Text mining-based analysis of online comments for skincare e-commerce

Wen Hong¹, Yiping Wu², Shangze Li³, Ying Wu⁴, Zehai Zhou⁵, Yanbin Huang¹*

¹ Management science and Engineering, Fuzhou University, Fuzhou, Fujian, 350108, China
² Accounting, Fuzhou University, Fuzhou, Fujian, 350108, China
³ Accounting, Fuzhou University, Fuzhou, Fujian, 350108, China
⁴ Accounting, Fuzhou University, Fuzhou, Fujian, 350108, China
⁵ Statistics, Fuzhou University, Fuzhou, Fujian, 350108, China
*Corresponding author’s e-mail: 1153435445@qq.com

Abstract: This article uses online review data of skincare products on e-commerce platforms to summarize the consumer demand characteristics through text clustering analysis and offer common marketing proposals for skincare merchants. From the role of review text clustering analysis, this paper derives two dimensions from e-commerce platforms and skin care product categories, and through feature extraction and lexical item clustering analysis of consumer online review information on different platforms, the focus of attention and characteristic tendencies of consumers on skin care products on different platforms are mined. In turn, the review information can be mined and analyzed to obtain information with business value, and relevant measures can be taken to improve the platform’s services, promote business growth, enhance customer satisfaction, etc. First, develop a reasonable marketing strategy. Second, strengthen product branding.

1. Introduction
At present, the research on consumer comment mining has achieved some results. Xuan Hu [1] analyzed the knowledge of text clustering technology for the characteristics of microblog information, designed and completed the process of microblog information analysis based on text clustering. Li Renyi [2] started from the definition of data mining and cluster analysis, briefly clarified the main methods and basic process, focused on the classification and the typical clustering algorithms of cluster analysis, and did experimental analysis for certain clustering algorithms combined with applications, and came up with some valuable results. Chen [3] constructed a recommendation model based on online reviews by taking clothing, shoes and hats products as an example, analyzing the sentiment of online reviews and calculating the similarity between users by adopting collaborative filtering recommendation algorithm after reasonably quantifying the ratings. Li Yanfeng [4] analyzed the sentiment tendency embodied by users in reviews into sentiment ratings, and then applied the similarity based on user ratings, user sentiment tendency and temporal order as a comprehensive similarity to a personalized recommendation model.

Online reviews are direct feedback from consumers on the process and content of consumption on online platforms after purchasing goods or services. It is an important source of information with
sufficient and rich data, featuring initiative, effectiveness, accuracy, and authenticity. Therefore, it is significant for e-commerce merchants to identify high-quality reviews from uneven ones and obtain accurate and valuable information from them.

This article uses online review data of skincare products on e-commerce platforms to summarize the consumer demand characteristics through text clustering analysis and offer common marketing proposals for skincare merchants.

2. Research content
The article is structured as follows. In the beginning, we briefly describe the purpose and significance of text mining. We then introduce the research framework structure and method adopted in the article.

The body of the article uses R language to cluster and analyze the text data of skincare product reviews. Based on the clustering results of both platform and category dimensions, the article represents the demand characteristics of consumers and suggests improvements for merchants' accurate marketing from a qualitative perspective.

Last, based on the foregoing analysis and conclusions, the significance and value orientation of this article are further clarified.

3. Data text clustering analysis

3.1. Feature word selection
TF means Term Frequency and IDF means Inverse Document Frequency. TF-IDF method calculates the weight of a word in the whole text collection based on the frequency of the word and the text it appears in, and then selects features. The higher the weight, the stronger the word's ability to distinguish the text, otherwise the weaker it is.

For the word $t$, the weight is calculated as shown in Equation 1

$$\text{Tfidf}(t) = \text{tf}(t) \times \log \left( \frac{N}{df(t)} \right)$$

After preprocessing and text representation of the captured comment text data, we extract feature items from Tmall, Jingdong and Vipshop datasets respectively. The results are shown in Figure 1 to 3.

![Figure 1 Feature items and weight values of Tmall dataset](image1)

As shown in Figure 1, according to the feature item results of the Tmall dataset, the feature words "Absorb", "good", "effect", "Moisture" and "nourishing" ranked top five in weight, which explains that consumers paid more attention to the efficacy of skincare products when purchasing them on Tmall. The word "package" ranked sixth. Also, it can be found that consumers had both positive and negative comments on the packaging, indicating that when buying skincare products on the Tmall, packaging was the second factor that consumers consider after efficacy.

