Applied Mathematics, Modeling and Computer Simulation
C.-H. Chen et al. (Eds.)

doi:10.3233/ATDE231002

Research on Consumer Preferences and Potential Users of Vacuum Cleaning Robots -Based on Text Mining and Questionnaire Surveys

Jia JIA [a,1], Hong TANG[a], Chong LEI[a], Chunrong PU[a]
[a] *Chengdu University of Information Technology, Chengdu, Sichuan, China*

Abstract. With the advancement of technology, online shopping has gained immense popularity, resulting in the generation of copious amounts of e-commerce data. Previous studies have delved into the examination of online reviews in relation to consumer behavior. However, there exist significant challenges in selecting appropriate methodologies and comparing them to traditional market research, as well as assessing their accuracy and representativeness. In light of these challenges, we conducted an empirical study that analyzed consumer preferences for robotic vacuum cleaners by employing text mining techniques and questionnaire surveys. By utilizing data extracted from a prominent Chinese online shopping platform and conducting a team survey, we were able to ascertain the specific preferences of consumers regarding floor sweeping robots and segment the potential user market accordingly. The empirical research findings indicate that consumer attention towards product characteristics varies among different brands. While consumers generally prioritize product performance in their preference for products within the same category, there are still differences in attention given to products from distinct brands. Furthermore, consumers with lower personal characteristics have higher requirements for product purchases and lower corresponding purchase desires. Our research underscores the importance of understanding consumer preferences in the vacuum cleaning robot market and the potential for exploring untapped markets, while integrating new data elements with traditional statistical research methods.

Keywords. Sweeping robot, consumer preference, text mining, questionnaire survey, LDA, TF-IDF, logic return, cluster analysis

1. Introduction

Consumer preferences encompass the deliberate and purposeful decisions made by individuals in selecting goods and services. However, it is important to note that consumer preferences are not fixed; they are subject to influence from various factors and evolve in tandem with contemporary trends [1]. To gauge consumer preferences, there exists a plethora of measurement techniques. While traditional questionnaire surveys have been commonly employed in such research, they suffer from limitations such as protracted timelines, exorbitant costs, and susceptibility to subjective biases of

1 Corresponding Author, Jia JIA, Chengdu University of Information Technology, Chengdu, Sichuan, China; E-mail: 18192542108@163.com.

respondents. Fortunately, the advent of novel data elements has bestowed statistical research with more efficient avenues [2].

With the rapid advancement of technology, intelligent products have assumed an increasingly pivotal role in our daily lives. Among these products, robot vacuum cleaners stand out as a prime example [3]. They have seamlessly integrated into our routines and are readily available for purchase in brick-and-mortar stores, supermarkets, as well as online shopping platforms, and there are a lot of online product reviews. Simultaneously, in the age of the Internet, the Online-to-Offline (O2O) Commerce has gained considerable traction, making it imperative to investigate consumer preferences within this model [4].

Electronic commerce data holds significant value, and electronic word of mouth (eWOM) plays a vital role in shaping consumers' online purchasing decisions. User online reviews (OCR) is a kind of data that can represent the overall word of mouth [5]. These reviews, in turn, exert influence on consumers' decision-making processes, consequently impacting potential users. Moreover, it is worth noting that positive and negative online user reviews have varying degrees of impact on consumers' perceptions [6].

However, the new data factors have a limited timeframe and numerous uncertainties. Can user-generated content serve as a substitute for traditional market research? Assessing the accuracy of e-commerce data research, such as online product reviews, is currently a prominent area of interest [7]. In order to further investigate the applicability of this approach, the research team conducted a comprehensive questionnaire survey and meticulously analyzed the collected data, thus allowing for cross-validation of the findings from the two studies. Despite its inherent limitations, survey data remains a prominent research method and has made significant contributions as a robust and adaptable approach [8].

