Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/47/20241294

Statistical and sentiment analysis based on comments of skin care products

Yunhan Jiang

School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian, Liaoning, 116025, China

fport62169@student.napavalley.edu

Abstract. With the continuous development of the Internet economy and the strong promotion in various aspects, the rise of new domestic brands has led to an increasing number of consumers paying attention to and purchasing cosmetics online. Grasping the pulse of the times in this context is a crucial key to understanding the Internet economy. This paper first crawled the product links of the JD.com skin care essence category, randomly selecting 100 products. Subsequently, a total of 49,560 comments were crawled for the selected products. Furthermore, employing text mining techniques, this paper conducts word frequency statistics, generates word clouds, performs cluster analysis, and analyzes the emotions expressed in the JD.com skin care essence product comments, aiming to achieve comprehensive mining and analysis of the comment content from multiple perspectives. Additionally, descriptive statistics and visual representations are utilized to enhance the accuracy and intuitiveness of the text processing. Finally, this paper summarizes the insights gained from the product comments on the internet e-commerce platform and provides some prospects for future development.

Keywords: Sentiment Analysis, Cluster Analysis, Text Mining.

1. Introduction

Based on the data disclosed by major shopping platforms during shopping festivals, it is evident that this year's transaction volume has maintained an increasing trend compared to last year. Despite the ongoing decline in online consumption due to the relaxation of the epidemic, Alibaba's "Double Eleven" product transactions still experienced an 8.45% year-on-year increase in 2022, reaching a transaction amount of 540.3 billion yuan. However, this growth rate fell compared to the 26% growth rate in 2021 during the same period. Similarly, JD's product transactions reached 349.1 billion yuan, but the 29% year-on-year growth this year was slightly lower compared to the 33% growth rate last year.

In the flourishing online shopping carnival, cosmetics and skincare products have demonstrated a leading advantage. With the continuous development of the economy and the improvement of people's living standards, Chinese consumers have shown an increasing trend in the consumption of cosmetics and skincare products, with higher unit prices. According to public data, China's cosmetics retail sales reached 204.9 billion yuan in 2015 and 340 billion yuan in 2020. It is expected that the market size of China's beauty industry will reach 455.3 billion yuan by the end of 2021. The expanding market for cosmetics and skincare products also reveals a secret: studying the core features of skincare products is

DOI: 10.54254/2755-2721/47/20241294

not only beneficial for selecting superior products from a wide range of options but also for understanding industry trends and capturing the pulse of internet consumption.

The internet consumption industry exhibits certain differences between supply and demand, creating a new gap in e-commerce. For major platforms, conducting text mining analysis on e-commerce reviews can provide more accurate information about user needs, assisting the platform in developing sales plans that align with those needs. For merchants, analyzing their own product reviews can provide a deep understanding of the industry's development situation and background, serving as a useful reference for their own development planning.

In the era of big data, major e-commerce websites have become crucial online shopping platforms, and the comment information released by these platforms can objectively and faithfully reflect the current demand and salary levels of relevant positions and industries. This helps buyers obtain relevant information, enhance shopping satisfaction, and improve their overall shopping experience.

2. Related work

In the current era of big data, a massive amount of data is generated every day, containing important and valuable information that is deeply buried within this unstructured data. Effectively extracting this information is crucial, and text mining has played a significant role in this process. Text mining focuses on massive text data as the object and performs automatic processing to extract textual information from online reviews, thereby obtaining important insights implied in the textual data. In recent years, research on text mining has yielded remarkable results, with applications utilizing sentiment dictionaries for classification and machine learning algorithms for sentiment analysis.

2.1. Data crawling and cleaning

The concept of text mining was first proposed in 1995. It is a natural language processing technology based on machine information retrieval and extraction, combined with statistical ideas and machine learning theory for text analysis. Among them, the word segmentation of text and mining and application are relatively mature.

In the study of text word segmentation, it mainly focuses on the identification of ambiguous words and unregistered words, as well as the accuracy and efficiency of word participle. Zhang et al. proposed a Chinese participle method based on the converging nature of the participle [1].

