Research on Topic Mining of Medical Surgical Mask Reviews Sold on E-commerce Platform

Jiao Qidi

School of International Education, Wuhan University of Technology, Wuhan, Hubei, 430070, P.R. China

*Author’s e-mail: jiaoqidi@whut.edu.cn

Abstract. Since the outbreak of the epidemic in 2020, the development potential of pharmaceutical e-commerce is huge. Mining and studying the review information of e-commerce platform to improve the dimensions of products is an important way for the healthy development of medical e-commerce. In this paper, the review information of medical surgical mask, a hot sales product of medical e-commerce, is extracted. TF-IDF algorithm, dependency syntax and other models are used to mine the topic information contained in the review, and then python is used to analyze the emotional polarity of the review, and finally the emotional assignments of the main review topics are obtained. The results show that consumers have higher satisfaction with the appearance of goods, but lower satisfaction with packaging. On the basis of the research results, corresponding reference suggestions are proposed for medical e-commerce.

1. Introduction

In recent years, the proportion of e-commerce market scale in the national economy has been increasing. With the outbreak of the Covid-19 pandemic in 2020, the whole country has taken measures, such as community closure management and traffic control, therefore online consumption has began to get more people’ s favor[1]. During this period, the e-commerce economy highlights its unique advantages and has become an important force in stimulating China's economy and made great contributions to the recovery of China’s economy. According to the 46th CNNIC Internet Use Report released by China Internet Information Center, as of June 2020, China’s online shopping users reached 749 million, an increase of 39.12 million compared with March 2020, accounting for 79.7 % of the total Internet users. During the epidemic period, in order to cope with the highly infectious Covid-19, the demand for epidemic prevention products represented by medical surgical masks sold online in China surged. Sales of masks on e-commerce platforms rose 68 times between January and May, according to Selixin's WPOS e-commerce monitoring system. Facing the opportunities created by the epidemic, medical e-commerce has great potential in the e-commerce field. In order to further understand the advantages and opportunities of medical e-commerce, this paper analyzes the focus of consumers and the main reasons affecting the sales of masks by mining the online review data of medical masks in Jingdong, a large e-commerce platform, in order to put forward feasible suggestions for medical e-commerce and e-commerce platform, and promote the healthy development of related industries.
2. Literature review
B2C pharmaceutical e-commerce relies on network technology between online pharmacies and consumers[2]. Since this year, the outbreak of the Covid-19 pandemic has triggered people’s health awareness, and the online trading scale of medical and health products has also increased greatly. Affected by the Covid-19 disease, the enthusiasm of medical e-commerce is rising. It is expected that the scale of medical e-commerce transactions will directly force 200 billion yuan in 2020, reaching 187.64 billion yuan at increase rate 94.58 %[3]. Rapid intelligent medical and telemedicine have dramatically developed[4]. During the epidemic prevention and control period, online consumption demand increased significantly, entertainment, medical, education, office and other activities are generally carried out online[5]. The epidemic has produced different degrees of positive stimulus effects on pharmaceutical e-commerce and contact-less distribution industries[1]. E-commerce companies can improve their product or services on the basis of people’s opinion and current trends[6].

3. Research Models and Methods

3.1. Research object
Jingdong Mall is a large comprehensive platform for the early development of pharmaceutical e-commerce in China. It has a large number of merchants and huge turnover, and has accumulated rich user reviews. During the outbreak, the sales of medical surgical masks have increased dramatically, and masks have become new social needs. Therefore, the study of consumers' comments on medical surgical masks is the key to the market share of medical e-commerce in the post-epidemic era. This paper selects the online reviews of medical surgical masks sold in Jingdong Mall since January 2020 as the research object, and studies consumers’ comments on online medical surgical masks under the background of the epidemic, so as to conduct in-depth analysis of user needs and behaviors.

3.2. Research models

3.2.1. Topic construction of product evaluation. By Pycharm, TD-IDF model is used to extract high-frequency words from comment text information, and preliminarily finds consumers’ attention hotspots and emotional tendencies. TF-IDF (Term Frequency-Inverse Document Frequency, word frequency-inverse file frequency) It is a method to evaluate the importance of words in the current article, if a word appears only in a certain article with a high frequency, it is considered to be more important for the article[7].

\[
TF - IDF = TF \times IDF
\]  

(1)

According to the characteristics of mask products, the top 500 high-frequency words after TF-IDF were screened and manually classified, then combined with feature-emotional words extracted from comment set, the product index system was summarized and constructed, such as ' comfort ', ' quality ' and ' price '.

3.2.2. Feature - emotional word extraction. Based on the dependency parsing proposed by French linguist L. Tesniere, the opinion-emotional word pairs in the comments are extracted to construct the feature-emotional word library. Dependency parsing is based on analysis and extraction of syntactic structure and grammatical components[8]. This study uses dependency parsing to extract feature words. According to the selection of feature words, the extracted results are standardized, such as ' rope ' and ' ear rope ' belong to the ' comfort ' feature of mask, so these two feature words should be classified as the evaluation topic ' comfort '. After the above process, finally get the < topic, feature words, emotional words > set of the review text, such as < price, cost-effective, high >, < comfort, earrings, tight >.