For online comments, current research focuses on how to effectively leverage the potential value of behavioral data. Decker [9] employed natural language processing techniques to preprocess online product reviews, identify potential correlations within unstructured data, and estimate consumers' overall preferences based on these reviews. Tobias [7] analyzed collected online reviews using attribute extraction and evaluation methods, aiming to explore the disparities between electronic word-of-mouth data and traditional market research. Xiao [10] utilized a combined model of TF-IDF and Logistic regression to forecast consumption behavior using online review data. Ye [11] employed a log-linear regression model to analyze consumer online reviews on hotel reservations and investigate its impact on the online reservation sector of the hotel and tourism industries.

The robustness of survey data, such as questionnaire surveys, is primarily reliant on the respondents. Currently, the most commonly employed investigation methods include traditional paper-and-pencil offline interviews and online questionnaire surveys. The analysis of survey data has a rich historical background. Gorton [12] utilized a questionnaire survey to examine the utilization, comprehension, and preferences of shoppers from diverse ethnic groups in New Zealand regarding nutrition labels. They employed multiple logistic regression analysis to identify the predictors. Vosbergen [13] conducted online surveys to investigate patients' preferences for self-management-related information. When addressing binary classification issues in questionnaires, i.e., the analysis of binary dependent variables, researchers often employ binary choice models [14]. In terms of cluster analysis, Carlyle [15] employed hierarchical clustering

to explore the impact of users' clustering behavior on the design of information display systems. Wen [16] utilized user logs and similar queries to retrieve information.

The objective of this study is to investigate consumer preferences for robot vacuum cleaner products and identify potential user markets through an empirical analysis that combines new data elements with traditional statistical research methods. Our research utilizes text mining data and questionnaire survey data. We collect and analyse online reviews of popular robot vacuum cleaner brands from Chinese e-commerce platforms, preprocess the text mining data, and examine the sentiment trends and potential review topics. Additionally, we conduct an online survey to complement the text mining research and segment the potential user market for robot vacuum cleaners. This study aims to contribute to more efficient empirical research on consumer preferences for products and fill the research gap in integrating new data elements with traditional market research methods.

The remaining sections of this paper are structured as follows. Section 2 provides an overview of the data source and details the process of data preprocessing. Furthermore, it presents the findings derived from the data analysis conducted in this study. Subsequently, section 3 performs an empirical analysis, elucidates the research model, establishes the model, and comprehensively examines the obtained results. Lastly, the concluding remarks are presented in the final section..

2. Data Source and Model Method

2.1. Data Source

The data for this study was sourced from two distinct channels. Firstly, the text mining data was gathered from product reviews of robot vacuum cleaners on renowned e-commerce platforms like JD.com and Suning. This data was collected using Python software to crawl and extract online comments related to the products. Secondly, the survey data was collected through a combination of online questionnaires and paper questionnaire survey. The questionnaire design primarily employed closed-ended questions to acquire specific user information and opinions, which served as valuable support and supplementary data for the text mining analysis. The questionnaire design framework is shown in the Table 1.

Table 1. Questionnaire design framework table

Survey items	Specific content
Cognition and feeling	Cognitive situation and cognitive channels
	Understanding of the situation and the degree of understanding
Consumption situation	Willingness to use the product
	Reasons for choosing and not choosing
	Influencing factors of whether to buy or not
	Importance of each function of the product
	Expected price is higher than expected function.
	Ways and means of purchase

User experience	Frequency of utilization
	Use feeling
	Usage satisfaction
	Use blocked situation
User info	Basic information (gender, age)
	Personal background (occupation, annual income, education level)
	Living conditions (living conditions, living area)

2.2. Data Capacity and Data Processing

By utilizing Python-based text mining techniques, a comprehensive collection of 46,515 online comment data was acquired, out of which 44,771 were deemed valid following rigorous cleansing procedures. Considering the unstructured nature of the collected data, an initial pre-processing phase was conducted to eliminate redundant linguistic information through methods such as text deduplication and Mechanical compression de-wording, among other advanced techniques [17]. The detailed procedure is illustrated in figure 1.

