; Siering et al., 2018).

The second component affecting readers in their evaluation of a review’s content is the sentiment of the review or text readability (Korfiatis et al., 2012; F. Wang & Karimi, 2019). Previous literature has shown that a message generated by a person with high credibility is more persuasive, leading to greater behavioral compliance (Butler & Wang, 2012; Crisci & Kassinove, 1973; Pavlou & Dimoka, 2006) and a higher level of acceptance of the suggestions offered by the person (Suzuki, 1978). Ghose et al. (2012) proposed a new ranking system for hotel search engines that mine consumer reviews and considered the readability of the review as a major factor. Their study found that content was the most critical aspect of a review. To present effective information, a review should be precise or easy to understand and without ambiguity, which refers to how easily a text can be understood by readers (Klare, 1974). Recently, a study by Yang et al. (2020) indicated reviews with similar text in title and content as more helpful; title sentiment positively affects review helpfulness and negatively moderates the relationship between text similarity and review helpfulness. In addition, informativeness, as a variable for identifying helpful reviews, significantly improves the accuracy of predicting review helpfulness compared with review length (Sun et al., 2019). Reviewers’ linguistic styles (e.g., shouting, immediacy, and evidence-based) are significantly important for product search and for product experience capitalization, and are evidently positively correlated to their helpfulness ratings (T. S. Li et al., 2019; Ludwig et al., 2013).

In line with our objective, the study used XGBoost to extract the important word features that reviewers use in their reviews based on importance score ranking. It then employed the skip-gram model to identify semantic words according to these important features to predict review popularity; key reviewers were identified based on the frequency of using important words in their reviews.

**Research Methodology**

Based on the objective of the study, the proposed method has five major stages, as presented in Figure 1. Following data collection from the website Urkosme (https://www.urkosme.com/), preprocessing was performed to eliminate redundancies, irrelevant information, and punctuation. Term frequency–inverse document frequency (TF-IDF) was then utilized to assign a weight to each word for input to the prediction model, depending on its value as an independent variable. To predict review popularity, two machine learning algorithms, XGBoost (T. Chen & Guestrin, 2016) and ridge regression (Y. R. Chen et al., 2018), were used to construct two classifiers individually for comparison. Furthermore, a fivefold cross-valid strategy was applied to train and test classifiers. In the fivefold cross validation, the original data set was randomly partitioned into five equal-size subsamples. Four folds were merged to form a training data set. The fifth fold was used as the testing data set. This process was then repeated five times. The outcomes of the above prediction models were compared with the objective function of the Root Mean Squared Logarithmic Error (RMSLE) to select the model with the optimal performance and minimal values of RMSLE. We next selected the model with the best performance in feature selection to identify key features from textual and non-textual variables that significantly contributed to creating review popularity. We used the skip-gram model (W. Zhang et al., 2018) to determine semantic words related to the crucial textual reviews...
and to identify key reviewers based on frequently used words in their reviews.

**Data Collection**

The data in this study were collected from urcosme.com, which is one of the most famous cosmetics and skincare product review websites in Taiwan. The website provides user-generated content regarding experiences using the products, product opinions, as well as a shopping cart. Skincare products are chosen as the focal items in our study as it takes up the highest percentage rate with 52.4% in the total Taiwan Cosmetic Market (Taiwan Cosmetic Market, 2018). Web crawler and R programming were selected to scrape data. Data scraping is a technique performed by a computer program to extract data from the human-readable output from a platform. A total of 540 reviewers who wrote reviews on the product name WHITIA lotion in the category of skincare products were chosen, yielding 127,135 valid reviews in total. The TF-IDF algorithm produced columns of 50,000 variables of preprocessed words.

User information published on the website includes reviewer gender, skin type, age, star sign, topics, index, article type, and article number. We distinguished between two types of review features: features of the review text and non-textual variables related to user and product information. Table 1 lists the details of our study’s review features. We investigated the effects of these factors on review popularity, that is, the dependent variable measures the number of people with access to read the review content; high popularity indicated that the review content comprised helpful information and consequently attracted relatively many viewers (J. Wu, 2017).

**Data Processing**

**Chinese word segmentation.** Chinese word segmentation is a key step to improve the accuracy of various types of Chinese text processing, such as speech recognition, machine translation, and language understanding (H. Li & Yuan, 1998). This study utilized “Jieba” for Chinese word segmentation. To the best of our knowledge, “Jieba” is the most general, stable, and easy-to-use technique for segmenting Chinese words.

**Stop word removal.** Stop words are frequently used in descriptions, such as “the,” “are,” “and,” and “a.” These words have little or no relevant meaning and thus limit the efficiency of information retrieval (IR). The Sklearn package in Python was used to eliminate such words.

**TF-IDF.** TF-IDF is often used in IR and text mining to determine the significance of a word in a corpus of documents (Ramos, 2003). In our data framework, we defined the TF-IDF weight \( w_{tr} \) for term \( t \) in review \( r \) in Equation 1.

\[
w_{tr} = tf_{tr} \times idf_{tr},
\]

where \( tf_{tr} \) is the term frequency of term \( t \) in review \( r \); \( idf_{tr} \) is equal to \( \log(R/df_t) \), where \( R \) is the total number of collected reviews and \( df_t \) is the review frequency with which term \( t \) occurs in the collected reviews. If term \( t \) occurs frequently in review \( r \), then, the score of \( tf_{tr} \) increases; low \( df_t \) results in high \( idf_{tr} \). A relatively high TF-IDF weight indicates that term \( t \) is key in review \( r \). We used the weight as the term feature to identify the significance of terms in each review.

**N-gram.** We used the \( n \)-gram model to evaluate “\( n \)” continuous words in a review text and predict the subsequent item in a sequence. In sentiment analysis, the \( n \)-gram model helps analyze the sentiments of texts or documents. Unigram refers to an \( n \)-gram of size 1; bigram refers to an \( n \)-gram of size 2; trigram refers to an \( n \)-gram of size 3; any higher \( n \)-gram term refers to some \( n \)-gram of size 4 or higher. Examples of unigrams include “the,” “food,” “is,” “not,” and “delicious,” where a single word is considered, and its bigrams are “the food,” “food is,” “is not,” and “not delicious,” where word pairs are considered.
Among various machine learning methods, we selected ridge regression and XGBoost for predictive model as a robust and effective machine learning method in prediction. XGBoost is a boosting algorithm belonging to supervised learning, which is an ensemble algorithm based on gradient boosted trees. It integrates the predictions of “weak” classifiers to achieve a “strong” classifier (tree model) via a serial training process, and it can avoid overfitting by adding a regularization term. Parallel and distributed computing hastens the learning process to enable a quicker modeling process (T. Chen & Guestrin, 2016). Ridge regression was used to enhance extreme learning machine as proposed by G.-B. Huang et al. (2006), which is a novel single hidden layer feed forward neural network, where the input weights and the bias of hidden nodes are generated randomly without tuning and the output weight is determined analytically. It is also an extreme learning machine as it owns an extremely fast learning algorithm and good generalization capability (G. Li & Niu, 2013).

