International Journal of Managerial Studies and Research (IJMSR)
Volume 11, Issue 1, January 2023, PP 22-31
ISSN 2349-0330 (Print) & ISSN 2349-0349 (Online)
https://doi.org/10.20431/2349-0349.1101003
www.arcjournals.org

Quality Improvement Strategies of Mobile Phone Product Based on Text and Sentiment Analysis

Xiaoxiao Qin

School of Management, Tianjin University of Technology, Tianjin 300384, China

Corresponding Author:** ***Xiaoxiao Qin, *School of Management, Tianjin University of Technology, Tianjin 300384, China*

Abstract: *The development of the Internet has brought people the opportunity to communicate online, and user reviews have appeared under the review interface of various brands of hot products, and these reviews containing emotions have also intensified the competition among products. This paper takes cell phone background as an example, through data mining cell phone user reviews, using word frequency and word cloud methods for text analysis of user reviews, extracting two types of hot review words-[service features] and [cell phone features], exploring whether each hot review word behind can significantly affect the favorable rating of cell phones by establishing a random forest regression model, and calculating the corresponding sentiment score of each hot review word according to the sentiment tendency calculation method. Finally, this paper analyzes each hot review term in depth and designs a specific evaluation system, which can be used to paint an overall and detailed portrait of the cell phone, thus helping merchants to quickly find the shortcomings of the product and determine the direction of improvement.*

Keywords: *Random forest regression; sentiment analysis; user comments; text analysis*

1. Introduction

The era of rapid development of e-commerce platforms, more and more consumers choose online shopping, under the influence of the epidemic, the sales of products in e-commerce channels have gradually increased in recent years, the lack of offline store user experience, consumers are more based on the platform product reviews to make decisions on whether to buy, for merchants, the way to sell goods exists in the user reviews on e-commerce platforms.

In terms of the relationship between the e-commerce platform and the merchant, the e-commerce platform selling mechanism and the provision of selling services are closely linked to the quality of merchants selling, merchants generally choose an after-sales service system to do more perfect e-commerce platform, which can be more convenient to communicate with customers, to give users a good sales attitude, thereby improving the quality of their own product sales. Relatively, the e-commerce platform because the merchant sales increase, their own platform also got publicity, in the netizens appear rate increased. Between the two, the user's online review data is one of the feedback factors to measure the quality of the partnership between the two.

This paper uses regression analysis and sentiment analysis as the theoretical basis, and combines data mining technology, text analysis methods and customer satisfaction evaluation models to analyze consumers' concerns and satisfaction with cell phone brands, which is of research significance to e-commerce platforms and merchants for the analysis of user reviews.

2. Literature Review

Textual data differs from structured data in that it carries an affective influence, and many scholars have explored research methods for sentiment analysis in natural language processing research. Pengwei Zhang (2020) integrated several well-known sentiment lexicon libraries and a collection of words related to e-commerce reviews to build a new sentiment lexicon, combined with python language to traverse degree adverbs, negation words, etc., to calculate the sentiment value of each review sentence, so as to realize the sentiment tendency classification of e-commerce products. Zul

(2021) analyzed the affective tendencies of citizens of Pekanbaru on social and political issues based on the sentiment analysis method of sentiment dictionary and the results showed that education was the most positive topic (42%), political figures was the most neutral topic (65%) and environment was the most negative topic (56%) [2].

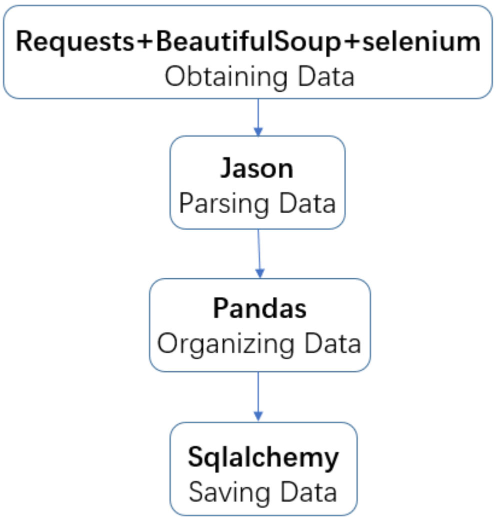
Based on the content of online review text data, many scholars have conducted many studies on the factors affecting consumer satisfaction. Lucini Filipe R (2020) used online reviews covering more than 400 airlines from 170 regions to explore the dimensions of airline customer satisfaction, and regression analysis revealed that flight attendants, in-flight service and value for money were the three highest dimensions of satisfaction[3].Wei and Wu(2020) used stepwise regression analysis to construct an investigation of the influence of IWOM on online booking of group travel goods from five aspects in two dimensions: product characteristics and consumer reviews. Based on review text data, Hong W (2019) used regression and quadratic graph models to explore satisfaction factors affecting T mall fresh food sales, and the results of the study yielded key influencing factors of consumer satisfaction[5].