In terms of the specific application of text mining, it mainly realizes the information extraction, association analysis, classification and clustering of text in various fields. Donghua Zhu et al. introduced network analysis and neural network methods to visualized information [2]. Wang Dizhi et al. (2020) used k-means clustering method to cluster government document summary, text, and other contents, so as to realize the application of text mining in government text classification [3].

2.2. Sentiment analysis

The online comments text contains consumers' attitudes towards the product and emotional tendencies, emotional analysis. The research method is mainly based on the emotion dictionary, machine learning algorithm and deep learning algorithms.

Foreign scholars mainly focus on English text analysis. Turney P Using the mutual information between the pending word and the positive seed word, minus the mutual information with the derogatory seed word, to get the emotional tendency weight of the pending word [4]. Based on a corpus of 40 0,000 daily events, LiuH uses four language models to extract emotional words and construct an emotional dictionary for emotion analysis [5]. Paltoglou Combining with the linguistic emotion classification method of emotion intensity and negative words, they classified Twitter and other social media comment texts, with an accuracy rate of 86.5% [6].

Emotional tendency analysis based on machine learning and deep learning methods, which does not rely on emotional dictionary and semantic information, generally has high accuracy and is widely used. Pang took the lead in using machine learning method in emotion analysis [7]. He used three models of maximum entropy, naive Bayes and support vector machine to classify film and television data, and

DOI: 10.54254/2755-2721/47/20241294

found that support vector machine had the best effect. Based on the principle of advance and AdaBoost idea, Tang Xiaobo et al. proposed a classification algorithm of regression support vector machine, AdaBoost with support vector machine (SVM)-L, to realize the emotional polarity judgment and threshold visualization [8].

2.3. Cluster analysis

Hao introduced machine learning algorithms such as BP neural network and support vector machine into the practice of book classification, proposing the method of feature weighting design, focusing on the construction of a shallow classification system [9]. Wu Yan-chao proposed a support vector machine (SVM) algorithm based on hierarchical clustering and local learning at fixed layer (HCFLL) to address large-scale classification issues [10]. Gang Lu put forward an accurate and scalable traffic classifier based on Naive Bayesian algorithm for Tibetan text classification [11].

2.4. Research focus

From a review of the literature, it could be found that many articles in text mining have employed traditional and single algorithms such as the Naive Bayes algorithm and Support vector machine algorithm, among others. Two examples of these methods are SVM and K-Means cluster analysis. Additionally, most studies on comment data in the existing literature, both domestically and internationally, have compared the comment time and attitude with online user comment data. However, emotional, and clustering analysis of user comment text content is rare. It is evident that text mining of user comment text content can fully delve into the information within the comments. Whether the latent information in the comment information can be fully tapped to discover users' concerns and attitudes towards the product, improve product quality, optimize services, and enhance market competitiveness of products is very important. Therefore, text mining is an aspect that needs to be emphasized.

3. Method

3.1. Data crawling and cleaning

Web crawler is a technology that automatically downloads the required information from the website according to specific rules, and realizes quick and regular access and capture by writing programs or scripts. Its working principle is to first select a webpage as the seed webpage, and download it to the crawler to process, and the crawler identifies all hyperlinks and adds them to the queue. Next, the URL in the queue is sent back to the crawler, the valid information is stored in the memory, and the hyperlink in the URL is identified again and added to the pending queue, so that all URLs are crawled.

Text data cleaning includes steps such as text weight removal and text denoising. First of all, for the collected comment text data, there is often a certain amount of duplicate data, which needs to be reprocessed. In this data set, there is duplication between different lines of comments. In order to remove this redundancy, only the earliest published comment is retained. Secondly, in Chinese text, meaningless punctuation marks, expressions and numbers need to be removed. If these text noise is removed, regular expressions can be used to search for and remove them, or a list of stop words can be used to eliminate them. Finally, for some abnormal comments that are not duplicates but have irrelevant content, they also need to be removed.