Ridge regression is a statistical method for modeling a linear relationship between a dependent variable and independent variables. This technique plays a major role in various machine learning algorithms, especially in the field of e-commerce, in which ridge regression is utilized to analyze a data set on the basis of user profiles and comments; and it summarizes them in a compact form to investigate item relevance for users (Y. R. Chen et al., 2018). Researchers recently conducted a study on the basis of feature selection in document retrieval. Findings showed that the ridge regression model has given the same or even better performance results compared with logistic regression and support vector machine algorithm (Aseervatham et al., 2011; T. Zhang & Oles, 2001); those methods are extensively adopted in previous research to assess the effect of online reviews.

The boosting method has been remarkably successful in terms of predictive performance (Zięba et al., 2016). For the XGBoost model parameters, we used \(x_i\) and \(y_i\) as independent variables to represent a total of 50,000 words, with \(k\) features and \(n\) observation values. Next, following the method of another study (T. Chen & Guestrin, 2016), a tree ensemble mode was constructed using \(m\) additive functions to establish the regression model for the current study and predict review popularity (\(\hat{y}_i\)), as demonstrated in Equation 2.

\[
E(Y|X = x) = \hat{y}_i = \sum_{m=1}^{N} f_m(x_i), f_m \in F,
\]

where \(F = \{f(x) = w_{h(x)}\} (h : \mathbb{R}^k \rightarrow T, w \in \mathbb{R}^T)\) is the space of the regression trees, and \(h\) represents the regression tree of each model. The leaf index \(T\) denotes the number of branches of each regression tree, each \(f_m\) represents an independent regression tree structure \(h\), and the leaf weight is denoted by \(w\).

In the tree, \(w_{h(x)}\) represents the score of \(h(x)\) independent branches, and the total scores of the branches and leaves are summed to obtain the value for review popularity (\(\hat{y}_i\)).

### Key Feature and Word Feature Weight of the Fractional Mode

After creating the basic XGBoost regression tree hypothesis, we can define the loss function as Equation 3 for the regression model of this study.
\[ L(\theta) = \sum_{i=1}^{n} \left[ \gamma \left( \hat{y}_i - y_i \right)^2 + f_i(x_i) \right] + P_\lambda(f_i). \]  

where \( P_\lambda(f) = \gamma T + (1/2)\lambda \|w_m\|^2. \]  

The loss function Equation 3 is set to a differentiable function, and the first term of Equation 3 is used to estimate the difference between the predicted review popularity \( \hat{y}_i \) and the actual review popularity \( y_i \). The second term of Equation 3 is a regularization function, which is used to avoid the problems of multicollinearity and overfitting. The regularization term in Equation 4 is \( P_\lambda \), where \( \gamma \) controls the value of the leaf nodes in the regression trees. If the value is large, then, the penalty for \( P_\lambda \) is more conservative; \( \lambda \) controls the regularization function of the size \( \|w_m\|^2 \), and if the value is large, then, both \( \gamma \) and \( \lambda \) can be manually controlled.

We added the regression model to the regression tree of the previous \((t-1)\) iteration and then added the \( f_i \) additively function of the \( t \) iteration, as demonstrated in Equation 5, thereby reducing the value of the loss function for the model and identifying the most suitable \( f_i \) additive function using the gradient descent method (T. Chen & Guestrin, 2016).

\[ L((\theta))^t = \sum_{i=1}^{n} \left[ \gamma \left( \hat{y}_i - y_i \right)^2 + f_i(x_i) \right] + P_\lambda(f_i). \]  

The second-order approximation identified in a study conducted by Friedman et al. (2000) was then used to implement the Taylor series from Equation 5 and obtain \( g_i = \partial_{\hat{y}_i} L(\hat{y}_i, y_i^{(t-1)}) \), which is the first-order differential loss function, and \( d_i = \partial_{\hat{y}_i}^2 L(\hat{y}_i, y_i^{(t-1)}) \), which is the second-order differential loss function. Thus, we can obtain Equation 6:

\[ L((\theta))^t = \sum_{i=1}^{n} \left[ \gamma \left( \hat{y}_i - y_i \right)^2 + g_i f_i(x_i) + \frac{1}{2} d_i f_i^2(x_i) \right] + P_\lambda(f_i). \]  

After removing the constant term, Equation 7 is obtained:

\[ L((\theta))^t = \sum_{i=1}^{n} \left[ g_i f_i(x_i) + \frac{1}{2} d_i f_i^2(x_i) \right] + P_\lambda(f_i). \]  

Expanding the penalty term enables the calculation of Equations 8 and 9.

\[ L((\theta))^t = \sum_{i=1}^{n} \left[ \gamma \left( \hat{y}_i - y_i \right)^2 + f_i(x_i) \right] + P_\lambda(f_i). \]  

where \( j \) is the branch of each regression tree. By letting \( G_i = \Sigma_i(g_i) \) and \( D_i = \Sigma_i(d_i) \), we can obtain Equation 10

\[ \gamma T = \sum_{j=1}^{T} G_j w_j + \frac{1}{2} \left( D_i + \lambda \right) w_j^2 + \gamma T. \]  

Then, we differentiated the weight \( w_j \) with respect to the loss function \((L, j)^t\) to obtain Equation 11.

\[ \frac{\partial L((\theta))^t}{\partial w_j} = G_i + (D_i + \lambda) w_j = 0. \]  

Weight \( w_j \) can then be calculated with Equation 11 to obtain Equation 12.

\[ w_j = - \frac{G_i}{D_i + \lambda}. \]  

Next, \( w_j \) can be substituted into the loss function. The optimal loss function value presented in Equation 13 was used in this study.

