JARQ 59 (2), 139-153 (2025) https://www.jircas.go.jp

Consumer Evaluation of Green Foods in China: An Approach from Text Mining and Random Forest on Online Consumer Reviews

Yili YANG and Shinsaku NAKAJIMA*

School of Agriculture, Meiji University, Kawasaki, Japan

Abstract

This study aims to identify consumer evaluation of green-labeled rice in China using e-commerce review data. It also proposes a random forest model to predict consumer evaluation of green-labeled rice. First, using text mining techniques, we summarized the tf-idf scoring for each of the two products reviewed. Second, we constructed a random forest model to find the important words affecting the rating of green-labeled rice. Finally, we used co-occurrence networks to clarify the relationship among keywords that influence consumer evaluations and whether the influence is positive or negative. We found that consumers placed importance on the packaging, texture, taste, price, quality, and likes of green-labeled rice at the time of purchase and after purchase. Moreover, we found that the words green food, traceability, and quality led to good evaluations of green-labeled rice. Chinese consumers were found to be more likely to purchase products with quality certification labels, but it was also found that green food certification is not necessarily an attribute that consumers value most.

Discipline: Social Science
Additional key words: co-occurrence network, e-commerce, machine learning, web crawler

Introduction

The Sustainable Development Goals (SDGs) are currently being highlighted as important goals around the world. In order to achieve these SDGs, it is important to include ethical aspects in the value of food, and measures include the promotion of organic agriculture and the expansion of organic food. Green and organic foods are also attracting attention in China. Indeed, as China's economy grows steadily and consumer incomes increase dramatically, Chinese consumers are increasingly concerned about food safety and social trust.

The China Green Food Development Centre defined green food as follows: "Under strict supervision, control, and regulation in production, processing, packing, storage, and transportation, Green Food adopts the whole-some quality control from field to table, while it requires reasonable applications of inputs, including pesticide, fertilizer, veterinary drug, and additive, etc. to prevent any pollution of toxic and harmful matters to produce and links in food processing so as to ensure environmental and product safety" (Yu et al. 2014, pp. 80-81). Green food certification can be divided into two different levels: Grade A (which allows the use of a certain amount of chemicals) and Grade AA (which is equivalent to "organic food[1]") (Sanders 2006). Noticeably, however, the China Green Food Development Centre officially suspended the certification of Grade AA green food in June 2008 in response to the strengthening of organic food certification (Yu et al. 2014, p. 81).

The market for green and organic foods in China has grown rapidly. Figure 1 shows the number of certified green and organic food companies and products from 2001 to 2022. The number of green food-certified

[1] "Organic food" is defined by the China Organic Food Certification Center as "agricultural products and processed products that are produced using organic production farming methods, processed according to organic farming production requirements and standards, and certified by a legitimate organic food certification body" (Zhao 2009, pp. 15-16).

*Corresponding author: anakajim@meiji.ac.jp
Received 9 May 2024; accepted 5 August 2024; J-STAGE Advanced Epub 30 January 2025.
https://doi.org/10.6090/jarq.24J03

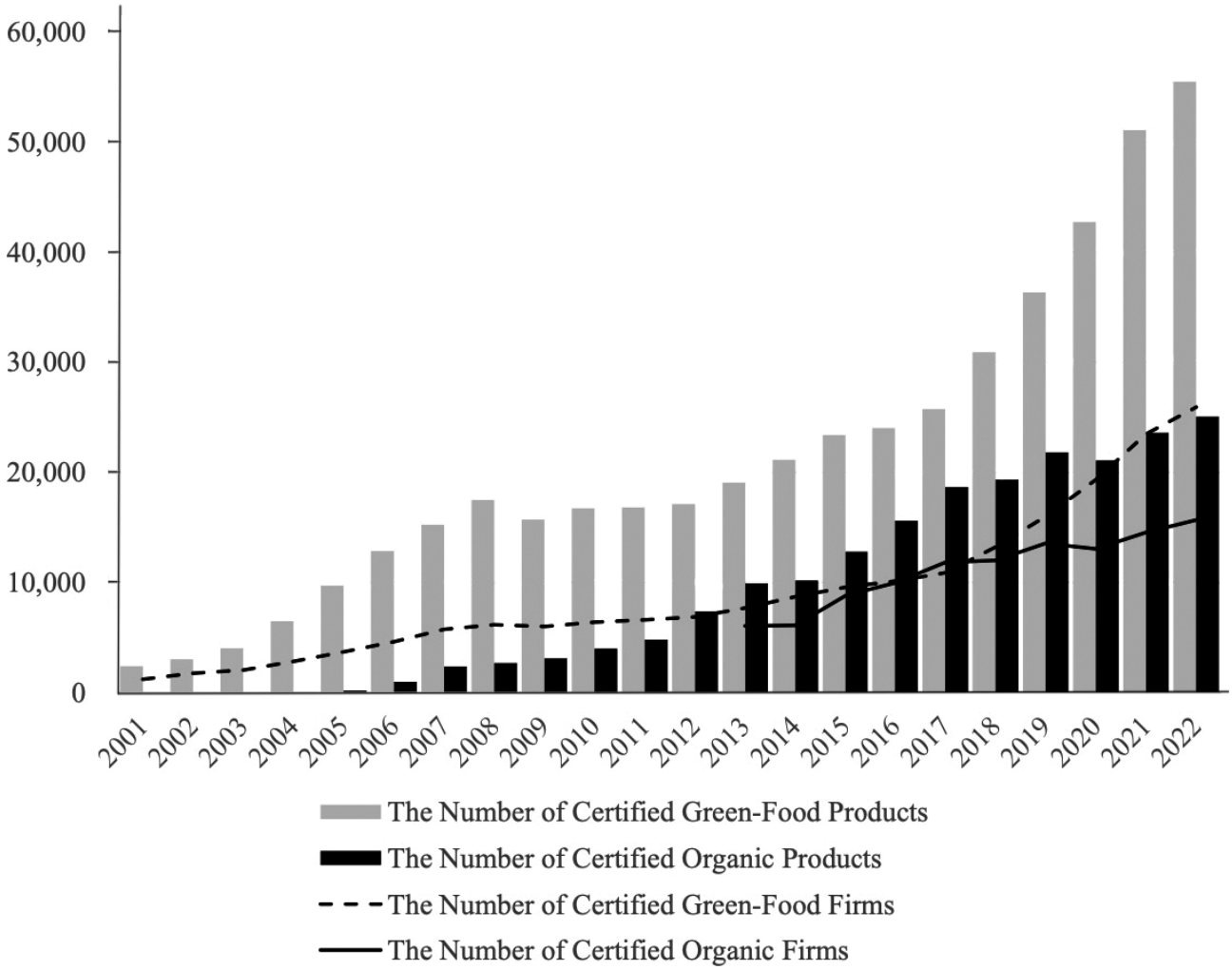

Fig. 1. The number of certified green and organic food firms and products (2001-2022)
Source(s): China Green Food Development Center (CGFDC) (2022), Green Food Statistical Yearbook, Beijing, China.
China's National Certification and Accreditation Administration (CNCA) (2022), Zhongguo Youjichanpin Renzheng yu Youjichanye Fazhan Baogao [China Organic Product Certification and Organic Industry Development Report]. Available at: https://www.samr.gov.cn/rzjgs/xczl/art/2022/art_667c5245ba6046679105613535858c0c.html (Accessed 15 January 2024).

products has grown approximately 23 times since 2001. Moreover, the number of firms that have received green food certification is 21 times that of 2001. Similarly, the organic food industry has grown dramatically, with the number of organic food products growing from 22 in 2004 to 25,063 in 2022. In addition, the number of firms with organic certification has grown from 6,051 in 2013 to 15,676 in 2022.