Sentiment analysis. Sentiment analysis is a technique that automatically analyzes a text content and then discovers the emotional attitude contained in the text content[9]. According to the above feature-emotional words extraction results < topic, feature words, emotional words >, using python’s snownlp
library, assign the feature emotional words, get the four-dimensional array \(<\text{topic}, \text{feature words}, \text{emotional words}, \text{score}>\). After traversing the entire text library, calculate the average emotional score for Topic \(i\).

\[
\text{Score}(T_{ki}) = \left[ \sum_{j=1}^{n} \text{score}(T_{kj}) \right]^{-1} n
\]

\(i\) : Topic \(i\)  
\(k\) : Merchant \(k\)  
\(j\) : Total frequency of feature words under \(T_{ki}\)

4. Empirical Analysis

The methods and steps of this study mainly include data acquisition, data preprocessing, topic system of product construction, feature-emotional words extraction, emotional analysis and establishment of multiple regression model.

4.1. Data acquisition

Octopus data acquisition system can simply and efficiently crawl online comment information. Therefore, this experiment selected Octopus collector to collect medical mask review information in Jingdong. To get a more general conclusion, this article crawls 80,200 comments and sales data from 189 merchants respectively.

4.2. Data preprocessing

The repeated comments of the same user, garbage comments and irrelevant comments in the commodity review area will affect the accuracy of the experiment, so the original comment corpus is artificially denoised to ensure the validity of the data. After preprocessing, 73,363 comments. Secondly, after cleaning the document, word segmentation tool Jieba system is used for text word segmentation processing. Then, by updating the stop words dictionary, the table removes words and punctuations that have no practical significance for the experiment, and obtains the final normative review data.

4.3. Construct topic system of product

In this paper, the TF-IDF model is used to extract the top 500 high-frequency words from the standard review data, and part of high-frequency words and their TF-IDF results are showed in table 1. According to the extracted product feature results for manual screening, delete such as ' feel ', ' beautiful ', retain product feature words, like ' quality ', ' price '. Then, artificially classify these feature words and sum up the product evaluation topics, which are quality, Packaging, Logistics, Price, Appearance, Comfort and Service. For example, feature words ' quality ', ' filtering ' belong to the quality topic; ' permeability ' and ' ear rope ' belong to the Comfort topic. Some feature words and their topics are illustrated in table 2.

| Table 1. Part of TF-IDF results |
|---------------------------------|
| Quality | 1.0 | Price | 0.41052508589919157 |
| Packaging | 0.7807340296457397 | Independence | 0.3527618446729316 |
| Pandemic | 0.3302003996072771 | Logistics | 0.3176606722061638 |
| Satisfy | 0.3131341261877388 | Delivery | 0.24324284094968762 |
| Economical | 0.23184200008646993 | Speed | 0.2222878755505287 |

| Table 2. Part of topic - feature wordlists |
|-------------------------------------------|
| Topic | Feature Words |
| Quality | quality, brand, protection, smell, safety... |
| Price | price, discount, cost-effective, raise, inexpensive... |
| Appearance | color, style, pattern, look, white... |
4.4. Feature - emotional word extraction

4.4.1. Syntactic path summary. In the human language structure, there are mainly 14 structures among the sentence elements, including subject-predicate relationship (SBV), definite-center structure (ATT), verb-object structure (VOB), predicate-complement structure (VCS), and adverbial-mid structure (ADV). For example, the structure of comment “Quality of mask is great,” can be summerized as ‘n SBV> a’. Combing the syntactic structure of the comments can help us locate the feature words in the comments.

4.4.2. Dependency parsing. In this paper, the LTP natural language processing tool Pytlp of Harbin Institute of Technology is used to analyze the dependency syntax of sentences, and the feature-emotional words are extracted according to the specific syntactic path form. Some examples are as follows: < quality, good >, < fit, good >.

4.4.3. Feature standardization processing. Since different people have different expressions for the same feature, and there are lower-level attributes under the same feature, such as color and size of appearance, the extracted results can be standardized according to the topic system of product. At the same time, the noise data extracted due to the diversity of Chinese grammar ( eg: yes & relatively coarse ) can be filtered out to form standard results, such as < quality, function, good >, < appearance, color, fashionable >.