\[ L((\theta))^t = - \frac{1}{2} \sum_{j=1}^{T} \left( \frac{G_i^2}{D_i + \lambda} + \gamma T. \right. \]  

Additional accurate predictions are produced when the value of \( L((\theta))^t \) is small. By extending the gain of the tree nodes of each regression model by \( L((\theta))^t \), we can obtain Equation 14, where the total score of different branches can be represented by the following summation: \( \Sigma_{u=1} \left( G_u^2 / D_u + \lambda \right) \). The total score of undivided branches is \( \left( \Sigma_{u=1} G_u^2 / D_u + \lambda \right) \). In this study, we maximized \( \Sigma_{u=1} G_u^2 / \Sigma_{u=1} D_u + \lambda \) using the gradient descent method. We eventually optimized the score for word feature weight (Equation 16) from Equation 15 and selected key variables according to Equation 17.

\[ \text{Gain} = \frac{1}{2} \sum_{k=1}^{K} \frac{G_k^2}{D_k + \lambda} - \frac{\sum_{k=1}^{K} G_k^2}{\sum_{k=1}^{K} D_k + \lambda} - \gamma, \]  

\[ \text{Optimal gain score} = \max \left\{ \frac{1}{2} \sum_{k=1}^{K} \frac{G_k^2}{D_k + \lambda} - \gamma \right\}, \]  

\[ \text{Text gain} = \frac{G_k^2}{D_k + \lambda}. \]
Word feature weight score
\[ f(x_j) = \begin{cases} \frac{1}{2} \left( \sum_{u=1}^{S} G_u^2 D_u + \gamma \right) - \frac{1}{2} \left( \sum_{u=1}^{S} G_u^2 D_u + \gamma \right) & \text{if Word feature weight score} > 0 \\ \text{Do not choose any } x_j. & \end{cases} \] (17)

For example, identifying the right split (R: word-skincare) and left split (L: word-moisturizer) branches as optimal for minimizing the undivided branch and leaf scores relative to other variables, as in Equation 18, would indicate that the weight score of R in Equation 19 is greater than that of L in Equation 20. We can then obtain the word feature using Equation 21 with word-skincare and word-moisturizer as the first and second features selected.

\[ \text{Optimal Gain score} = \max \left\{ \frac{1}{2} \left[ \sum_{u=1}^{S} G_u^2 D_u + \gamma \right] - \frac{1}{2} \left[ \sum_{u=1}^{S} G_u^2 D_u + \gamma \right] \right\}, \]
\[ = \frac{1}{2} \left[ \frac{G_r^2}{D_r + \lambda} + \frac{G_l^2}{D_l + \lambda} - \frac{(G_r + G_l)^2}{D_r + D_l + \lambda} \right], \quad \gamma \leq 0, \quad \gamma \equiv \gamma \quad \text{(18)} \]

Weight score for word-skincare = \[ \frac{G_r^2}{D_r + \lambda} \] (19)

Weight score for word-moisturizer = \[ \frac{G_l^2}{D_l + \lambda} \] (20)

\[ f(x) = \{x_{\text{maintenance}}, x_{\text{moisturizing}}\}. \] (21)

**Identify Semantic Words**

After obtaining keyword features with \( f(x) \) in section “Key Feature and Word Feature Weight of the Fractional Mode,” we used the skip-gram model as suggested by W. Zhang et al. (2018) to fine-tune the word vectors. Thus, the skip-gram model required to maximize the average log probability given a sequence of training words \( W_{i+1} \) is in Equation 22, \( m \) is the maximum distance size from word \( W_i \), which here denotes \( f(x) \). The larger the \( m \), the more training word examples are involved in the model causing a higher accuracy of the model with more computation complexity.

\[ \text{obj} (\theta) = \frac{-1}{T} \log L (\theta) = \frac{-1}{T} \sum_{t=1}^{T} \sum_{-m \leq j \leq m} P(W_{t+j} | W_t; \theta). \] (22)

We predicted word \( W_i \) for five maximum distance words \( (W_{i-2}, W_{i-1}, W_{i+1}, W_{i+2}) \) where \( W_i \) refers to the word at \( t \) position in a document, and \( W_{i-2}, W_{i-1}, W_{i+1}, W_{i+2} \) are all continuous numerical word vectors. Figure 2 depicts the skip-gram model to identify keywords and semantic words for a better understanding of the mechanism.

The skip-gram model defines \( P(W_{t+j} | W_t) \) using softmax function in Equation 23, where \( u_w \) is the output word vector for word \( u_w \), and \( w \) is the size of words in the document, \( v_c \) is the word vector of our selected \( W_t, u'_c \) is the output word vector for \( u_w \), and \( o \) is the size of our five maximum distance words.

Negative sampling is suggested by W. Zhang et al. (2018) in solving the problem to calculate \( P(W_{t+j} | W_t) \) within the scope of \( k \) native samples in Equation 24, where \( k \) is set to be five for our large training data sets.

\[ P(W_{t+j} | W_t) = P(o | c) = \frac{\exp(u'_o v_c)}{\sum_{w \in \text{sample}} \exp(u'_w v_c)}, \] (23)
\[ \text{obj}_{\text{negative-sample}} (o, v_c) = - \log \left( \frac{\exp(u'_o v_c)}{\sum_{w \in \text{sample}} \exp(u'_w v_c)} \right), \]
\[ = - \log \left( \sigma(u'_o v_c) - \sum_{k=1}^{K} \log \left( \sigma(-u'_o v_c) \right) \right). \] (24)

Finally, cosine similarity was used to identify semantically related words through every word \( (W_1, \ldots, W_n) \) with \( v_c \) in Equation 25, and the top 10 most semantically related words would be selected.

\[ \text{Cosine similarity} = \frac{v_c \cdot u_w}{\max (\|v_c\|, \|u_w\|)}, \text{for win } W_t \text{ to } W_n. \] (25)

**Key Reviewer**

Key reviewers were identified on the basis of the frequency in which they used keywords and semantic words in their reviews. For example, the word “effect” and its semantic words have a highly important ranking; the user who has frequently used effect and its semantic words in their comment is a key reviewer related to these important terms.

**Results**

**Prediction Model**

After eliminating the stop words and calculating the word vector for each word, we used these word vectors as inputs in training the models to predict review popularity. To develop the prediction model, we used 80% of the data as the training set and the remaining 20% as the test set. The trained model was then tested on the test set to evaluate its performance.