Previous studies on green food consumption, using choice experiments, have clarified that people are generally willing to pay a premium for green food (Yu et al. 2014, Zhou et al. 2017, Zhu et al. 2013). However, few studies have explicitly investigated customer satisfaction using actual green food purchase data. The current development of the Internet provides new opportunities to fill this gap since an ever-growing amount of word-of-mouth data that documents users' ideas and emotions can be found online. Indeed, research using word-of-mouth data has progressed remarkably in recent years, and coupled with the sophistication of machine learning methods, it has demonstrated the power to analyze various types of economic problems (Storm et al. 2020). In a study analyzing such word-of-mouth data on green food, Huang et al. (2022) used text mining techniques and neural networks to conduct sentiment analysis to study Chinese consumers' green consumption. Xu et al. (2023) mined e-commerce review data and used topic modeling and sentiment analysis to investigate differences between consumers' green-labeled rice consumption and conventional rice consumption.

This study aims to add to the existing literature by identifying and predicting consumers' evaluations of green-labeled rice in China. To do so, we adopt methods not used by Huang et al. (2022) and Xu et al. (2023), such as the tf-idf score, co-occurrence network, and random forest model. E-commerce was selected to collect textual review data. Moreover, for a more nuanced understanding of consumers' perceptions of green-labeled products, we compared the evaluation of green-labeled rice from different production sites, namely one product from Heilongjiang Province and one product from Hubei Province. The rationale for selecting these products is as follows. Firstly, Heilongjiang was chosen due to its premier status in China's rice production and its image as a dynamic rice production area. In contrast, Hubei ranks as the nation's fifth-largest rice producer and is recognized as a traditional rice cultivation area. By selecting these different production areas, we aim to identify if and how the geographical indication interacts with consumers' perceptions of green-labeled products. Secondly, we used JD.com's product search results to select the products with the highest sales and word-of-mouth data for green-labeled rice from Heilongjiang and Hubei provinces.

Online customer reviews provide considerable information but cannot be read directly. Therefore, this study also aims to identify consumer concerns about green-labeled rice based on tf-idf scores. In addition, a classification model based on the random forest algorithm will be conducted to investigate the relationship between actual green food consumption and satisfaction.

The practical significance of this study is that the analysis of consumer comments provided a new research method in the development of green food and other safety-certified food products to study consumer perceptions and emotions toward green-labeled rice. Since consumers can use online reviews to freely express their opinions about a product, the word-of-mouth data allows us to analyze genuine consumer feedback, and the application of text mining and a random forest model to e-commerce word-of-mouth data offers more realistic and objective information for evaluating consumer perceptions of green-labeled rice compared to previous studies (Yu et al. 2014, Zhou et al. 2017, Zhu et al. 2013) that relied on questionnaires or interviews.

Methods

1. Data collection

JD.com, one of the three largest integrated e-commerce companies in China, was selected for this study as a source of consumer review data. The Jingdong platform allows consumers to freely review and rate products on a scale of 1-5. In this study, we searched for "green-labeled rice 5kg" to reach the product list page. We selected two products that ranked in the top 10 and came from different origins with significant price differences, as follows: Product F originating in Heilongjiang province (price: 99.90 yuan, At the time of this study, 1 yuan = 0.14 USD) and Product G originating in Hubei province (price: 57.90 yuan). These two commodities are green food-certified rice and national geographical indication-certified rice. In particular, Product F is a renowned corporate-branded rice in Heilongjiang Province, a significant Chinese rice production area. Its geographical indication as *Wuchang rice* [五常大米] is esteemed nationwide.

In this study, data extraction was performed from August 18, 2023 to August 27, 2023, using web crawler technology with Python 3.8.8 in an Anaconda 3 environment. The data contained user ID, posting time, purchasing time, user location, whether the user is a member of the e-mall or not, user's product reviews, and user's product ratings. A total of 2,483 from the available public reviews for Product F and 1,949 for Product G were obtained and saved in CSV format for data processing and analysis. In the preprocessing of data analysis, we first removed duplicate data, which we considered to be the same content posted by the same user ID. We removed reviews with no review content and only product ratings. We then deleted text with confusing format, missing content, and irrelevant text. Finally, we cleaned up some meaningless information in the text, including special symbols and emoticons.

We performed word frequency statistics and word separation steps several times to build a customized word list. We found that some words could not be recognized during the word separation process, which affected the result of word separation. With reference to Xu et al. (2023), we developed a customized dictionary and added 54 words to improve the accuracy of word segmentation (Table 1).

Finally, we collected 53,674 words (of 4,205 types) in the customer reviews of Product F and 29,572 words (of 2,685 types) in the customer reviews of Product G to be set as target words for analysis.

2. Analyzing datasets and feature extraction: an approach based on the tf-idf score

Before the text mining analysis of the customer reviews, words had to be extracted from the sentences of each analyzed customer review, and their parts of speech identified. This procedure, known as morphological analysis, was performed using the jieba package in

Table 1. Custom dictionaries

No.	Word (Chinese)	Word (English)
1	不好吃	Not good
2	不香	Not fragrant
3	不喜欢	Don't like
4	不划算	Not a good deal
5	米香	Rice fragrant
6	新米	New rice
7	陈米	Stale rice
8	长粒香	Long-grain rice
9	珍珠米	Pearl rice
10	旧米	Old rice
11	五常大米	Wuchang rice
12	颗粒均匀	Uniform grains
13	颗粒饱满	Full of grains
14	晶莹剔透	Crystal clear
15	软糯	Soft and sticky
16	送货上门	Home delivery
17	物流速度	Logistics speed
18	发货速度	Shipping speed
19	京东快递	Jingdong express
20	真空包装	Vacuum packed
21	包装破损	Packaging broken
22	漏气	Leaking
23	原产地	Origin
24	绿色食品	Green food
25	生产日期	Production date
26	煮粥	Cooked porridge
27	降价	Price reduced
28	性价比	Value for money
29	发霉	Mouldy
30	国家地理标志	National geographical indication
31	南方大米	Southeast rice
32	京山桥米	Qiaomi rice of Jingshan
33	官方溯源	Official traceability
34	稻花香	Daohuaxiang rice
35	东北大米	Northeast rice
36	给力	Awesome
37	普通大米	Conventional rice
38	F 商品名称	Name of product F
39	防伪标识	Anti-forgery label
40	很好	Very good
41	很香	Smell very good
42	香糯	Fragrant and sticky
43	挺好	Well
44	网购	Online shopping
45	嚼劲	Chewiness
46	点赞	Press the good button
47	五星好评	Five-star review
48	五常米	Wuchang rice
49	五常稻花香大米	Wuchang daohuaxiang rice
50	快递员	Courier
51	京东自营	Jingdong self-supported
52	东北人	Northerner
53	长粒米	Long-grain rice
54	G 商品名称	Name of product G

Python 3.8.8. The tf-idf scoring technique was then used to extract identifiable words from the sample.