Based on previous studies by scholars, this paper takes cell phone favorable rating as the main observable, and chooses the random forest regression method to explore the degree of influence of each factor on favorable rating as evaluation index one, and then scores the sentiment of each category of factors' reviews according to the lexicon-based sentiment analysis method as evaluation index two. Finally, a mean four-quadrant graph evaluation model is established to analyze in depth the influencing factors in each quadrant, and from the conclusions obtained, suggestions are made to merchants of e-commerce platforms, thus helping them to give full play to their advantages and make up for their shortcomings, which is of great significance to both e-commerce platforms and online merchants.

3. Text Data Processing and Analysis

3.1. Introduction to the Data

This paper selects the highest selling products of a brand cell phone flagship store on JD.com for data crawling, and crawls a total of 1925 data as of January 4, 2023 through the python program written, and the crawling process is shown in Figure 1.The crawled fields are saved to MySQL, including user id, rating star, time, follow-up review time, follow-up review content, product attributes, positive rating, and web page url.



3.2. Data Pre-processing

Data pre-processing of text data includes data de-duplication, removal of punctuation, useless words and word separation operations to facilitate model building analysis below.

In this paper, a comparative de-duplication method was used for the initial processing of the comment data, and some of the data are shown in Table 1.

Table1. *Part of the data after removing duplicate values*

id	star	content(Chinese text)	time	praise degree
1	5	手机不错外观时尚内存大流畅	2023-01-04 16:26	94%
2	5	这款非常好很流畅，用了好几天才来评价。已经买了了。售后也非常好。物流也很好。还会继续购买的。	2022-12-30 21:43	94%
3	5	经常光顾质量很好,服务一流	2022-12-29 23:25	93%
4	5	商品设计完美，外观也很高大上，让人爱不释手。客服更是热情的没话说，这次购物很满意，东西已经收到！	2022-12-29 20:20	96%

Useless words are mostly words of unknown meaning, such as intonation words and prepositions, and words such as "ha" and "though" cannot provide mining value for text analysis, so they need to be eliminated.

Before text analysis, it is an essential task to split the text data into words. This article uses a Chinese word splitting tool (jieba) written in python to split words and remove useless words and punctuation marks, and choose to create a useless word list and a custom word list (including punctuation marks). Jieba Chinese word splitting tool has the feature of fast and accurate word splitting, and some of the results obtained by using its word splitting are shown in Table 2.

Table2. *Part of the data after word segmentation*

content(Chinese text)	word segmentation(Chinese text)
手机不错外观时尚内存大流畅	'手机', '不错', '外观', '时尚', '内存', '大', '流畅'
这款非常好，很流畅，用了好几天才来评价。已经买了了。售后也非常好。物流也很好。还会继续购买的。	'这款', '非常', '好', '很', '流畅', '用', '了', '好几', '天', '才', '来', '评价', '已经', '买', '了', '了', '售', '后', '也', '非常', '好', '物流', '也', '很', '好', '还', '会', '继续', '购买', '的'
经常光顾质量很好,服务一流	'经常', '光顾', '质量', '很', '好', '服务', '一流'
商品设计完美，外观也很高大上，让人爱不释手。客服更是热情的没话说，这次购物很满意，东西已经收到！	'商品', '设计', '完美', '外观', '也', '很', '高大', '上', '让', '人', '爱不释手', '客', '服', '更', '是', '热情', '的', '没话', '说', '这次', '购物', '很', '满意', '东西', '已经', '收到'

From the results, it can be seen that the Jieba Chinese word separation tool is effective, and there are 1756 valid data obtained after data pre-processing.

3.3. Word Frequency Graph and Word cloud Graph Analysis

Word frequency graph and word cloud graph analysis of text data after word separation is an important tool for data mining. Word frequency analysis determines hot words and their degree of variation by counting the number of occurrences of words in the text. The word cloud map is a more intuitive way to render the hot words and determine their importance based on the size of the hot words. Combining word cloud word cloud module and collections module (Python) to analyze the data for word cloud map and word frequency map, the results are as follows (Figure 2 and Figure 3).