We compared the performance of the two approaches, the ridge regression and XGBoost methods, using RMSLE as
the optimization metric. The goal was to identify which model produced minimal error prediction; that is, the model with the lower RMSLE. The results revealed that the XGBoost model with 50,000 predictors achieved an RMSLE value of 1.168, outperforming the ridge regression model, which had an RMSLE of 1.471. Thus, the XGBoost model demonstrated high prediction accuracy relative to the ridge regression model.

Textual and Feature Selection Analysis

XGBoost is selected to extract important non-textual and textual reviews based on their importance ranking scores. After deriving 501 features extracted from textual and non-textual feature variables, we selected the top 18 textual review features and the top 10 non-textual feature variables according to the contribution of key terms and weighted rankings of these features to review popularity. In the comparison of weight contribution, the analysis results indicated that review textual features obtained a majority at the top of key features that contributed to review popularity. Therefore, the textual review showed a higher contribution to review popularity than the non-textual variables. Figures 3 and 4 display the weightings of significance.

Figure 3 presents the details of textual features and the contribution of its significance weighting to review popularity. These top 18 key features were as follows: makeup, feeling, effect, skin, starting, moisturizing, product, sense, like, color, time, recommend, suitable for, repurchase, concealer, basic maintenance, acne, and refreshing. High weightings indicate that a term contributed highly to the review popularity. For example, the term “makeup” contributed a value of 0.039 to review popularity, and the term “feeling” contributed a value of 0.026 to popularity concerning textual terms.

Accordingly, non-textual variables with significant weightings to review popularity were selected, with the author_age variable contributing the highest value of 0.023 to review popularity creation. Non-textual variables with the next highest values were as follows: comment_exp_real product refers to reviewer experience regarding the real product, author_index measures the frequency of user activity on the online platform, attribute_lotion represents the contribution of lotion quality to the product review, author_skin_combination measures the combined skin type of the reviewer, author_article_num measures the number of reviews written by the reviewer, comment_exp_sample is user experience of sample products provided by the supplier,
attribute_butter is related to the contribution of the butter factor to the product review, attribute_essence measures the contribution of the essential factor to the product review, and attribute_cream is the influence of the cream factor on the product review. Figure 4 presents the weightings of these features.

After selecting key terms, the skip-gram model was employed to predict the semantic terms surrounding the key feature words. Table 2 presents the results.

For example, the key textual feature “feeling” was predicted from semantic words: smell, look, clean, feel, obvious, skin, absorb, refresh, same, and effect. For the term “product,” the list of semantic words are brand, series, ingredients, merchandise, skincare product, lotion, effect, stuff, loose powder, and sunscreen. These semantic words are highly correlated to important words and contribute highly to review popularity.

**Key Reviewer Detection**

Key authors were selected based on the frequency with which they used key or semantic words in their reviews according to the ranking of significance in our proposed method. In Table 3, we randomly selected 10 reviewers to determine the frequency with which terms occurred in the text of their reviews. For example, when reviews included the term “effect” or its semantic words, Author A with ID 379102, who used “effect” has the highest frequency of 31 counts of “effect” compared with other authors. Therefore, Author A was selected as the first key author for the term “effect.” However, for the term “feeling,” Author B used it 17 times in reviews, which are several times more than other author used the term. Therefore, Author B was considered the key reviewer for the term “feeling.”

**Evaluating Prediction Model**

To increase the reliability and effectiveness of the proposed method in selecting keywords for predictions, we randomly selected 100 reviews to predict the review popularity measure for a WHITIA lotion product. Our proposed method was designed to select key features according to their weighted ranking in terms of their contribution to review popularity by using the XGBoost algorithm and the skip-gram model for the selection of semantic words in textual terms. The single-term selection method (STSM) and the random-term selection method (RTSM) were used for comparison with our proposed method concerning review popularity prediction. Table 4 presents the statistical details of this comparison.

Apart from the 18 important word features and semantic words, the table also presents a total count, which combines the top 18 phrases to create a word group. The mean was calculated using a logarithmic transformation of popularity. Given that popularity had a right-skewed distribution of 17.76, it was logarithmically transformed into a normal distribution of 0.26. We used the $t$ test to assess the performance of the proposed method, with $t > 0$ indicating that the proposed method exhibited superior prediction performance compared with the other methods. A significantly better performance by the proposed method was indicated by $p < .05$.

A comparison of the performance of the proposed method with the STSM in terms of the $t$ statistic value revealed that only the term “basic maintenance” had $t < 0$. Thus, 95% of the feature terms using the proposed method outperformed...
those selected using the STSM. Only two features, “makeup” and “basic maintenance,” did not exhibit significantly better performance \( (p > .05) \), whereas the remaining terms did \( (p < .05) \). Therefore, 89% of the terms selected using the proposed method were more significant than those selected using the STSM. Figure 5 shows the performance comparison between the proposed method and the STSM in the term “makeup.” However, a comparison of our proposed method with the RTSM in terms of the t-test and p-values indicated that our model significantly outperformed the RTSM for all terms, and we could observe that the distributions of the proposed method compared with those of the RTSM were significantly different in the example for the term “makeup” shown in Figure 6. Thus, our proposed method was demonstrably more effective in predicting the review popularity for the WHITIA lotion product compared with the STSM and the RTSM.

### Discussion

This study aims to investigate the impact of non-textual and textual reviews on online review popularity and to identify the most appropriate semantic terms and key reviewers in terms of key textual variables; the study employs a new method that uses a combination of the XGBoost and skipgram models. The findings indicate that review textual features are the most significant parameters to determine review popularity (compared with the important scores in Figures 3 and 4). This outcome emphasizes the importance of review content and suggests that a customer who uses textual reviews for potential buyers and expresses them through meaningful and concise wording will contribute to increasing online review popularity. In line with previous studies, this study verifies the importance of textual characteristics (Ghose & Ipeirotis, 2010; Sun et al., 2019; F. Wang & Karimi, 2019; Yang et al., 2020) in customers’ evaluation of the effectiveness of online reviews.