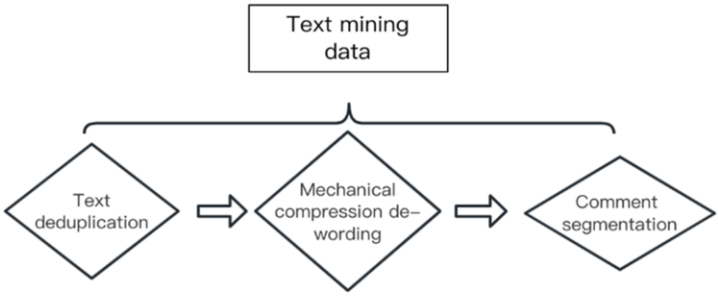

Figure 1. Data specific process

In summary, the final result of the text mining data is shown in table 2:

Table 2. Changes in data volume and cleaning for each brand

Brand	Crawled Comments	Cleaned Comments	Data Validity Rate
ECOVACS	8338	7636	91.58%
MI HOME	9078	8098	89.20%
ROBOROCK	3200	3199	99.97%
NARWAL	2470	2463	99.72%
MIDEA	2520	2513	99.72%
HAIER	6646	6644	99.97%
IROBOT	7160	7160	100.00%
DREAME	3023	3023	100.00%

WHIRLPOOL	1080	1035	95.83%
360	3000	3000	100.00%

A total of 1,300 survey questionnaires were distributed, of which 1,155 valid questionnaires were collected, yielding an impressive response rate of 88.85%. The questionnaire scales demonstrated strong reliability and validity, with Cronbach's α coefficients of 0.872 and 0.797, respectively. Additionally, the KMO coefficient of 0.881 further confirmed the questionnaire's reliability and validity. These satisfactory results allow for the utilization of the questionnaire in subsequent modeling and analysis [18]. Figure 2 presents an overview of the respondents' demographics as captured in the questionnaire survey.

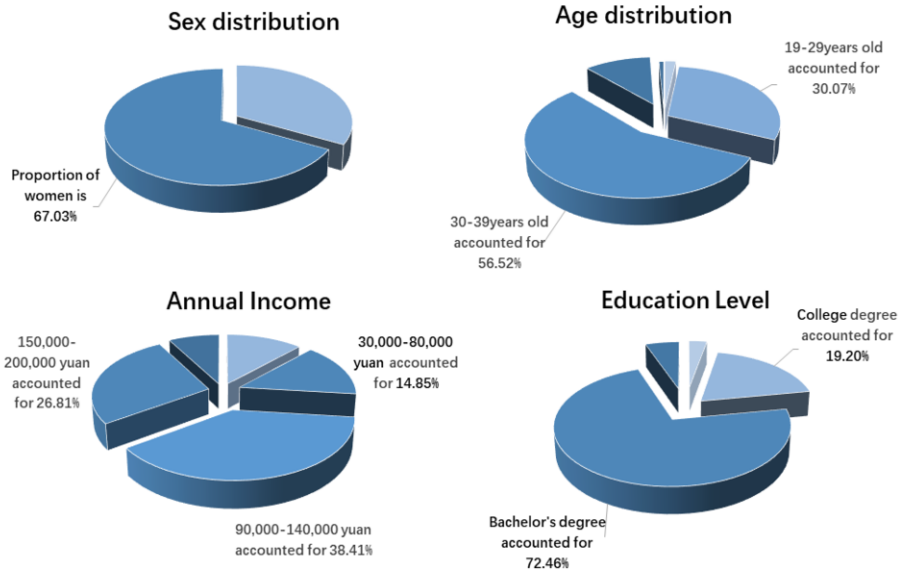

Figure2. Basic situation of respondents in the questionnaire survey

3. Empirical analysis

3.1. Analysis of Consumers' Emotional Tendency

Currently, there are two methodologies employed for analyzing consumers' emotional tendencies: machine learning-based and semantic dictionary-based approaches [19]. The former utilizes various machine learning classification techniques to identify emotions, while the latter involves constructing an emotional dictionary and utilizing it to assess emotional tendencies. In this study, the snownlp library in Python is utilized to establish a dedicated emotional dictionary for the preprocessed textual data. Subsequently, the emotional tendencies of ten brand products are analyzed using this emotional dictionary, resulting in emotional scores as presented in table 3. Finally, the emotional scores of different brand products are evaluated and examined in subsequent articles. The scores range from 0 to 1, with values closer to 1 indicating positive emotions and values closer to 0 representing negative emotions.