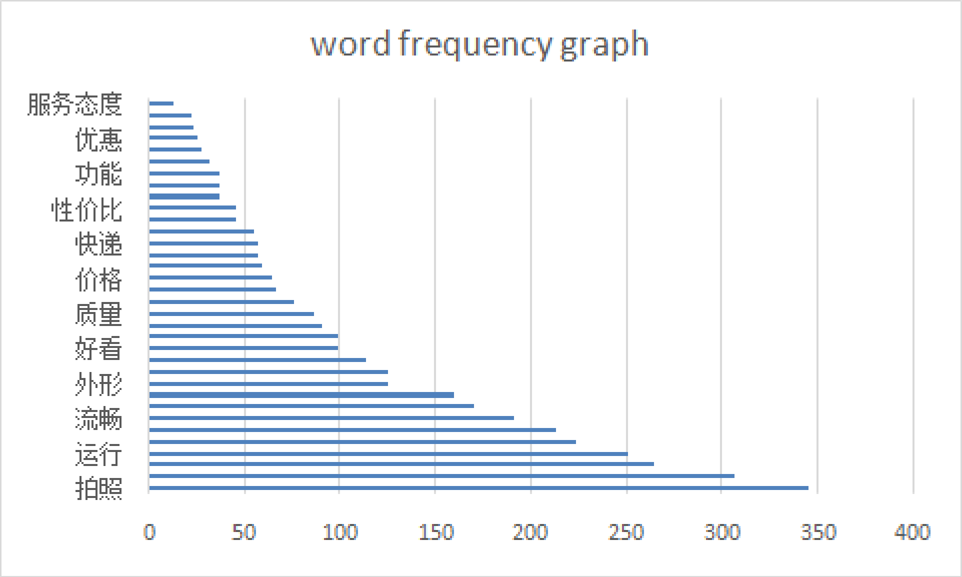


Figure2. *Word frequency graph (Chinese text)*

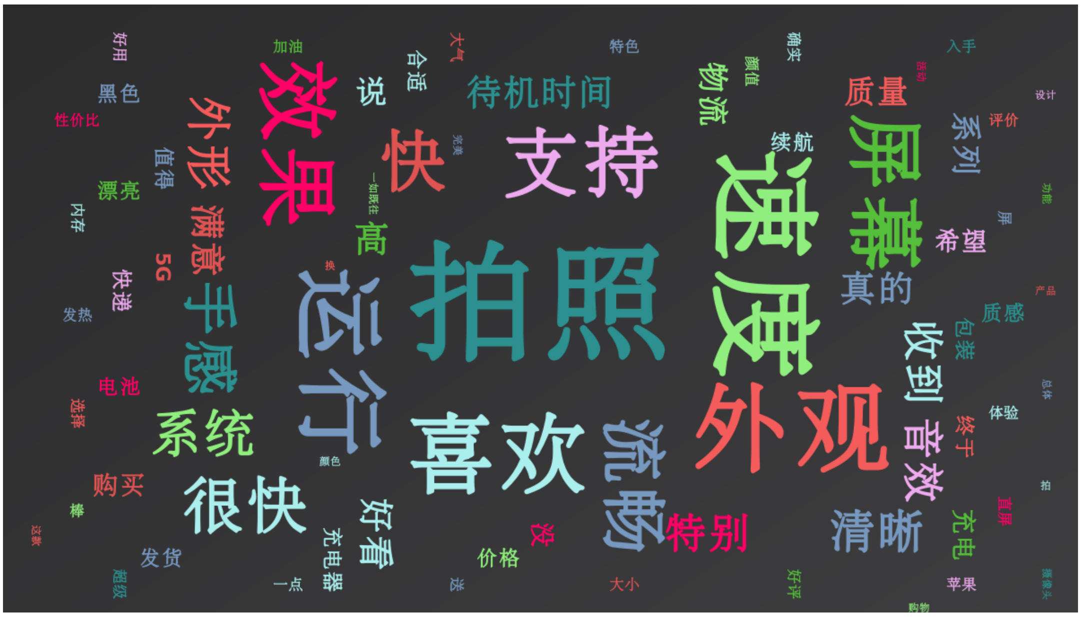


Figure3. *Word cloud graph (Chinese text)*

The top twelve hot words are selected for frequency statistics, with the horizontal axis indicating the frequency and the vertical axis decreasing from bottom to top. The first word frequency is "拍照", it can be seen that most consumers are most concerned about whether the phone photo is clear cell phone performance. Next is "运行" and "流畅", which can be combined into a single term to describe the performance of the mobile system in terms of smoothness of use. Similarly, "外形" and "好看" are combined to form "good-looking" to describe the nature of the phone's appearance. The frequency of the words "质量" and "功能" is 125 times, which means that consumers value the quality and function of cell phones. The three hot words "价格", "优惠" and "折扣" show that consumers are concerned about whether the price of the phone is reasonable."快递" is used to describe the speed of arrival of goods, which is one of the reasons why many consumers choose the JD.com, known for its fast delivery speed. Finally, the "服务态度", after-sales service is also a major feature of the JD.com, timely response to consumer inquiries, as well as product after-sales warranty and other services to improve, is also one of the feedback evaluation of the impact of consumers after the purchase. In the word cloud diagram, there are some fragmented words that do not appear in the word frequency diagram, such as "包装" and "充电器", which can be combined together as the nature of product configuration.