Moreover, the results indicate the relevance of semantic words associated with textual reviews in the contribution to online review popularity. Consistent with previous studies, this study emphasizes the importance of information value (words used in textual review) in online product reviews; readers of product reviews do not merely count the number of stars but also read the corresponding reviews (Butler & Wang, 2012; Pavlou & Dimoka, 2006). Rich content of feedback text comments plays a significant role in building a buyer’s trust in a seller (Hennig-Thurau et al., 2004; Yoo & Gretzel, 2011). When writing an online review, like any other textual message, reviewers not only decide on what to write but also how to write it. These decisions define the content review and linguistic characteristics (Huffaker et al., 2011; Ludwig et al., 2013). In line with these suggestions, our study emphasizes the importance of key reviewers in creating the helpfulness of reviews that lead to high review popularity. Furthermore, through our model, we can identify key reviewers, who are those that consistently employ important textual terms in their reviews to attract review popularity.

Consistent with previous studies, the results regarding the non-textual review variables highlight the importance of reviewer age, reviewer experience with real products (Bickart & Schindler, 2001; D. H. Park et al., 2007), and the time they spent on websites (Hu & Chen, 2016) in creating online

### Table 3. Frequency of Key Textual Appear in Reviews.

| Feature textual review | Author A | Author B | Author C | Author D | Author E | Author F | Author G | Author H | Author I | Author J |
|------------------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Makeup                 | 34      | 3       | 2       | 2       | 5       | 1       | 1       | 1       | 3       | 0       |
| Feeling                | 13      | 17      | 2       | 4       | 7       | 1       | 8       | 4       | 0       | 9       |
| Effect                 | 31      | 30      | 3       | 4       | 3       | 0       | 18      | 9       | 0       | 20      |
| Skin                   | 12      | 7       | 20      | 6       | 31      | 0       | 25      | 28      | 0       | 14      |
| Starting               | 30      | 16      | 6       | 19      | 26      | 5       | 10      | 17      | 4       | 10      |
| Moisturizer            | 21      | 7       | 6       | 1       | 4       | 0       | 4       | 6       | 0       | 4       |
| Product                | 11      | 5       | 3       | 2       | 4       | 0       | 1       | 5       | 0       | 3       |
| Sense                  | 7       | 1       | 3       | 4       | 9       | 0       | 8       | 8       | 1       | 3       |
| Like                   | 9       | 9       | 4       | 3       | 4       | 4       | 2       | 7       | 2       | 7       |
| Color                  | 35      | 18      | 4       | 1       | 2       | 5       | 2       | 4       | 4       | 32      |
| Time                   | 23      | 8       | 6       | 11      | 28      | 0       | 9       | 6       | 4       | 13      |
| Recommend              | 5       | 7       | 7       | 2       | 1       | 0       | 0       | 9       | 0       | 6       |
| Suitable for           | 12      | 7       | 4       | 3       | 1       | 1       | 2       | 2       | 2       | 6       |
| repurchase             | 6       | 2       | 2       | 7       | 3       | 4       | 0       | 6       | 3       | 1       |
| Concealer              | 66      | 22      | 0       | 0       | 4       | 0       | 6       | 3       | 1       | 40      |
| Basic maintenance      | 0       | 0       | 0       | 0       | 0       | 0       | 0       | 1       | 0       | 0       |
| Acne                   | 6       | 20      | 14      | 2       | 12      | 0       | 19      | 6       | 0       | 11      |
| Refreshing             | 17      | 9       | 5       | 3       | 5       | 0       | 5       | 9       | 0       | 7       |
In the cosmetic field, most consumers are young or middle aged; therefore, different age groups will have different styles of review, leading to different contributions to review popularity. Regarding product experience, reviewers who have used the real product will know more about its quality and effectiveness than customers who have only used suppliers’ sample products; their review will therefore attract more readers. Regarding product factors, previous studies have suggested that uncertain product information in a review leads to consumers’ difficulty in evaluating product attributes (Kim, 2014; T. S. Li et al., 2019). Hence, we suggest that a reviewer’s use of the most relevant product distribution in writing a review will significantly contribute to review popularity.

Table 4. Comparison Between Difference Selection Methods.

| Features         | Semantic words                                                                 | Proposed method | STSM | RTSM | Proposed vs. STSM | Proposed vs. RTSM |
|------------------|--------------------------------------------------------------------------------|-----------------|------|------|-------------------|-------------------|
| Makeup           | How, hand, dark circles, remove, gloss, open frame, air cushion, activity, brand, habit | 6.97 1.3        | 6.71 | 1.41 | 5.02 1.77        | 1.34 .092         |
| Feeling          | Smell, look, clean, feel, obvious, skin, absorb, refresh, same, effect          | 7.48 1.19       | 5.88 | 1.89 | 1.85             | 7.15 .000         |
| Effect           | Feeling, level, concealer, pores, skin, refresh, flavor, function, obvious, product | 7.63 1.20       | 5.45 | 1.89 | 1.75             | 9.72 .000         |
| Skin             | Oily face, acne hair, face, skin texture, pores, dryness, on the face, scalp, skin, acne | 7.60 1.15       | 5.93 | 1.72 | 1.95             | 8.08 .000         |
| Starting         | Once, just, now, one, always, discovery, month, end, before, every time         | 7.67 1.06       | 5.84 | 1.74 | 5.47 1.86        | 8.96 .000         |
| Moisturizer      | Moist, sunscreen, whitening, ingredients, refreshing, oil control, lotion, concealer, effect, moisturizer | 7.24 1.31       | 5.85 | 1.63 | 5.34 1.76        | 6.65 .000         |
| Product          | Brand, series, ingredients, merchandise, skin care product, lotion, effect, stuff, loose powder, sunscreen | 7.58 1.12       | 6.24 | 1.67 | 5.19 1.80        | 6.66 .000         |
| Sense            | Fragrance, light, refreshing, smell, comfort, moisture, feeling, effect, price, apply | 7.26 1.37       | 5.29 | 1.86 | 5.47 1.91        | 8.54 .000         |
| Like             | Cute, love, suitable, satisfaction, natural, convenient, flavor, color, promote, same | 7.49 1.11       | 5.76 | 1.79 | 5.16 2.02        | 8.23 .000         |
| Color            | Nature, color, skin color, color degree, look, pearlescent, transparent, makeup, smell, pigmented | 7.63 1.04       | 5.93 | 1.64 | 5.44 1.87        | 8.73 .000         |
| Time             | Hour, later, after, people, stuff, subsequently, if, on the face, immediately, long | 7.83 0.94       | 5.69 | 1.81 | 5.41 1.77        | 10.46 .000        |
| Recommend        | Friend, give, introduce, closet sister, seek, purchase, counter, promote, select, share | 7.53 1.12       | 5.96 | 1.63 | 5.28 1.91        | 7.93 .000         |
| Suitable for     | Summer, recommendation, in general, winter, nature, itself, nowadays, demand, appropriate, time | 7.14 1.26       | 6.01 | 1.83 | 5.11 1.88        | 5.12 .000         |
| Repurchase       | Buy, purchase, pay, everyone, satisfaction, acquire, experience, introduce, lavish, useful | 7.45 1.08       | 6.17 | 1.53 | 5.25 1.86        | 6.85 .000         |
| Concealer        | Cover, modification, enduring, effect, pores, powder foundation, oil control, makeup, loose powder, foundation | 7.56 0.89       | 6.12 | 1.76 | 5.21 1.84        | 7.29 .000         |
| Basic maintenance| Wash face, apply first, bath, moderate, evenly apply, soft, wet, moist, between, cyclograph | 6.39 1.71       | 6.54 | 1.33 | 5.31 1.98        | −0.44 .329        |
| Acne             | Pimple, pores, smallpox, allergies, fine line, redness, skin, oil, peeling, sputum | 7.26 1.19       | 5.91 | 1.84 | 4.60 1.91        | 6.18 .000         |
| Refreshing       | Moist, moisturizing, mild, flavor, non-sticky, absorb, comfortable, natural, light, degree | 7.55 1.32       | 5.57 | 1.88 | 5.19 1.90        | 8.63 .000         |
| Total counts     |                                                                           | 7.87 1.09       | 5.43 | 1.63 | 5.39 1.98        | 12.43 .000        |