Table 3. Sentiment analysis of product reviews

Product	Positive Sentiment	Neutral Sentiment	Negative Sentiment
ECOVACS	74.91%	9.24%	15.85%
MI HOME	76.77%	8.53%	14.70%
ROBOROCK	76.06%	7.53%	16.41%
NARWAL	71.57%	8.57%	19.86%
MIDEA	81.29%	7.95%	10.77%
HAIER	75.16%	10.49%	14.36%
IROBOT	70.22%	11.15%	18.63%
DREAME	72.31%	7.92%	19.76%
WHIRLPOOL	72.95%	17.81%	9.25%
360	82.10%	4.40%	13.50%

According to the data presented in Table 3, it is evident that the products affiliated with the 360 brand exhibit the highest level of consumer satisfaction, boasting an impressive positive emotion rate of 82.10%. Conversely, the products associated with the iRobot brand demonstrate the lowest positive emotion rate, merely reaching 70.22%.

However, it is worth noting that both brands receive a considerable proportion of highly positive and highly negative emotional evaluations, indicating a significant presence of extreme emotions among consumers. Nevertheless, when considering the overall evaluation score, the positive emotional assessment surpasses 80%, leading to the conclusion that the prevailing consumer sentiment leans predominantly towards positivity.

3.2. Topic analysis of potential advantages and disadvantages of comments using the Latent Dirichlet Allocation (LDA) model

To identify significant aspects within the data, we implemented the following procedures:

The classification quality of new topics in the subsequent steps is influenced by the number of potentially relevant topics. To evaluate the impact of different topics, we calculated the perplexity score. Firstly, we imported the Gensim natural language processing package into Python. Then, we computed the perplexity scores for different topic numbers. The perplexity score is employed to measure the accuracy of the model, with lower scores indicating higher accuracy. Common methods for calculating the perplexity score include:

$$\text{Perplexity}(D) = \exp\left\{\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{N} N_d}\right\} \quad (1)$$

Firstly, we conducted an extensive analysis to determine the optimal number of potential tendency topics. Our findings revealed that when the number of topics was set to 5, the impact on the quality of new topic classification was minimal, resulting in the most accurate results.

Next, we employed the widely-used standard implementation of LDA, as proposed by Brody [20]. To ensure a focused analysis, we treated each sentence as an individual

document. The output of the model provided us with the distribution of topic inference for each sentence in the dataset.

To visualize the results, we utilized the LDAvis tool, which allowed us to assess the topic Perplexity Score (with a fixed topic count of 5) and employed standard parameters (α =0.1, β =0.1, and 500 iterations). Subsequently, we extracted the top 10 keywords with the highest probability from the topic classification results and aggregated them in clusters. The final outcomes are presented in table 4.

Table 4. Summary of thematic keywords

Theme 1	Theme 2	Theme 3	Theme 4	Theme 5
"Cleanliness"	"Cleanliness"	"Cleanliness"	"Cleanliness"	"Cleanliness"
"Suction"	"Time"	"Functionality"	"At home"	"Sensitivity"
"Intelligence"	"Obstacle avoidance"	"Mopping"	"Noise"	"Time"
"Mopping"	"Satisfaction"	"Intelligence"	"Liberation"	"Functionality"
"Sweeping"	"Sensitivity"	"Noise"	"Suction"	"Liberation"
"Hands-free"	"Hands-free"	"Satisfaction"	"Size"	"Obstacle avoidance"
"At home"	"Capability"	"Hands-free"	"At home"	"Size"
"Sweeping"	"Sound"	"Time"	"Functionality"	"At home"
"Liberation"	"Automatic"	"Endurance"	"Mopping"	"Hands-free"
"Capability"	"Particularly"	"Mop"	"Intelligence"	"Sensitivity"

Among the five new themes, the foremost keyword is "cleanliness," underscoring the paramount importance consumers attach to the sweeping robot's cleaning efficacy. Each of these themes centers on distinct aspects. The initial theme accentuates suction strength power, while the second theme emphasizes obstacle-crossing capabilities. The third theme prioritizes the practical functionality of the sweeping robot, whereas the fourth theme centers on the noise emissions during operation. Lastly, the fifth theme delves into the sweeping robot's sensitivity and adeptness in obstacle avoidance.