The tf-idf scoring technique is a method used to represent the importance (weight) of words in a text document. Here, a high tf-idf score means high importance, where a word identified as such can be used as the feature word of the text. Conversely, words with low importance can be considered to have no significant impact on classification (Gentzkow et al. 2019, p.538). The tf-idf score is the value obtained by multiplying the two metrics, the term frequency (tf) and the inverse document frequency (idf). The tf-idf score equation is as follows:

$$\text{tf-idf}_\text{i} = tf_i \times idf_i$$
$$= tf_i \times (log(N/df_i) + 1),$$

where subscript i and N represent each word and the total number of customer reviews used in our analysis, respectively. In addition, df_i is the document frequency of word i expressing the number of documents in which word i appears.

In this research, we used the TfidfTransformer of feature_extraction.text module from the scikit-learn package in Python to calculate the tf-idf score of review contents. The top 300 highest-scoring words were extracted as features for the construction of the random forest model.

3. Data analysis: random forest algorithm and co-occurrence network graph

The random forest algorithm is a supervised learning technique for regression and classification problems (Hastie et al. 2009). This technique builds decision trees on different subsets of a given dataset and classifies them by majority vote to improve the predictive reliability of

that dataset and predict the final output. As the number of decision trees in the forest increases, the accuracy also increases, thereby avoiding the problem of overfitting. The random forest algorithm is used to predict scores for large-scale datasets because they are highly accurate.

After preprocessing the data through morphological analysis and preparing the target words using tf-idf scoring, the tf-idf scoring for the top 300 words was introduced at the analysis stage as the variable name of the feature during the construction of the random forest model. In this study, we used the RandomForest package in R to perform the analysis. The number of occurrences of tf-idf scoring for the top 300 words was assigned as an independent variable. There are five levels of scores (1, 2, 3, 4, and 5), with classification divided into five levels and assigned as response variables.

Furthermore, we created a co-occurrence network graph connected with the keyword "green consumption" to visualize the words specific to green food evaluation. This co-occurrence network graph was drawn using Gephi software based on the Python language. This quantified the co-occurrence relationships between variables in customer dictionaries and each word. We thereby analyzed the impact of important terms related to green food on consumer ratings.

Results

1. Descriptive statistics on consumer review data

We counted review data of green-labeled rice between August 18, 2023 and August 27, 2023. The final obtained data of 2,483 reviews for Product F spans over a period of October 21, 2019 to August 20, 2023, whereas the data of 1,949 reviews for Product G are from September 6, 2016 to August 18, 2023. The ratings in the data of Product F and Product G can be summarized as follows: Product F: score 1 (173 reviews), score 2 (72 reviews), score 3 (188 reviews), score 4 (17 reviews), and score 5 (1,948 reviews); Product G: score 1 (68 reviews), score 2 (29 reviews), score 3 (67 reviews), score 4 (20 reviews), and score 5 (1,631 reviews). The JD.com review data includes the location of consumers; Figure 2, therefore, shows the geographical distribution of consumers of green-labeled rice obtained from the review data. From Figure 2, comparing the locations of consumers of Product F and Product G, consumers of Product F produced in Heilongjiang Province are more widely distributed geographically; in particular, many consumers are from Guangdong Province. In contrast, most purchasers of Hubei Province's Product G are from that province.

2. Summary of consumer evaluation of green-labelled rice

Table 2 and Table 3 list the top 100 words based on the tf-idf score.

First, from Table 2, it can be seen that consumers of Product F rate the attributes of rice in the following order: national geographical indication *Wuchang rice* [五常大米] (rank 1), *taste* [口感] (rank 3), *packaging* [包装] (rank 6), *aroma* [香味] (rank 12), *logistics* [物流] (rank 15), *place of origin* [五常 (Wuchang)] (rank 24), *variety of rice* [稻花香 (Daohuaxiang rice)] (rank 26), *quality of rice* [新米 (new rice)] (rank 28), *quality* [品质] (rank 40), and *price* [价格] (rank 29). Second, numerous words pertaining to certification and quality are mentioned as follows: *trust* [信赖] (rank 74), *organic* [有机] (rank 75), *genuine* [正品] (rank 125), *traceability* [官方溯源] (rank 153), and *green* [绿色] (rank 288).

Next, in Table 3, we discover that Product G customers prioritize the following attributes of