Note. STSM = single-term selection method; RTSM = random-term selection method.
reviewer should focus on content. In addition, previous studies have suggested that the review publish day affected the review effectiveness evaluation when the number of days of the review was shown on the first page and the number of reviews with the same rating at the time that the review was posted was included in the prediction model (Hu & Chen, 2016). Therefore, this research question is recommended for future research.

In contrast to a previous study by Malik and Hussain (2018) that used different machine learning methods to analyze review content and reviewer variables that contribute to review helpfulness, this study’s results indicate that the stochastic gradient boosting learning method yields the most accurate prediction of helpfulness. The current study not only applies the XGBoost to analyze non-textual and textual contents based on importance ranking but also incorporates the skip-gram model to identify semantic words related to key terms, which provides more robust results. Moreover, the XGBoost produces a more accurate prediction compared with the ridge regression model; our method reports better performance in selecting important texts and their semantic words compared with the STSM and the RTSM. Therefore, the study’s techniques are appropriate for estimating complex relationships among variables using multiple learning algorithms to obtain better predictive performance than can be achieved with any of the constituent learning algorithms (T. Chen & Guestrin, 2016).

**Implications**

The implications of this study can be explained from two perspectives: theoretical and managerial aspects. The research takes a major step toward identifying the factors that drive customers to read reviews. First, the findings highlight that the qualitative characteristics of the content of reviews (i.e., important texts used to describe products and their semantic words) contribute more to explaining review popularity than other characteristics, such as reviewers’ and
product factors. Moreover, this study assesses a combination of all the factors related to online reviews, thus allowing for development of a comprehensive model to predict review popularity. In terms of the theory of eWOM communication, which refers to the extent to which consumers write online reviews to share experience and to support future customers in product evaluation (Godes & Mayzlin, 2004; T. S. Li et al., 2019; D. H. Park et al., 2007), online WOM affects review helpfulness assessment (S. Lee & Choeh, 2016; J. Wu, 2017; Yang et al., 2020) and product sales (Chevalier & Mayzlin, 2006; Zhu & Zhang, 2010). The findings of this research shed light on multiple aspects of online review that attract readers to reviews and lead to review popularity. In the online environment, where consumers have limited ability to assess the integrity of product information, especially with regard to cosmetic goods, the information reviewer uses review content, word description, experiences with real products, and messages related to product attributes to improve consumers’ perceived usefulness, which leads to high review popularity.

Electronic commerce merchants and sellers provide a platform for internet users to leave their opinions and post valuable feedback related to online product. This feedback by customers assists in building customer trust and supports customers’ purchase decisions. From a practical perspective, this study has several major implications for research on increasing product review readership and thus future sales (R. Huang & Sarigollu, 2012; Luan et al., 2019).

Online marketing could develop online marketing strategies. First, online vendors can use this method of analyzing customers’ online reviews to achieve high review popularity. Online retailers could generally advise their customers to describe product-related aspects while avoiding uncertain language, and to express their opinions using strong words. Similarly, online marketers need a word suggesting system to support online reviewers while writing product reviews. For instance, for the word “feeling” used in a review, the system would automatically suggest related semantic words, such as smell, clean, feel, obvious, skin, absorb, refresh, same, and effect, to describe the term, which would lead to increased review popularity. Online retailers might also automatically select key reviewers with high review popularity based on our study proposed model for corporate product advertising. Finally, these outcomes can also be effectively applied on websites to analyze customer reviews where the helpfulness vote system is ineffective or the system does not have the vote functionality at all. Online cosmetic marketers need to find reviews that attract readers and delight them, and that are easy to read, and regularly update them on the front pages.

### Conclusion

The market for cosmetic products is highly competitive, which means that text mining research must comprehensively evaluate the online reviews of cosmetics users. We developed a new model to obtain important variables related to textual and non-textual features that contribute to review popularity. By considering keywords and semantic properties, online vendors can discern which features mentioned in reviews are more helpful, and users can learn how to write better reviews and subsequently attract more people to read their comments and provide helpful information for other buyers. In addition, the study also proposes a method of selecting a key reviewer based on the frequency with which important words and semantic words are used in their review; online vendors can select key specific textual reviewers for their promotional campaigns by identifying reviewers who frequently use key feature words.

Another important finding of this research is that the XGBoost performs better than the ridge regression technique. The RMSLE obtained when testing the review data on the WHITIA lotion product using the gradient boosting technique, a type of ensemble learning, is much less than that obtained using linear regression.

### Limitations and Future Research

Despite its contribution to online review research, this study has certain limitations. First, many prior studies found bias in online review helpfulness ranking as affected by review publishing order. Owing to the Matthew effect, early posted most helpful reviews typically attracted a disproportionally large number of review popularity and votes than later posted reviews, thus defeating the original review popularity and reinforcing the sequential bias (Wan, 2015). This research did not address this sequential bias in popular prediction model development. Hence, future study can address this issue. Second, previous research employed different machine learning methods in predictive model (Malik & Hussain, 2018). The current study used XGBoost and ridge regression for review popularity prediction. Therefore, future studies can encompass the machine learning methods such as Deep Learning or Genetic Algorithm to further enhance the current findings.