3.2. SVM-based sentiment analysis

The principle of SVM classification is that there is such a hyperplane in the training set, which can separate the data into different categories. As there are many hyperplanes that can be used to separate the data into different categories, it is necessary to find the one with the best performance. SVM finds this hyperplane by using support vectors.

DOI: 10.54254/2755-2721/47/20241294

3.3. K-Means clustering sentiment analysis

The data in this article is unstructured, but research requires structured data for model analysis. Therefore, the first step is to perform data transformation. This article uses the vector space model as a method for transformation. The vector space model assumes that each document can be represented by a collection of feature terms and weight vectors. The set of feature values t represents a text, and the n-dimensional vector D(T1, T2, ..., Tn) represents the feature terms. Each feature term has a weight that represents its importance in a document. Therefore, a document can be represented by the combination of feature terms and weights, denoted as D = D(W1, W2, ..., Wn), abbreviated as D = D(Wi,i=1,2,...,n). Among them, Wk is the weight of Tk. In summary, each document can be represented by any vector in an n-dimensional space. How to assign feature weights is crucial. Boolean model and TF-IDF weight are commonly used two ways of weighting. This article chooses to use TF-IDF weights for weighting. The core of TF-IDF is to effectively distinguish keywords in different documents, which can clearly show the importance of a word in a document through TF-IDF.

4. Result

4.1. Result of data crawling and cleaning

The data for this project were collected from the customer review section of relevant products on JD.com. Firstly, this work collected all the product links for the search results of "essence" on JD.com and randomly selected 100 products from the collected links. Representative examples are displayed in Table 1. The main modules used in the collection process were third-party modules: selenium, time, and random. The method involved defining a function to gather data and relied primarily on the xpath helper to locate and obtain product information.

Table 1. Examples of crawled information

Name	Price	Link
La Mer The Hydrating Infused Emulsion	2000	https://item.jd.com/100024892070.html
Cle De Peau Beaute The Serum	1880	https://item.jd.com/100013377422.html
Cle De Peau Beaute Protective Fortifying Emulsion	1150	https://item.jd.com/3543774.html
SkinCeuticals H.A.Intensifier	980	https://item.jd.com/100022089050.html
SkinCeuticals Phyto Corrective Gel	585	https://item.jd.com/2987587.html
Kiehl's Dermatologist Solutions Clearly Corrective Dark Spot Solution	540	https://item.jd.com/100023662412.html
Olay Professional Pro-X Spot Fading Treatment	399	https://item.jd.com/100019789100.html

After obtaining the required product links for crawling, this work utilized the requests library to retrieve web page information, processed strings using the re module, stored files using the csv module, and formatted and processed time information using the time module. In the event of unexpected errors during the crawling process, this work handled exceptions using the traceback module.

Regarding the crawled review data, the text is initially decoded. However, the data returned was not in standard JSON format and contained unnecessary characters. Therefore, string manipulation was performed to obtain the desired result. Subsequently, this work converted the strings to JSON format. Finally, the data was separated into three variables: reply content, timestamp, and user nickname. The frame of the data spanned throughout 2022, comprising a total of 49,560 comments. Representative comments are illustrated in Table 2.

DOI: 10.54254/2755-2721/47/20241294

Table 2. Examples of representative comments.

Contents	Time
Xiu Li can color repair has always been his main product has been the major bloggers grow grass this time taking advantage of the double 11 huge discount margin finally cut down! Use a week or so, feel the skin state is in good, obviously can see the face red began to fade, but red blain print temporarily haven't see what improvement, continue to stick to use estimates will have effect, essence of texture is thin, good push, absorption is fast, taste is really good smell, feeling is the taste of vegetation	2022/11/27 13:21
Girlfriend suddenly said that he wanted to buy this, so he bought it for him, the price is beautiful, she said the effect is good, then try, when she received the express, happy. Expect the effect.	2022/11/27 11:14
Insist on a period of time	2022/11/27 11:09
Just finished, the day price, garbage, price protection does not admit, because it is a combination to buy? And member and non-member two attitude, two ways of treatment, angry I charge a member on the same day, really is speechless	2022/11/27 10:22
At the beginning, the effect is very good, after the use of a rash, I don't know what happened.	2022/11/27 8:23
Use a few days to temporarily still have no feeling, estimated to pass a period of time to have an effect	2022/11/27 0:11
It does work, and I will stick to it	2022/11/26 23:49