From the above conclusions, two types of characteristics can be summarized - [service features] and [cell phone features].Service features include three indicators of delivery speed, mobile phone equipment, and service attitude. Cell phone features include five indicators: photograph clearly, external appearance, system fluency, endurance quality (battery), and price favorable.

4. Basic Model

4.1. Random Forest Regression

Random Forest (RF) algorithm is a classification and regression algorithm that forms a forest by constructing multiple decision trees. It takes the decision tree as the basic unit, selects the bootstrap resampling method to randomly obtain several different subsets of samples, and uses the random subspace division method to build the decision tree based on each subset of samples. When the decision tree is split by nodes, the best features in the randomly generated feature subset are selected for splitting. Finally, the prediction results of all decision trees are averaged to obtain the final prediction results of the random forest. This paper uses random forest regression to explore the importance of the eight factors summarized above on the favorable rating, and the model study variables are defined in the following table (Table 3).

Table3. *Selection of research variables*

Research Variable	Variable Symbol
photograph clearly	x1
external appearance	x2
system fluency	x3
endurance quality	x4
price favorable	x5
delivery speed	x6
mobile phone equipment	x7
service attitude	x8
favorable rate	y

The variables are first analyzed for correlation, and the following heat map is obtained by combining R software with python software (Figure 4).

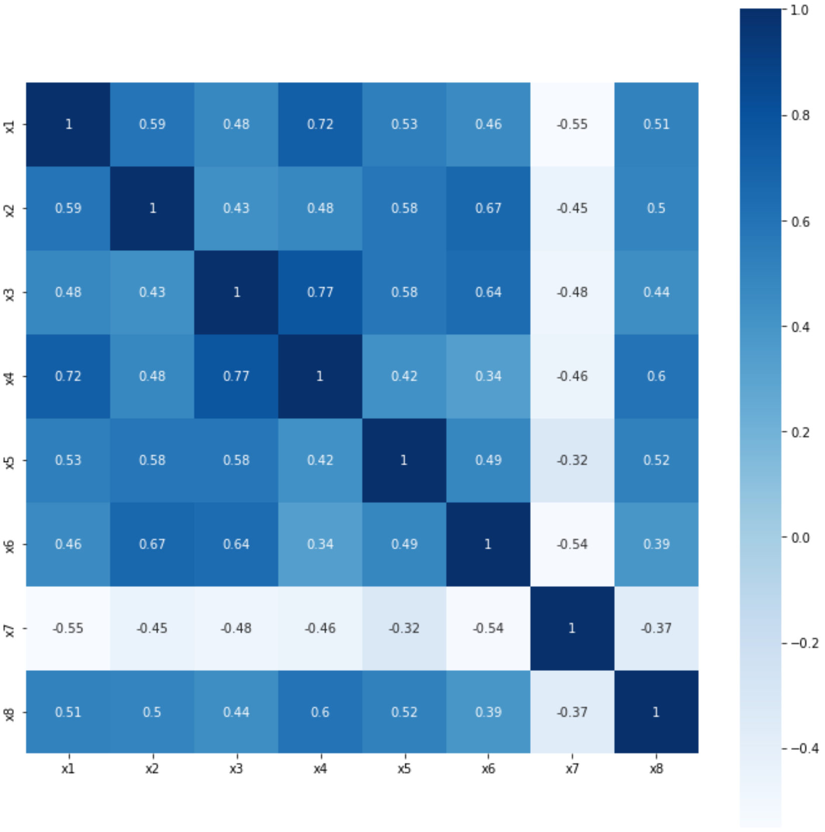


Figure4. *Heat map of correlation coefficients between variables*

The heat map of correlation coefficients shows that the correlation coefficients among the variables are basically distributed between 0.3 and 0.8, and it is tentatively concluded that there is no problem of multicollinearity, and it is found that all variables are positively correlated except for the variable of mobile phone equipment, which is negatively correlated with other variables. Further tests to check the variance inflation factor are shown in the table below (Table 4), and it is found that the mean value of vif between variables is 2.38 less than 3. Further tests do not have multicollinearity and meet the regression model construction conditions.