3.3. Analysis of preference characteristics of commodity attributes based on TF-IDF model

In this study, we employ the TF-IDF model to analyze the preference characteristics of commodity attributes. Building upon the findings of the LDA model analysis, we delve deeper into the keywords associated with various products to gain insights into the distinctive product features of specific sweeping robot brands [21]. Consequently, we utilize the TF-IDF algorithm to determine the weightage of each product attribute. The specific calculation formula is as follows:

$$TF - IDF_i(d) = tf_i \times \ln\left(\frac{N}{n_i+1}\right) \tag{2}$$

To begin with, the LDA model is utilized to extract keywords, which are then subjected to statistical aggregation. This process yields a total of 26 initial keywords, with a cumulative occurrence of 62,952. Due to the excessive quantity and dispersed distribution, these original keywords are subsequently reorganized based on the principle of similarity. Consequently, six distinct sets of novel feature factors are derived. Please refer to table 5 for the comprehensive outcome of this amalgamation.

Table 5. New feature factors of high-frequency keywords

High-Frequency Keywords Summary	New Feature Factors
Price, Ease Of Use, Durability, Value, Good Quality And Low Price, Cost-Effectiveness	Cost-effectiveness
Cleanliness, Suction Power, Mopping, Cleaning Effectiveness	Strong Cleaning Power
Intelligence, Functionality, Automation, Convenience, Time-Saving, Ease Of Control	Intelligence
Appearance, Aesthetics, Style, Design	Aesthetics
Logistics, Warranty, Installation	Service
Sound, Household, Home Use	Performance

The calculation of TF-IDF weights is based on the feature factor set derived from Table 5. The corresponding results are presented in table 6, with the most concerned items of the brand highlighted in red and the least concerned items marked in green.

Table 6. Summary table of TF-IDF weights scores

Keywords	Cost-Effectiveness	Strong Cleaning Power	Intelligence	Aesthetics	Service	Performance
Ecovacs	0.005	0.057	0.034	0.004	0.003	0.006
Mi Home	0.012	0.051	0.028	0.004	0.003	0.006
roborock	0.006	0.054	0.029	0.003	0.003	0.006
narwal	0.003	0.056	0.019	0.003	0.002	0.005
Midea	0.010	0.050	0.032	0.008	0.006	0.005
Haier	0.014	0.057	0.022	0.005	0.006	0.007
iRobot	0.006	0.047	0.026	0.003	0.003	0.009
dreame	0.008	0.054	0.028	0.003	0.002	0.005
Whirlpool	0.037	0.047	0.010	0.005	0.002	0.007
360	0.007	0.045	0.025	0.004	0.002	0.006

The sweeping robot brands ECOVACS, Haier, iRobot, ROBOROCK, and Whirlpool exhibit overall stability when combined with the six attribute values. However, MI HOME and NARWAL fall short in all aspects. Additionally, consumer attention varies across different aspects. For instance, consumers show the least interest in the services provided by ECOVACS products. On the other hand, aesthetics (appearance) is the least important factor for consumers when considering Haier and iRobot products. As for MIDEA products, consumers prioritize comprehensive performance the least.

3.4. Market segmentation of potential users based on Binary selection model and K-means clustering analysis

Based on the data from the questionnaire survey, respondents showed a preference for sweeping robot products that have superior cleaning ability and higher cost performance, as opposed to focusing solely on appearance and intelligence. This finding is also supported by the text mining data. The results obtained from both methods are consistent, indicating that the prerequisite for the existence of consumer preferences is a basic understanding of the product, and that consumers' basic understanding and interest in the product is the key to expanding the product market. Therefore, this section will examine the characteristics of individuals who are familiar with sweeping robot products.