4.2. Visualization analysis of product review frequency

Figure 1 below shows a pie chart of the product review frequency distribution over the four quarters of 2022. It isn't hard to find that the third and fourth quarters were nearly half of the whole year. This is due to a surge in express deliveries at the shopping festival. This picture also shows the important influence of the shopping festival on product sales volume.

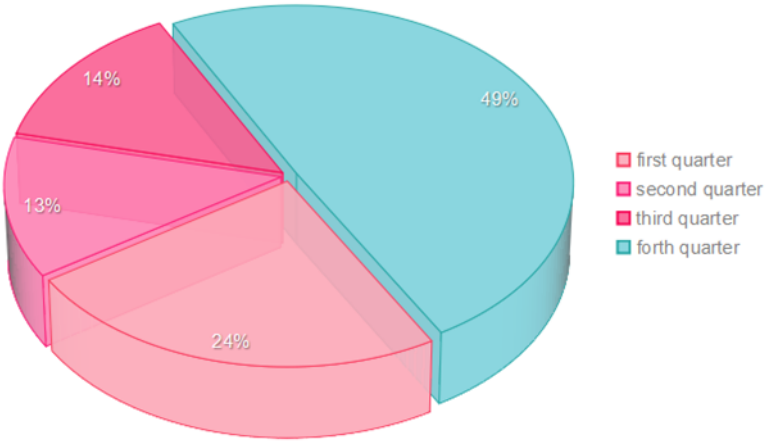

Figure 1. Product review frequency distribution (Figure Credits: Original).

4.3. Visualization analysis of product reputation

Figure 2 shows the content pie chart of the positive and negative reviews.

DOI: 10.54254/2755-2721/47/20241294

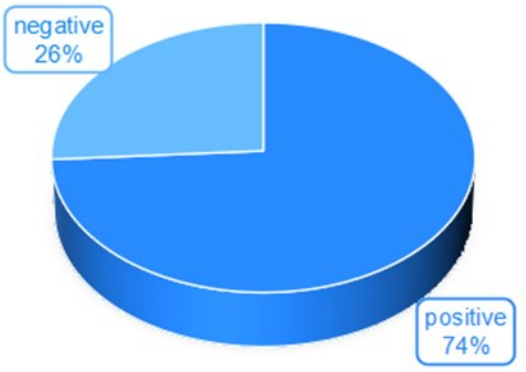

Figure 2. Positive and negative review distribution (Figure Credits: Original).

As can be seen from the content in the picture, the cosmetics categories selected in this paper have the characteristics of being widely recognized and having a high praise rate. In other words, mining the comment text of products can not only grasp the hot trend of today's beauty makeup category, but also analyze the consumption view of the people in today's Internet era.

4.4. Word frequency statistics analysis

Word cloud is a type of visualization that is used to display high-frequency keywords. It combines text, color, and graphics to create a prominent visual effect and communicate valuable information.

First, import the necessary libraries and load the extracted text data. Since the comment data is in the form of a whole paragraph, it is necessary to classify the comment text before displaying it in a word cloud. This paper utilizes the thesaurus provided by the jieba library, and the stop words list is employed to remove irrelevant words such as symbols, adverbs, and so on.

After the completion of word segmentation, it is necessary to calculate the frequency of word segmentation results. In this paper, the collections module's counter method is used. After using the count counting method, the counter is converted into a list, and the top 500 words with word frequency are selected, as displayed in Figure 3.