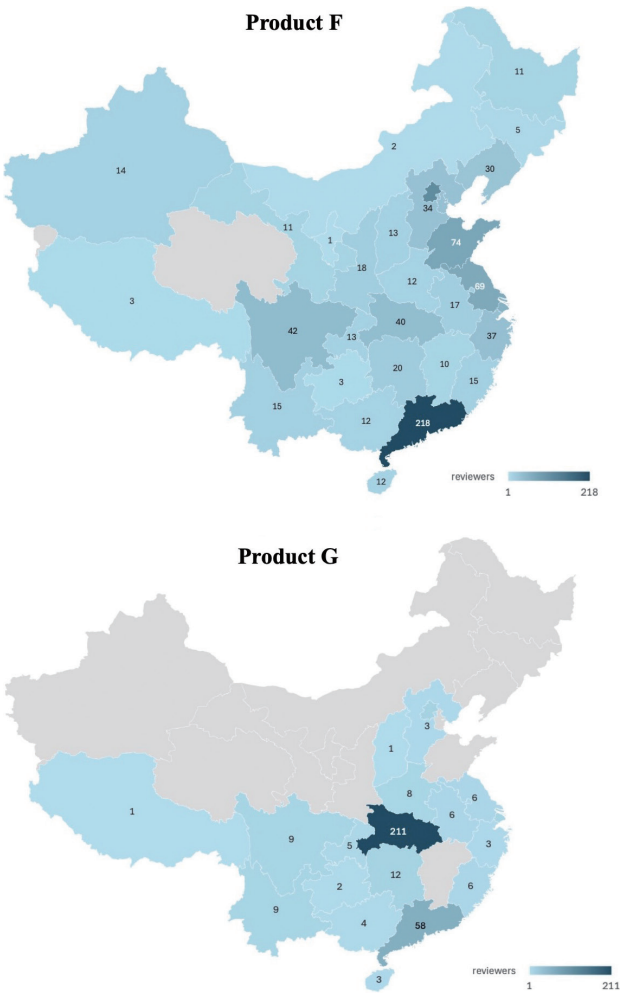

Fig. 2. Number of reviewers by geographical region

Table 2. tf-idf scoring for top 100 words: Product F

Rank	Word (Chinese)	Word (English)	tf-idf	Rank	Word (Chinese)	Word (English)	tf-idf
1	五常大米	Wuchang rice	0.261	51	嚼劲	Chewiness	0.023
2	大米	Rice	0.208	52	打开	Open	0.023
3	口感	Taste	0.177	53	送货上门	Delivered to your door	0.022
4	好吃	Yummy	0.156	54	购物	Shopping	0.022
5	京东	Jingdong	0.114	55	下次	Next time	0.022
6	包装	Packaging	0.106	56	多次	Multiple times	0.022
7	不错	Not bad	0.105	57	正宗	Authentic	0.021
8	非常	Very	0.102	58	感觉	Feeling	0.021
9	产品包装	Product packaging	0.087	59	颗粒饱满	Full of grains	0.021
10	很好	Very good	0.086	60	可以	Possible	0.021
11	米饭	Rice	0.079	61	快递小哥	Courier	0.021
12	香味	Aroma	0.075	62	品牌	Brand	0.021
13	物流速度	Logistics speed	0.074	63	给力	Awesome	0.020
14	很香	Very fragrant	0.067	64	挺好	Well	0.020
15	物流	Logistics	0.061	65	香甜	Sweet	0.019
16	购买	Purchase	0.061	66	乔府	Qiaofu	0.019
17	香气	Aroma	0.056	67	色泽	Color	0.019
18	真空包装	Vacuum-packed	0.052	68	实惠	Affordable	0.019
19	这个	This one	0.045	69	饱满	Full	0.018
20	味道	Flavor	0.045	70	粒粒	Grain	0.018
21	软糯	Soft and sticky	0.045	71	推荐	Recommended	0.018
22	风味	Flavor	0.045	72	还是	Still	0.018
23	喜欢	Favorite	0.044	73	就是	Just	0.017
24	五常	Wuchang	0.042	74	信赖	Trust	0.017
25	很快	Very fast	0.042	75	有机	Organic	0.017
26	稻花香	Daohuaxiang rice	0.039	76	性价比	Value for money	0.016
27	回购	Repurchase	0.039	77	比较	Compare	0.016
28	新米	New Rice	0.037	78	发货	Shipping	0.016
29	价格	Price	0.036	79	第一次	First time	0.016
30	产品	Products	0.036	80	牌子	Brand	0.016
31	收到	Received	0.035	81	一袋	One bag	0.016
32	东北大米	Northeast rice	0.034	82	大院	Compound	0.016
33	米香	Rice aroma	0.034	83	京东快递	Jingdong Express	0.016
34	煮饭	Cooking	0.033	84	浓郁	Strong	0.015
35	质量	Quality	0.031	85	第二天	Next day	0.015
36	满意	Satisfied	0.031	86	起来	Get up	0.015
37	京东自营	Jingdong self-support	0.029	87	一次	Once	0.015
38	米粒	Grain of rice	0.029	88	这次	This time	0.015
39	下单	Place an order	0.029	89	超市	Supermarket	0.015
40	品质	Quality	0.029	90	闻到	Smell it	0.015
41	送货	Delivery	0.028	91	漏气	Leaking	0.015
42	好评	Good reviews	0.027	92	煮粥	Cooking porridge	0.014
43	特别	Special	0.027	93	一般	General	0.014
44	一直	Always	0.026	94	一斤	One catty	0.014
45	产地	Origin	0.026	95	点赞	Press the good button	0.014
46	值得	Worth it	0.026	96	速度	Speed	0.014
47	快递	Express	0.024	97	到货	Arrived	0.014
48	晶莹剔透	Crystal clear	0.023	98	确实	Indeed	0.014
49	出来	Come out	0.023	99	活动	Promotion	0.013
50	没有	No	0.023	100	两袋	Two bags	0.013

Table 3. tf-idf scoring for top 100 words: Product G

Rank	Word (Chinese)	Word (English)	tf-idf	Rank	Word (Chinese)	Word (English)	tf-idf
1	口感	Taste	0.243	51	比较	Comparison	0.024
2	好吃	Yummy	0.185	52	国宝	National Treasure	0.024
3	G 商品名称	Name of product G	0.166	53	下次	Next	0.024
4	京东	Jingdong	0.157	54	京东快递	Jingdong Express	0.024
5	大米	Rice	0.150	55	配送	Delivery	0.023
6	很好	Very good	0.130	56	可以	Can	0.023
7	不错	Not bad	0.123	57	一袋	One bag	0.023
8	非常	Very	0.100	58	品质	Quality	0.022
9	真空包装	Vacuum-packed	0.097	59	真空	Vacuum	0.022
10	包装	Packaging	0.085	60	这次	This time	0.022
11	物流	Logistics	0.081	61	给力	Awesome	0.021
12	产品包装	Product Packaging	0.076	62	两袋	Two bags	0.021
13	购买	Purchase	0.073	63	便宜	Cheap	0.021
14	一直	Always	0.070	64	京东自营	Jingdong Self-support	0.021
15	送货	Delivery	0.067	65	日期	Date	0.020
16	这个	This one	0.066	66	经常	Often	0.020
17	物流速度	Logistics speed	0.066	67	家乡	Hometown	0.020
18	软糯	Soft and sticky	0.064	68	活动	Promotion	0.020
19	喜欢	Favorite	0.062	69	牌子	Brand	0.020
20	湖北	Hubei	0.058	70	好评	Good reviews	0.020
21	很快	Very fast	0.054	71	米香	Rice aroma	0.019
22	回购	Repurchase	0.049	72	东西	Objects	0.019
23	风味	Flavor	0.049	73	新米	New Rice	0.019
24	多次	Multiple times	0.049	74	速度	Speed	0.019
25	送货上门	Delivered to the door	0.047	75	清香	Fresh aroma	0.018
26	很香	Very fragrant	0.046	76	点赞	Press the good button	0.018
27	香气	Aroma	0.045	77	服务态度	Service Attitude	0.018
28	米饭	Rice	0.043	78	产地	Place of origin	0.018
29	收到	Received	0.040	79	小哥	Delivery personnel	0.018
30	漏气	Leaking	0.039	80	煮粥	Cook porridge	0.017
31	快递小哥	Courier	0.038	81	性价比	Value for money	0.017
32	满意	Satisfied	0.037	82	推荐	Recommended	0.017
33	味道	Taste	0.037	83	品牌	Brand	0.017
34	京山桥米	Qiaomi rice of Jingshan	0.036	84	宝贝	Product	0.016
35	下单	Place an order	0.034	85	实惠	Affordable	0.016
36	质量	Quality	0.034	86	出来	Come out	0.016
37	快递	Express delivery	0.033	87	就是	Just	0.016
38	产品	Product	0.033	88	感谢	Thank you	0.016
39	价格	Price	0.033	89	超级	Super	0.015
40	京山	Jingshan	0.032	90	服务	Service	0.015
41	值得	Worth	0.031	91	米粒	Grain of rice	0.015
42	煮饭	Cooking	0.031	92	还是	Still	0.015
43	超市	Supermarket	0.030	93	一次	Once	0.015
44	香味	Aroma	0.029	94	谢谢	Thank you	0.015
45	特别	Special	0.029	95	发货	Shipping	0.015
46	送到	Delivery	0.025	96	五星好评	Five-star review	0.014
47	方便	Convenient	0.025	97	物美价廉	Good value for money	0.014
48	购物	Shopping	0.025	98	不是	No	0.014
49	家里	Home	0.025	99	没有	No	0.014
50	新鲜	Fresh	0.024	100	辛苦	Hard work	0.014