The Binary selection model is a model that analyzes an individual's choice behavior between two options. It is highly accurate in dealing with binary classification problems [14]. There are various forms of the binary choice model based on different distribution

functions of each random error term. In this study, Logistic regression analysis is employed due to its wide applicability and flexible coefficient interpretation. It is used to analyze the factors influencing the understanding of sweeping robot products. The study utilizes dummy variables for the categorical variables in the questionnaire and simplifies the model based on the analysis results.

The binary selection behavior of whether the ith individual is familiar with the sweeping robot is expressed as a dependent variable. When the individual is deemed to be familiar, the value of Y_i is set to 1, and the corresponding probability is denoted as $P_i = Pr(Y_i = 1)$. Conversely, when the individual is not familiar, the value of Y_i is set to 0, and the probability is represented as $1 - P_i$. Therefore, we can summarize the relationship as follows:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu_i \quad (3)$$

After obtaining the regression coefficient for each index, the impact of each variable on the dependent variable is compared and analyzed based on its sign, magnitude, and statistical significance. The evaluation index values of the Logistic regression model are all above 0.5, and the AUC value is 0.705, indicating a high degree of proximity to 1. Thus, the logistic regression model demonstrates a robust classification performance.

Table 7. Final regression results for the logistics model

Variable	Regression Coefficient	Standard Error	P-Value
Gender	0.066	0.171	0.702
Occupation	0.051	0.048	0.288
Annual Income	0.22	0.089	0.014
Education Level	0.266	0.121	0.028
Living Area	0.214	0.135	0.113
Living Condition 1	0.691	0.648	0.286
Living Condition 2	0.741	0.69	0.283
Living Condition 3	0.402	0.642	0.532
Living Condition 4	-0.227	0.682	0.740
Living Condition 5	-0.586	0.647	0.365
Living Condition 6	-0.335	0.643	0.603
Age 1.0	-0.573	0.538	0.287
Age 2.0	-0.511	0.559	0.360
Age 3.0	0.499	0.59	0.398
Age 4.0	-1.439	1.193	0.228
Age 5.0	-16.937	3812.404	0.996

Based on the findings in Table 7, it is evident that factors such as gender, age, and occupation have minimal impact on users' awareness of sweeping robots. However, education level and annual income do influence users' understanding to some extent. The most influential factors, however, are the living area and living conditions, which significantly affect users' knowledge of sweeping robot products.

Therefore, a more nuanced segmentation of the target consumer group for robotic vacuum products can be achieved by considering factors such as the consumer's residential area, living situation, educational attainment, and annual income. By studying how consumers' purchase intentions for products are affected by different degrees of personal characteristics, the target population that is most receptive to the robot vacuum cleaner market is determined.

Clustering analysis is a technique that simplifies data by categorizing it into different clusters based on its characteristics. In this study, we employ the K-means model to classify potential users and explore the potential market for sweeping robots. We select three variables (annual income, education level, and family living area) that have a

significant influence on users' understanding of sweeping robots. These variables, along with age and occupation, which are two easily distinguishable characteristics, are combined with the question of "the importance of sweeping robots" from the questionnaire survey. This combined index, referred to as the APMEAI model, serves as a means to identify customer value. By clustering customers into six categories, we can determine the clustering center of potential user types.

Table 8. Cluster Centers of Potential User Types

Potential User Type	Clustering Attribute	Attribute Categories (Corresponding Respectively)
1	19-29, Employee, 30,000-80,000, College, 60-100 Square Meters, Three Items	Age
2	19-29，Student，Below 30,000, Bachelor's Degree，60-100，Four Items	Occupation
3	30-39, Manager, 150,000-200,000, Master'S Degree Or Above, 100-200, Four Items	Annual Income(Yuan)
4	30-39, Manager, 90,000-140,000 , Master's Degree Or Above, 100-200, Three Items	Education Level
5	19-29, Self-Employed,30,000-80,000, College Degree, 60-100, Four Items	Family Living Area(Square Meters)
6	30-39, Civil Servant, 90,000-140,000, Bachelor's Degree, 100-200, Five Items	Importance Of Sweeping Robot Project

Based on the findings presented in Table 8, the consumer groups for sweeping robots exhibit a certain level of concentration. These groups primarily consist of middle-aged and young individuals who possess a high level of education. Furthermore, each categorized customer group displays unique performance characteristics. To further analyze their purchasing intentions, these potential users can be divided into four groups: purchase customers, important potential customers, wait-and-see customers, and low-value potential customers.