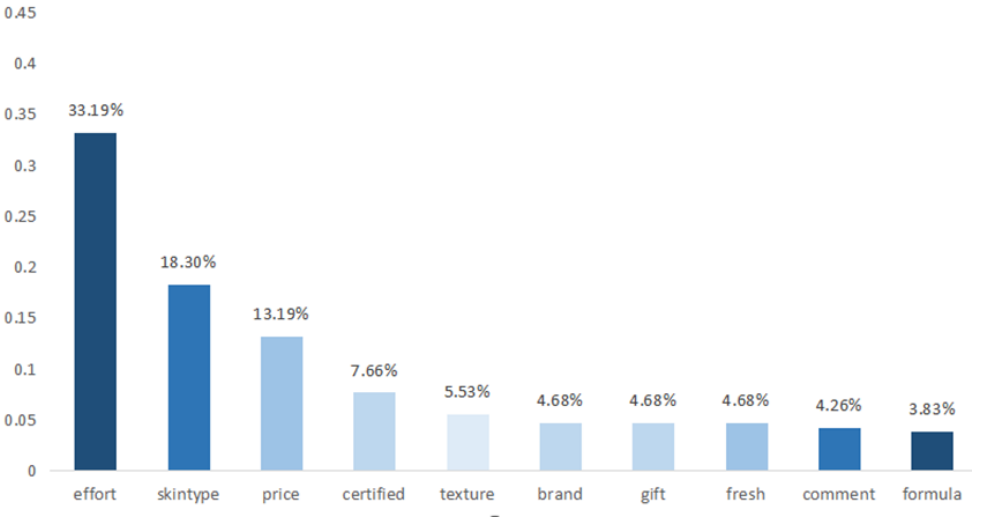

Figure 3. Keywords frequency distribution (Figure Credits: Original).

When creating the word cloud map, this paper chooses to use the word cloud module to visually display the sorted data. The main details include setting the word gap, size, font, and the word cloud shape map, etc. After writing the processing function, the results are executed as shown in Figure 4.

DOI: 10.54254/2755-2721/47/20241294

Figure 4. Word cloud of key words (Figure Credits: Original).

4.5. Sentiment analysis

To gain a better understanding of the emotional sentiment expressed in the comment content and enhance the comprehension of the selection process, it is necessary to conduct emotional classification analysis on the comment data. The main steps for implementation are as follows: First, pre-process the text comment data. Since the text comment data may contain irrelevant items, such as "the user does not evaluate," these elements can significantly impact the results of word frequency statistics or emotion analysis. Therefore, text preprocessing is essential to improve the quality of the outcomes.

This paper utilizes the SVM method to construct a supervised learning-based training model for emotion analysis. The decision to use SVM was primarily driven by the similar content characteristics of the products considered in this paper's topic selection, which required learning from comments. Consequently, an SVM training model is constructed, utilizing multiple existing review data for emotion training. Upon separating the training and test data, this study employs the snownlp library to facilitate emotional training processing. During the content processing stage, the sklearn library is employed to extract data features and calculate the evaluation of the final emotional content, leveraging a machine learning approach. The resulting emotion analysis outcomes are presented in Table 3.

Table 3. Result of sentiment analysis.

Name	Value
accuracy	0.81
precision	0.82
recall	0.80
F-score	0.81

Based on the results, it is evident that the model achieves scores exceeding 0.8 in accuracy, precision, recall, and F-score, indicating a high level of accuracy. To further evaluate the model, positive and negative comments were selected for testing. Firstly, the accurate word segmentation demonstrates precise text division. Secondly, the text exhibits strong discrimination in terms of emotion estimation, indicating the accuracy of the emotion analysis model.

4.6. Clustering analysis

Beauty makeup essence has the characteristics of differentiation and diversification in terms of product performance. By analyzing products with different attributes and different price ranges, better grasp of the secrets behind Singles' Day sales could be achieved. Text clustering is particularly suitable for achieving this goal.

DOI: 10.54254/2755-2721/47/20241294

This work used K-means clustering from the cluster module in nltk. After that, the same words of the same category are used for text subject statistics by printing the general word content, and finally selected the words with high word frequency as the topics according to the classified documents.