Table4. *multicollinearity test.*

Variable Variance	Inflation Factor
x1	3.54
x2	2.33
x3	3.10
x4	2.42
x5	2.55
x6	1.98
x7	2.15
x8	1.00
vif mean	2.38

There are two main factors affecting the performance of the random forest algorithm: the number of features used to construct the decision tree and the number of decision trees in the random forest, and different parameter choices will yield different prediction results and accuracy.

After continuous adaptation, this paper chooses to divide the data set according to 4:1 ratio, and finally determines the number of decision trees to be 100, at which time the model fitting has been more stable, and the maximum number of features is 25, which gets to be 0.82, and the fitting effect is better. The derived significant features diagram is shown in the following figure (Figure 5).

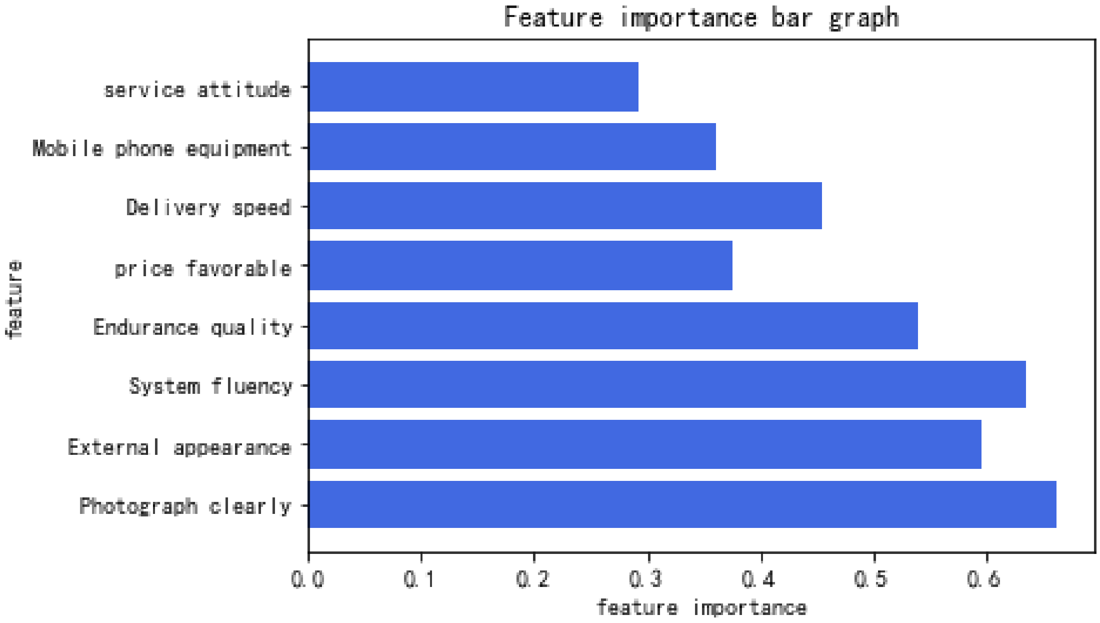


Figure5. *Feature importance bar graph*

4.2. Sentiment Analysis Based on Sentiment Dictionary

The research methods for sentiment analysis in the last decade or so have been based on two main categories. Machine learning-based sentiment analysis and sentiment analysis based on sentiment dictionaries. Based on the sentiment dictionary , words with sentiment are grouped into one category after word division, such as adverbs and negation words with strong sentiment, and are assigned a score and multiplied by the corresponding weight, and finally a sentiment score is assigned to each comment.

In this paper, the natural language processing tool Snow NLP library (Python) is used to process the comments: the data set is first divided according to the classification to which each comment belongs (based on the eight influencing factors mentioned above), then the lexicality is labeled with s.tags, and finally the sentiment tendency of the comments is scored using the sentiment method. The scores are distributed in the interval [0,1], the closer to 1 represents the more positive sentiment tendency, and conversely close to 0 shows negative sentiment, after scoring the average score of each type of data set to indicate consumer satisfaction, the results are shown in the figure (Figure 6).