Third, the data set used focused on words used in cosmetic product categories. Consequently, the generated results are only applicable to cosmetic products. Other fields might have different words emphasizing their product features. Therefore, the application of the study results is limited. Future research can replicate the method analysis in different product categories. Fourth, we measured review popularity only with respect to the number of people who accessed reviews and used it as the dependent variable in this study. However, users who are more active receive more feedback on their comments than their less active counterparts and thus interact more with other users, which leads to their increased review popularity. Therefore, a more detailed analysis is required to explain the findings of this research in the future. Finally, future investigation of review popularity in different product categories and of
the relationship between review popularity and company product sales is warranted.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors wish to thank the Ministry of Science and Technology of Taiwan for providing financial support (MOST 106-2410-H-011-006).

**ORCID iDs**

Lien Thi Kim Nguyen [https://orcid.org/0000-0002-6020-7519](https://orcid.org/0000-0002-6020-7519)

Kristine Velasquez Tuliao [https://orcid.org/0000-0003-0596-1479](https://orcid.org/0000-0003-0596-1479)

**References**

Al-Natour, S., & Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. *International Journal of Information Management, 54*, Article 102132.

Álvarez, L., Martín, A., & Casielles, R. (2007). Relationship marketing and information and communication technologies: Analysis of retail travel agencies. *Journal of Travel Research, 45*(4), 453–463.

Aseervatham, S., Antoniadis, A., Gaussier, É., Burlet, M., & Denneulin, Y. (2011). A sparse version of the ridge logistic regression for large-scale text categorization. *Pattern Recognition Letters, 32*(2), 101–106.

Bickart, B., & Schindler, R. M. (2001). Internet forums as influential sources of consumer information. *Journal of Interactive Marketing, 15*(3), 31–40.

Bjereng, E., Havro, L. J., & Moen, O. (2015). An empirical investigation of self-selection bias and factors influencing review helpfulness. *International Journal of Business and Management, 10*(7), Article 16.

Brown, J., Broderick, A. J., & Lee, N. (2007). Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing, 21*(3), 2–20.

Butler, B. S., & Wang, X. (2012). The cross-purposes of cross-posting: Boundary reshaping behavior in online discussion communities. *Information Systems Research, 23*(3), 993–1010.

Chen, T., & Guenstir, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). Association for Computing Machinery.

Chen, Y. R., Rezapour, A., & Tzeng, W. G. (2018). Privacy-preserving ridge regression on distributed data. *Information Sciences, 451*, 34–49.

Cheng, Y. H., & Ho, H. Y. (2015). Social influence’s impact on reader perceptions of online reviews. *Journal of Business Research, 68*(4), 883–887.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345–354.

Chua, A. Y., & Banerjee, S. (2016). Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. *Computers in Human Behavior, 54*, 547–554.

Crisci, R., & Kassinove, H. (1973). Effect of perceived expertise, strength of advice, and environmental setting on parental compliance. *The Journal of Social Psychology, 89*(2), 245–250.

Dimoka, A., Hong, Y., & Pavlou, P. A. (2012). On product uncertainty in online markets: Theory and evidence. *Management Information System Quarterly, 36*(2), 395–426.

Doh, S., & Hwang, J. (2009). How consumers evaluate EWOM (electronic word-of-mouth) messages. *Cyberpsychology and Behavior, 12*, 193–197.

Eslami, S. P., Ghasemaghaei, M., & Hassanein, K. (2018). Which online reviews do consumers find most helpful? A multi-method investigation. *Decision Support Systems, 113*, 32–42.

Fan, Z. P., Li, G. M., & Liu, Y. (2020). Processes and methods of information fusion for ranking products based on online reviews: An overview. *Information Fusion, 60*, 87–97.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research, 19*(3), 291–313.

Friedman, N., Linial, M., Nachman, I., & Pe’er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology, 7*(3–4), 601–620.

Gao, B., Hu, N., & Bose, I. (2017). Follow the herd or be myself? An analysis of consistency in behavior of reviewers and helpfulness of their reviews. *Decision Support Systems, 95*, 1–11.

Ghose, A., & Ipeirotis, P. G. (2010). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498–1512.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science, 31*(3), 493–520.

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science, 23*(4), 545–560.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing, 18*(1), 38–52.

Hu, Y. H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. *International Journal of Information Management, 36*(6), 929–944.

Huang, A. H., Chen, K., Yen, D. C., & Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior, 48*, 17–27.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing, 70*, 489–501.

Huang, L., Chou, Y., & Lan, I. (2007). Effects of perceived risk, message types, and reading motives on the acceptance and
transmission of electronic word-of-mouth communication. *Contemporary Management Research*, 3, 299–312.

Huang, R., & Sarigollu, E. (2012). How brand awareness relates to market outcome, brand equity, and the marketing mix. *Journal of Business Research*, 65, 92–99.

Huffaker, D. A., Swaab, R., & Diermeier, D. (2011). The language of coalition formation in online multipartiy negotiations. *Journal of Language and Social Psychology*, 30(1), 66–81.

Hung, C. (2017). Word of mouth quality classification based on contextual sentiment lexicons. *Information Processing & Management*, 53(4), 751–763.

Johanes, R., Pinem, A. A., Hidayanto, A. N., & Shihab, M. R. (2016, October). What features of online review affects readers’ intention to travel? In 2016 international conference on advanced computer science and information systems (ICACSIS) (pp. 263–270). Institute of Electrical and Electronics Engineers.

Kim, J. B. (2014). *Impact of online customer reviews and incentives on the product sales at the online retail store: An empirical study on video game titles at Amazon.com*. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1071&context=amcis2014

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, 10(1), 62–102.

Korfiatis, N., García-A-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3), 205–217.

Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems With Applications*, 42(7), 3751–3759.

Kumar, N., & Benbasat, I. (2006). Research note: The influence of recommendations and consumer reviews on evaluations of websites. *Information Systems Research*, 17(4), 425–439.

Lee, M., Rodgers, S., & Kim, M. (2009). Effects of valence and extremity of eWOM on attitude toward the brand and the website. *Journal of Current Issues and Research in Advertising*, 31(2), 1–11.