4.5. Sentiment analysis
Snownlp sentiment analysis library is a common sentiment analysis tool, so this paper uses snownlp to analyze comments. In order to improve the accuracy of sentiment analysis of reviews, additional training was conducted after artificial classification of some mask reviews on the original sentiment analysis model. The specific processing steps are as follows. First, calculate the score of each feature word by snownlp of merchant k (k = 1, 2,... 188). Then, according to (2), calculate score of T\textsubscript{ki} (i=1, 2,..., 7) of merchant k. Finally can get the emotional score of each merchant. Some of the results are shown in table 3.

| Merchant | Price | Quality | Packaging | Service | Logistics | Comfort | Appearance |
|----------|-------|---------|-----------|---------|-----------|---------|------------|
| 1        | 0.910103785 | 0.765312912 | 0.620720354 | 0 | 0.807627367 | 0.772990575 | 0 |
| 2        | 0.774275846 | 0.759438911 | 0.685097246 | 0.683371586 | 0.704071014 | 0.760992759 | 0 |
| 3        | 0.843130099 | 0.76009669 | 0.727401646 | 0.695367745 | 0.760981837 | 0.822514792 | 0.807294927 |
| 4        | 0.813609683 | 0.75778263 | 0.660550262 | 0.678449192 | 0.782943508 | 0.774731025 | 0.810433751 |
| 5        | 0.801366497 | 0.754322006 | 0.666710114 | 0.66165787 | 0.779913838 | 0.786660257 | 0.81392801 |
| 6        | 0.80096816 | 0.748249072 | 0.685953183 | 0.661958641 | 0.788521269 | 0.819163502 | 0.842929745 |
| 7        | 0.80478333 | 0.742628507 | 0.695284276 | 0.659688156 | 0.785369553 | 0.80277049 | 0.845399866 |
| 8        | 0.802664023 | 0.743581571 | 0.692148797 | 0.659762174 | 0.787582892 | 0.801553258 | 0.846639021 |
| 9        | 0.809564081 | 0.741775927 | 0.690904631 | 0.66069587 | 0.78410561 | 0.793104352 | 0.850877169 |
| 10       | 0.806645853 | 0.740744323 | 0.677798581 | 0.650976101 | 0.781154216 | 0.790063469 | 0.852011204 |
After obtaining the emotional scores of each comment topic of each store, this paper conducts weighted processing according to the sales data of each store, and finally obtains the weighted emotional scores of each topic of all stores. The results are illustrated in table 4.

Table 4. Weighted sentiment scores of topics

|                       | Price       | Function     | Packaging    | Service     | Logistics   | Comfort     | Appearance  |
|-----------------------|-------------|--------------|--------------|-------------|-------------|-------------|-------------|
| Price                 | 0.803639707 | 0.728417935  | 0.65310446   | 0.667391432 | 0.757376009 | 0.775084115 | 0.865737323 |

5. Conclusions and Suggestions

According to the results of the emotional analysis module, it can be seen that the emotional scores of seven topics are all positive, indicating that the consumption experience of medical surgical masks is optimistic. Among the seven topics, appearance scored the highest, followed by price, comfort, logistics, quality, service, and in all topics, packaging scored the lowest. Therefore, based on the statistics of high-frequency words and emotional scores, this paper provides the following suggestions for medical e-commerce. First, in the sales process of products, pharmaceutical e-commerce should focus on strengthening the evaluation plates such as packaging and services with relatively low consumer satisfaction. For example, in the process of high-frequency word statistics, it is found that the word 'independence' under the packaging topic has the highest frequency, so it can be speculated that consumers attach great importance to the packaging of masks is independent packaging. Therefore, in the future, merchants can adopt safer independent packaging for masks. Second, in the high-frequency word statistics module, the frequency of feature words that reflect the quality topic accounted for the highest proportion, so it can be known that consumers pay more attention to the quality of masks. However, in the process of emotional analysis, it is found that the score of quality topic is low, indicating that the quality of masks sold online is uneven. Therefore, merchants should enhance their sense of social responsibility and strictly control the quality in the process of mask manufacturing, which not only provides goods to satisfy consumers, but also makes positive contributions to the prevention and control of epidemic in China.

References

[1] Liao, M. (2020) Influence of coronavirus pneumonia epidemic on China’s consumption market and response to fiscal and tax law. The Journal of Humanities, 20–29.
[2] Erdem S.A., Chandra A., (2003) E-commerce in healthcare and pharmaceutical marketing—opportunities and concerns. Clin Res Regul Aff, 20(4):399–407.
[3] Hu, W., (2020) High heat of medical e-commerce in post-epidemic era. China's Foreign Trade, 64–65
[4] An, G., Jia, F., (2020) Analysis on the influence of new crown epidemic on economy and countermeasures. Financial Theory and Practice, 45–51.
[5] Zhang, Y., Zeng, J., (2020) Significance and measures of promoting consumption in the context of coronavirus pneumonia. Journal of Graduate School of Chinese Academy of Social Sciences, 53–64.
[6] Agarwal, B., Mittal, N., Bansal, P., et al. (2015) Sentiment analysis using common-sense and context information. Computational Intelligence & Neuroscience, 2015: 30.
[7] Wang, L., Xiao, X., Zhang, L., (2020) A comparative study of TF-IDF and Word2vec in news text classification. Computer Knowledge and Technology, 220–222.
[8] Wang, H., (2019) Research on user requirements mining based on feature extraction and sentiment analysis. M. S. thesis, Zhejiang Sci-tech University, Hangzhou, China.
[9] Zhang, Z., Ye, Q., Li, Y., (2010) Review on emotional analysis of Internet commentary. Journal of Management Sciences in China, 13(6): 84–96.