After product clustering, the main skin problems reflected in the product reviews include: oily skin, dry skin, sensitivity, anti-aging, etc. This is a reflection of the higher development degree of today's society and people's increasing attention to their own skin. This also shows that in today's era of consumerism, by hitting the main points could be leveraged to seize the lead in the high-speed product changes.

5. Discussion and suggestion

According to the analysis results, the quality of the product is mostly positive, but there are also some negative aspects related to the lack of subsequent logistics progress, bad customer service attitude, and decrease. This kind of phenomenon indicates that Internet shopping should not only focus on the form, but also requires efforts from the relevant derivative industries to form a better and harmonious Internet ecological chain.

With the continuous transformation and development of consumption patterns, the e-commerce industry will continue to expand, bringing more shopping choices to consumers. As consumers, they should make reasonable use of their information advantages for shopping decisions. When faced with commodity selection, they should give priority to products with similar efficacy to their needs. In the face of different platforms, more attention should be leveraged for choosing platforms with more preferential benefits and more genuine product guarantees to obtain better shopping experiences.

The analysis of the influencing factors of good reviews and bad reviews requires various and deep-level data processing, such as the background of consumers themselves (consumption views, understanding of products, etc.), which cannot be reflected in the comment content information of e-commerce platforms. If only relying on the use of models may cause errors and uncertainty in the results of data analysis, which needs special attention and evaluation.

The model needs to be combined with personal consumption concepts and consumers' own situations to form a comprehensive analysis system, so as to maximize the advantages of various text mining methods. In practical applications, it is necessary to choose appropriate methods according to specific situations to achieve the best analysis results.

6. Conclusion

This paper focuses on analyzing the comment text data of JD's skincare essence products and employs selected text mining methods to illustrate the primary content focus and emotional attitudes expressed in the comments. The analysis reveals the following findings. (1) Currently, there is a significant emphasis on product usage. For example, when it comes to essence products, people have high expectations and requirements. (2) Online shopping often involves preferential policies, and consumers have a specific demand for price concessions and bundled gifts. (3) Many keywords used in product evaluations are related to brands, indicating a certain degree of distrust in online shopping channels. This highlights the need for major platforms to enhance their credibility.

By comparing the conclusions and methodologies presented in this paper, the author not only draws conclusions about the efficacy focus and product satisfaction across different products but also identifies the distinct preferential mechanisms employed by different platforms for the same product. This, in turn, enables consumers to better grasp market trends and conduct comprehensive analysis and comparisons.

References

[1] Zhang, M. Y., Lu, Z. D., & Zou, C. Y. (2004). A Chinese word segmentation based on language situation in processing ambiguous words. Information Sciences, 162(3-4), 275-285.

[2] Zhu, D., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. Technological forecasting and social change, 69(5), 495-506.

DOI: 10.54254/2755-2721/47/20241294

[3] Wang, D., Li, J., Shi, Y., (2020) Methods of Government Document Clustering Based on K-means Algorithm. Software Guide, 19(06), 201-204.
[4] Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 417- 424.
[5] Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In Proceedings of the 8th international conference on Intelligent user interfaces, 125-132.
[6] Paltoglou, G., & Thelwall, M. (2012). Twitter, myspace, digg: Unsupervised sentiment analysis in social media. ACM Transactions on Intelligent Systems and Technology, 3(4), 1-19.
[7] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Empirical Methods in Natural Language Processing, 79-86.
[8] Tang, X., Yan, C., (2013) Research on text sentiment analysis based on precession principle and support vector machine. The Italian journal of urology and nephrology, 36(1), 98-103.
[9] Wang, H., Yan, M., & Su, X. (2011). An investigation of machine learning based automatic classification of Chinese books. Journal of Library Science in China, 3(0), 114-130.
[10] Wu, G., Xiao, F., (2012) A Hierarchical Clustering and Fixed-Layer Local Learning Based Support Vector Machine Algorithm for Large Scale Classification Problems. Journal of Donghua University (English Edition), 29(01), 46-50.
[11] Lu, G., Guo, R., (2018). An Accurate and Extensible Machine Learning Classifier for Flow-Level Traffic Classification. China Communications, 15(06), 125-138.