High-value potential customers, i.e., Type 3 potential users, representing approximately 60% of the potential users, are individuals who have recently entered middle age. They enjoy relatively stable work and living conditions, possess a high level of education, reside in spacious households, and have a strong basis for purchasing and demand.

Important potential customers comprise the 2 and 5 categories of potential users. These individuals have recently entered society and fall within the age range of 19 to 29. They have received a fundamental higher education and exhibit a high willingness to purchase sweeping robot products.

Wait-and-see customers belong to the 4 and 6 categories of potential users. This group consists of individuals with stable jobs, earning an income that ranks second after the first cluster group. They have similar family living areas and fall within the age range of 30 to 39. However, due to external factors such as family influence, their purchasing time and willingness remain uncertain.

Low-value potential customers represent the 1 category of potential users. While they possess a high level of education similar to the 2 cluster, as well as a higher level of education, their desire to purchase products is hindered by their low annual income levels and living conditions.

Based on the aforementioned clustering results, we can establish a logical relationship. For instance, the first category of potential users, despite their aspirations

and high educational background, are constrained by their income levels and living environments. Consequently, they prioritize practicality over indulgence, making themselves more biased towards survival than enjoyment of life, leading to a lower desire to purchase sweeping robot products compared to other clusters.

4. Conclusion

For different product brands, although consumers are generally concerned about similar product characteristics, there are still variations. The analysis above reveals that across all the brands studied, consumers prioritize cleanliness, specifically the quality of cleaning power in cleaning robots, which directly influences consumer preferences. However, when it comes to factors that receive less attention, different consumers exhibit varying preferences for products from different brands. From the discrepancies in consumer attention towards different brands, we can infer that most consumers' preferences for the same product category primarily depend on product performance in key areas of use.

According to the binary selection model, living situations and living areas are important factors in defining the target population for robotic vacuum cleaners in terms of ease of use, followed by annual income and education level. K-Means clustering analysis suggests that vacuum sweeping robot products have a wide range of potential users, with significant potential customers and cautious customers, indicating a promising market potential. However, the conversion rate from awareness to purchase needs improvement. When consumers possess lower personal characteristics and attributes, their purchase requirements for the product are higher, resulting in a lower purchase desire. Additionally, the more extreme a consumer's own conditions tend to be, the more distinct the characteristics they seek in their target products. This is because when a consumer's demand for the product leans towards survival, there is a higher likelihood of dissatisfaction with the overall product if certain attributes do not meet ideal requirements, significantly reducing their preference for the product. Furthermore, consumers with different characteristics exhibit significant differences in their attention towards various product attributes, with varying requirements for these attributes. Limited product design makes it challenging to meet the complex needs and preferences of current consumers. While products may have standout features, their shortcomings cannot be overlooked.

In general, the current market for sweeping robots is effective in capturing consumer interest in the product. However, for various reasons, consumers struggle to achieve a higher level of realization and satisfaction of their personal preferences when making purchases. Even if most consumers are reasonably satisfied with the products they own, it is difficult to maintain the stability of consumer groups.