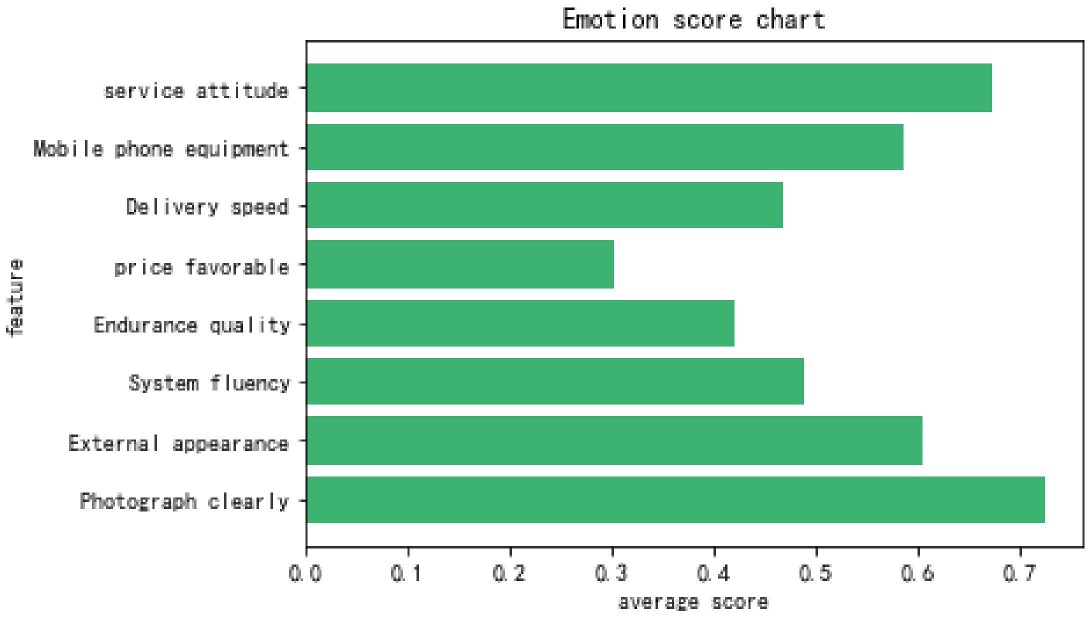


Figure6. *Emotion score chart*

Emotional scoring results can be observed that consumers for most indicators scored greater than 0.5, can be considered positive, the lowest score is the nature of the phone price concessions, it can be seen that most consumers are not very satisfied with the strength of the product concessions, which is also a point that businesses need to pay attention to.

4.3. Mean Four-quadrant Graph Evaluation Model

The Importance Factor Derivation Model is a qualitative research diagnostic model that makes a four-quadrant chart of two indicators of customer satisfaction (horizontal axis) and the company's evaluation of customer satisfaction (importance, as the vertical axis), and the company will analyze these classified indicators according to the quadrant in which they fall.

In this paper, the importance of the features obtained from the random forest regression model is used as the importance score, and the sentiment scoring based on the sentiment dictionary is used as the satisfaction score, and the following table is compiled (Table 5).

Table5. *Importance and Satisfaction Scores of Eight Factors*

Feature	feature word(Chinese text)	importance	satisfaction
photograph clearly	像素高，清晰	0.664	0.726
external appearance	简单大方，手感	0.597	0.604
system fluency	系统完善，流畅	0.636	0.489
endurance quality	便捷，快速，充电时间短	0.541	0.421
price favorable	性价比，便宜，优惠	0.376	0.302
delivery speed	送货速度	0.455	0.467
mobile phone equipment	充电器，耳机，包装	0.362	0.587
service attitude	客服，售后	0.293	0.674

After dividing the eight factors with satisfaction and importance as the coordinates and the mean value of both as the dividing line, the graph is drawn as shown in Figure 7.It can be observed that photograph clearly and external appearance are in zone A (advantage zone), indicating that both indicators are of higher importance and satisfaction and are the advantage for the commodity. System fluency and endurance quality are in zone B (repair zone), in this zone means that consumer satisfaction is low but the importance of favorable rate is high, and merchants need to find the gaps and raise consumer expectations. Price favorable and delivery speed are in Zone C (Opportunity Zone), which indicates that the merchant can still grasp the opportunity to improve without spending too much time and effort for the time being. Finally in the D zone (maintenance zone) is mobile phone equipment and service attitude, indicating that consumer satisfaction with these two factors, which are not very important, is still above average, and the business needs to maintain the advantage and does not need to spend too much time to improve.