![Figure 2 Feature items and weight values of Jingdong dataset](image2)

As shown in Figure 2, according to the results of characteristic items of the Jingdong dataset, the weight of "effect" ranked first, which indicates that consumers of skincare products on Jingdong cared
more about the efficacy of products. The term "Jingdong" ranked 4th. By observing the original review data, we can see that "Jingdong" often appeared in the words "Jingdong Mall", "Jingdong self-operated", "Jingdong delivery man" and "Jingdong Logistics"(e.g. "It's late, but the Jingdong delivery man is still delivering." "Trust Jingdong self-operated." "Special praise to Jingdong Express."). It indicates that consumers had strong platform awareness when shopping on Jingdong, and that identification with the platform may help improve their satisfaction with the products.

As shown in Figure 3, according to the results of the feature items of the Vipshop dataset, the words "effect" and "good" ranked in the top two, indicating that consumers of skincare products on the Vipshop were also most concerned about the effectiveness of the products. The word "Vipshop" ranked 3rd, we can find that Vipshop skincare consumers also had a strong sense of the platform and a high degree of recognition. "Brand" ranked 12th, combined with the context of the review text, we found that "brand" is mostly found in statements such as "always use this brand" and "like this brand", showing that consumers of skincare products on Vipshop had strong brand loyalty.

3.2. Word item clustering

According to the method described in the previous section, this article adopts the systematic clustering method (i.e., hierarchical clustering method) to cluster the word items on the datasets of Tmall, Jingdong, and Vipshop. The purpose is to make the filtered feature items aggregated into multiple clusters, with low similarity of feature items among the clusters and high similarity of feature items within the clusters. The sum of squares of deviations method is used for the hierarchical clustering of the feature terms retained in all three datasets. The results are shown as follows.

We can see that the most similar categories such as "good - Buy again", "Cost-effective-giveaway", "product - customer service", "Moisture - nourishing", and "good - effect" appear based on their similarity. The "product-customer service" combination shows that Tmall's customer service has a strong relationship with consumers' shopping satisfaction. From the combination of "Cost-effective-giveaway", it can be seen that skincare products with free giveaways may increase consumers' psychological satisfaction and make them think that the product is more cost-effective (e.g., "The giveaways are also Cost-effective", "Very good, they give a lot of giveaways, very practical and cheap" "There are giveaways, very Cost-effective", etc.), while for products that do not contain giveaways, consumers may even express disappointment (e.g., "It's OK, sad that there is not a single giveaway" "It is good to use, but there is no giveaway at all......" "It's okay, I thought there were giveaways, a little disappointed?" etc.), it can be seen whether the inclusion of giveaways has a certain impact on Tmall's consumer satisfaction.

The results of the hierarchical clustering on Jingdong showed the most similar classes such as "effect-good", "package-logistics", "refreshing-greasy" and "moisture-hydrate". From the "package-logistics" and Jingdong's original review data, we found that when consumers evaluate Jingdong's logistics, they will evaluate the quality of the logistics service with the consideration of the integrity of the packaging, in addition to their requirements for the speed of the logistics (e.g., "The logistics is fast, the packaging is excellent, no damage at all." "The goods are received, the logistics is very fast, the package is intact.", etc.), indicating that consumers had a high level of satisfaction with the speed and quality of Jingdong logistics.

After hierarchical clustering of the Vipshop dataset, the most similar classes were obtained, such as "moisture-hydrate", "absorb-greasy", "brand-trust", "product-trust", and "effect-good". It was found that consumers generally agree with the authenticity of brands on Vipshop and have a good impression.
of the platform as authentic (e.g., "Trusted big brand, many times Vipshop shopping, genuine products" "Trusted brand, Vipshop is authentic"). This shows that Vipshop's "brand name discount + limited time purchase + genuine guarantee" mode has gained consumers' brand trust. The "product-brand" combination represented that consumers had high recognition and loyalty to the brand of the products on Vipshop.

3.2.1. Comparative Analysis

(1) Similarity Analysis

By utilizing the hierarchical clustering approach to cluster the document terms, it is illustrated from the results that the word "effect" appears with high weight and similar clustering combinations such as "effect-good" are reflected in all three platforms. This shows that no matter which platform, the most important thing for consumers is still the efficacy, followed by the smell, the applicable skin type, etc., and the third involves the shopping experience, such as customer service quality, logistics service quality, etc.