References

[1] Mărcuță, L., Mărcuță, A., Mârza, B., 2014. Modern Tendencies in Changing the Consumers' Preferences. Procedia Econ. Finance 16, 535–539. https://doi.org/10.1016/S2212-5671(14)00835-1

[2] Buettner, R., 2017. Predicting user behavior in electronic markets based on personality-mining in large online social networks: A personality-based product recommender framework. Electron. Mark. 27, 247–265. https://doi.org/10.1007/s12525-016-0228-z
[3] Hendriks, B., Meerbeek, B., Boess, S., Pauws, S., Sonneveld, M., 2011. Robot Vacuum Cleaner Personality and Behavior. Int. J. Soc. Robot. 3, 187–195. https://doi.org/10.1007/s12369-010-0084-5
[4] Yao, P., Osman, S., Sabri, M.F., Zainudin, N., 2022. Consumer Behavior in Online-to-Offline (O2O) Commerce: A Thematic Review. Sustainability 14, 7842. https://doi.org/10.3390/su14137842
[5] Filieri, R., 2015. What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. J. Bus. Res. 68, 1261–1270. https://doi.org/10.1016/j.jbusres.2014.11.006
[6] Wang, F., Liu, X., Fang, E. (Er), 2015. User Reviews Variance, Critic Reviews Variance, and Product Sales: An Exploration of Customer Breadth and Depth Effects. J. Retail. 91, 372–389. https://doi.org/10.1016/j.jretai.2015.04.007
[7] Tobias, R.B., Johannes, H., Martin, K., 2023. Automated inference of product attributes and their importance from user-generated content: Can we replace traditional market research? Int. J. Res. Mark. 40, 164–188. https://doi.org/10.1016/j.ijresmar.2022.04.004
[8] Couper, M.P., 2017. New Developments in Survey Data Collection. Annu. Rev. Sociol. 43, 121–145. https://doi.org/10.1146/annurev-soc-060116-053613
[9] Decker, R., Trusov, M., n.d. Estimating Aggregate Consumer Preferences from Online Product Reviews.
[10] Xiao, S., Tong, W., 2020. Prediction of User Consumption Behavior Data Based on the Combined Model of TF-IDF and Logistic Regression. J. Phys.
[11] Ye, Q., 2011. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Comput. Hum. Behav.
[12] Gorton, D., Ni Mhurchu, C., Chen, M., Dixon, R., 2009. Nutrition labels: a survey of use, understanding and preferences among ethnically diverse shoppers in New Zealand. Public Health Nutr. 12, 1359–1365. https://doi.org/10.1017/S1368980008004059
[13] Vosbergen, S., Peek, N., Wiggers, A.-M., Kemps, H., Jaspers, M., Lacroix, J., Kraaijenhagen, R., 2014. An online survey to study the relationship between patients’ health literacy and coping style and their preferences for self-management-related information. Patient Prefer. Adherence 631. https://doi.org/10.2147/PPA.S57797
[14] Donkers, B., Franses, P.H., Verhoef, P.C., 2003. Selective Sampling for Binary Choice Models. J. Mark. Res. 40, 492–497. https://doi.org/10.1509/jmkr.40.4.492.19395
[15] Carlyle, A., 2001. Developing organized information displays for voluminous works: a study of user clustering behavior. Inf. Process. Manag. 37, 677–699. https://doi.org/10.1016/S0306-4573(00)00048-0
[16] Wen, J.-R., Nie, J.-Y., Zhang, H.-J., n.d. Clustering User Queries of a Search Engine.
[17] Xiao, S., Tong, W., 2020. Prediction of User Consumption Behavior Data Based on the Combined Model of TF-IDF and Logistic Regression. J. Phys.
[18] Vithayathil, J., Dadgar, M., Osiri, J.K., 2020. Social media use and consumer shopping preferences. Int. J. Inf. Manag. 54, 102117. https://doi.org/10.1016/j.ijinfomgt.2020.102117
[19] Birjali, M., Kasri, M., Beni-Hssane, A., 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowl.-Based Syst. 226, 107134. https://doi.org/10.1016/j.knosys.2021.107134
[20] Brody, S., Elhadad, N., n.d. An Unsupervised Aspect-Sentiment Model for Online Reviews.
[21] Zhou, Y., Yang, S., Li, Y., Chen, Y., Yao, J., Qazi, A., 2020. Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining. Inf. Process. Manag. 57, 102179. https://doi.org/10.1016/j.ipm.2019.102179