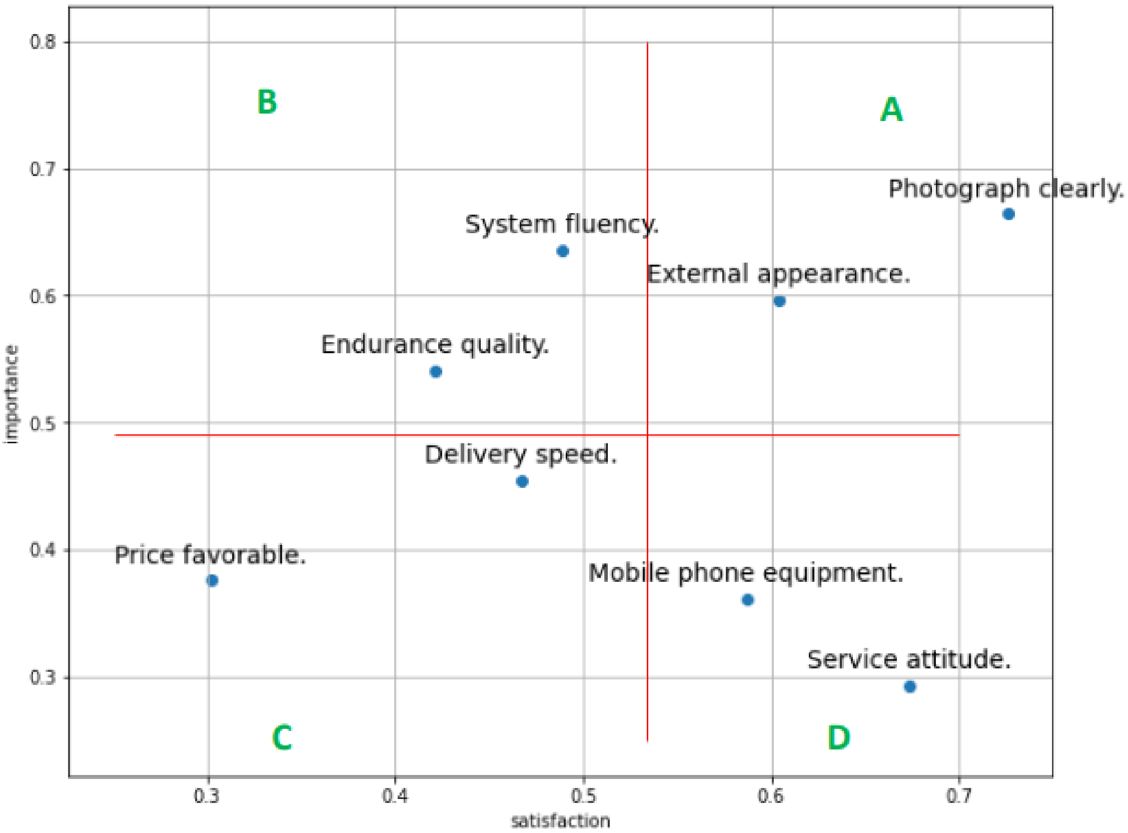


Figure7. *Mean four-quadrant evaluation chart*

5. Recommendations

Based on the above results, this paper gives advice to merchants from both overall and detailed perspectives.

Merchants can capture the key influencing factors based on the product's favorable rating metrics as a whole: By extracting the factors that consumers care most from online review data, and visualizing the data. It can initially find out what the influencing factors are. Based on the cell phone example in this paper, it is found that in the two categories of characteristics, the largest proportion is price favorable and photograph clearly, thus the businessman can initially focus on the cell phone price and cell phone pixels.

Thinking in terms of details, merchants can dig deeper, play to their strengths and make up for their weaknesses according to the mean four quadrant evaluation model. For factors in the dominant zone, photograph clearly and external appearance are important selling points for such goods, but with increasingly fierce competition, merchants also need to have their own characteristics in terms of technical and design power in order to steadily take advantage. For the factors in the repair zone, businesses need to pay extra attention to system fluency and endurance quality is an important technical force in the performance of cell phone features, but consumers are less satisfied with it, proving that businesses focus too much on the phone's photo function and do not grasp the hard conditions of the phone, so the next attention needs to be placed on improving the phone system and strengthening the battery life. In the opportunity zone, merchants can consider the logistics factor in the case of spare capacity to enhance their advantages. Cell phone price development, merchants can issue preferential prices on holidays, after-sales subsidies, thus improving customer satisfaction and enhancing their competitiveness. Finally, the factors in the maintenance area, merchants only need to maintain the status quo, considering the low satisfaction of service attitude, so a slight improvement in service attitude will bring consumers a better shopping experience.

6. Innovation and shortcomings

This paper takes cell phone products' favorable rating as the analysis point, not only considering the characteristics of the cell phone itself, but also combining with the characteristics of merchant services, so the research perspective is more comprehensive. This paper uses data mining and text analysis combined with econometric evaluation models to save research costs and improve information accuracy.