green-labeled rice: *taste* [口感] (rank 1), *packaging* [包装] (rank 10), *logistics* [物流] (rank 11), *place of origin Hubei* [湖北] (rank 11), *flavor* [风味] (rank 23), national geographical indications *Qiaomi rice of Jingshan* [京山桥米] (rank 34), *quality* [质量] (rank 36), and *price* [价格] (rank 39) of rice. The tf-idf score of corporate product brand product *G* [国宝桥米] (rank 3) is higher than that of the national geographic indication *Qiaomi rice of Jingshan* [京山桥米] (rank 34). This result differs from Product F. And the tf-idf scores for corporate product brand product *G* [国宝桥米] (rank 3) and national geographical indication *Qiaomi rice of Jingshan* [京山桥米] (rank 34) are higher than those for *green food* [绿色食品] (rank 268). Moreover, the reviews of Product G mention *repurchase* [回购] (rank 22), *hometown* [家乡] (rank 67), *fried rice* [炒饭] (rank 119), *southeast rice* [南方大米] (rank 135), *long-grain rice (Indica rice)* [籼米] (rank 137), *place of region* [老家] (rank 153), and *southerner* (people from China who live south of the Yangtze River) [南方人] (rank 196).

In summary, Product F customers are highly conscious of commodity labels and brands; on the other hand, a lot of consumers select Product G based on their eating preferences as well as their preferred texture, and a lot of people who purchase the rice are either from Hubei Province or are repeat customers.

3. Understanding consumer ratings of green-labeled rice from the random forest model

Next, we examine in detail the part these components play in how consumers assess green-labeled rice. Table 4 and Table 5 report the variable importance for the top 300 words of tf-idf scoring resulting from the random forest model[2,3].

Table 4 shows that the national geographical indication *Wuchang rice* [五常大米] (rank 14) is an important factor influencing the consumer rating of Product F. Also, words such as *taste* [口感] (rank 18), *price* [价格] (rank 23), *aroma* [香味] (rank 24), *rice quality* [新米 (new rice)] (rank 33), *very fragrant* [很香] (rank 37), *logistics* [物流] (rank 38), *make promotion* [搞活动] (rank 42), *flavor* [风味] (rank 47), and *quality* [质量] (rank 48) are highly predominant. On the other hand, *green* [绿色] (rank 284) has less of an effect on the consumer rating.

For Product G, Table 5 indicates that the corporate product brand product *G* [国宝桥米] (rank 41) has a higher influence than the *place of origin* [Jingshan [京山] (rank 72), 湖北 (Hubei)] (rank 83), and the national geographical indicator *Qiaomi rice of Jingshan* [京山桥米] (rank 129). In addition, *vacuum packed* [真空包装] (rank 4), *taste* [口感] (rank 8), *price* [价格] (rank 25), and *value for money* [性价比] (rank 50) are significant elements in the evaluation of consumers. The impact of *green food* [绿色食品] (rank 57) is higher than that of *place of region* [京山 (Jingshan), 湖北 (Hubei)]. This suggests that being a green food appears to be highly significant to Product G customers when buying Hubei green-labeled rice.

Figure 3 and Figure 4 show partial dependence plots. Figure 3 shows customer ratings in Product F reviews are positively impacted by the following words: *Wuchang* [五常], *northeast of China* [东北], *Wuchang rice* [五常大米], *northeast rice* [东北大米], and *traceability* [溯源], as well as *promotion* [活动], *new rice* [新米], *flavor* [清香 (fresh aroma)], and *promotion* [活动]. Figure 4 illustrates how customer ratings in Product G reviews are positively impacted by *green food* [绿色食品], *new rice* [新米], *flavor* [清香 (fresh aroma)], *long-grain rice (Indica rice)* [籼米], *promotion* [活动], *a little bit expensive* [小贵], *southeast rice* [南方大米], and *Hubei* [湖北]. It can be concluded that in the case of Product F, customers rate green-labeled rice higher when words related to the place of origin appear in reviews, while in the case of Product G, customers rate green-labeled rice higher when words related to green food appear in reviews. The two products differ greatly in the attributes of the green-labeled rice consumers evaluate, depending on whether it is from the place of origin or green food.

4. Understanding consumer perceptions included in reviews

We then use the co-occurrence network to specifically identify which elements *green food* [绿色食品] is related to in the reviews of the two commodities, as well as to discuss the reasons why *green food* [绿色食品]

[2] The total sample of Product F is split to obtain 1,439 training data (60%) and 959 test data (40%); the total sample of Product G is split to obtain 1,089 training data (60%) and 726 test data (40%). To start, we need to choose the optimal number of trees to use in the random forest algorithm in order to reduce misclassification errors. The minimum OOB error is found with a number of 714 trees for Product F and a number of 252 trees for Product G, which indicates that aggregation over this number of trees provides the most accurate predictions.

[3] The out-of-sample classification performance of the random forest model was evaluated using Out-Of-Bag (OOB) observations. The classification model for the Product F compound has a high accuracy rate of 0.961 and Cohen's kappa (0.873) stays between 0.81 to 1.00, indicating a high degree of consistency within classification results. On the other hand, for Product G, Cohen's kappa (0.698) is slightly lower, but the accuracy rate is as high as 0.956, indicating that both the model and the classification results are reasonably convincing.