(2) Difference Analysis

After comparing the feature term results of the three platforms, their respective differential feature terms (top 25) and differential term combinations were derived as shown in Figure 4-5. By observation, we can reasonably infer that consumers on the Tmall platform pay more attention to the cost-effectiveness of products (cost-effective, giveaway) and the quality of customer service (customer service), while the platform awareness is not strong. Consumers on Jingdong have high requirements for the quality of the products (delicate and tight) and the platform itself (logistics). Consumers on Vipshop have strong platform and brand recognition, brand loyalty, and word-of-mouth effects.

![Figure 4](image)

Figure 4 Comparison of feature terms and clustering results for each platform

We found that the reason for these differences lies in the different positioning of each platform (see Table 5). Tmall focuses on cost efficiency and customer service, Jingdong focuses on logistics service, and Vipshop places more emphasis on guaranteeing authentic products. Meanwhile, in terms of platform awareness, the weight of "Jingdong" and "Vipshop" is high on both platforms, indicating that the self-operated model has increased consumers' perception of the platforms. Furthermore, Jingdong and Vipshop are more involved with consumers while Tmall's reseller model allows consumers to be more store-specific, and merchants on Tmall have more options in terms of sales, logistics, and services.

![Figure 5](image)

Figure 5 Positioning characteristics of the three major platforms

3.3. Countermeasures and proposals

3.3.1. Develop product-platform marketing strategy

Consumers on Tmall pay more attention to cost performance and customer service quality, so merchants can choose products with higher cost performance and strengthen customer relationship management;
consumers on Jingdong have higher perceptions of logistics, so it is recommended that merchants actively seek cooperation with Jingdong logistics to improve the service.

**Strengthen product branding**

We can see that the weight value of "authentic" is high, indicating that consumers are very concerned about the authenticity of products. Consumers are accustomed to brand-oriented product selection to measure the quality of such expensive products as serums, lotions, and creams. Therefore, merchants should focus on branding their products.

**4. conclusion**

1. Finally, it can be concluded that for e-commerce platforms, on the one hand, they can focus on the promotion of online activities to create a pro-people atmosphere; on the other hand, they can improve the logistics and packaging aspects of skin care products to make customers have a pleasant consumption experience.

2. Suggestions:

   In the context of the era of big data, the shopping platform converts the review information into data and stores it in the data system, which can then be mined and analyzed to get information with commercial value, and take relevant measures for this information to improve the platform's services, promote business growth, enhance customer satisfaction, etc. First, develop a reasonable marketing strategy. Second, strengthen product branding.

3. Significance of this research:

   Based on the era of big data, this paper analyzes online reviews of skin care products on e-commerce platforms to obtain high-value business and social information, which helps merchants to develop more accurate marketing strategies.

   From the role of comment text clustering analysis, this paper derives two dimensions from e-commerce platforms and skin care product categories. First, by feature extraction and lexical item clustering analysis of consumer online comment information on different platforms, the focus of attention and characteristic tendencies of consumers on skin care products on different platforms are unearthed. With the development of e-commerce in China, more and more shopping platforms have emerged, and the competition between platforms has become more and more intense. The marketing models between different platforms have their own strengths, then both for platforms and merchants can find the importance consumers attach to customer service, logistics, packaging, etc. through the platform-based clustering experiments in this chapter, and when platforms or merchants make optimization choices, they can be considered according to the weight values to maximize optimization efficiency. Secondly, the clustering experiments on online reviews of skin care products of different categories help merchants selling skin care products of different categories and efficacy to derive the product features that consumers value most for such products, and dig out the hierarchical relationships between product features, and finally make a reasonable combination of sales to meet the diverse consumer needs of the public and finally gain loyal customers.

**References**

[1] Xuan Hu. Research on the analysis of microblog information based on text clustering [D]. Wuhan University of Technology. 2014.

[2] Li Renyi. Research and application of clustering analysis algorithm in data mining [D]. University of Electronic Science and Technology.2012.

[3] Chen Xiaoting. Research on recommendation model for customer purchase problem on e-commerce platform - taking clothing, shoes and hats as an example [D] Capital University of Economics and Business.2018

[4] Li Yanfeng. Research on personalized product recommendation system based on e-commerce reviews [D]. Harbin Institute of Technology.2016.