Lee, M., & Youn, S. (2009). Electronic word of mouth (eWOM): How eWOM platforms influence consumer product judgement. *International Journal of Advertising*, 28(3), 473–499.

Lee, S., & Choi, J. Y. (2016). The determinants of helpfulness of online reviews. *Behaviour & Information Technology*, 35(10), 853–863.

Li, G., & Niu, P. (2013). An enhanced extreme learning machine based on ridge regression for regression. *Neural Computing and Applications*, 22(3–4), 803–810.

Li, H., & Yuan, B. (1998, February). Chinese word segmentation. In *Proceedings of the 12th Pacific Asia conference on language, information and computation* (pp. 212–217). https://www.aclweb.org/anthology/Y98-1020.pdf

Li, S. T., Pham, T. T., & Chuang, H. C. (2019). Do reviewers’ words affect predicting their helpfulness ratings? Locating helpful reviewers by linguistics styles. *Information & Management*, 56(1), 28–38.

Li, X., Wu, C., & Mai, F. (2019). The effect of online reviews on product sales: A joint sentiment-topic analysis. *Information & Management*, 56(2), 172–184.

Li, Y. M., Lai, C. Y., & Chen, C. W. (2011). Discovering influencers for marketing in the blogosphere. *Information Sciences*, 181(23), 5143–5157.

Liang, Y. J. (2016). Reading to make a decision or to reduce cognitive dissonance? The effect of selecting and reading online reviews from a post-decision context. *Computers in Human Behavior*, 64, 463–471.

Liu, Y., Huang, X., An, A., & Yu, X. (2008, December). Modeling and predicting the helpfulness of online reviews. In *2008 eighth IEEE international conference on data mining* (pp. 443–452). Institute of Electrical and Electronics Engineers.

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, 47, 140–151.

Luan, J., Shan, W., Wang, Y., & Xiao, J. (2019). How easy-to-process information influences consumers over time: Online review vs. brand popularity. *Computers in Human Behavior*, 97, 193–201.

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzel, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, 77(1), 87–103.

Malik, M. S. I., & Hussain, A. (2018). An analysis of review content and reviewer variables that contribute to review helpfulness. *Information Processing & Management*, 54(1), 88–104.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon.com. *Management Information Systems Quarterly*, 34(1), 185–200.

Park, D. H., Lee, J., & Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, 11(4), 125–148.

Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of Tourism Research*, 50, 67–83.

Pavlou, P., & Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research*, 17(4), 392–414.

Racherla, P., & Friske, W. (2012). Perceived “usefulness” of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications*, 11(6), 548–559.

Ramos, J. (2003, December). Using TF-IDF to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning*, 242, 133–142.

Siering, M., Muntermann, J., & Rajagopalan, B. (2018). Explaining transmission of electronic word-of-mouth communication. In *Proceedings of the 12th Pacific Asia conference on language, information and computation* (pp. 443–452). Institute of Electrical and Electronics Engineers.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. (2017). Predicting the “helpfulness” of online consumer reviews. *Journal of Business Research*, 70, 346–355.

Srivastava, V., & Kalro, A. D. (2019). Enhancing the helpfulness of online consumer reviews: The role of latent (content) factors. *Journal of Interactive Marketing*, 48, 33–50.

Stargage. (2017). Influencer marketing in Taiwan. https://stargage.com/influencer-marketing-taiwan/

Sun, X., Han, M., & Feng, J. (2019). Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products. *Decision Support Systems*, 124, Article 113099.
Suzuki, K. (1978). Acceptance and rejection of a suggestion. *Japanese Psychological Research, 20*(2), 60–70.
Taiwan Cosmetic Market. (2018). https://www.export.gov.il/files/cosmetics/OverviewoftheTaiwanesecosmeticsmarket_2018.pdf
Vasconcelos, M., Almeida, J., & Goncalves, M. (2014, March). What makes your opinion popular? Predicting the popularity of micro-reviews in Foursquare. In *Proceedings of the 29th annual ACM symposium on applied computing* (pp. 598–603). Association for Computing Machinery.
Wan, Y. (2015). The Matthew effect in social commerce. *Electronic Markets*, 25, 313–324.
Wang, D., Zhu, S., & Li, T. (2013). SumView: A web-based engine for summarizing product reviews and customer opinions. *Expert Systems With Applications, 40*(1), 27–33.
Wang, F., & Karimi, S. (2019). This product works well (for me): The impact of first-person singular pronouns on online review helpfulness. *Journal of Business Research, 104*, 283–294.
Wu, J. (2017). Review popularity and review helpfulness: A model for user review effectiveness. *Decision Support Systems, 97*, 92–103.
Wu, X., Jin, L., & Xu, Q. (2020). Expertise makes perfect: How the variance of a reviewer’s historical ratings influences the persuasiveness of online reviews. *Journal of Retailing*. Advance online publication. https://doi.org/10.1016/j.jretai.2020.05.006
Yang, S., Yao, J., & Qazi, A. (2020). Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining. *Information Processing & Management, 57*(2), Article 102179.
Yin, D., Bond, S., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *Management Information Systems Quarterly, 38*(2), 539–560.
Yin, G., Wei, L., Xu, W., & Chen, M. (2014). Exploring heuristic cues for consumer perceptions of online reviews helpfulness: The case of Yelp.com. In *PACIS*, 52. https://aisel.aisnet.org/pacis2014/52/
Yoo, K. H., & Gretzel, U. (2011). Influence of personality on travel-related consumer-generated media creation. *Computers in Human Behavior, 27*(2), 609–621.
Zablocki, A., Makri, K., & Houston, M. J. (2019). Emotions within online reviews and their influence on product attitudes in Austria, USA and Thailand. *Journal of Interactive Marketing, 46*, 20–39.
Zhang, T., & Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information Retrieval, 4*(1), 5–31.
Zhang, W., Du, Y., Yoshida, T., & Wang, Q. (2018). DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management, 54*(4), 576–592.
Zhao, X., Wang, L., Guo, X., & Law, R. (2015). The influence of online reviews to online hotel booking intentions. *International Journal of Contemporary Hospitality Management, 27*(6), 1343–1364.
Zhao, Y., Yang, S., Narayan, V., & Zhao, Y. (2013). Modeling consumer learning from online product reviews. *Marketing Science, 32*(1), 153–169.
Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*(2), 133–148.
Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems With Applications, 58*, 93–101.