Table 4. Random forest model for top 100 words: Product F

Rank	Word (Chinese)	Word (English)	Mean Decrease Gini	Rank	Word (Chinese)	Word (English)	Mean Decrease Gini
1	一般	General	35.943	51	觉得	Feel like	3.158
2	没有	No	23.399	52	真空包装	Vacuum-packed	3.049
3	不值	Not worth	17.331	53	特别	Special	3.015
4	不错	Not bad	16.039	54	一直	Always	3.012
5	一般般	Average	13.488	55	有点	A little	2.996
6	不好吃	Not good for eating	12.746	56	产品包装	Product packaging	2.954
7	不是	Not	12.257	57	一样	Same	2.919
8	好吃	Yummy	11.236	58	回购	Repurchase	2.829
9	非常	Very	11.047	59	产品	Products	2.804
10	很好	Very good	10.369	60	知道	Knowing	2.761
11	区别	Difference	10.218	61	里面	Inside	2.722
12	一斤	One catty	9.652	62	差不多	Almost	2.690
13	不香	Unflavored	8.260	63	起来	Get up	2.556
14	五常大米	Wuchang rice	8.115	64	这么	So/Much	2.542
15	大米	Rice	7.792	65	出来	Come out	2.478
16	想象	Imagine	6.810	66	第一次	First time	2.472
17	块钱	Yuan (RMB)	6.325	67	活动	Promotion	2.448
18	口感	Taste	6.215	68	外包装	Outer package	2.433
19	感觉	Feeling	6.053	69	两袋	Two bags	2.419
20	购买	Purchase	5.950	70	米香	Rice aroma	2.352
21	漏气	Leaking	5.752	71	物流速度	Logistics speed	2.287
22	包装	Packaging	5.748	72	这次	This time	2.265
23	价格	Price	5.582	73	不会	No/Cannot	2.220
24	香味	Aroma	5.576	74	这种	This kind of	2.214
25	一点	A little	5.555	75	还是	Still	2.140
26	喜欢	Favorite	5.496	76	一个	One	2.116
27	京东	Jingdong	5.485	77	收到	Received	1.999
28	这个	This one	5.276	78	香气	Aroma	1.981
29	超市	Supermarket	4.992	79	软糯	Soft and sticky	1.977
30	陈米	Stale rice	4.930	80	打开	Open	1.941
31	普通	Ordinary	4.733	81	一次	Once	1.907
32	一袋	One bag	4.654	82	溯源	Traceability	1.904
33	新米	New Rice	4.646	83	商品	Commodity	1.855
34	以前	Previously	4.428	84	好评	Good reviews	1.832
35	之前	Before	4.384	85	品牌	Brand	1.809
36	很快	Very fast	4.333	86	评价	Evaluation	1.798
37	很香	Very fragrant	4.321	87	评论	Comments	1.796
38	物流	Logistics	4.213	88	其他	Other	1.780
39	米饭	Rice	4.145	89	生产日期	Date of manufacture	1.774
40	可以	Possible	3.824	90	而且	And	1.750
41	失望	Disappointment	3.771	91	发货	Shipping	1.694
42	搞活动	Promotion	3.638	92	满意	Satisfied	1.674
43	大家	Everyone	3.482	93	已经	Already	1.598
44	但是	But	3.418	94	品质	Quality	1.568
45	味道	Flavor	3.401	95	正宗	Authentic	1.529
46	客服	Customer service	3.390	96	五常米	Wuchang rice	1.525
47	风味	Flavor	3.263	97	那种	That kind of	1.507
48	质量	Quality	3.228	98	五常	Wuchang	1.477
49	快递	Express	3.172	99	小贵	A little bit expensive	1.466
50	就是	Just	3.160	100	更好	Even better	1.456

Table 5. Random forest model for top 100 words: Product G

Rank	Word (Chinese)	Word (English)	Mean Decrease Gini	Rank	Word (Chinese)	Word (English)	Mean Decrease Gini
1	一般	General	10.200	51	喜欢	Favorite	1.260
2	漏气	Leaking	9.874	52	那么	In that case	1.249
3	不是	No	9.218	53	第一次	First time	1.248
4	真空包装	Vacuum-packed	6.711	54	质量	Quality	1.240
5	不好吃	Not good for eating	6.041	55	实惠	Affordable	1.234
6	陈米	Stale rice	5.606	56	比较	Comparison	1.222
7	没有	No	4.670	57	绿色食品	Green food	1.221
8	口感	Taste	4.130	58	还是	Still	1.205
9	真空	Vacuum	3.806	59	但是	But	1.203
10	包装	Packaging	3.774	60	新鲜	Fresh	1.172
11	袋子	Bags	3.733	61	可以	Can	1.168
12	一袋	One bag	3.635	62	今天	Today	1.163
13	不错	Not bad	3.323	63	送货	Delivery	1.126
14	好吃	Yummy	3.217	64	已经	Already	1.111
15	产品	Product	3.119	65	真是	It's really	1.052
16	米虫	Rice worm	2.985	66	价钱	Price	1.045
17	知道	Knowing	2.959	67	很多	A lot	1.014
18	这个	This one	2.854	68	送到	Delivery	0.985
19	一样	Same	2.785	69	快递	Express delivery	0.978
20	问题	Problem	2.656	70	一次	Once	0.966
21	很好	Very good	2.628	71	一包	One bag	0.958
22	包装袋	Packaging bag	2.572	72	京山	Jingshan	0.875
23	就是	Just	2.415	73	配送	Delivery	0.868
24	客服	Customer service	2.365	74	保鲜膜	Plastic wrap	0.857
25	价格	Price	2.355	75	打开	Open	0.857
26	新米	New Rice	2.328	76	那种	That kind	0.842
27	一点	A little	2.317	77	很快	Very fast	0.826
28	收到	Received	2.253	78	物流速度	Logistics speed	0.819
29	米袋	Rice bag	2.251	79	希望	Hope	0.817
30	大米	Rice	2.227	80	方便	Convenient	0.789
31	两袋	Two bags	2.220	81	软糯	Soft and sticky	0.786
32	东西	Objects	2.214	82	风味	Flavor	0.771
33	京东	Jingdong	2.185	83	湖北	Hubei	0.766
34	以后	Later	2.024	84	这种	This kind of	0.761
35	非常	Very	1.987	85	每次	Every time	0.754
36	产品包装	Product Packaging	1.949	86	特别	Special	0.745
37	破损	Broken	1.937	87	好评	Good reviews	0.740
38	一直	Always	1.909	88	送货上门	Delivered to the door	0.735
39	塑料袋	Plastic bag	1.842	89	香味	Aroma	0.715
40	超市	Supermarket	1.755	90	多次	Multiple times	0.701
41	G 商品名称	Name of product G	1.733	91	体验	Experience	0.697
42	里面	Inside	1.732	92	购物	Shopping	0.691
43	味道	Taste	1.644	93	保存	Preservation	0.685
44	购买	Purchase	1.581	94	评价	Evaluation	0.679
45	一个	One	1.533	95	送来	Deliver	0.676
46	以前	Previously	1.521	96	米粒	Grain of rice	0.676
47	这次	This time	1.454	97	满意	Satisfied	0.653
48	物流	Logistics	1.420	98	一斤	One catty	0.638
49	速度	Speed	1.370	99	一下	One time	0.637
50	性价比	Value for money	1.326	100	有点	A little	0.631