In this paper, when analyzing the basic influencing factors, the results are only roughly derived from the word frequency graph word cloud map, the method is not comprehensive enough to consider, and there is no precise statistics and further processing of similar words with similar semantics. Later studies can combine clustering methods into word classification to further improve the accuracy of extracting key factors for more in-depth analysis.

References

[1] L. Yang, Y. Li, J.Wang, Sherratt, R. S.Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE access,2020, 8:23522-23530.

[2] Zul, I. ,Azhari, S. ,Yessi, J. Implementation of Machine Learning for Sentiment Analysis of Social and Political Orientation in Pekanbaru City. Journal of Physics: Conference Series, 2021, 1803(1):152-171.

[3] Lucini Filipe R., Tonetto Leandro M., Fogliatto Flavio S.,Anzanello Michel J. Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. Journal of Air Transport Management,2020,83: 101760.

[4] C. Zhou, M. Leng, Z. Liu, et al. The impact of recommender systems and pricing strategies on brand competition and consumer search. Electronic Commerce Research and Applications, 2022, 53: 1-15.

[5] W.Hong, C. Zheng, L.Wu, X.Pu. Analyzing the relationship between consumer satisfaction and fresh e-commerce logistics service using text mining techniques. Sustainability, 2019,11(13):3570.

[6] C. Zhou, N. Ma, X. Cui, Z. Liu. The impact of online referral on brand market strategies with consumer search and spillover effect. Soft Computing, 2020, 24(4): 2551-2565.

[7] W. Fu, Choi, E. K., Kim, H. S. Text mining with network analysis of online reviews and consumers' satisfaction: A case study in Busan wine bars. Information, 2022, 13(3): 127

[8] Z. Liu, et al. Impact of cost uncertainty on supply chain competition under different confidence levels. International Transactions in Operational Research, 2021, 28(3): 1465-1504.

[9] Ahmadi S , Amin S H . An integrated chance-constrained stochastic model for a mobile phone closed-loop supply chain network with supplier selection[J]. Journal of Cleaner Production, 2019, 226:988-1003.

[10] C. Zhou, G. Xu, Z. Liu. Incentive contract design for internet referral services: Cost per click vs cost per sale. Kybernetes, 2020, 49(2): 601-626.

[11] C. Li, M. Chu. Is it always advantageous to add-on item recommendation service with a contingent free shipping policy in platform retailing? Electronic Commerce Research and Applications, 2019, 37: 1-11.

[12] J. Yu, Z. Song. Self-supporting or third-party? The optimal delivery strategy selection decision for e-tailers under competition. Kybernetes, 2022, DOI: 10.1108/K-02-2022-0216.

[13] Ullah M , Sarkar B . Recovery-channel selection in a hybrid manufacturing-remanufacturing production model with RFID and product quality[J]. International journal of production economics, 2020, 219(Jan.):360-374.

[14] C. Zhou, W. Tang, R. Zhao. Optimal consumption with reference-dependent preferences in on-the-job search and savings. Journal of Industrial and Management Optimization, 2017, 13(1): 503-527.

[15] P. Chen, R. Zhao, Y Yan, et al. Promoting end-of-season product through online channel in an uncertain market. European Journal of Operational Research, 2021, 295(3): 935-948.

[16] J. Yu, J. Zhao, et al. Strategic business mode choices for e-commerce platforms under brand competition. Journal of Theoretical and Applied Electronic Commerce Research, 2022, 17(4): 1769-1790.

[17] M. Chu. The impact of online referral services on cooperation modes between brander and platform. Journal of Industrial and Management Optimization, 2022, DOI: 10.3934/jimo.2022174.

[18] CootesT. F., Ionita, M. C., Lindner, C., Sauer, P. Robust and accurate shape model fitting using random forest regression voting.European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 278-291.

[19] Grömping U. Variable importance assessment in regression: linear regression versus random forest. The American Statistician, 2009, 63(4): 308-319.

[20] Ren L , Meng Z , Wang X , et al. A Data-Driven Approach of Product Quality Prediction for Complex Production Systems[J]. IEEE Transactions on Industrial Informatics, 2020, PP(99):1-1.

AUTHOR'S BIOGRAPHY

Xiaoxiao Qin, a student, from School of Management, Tianjin University of Technology, Tianjin 300384, P.R. China.

Citation: *Xiaoxiao Qin. "Quality Improvement Strategies of Mobile Phone Product Based on Text and Sentiment Analysis" International Journal of Managerial Studies and Research (IJMSR), vol 11, no. 1, 2023, pp. 22-31. DOI: https://doi.org/10.20431/2349-0349.1101002.*