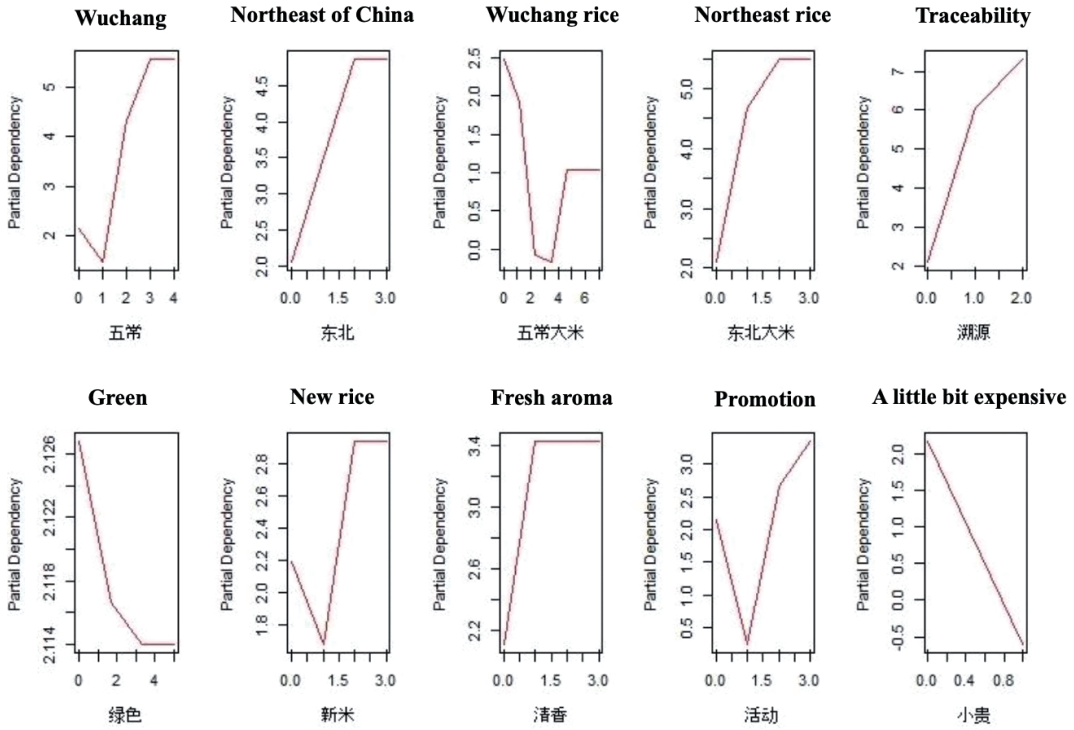

Fig. 3. Partial dependence of the rating of green-labelled rice (Product F)

Notes: The vertical axis represents the degree of influence on classification for each class, and the horizontal axis refers to the number of times the corresponding word appears in consumer reviews.

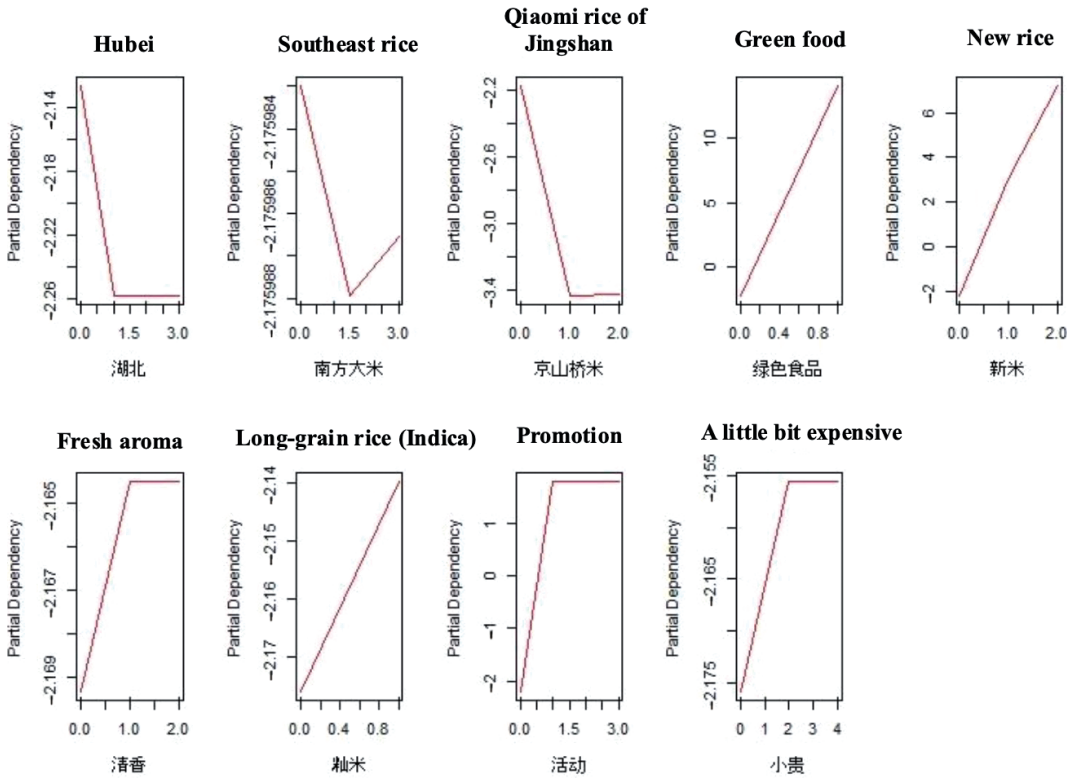

Fig. 4. Partial dependence of the rating of green-labelled rice (Product G)

Notes: The vertical axis represents the degree of influence on classification for each class, and the horizontal axis refers to the number of times the corresponding word appears in consumer reviews.

plays opposite roles in the reviews of the two commodities, because words related to *green food* [绿色食品] show opposite influence relationships in Figure 3 and Figure 4.

Figure 5 and Figure 6 show the co-occurrence network diagrams of Product F and Product G with green consumption as a keyword. First, we can see that Product F in Figure 5 has many intricate elements related to *green* [绿色], while Product G in Figure 6 has a simpler lexical connection for *green food* [绿色食品]. Second, the words *green* [绿色] and *organic* [有机] appear together in Product F, indicating that consumers are confused about the concepts of green and organic foods. In fact, the tf-idf score for product F shows that *organic* [有机] (rank 75) is higher than *green* [绿色] (rank 288), and similarly, *organic* [有机] (rank 129) is more important than *green* [绿色] (rank 284) in the random forest model.

This finding is interesting with regard to the Chinese food labeling system. As mentioned above, prior to 2008, green food certification in China was divided into two categories: Class A and Class AA (equivalent to organic food certification), but after 2008, the Class AA label was abolished in favor of a stricter certification as "organic." Organic foods are generally more expensive than green foods. In the case of Product F, the rice is expensive despite "only" being labeled as a green food. Therefore, the coexistence of organic and green food labeling at the co-occurrence network of Product F suggests that customers erroneously perceive Product F as an organic food rather than a green food and, accordingly, are willing to accept a higher price. Conversely, it is conceivable that Product G consumers are fully aware of the attributes of the rice. This could be the reason why the partial dependent plot of green food is positive in Product G but negative in the Product F plot.

Conclusion

This study aimed to explore the actual green food consumption experience and its relationship with green food satisfaction by performing text mining on online green-labeled rice reviews, ranking the top 300 words with the highest frequency and importance based on tf-idf scoring, and constructing a random forest model. The e-mall platform JD.com was used in this study as a source of research data. This study compared customer reviews of green-labeled rice of northeast origin and green-labeled rice of southern origin and analyzed the factors influencing consumer purchases of green-labeled rice of different origins. We find that consumer concerns when shopping for green-labeled rice lie in the areas of taste, aroma, price, place of origin, type of rice (consistent with Xu et al. 2023), the convenience of shopping access, shopping logistics (consistent with Zhu et al. 2013), and the presence or absence of food certifications. The presence of keywords related to food safety and social trust in the tf-idf scoring confirms the increased awareness of health and food safety among Chinese consumers. Still, it was also found that green food certification is not necessarily an attribute that consumers value most.

The findings of this study suggest that green food does not always represent a decisive attribute in the food choice process. We further find that this might be related to consumers' (mis)perceptions of which foods are "organic" and/or "green." Therefore, stricter labeling regulations for organic and green food, together with easier consumer identification and comprehension, are required to expand conscious green food consumption.

It should be noted that despite the rapid growth of online shopping in China, many consumers still use it infrequently or not at all. Consequently, e-commerce review data only capture the characteristics of a limited segment of consumers who purchase green food online. Despite these limitations, however, we believe this study's approach could be beneficial for marketers of agro-food products and researchers in food and agricultural economics.

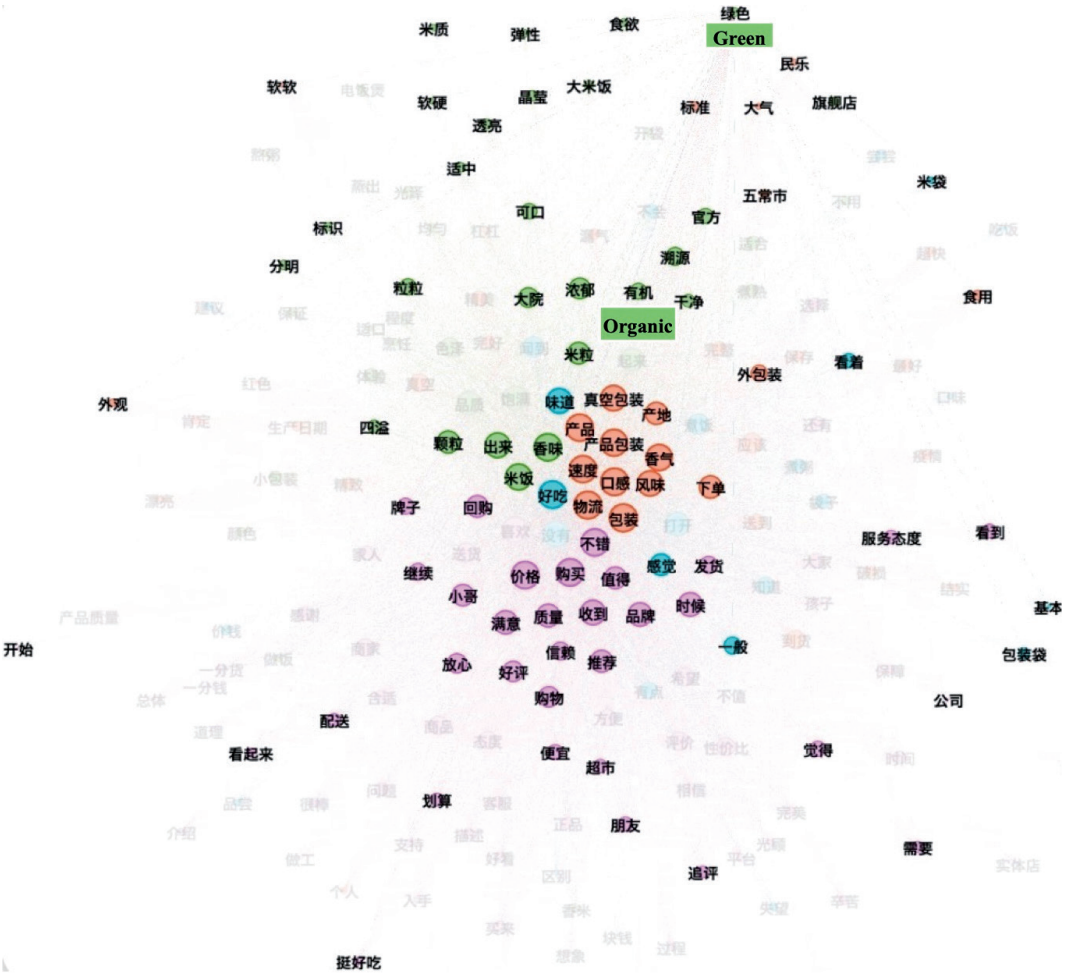

Fig. 5. Co-occurrence network graph connected with variable "Green" (Product F)

Notes: The keyword for this co-occurrence network is 绿色 (Green), and the words that co-occurred with 绿色 (Green) are as follows: 认证 (Authenticate), 米质 (Rice quality), 食欲 (Appetite), 弹性 (Elastic), 民乐 (Folk music), 软软 (Soft), 大米饭 (Rice), 晶莹 (Shiny), 软硬 (Soft or hard), 透亮 (Crystal), 标准 (Standard), 大气 (Decent), 旗舰店 (Flagship store), 适中 (It's just right), 米袋 (Rice bag), 五常市 (Wuchang), 可口 (Delicious), 官方 (Official), 标识 (Identification), 分明 (Clearly), 溯源 (Traceability), 粒粒 (Grain), 浓郁 (Strong), 大院 (Yard), 有机 (Organic), 干净 (Clean), 食用 (Edible), 米粒 (Grain), 看着 (Look at that), 外包装 (Outer package), 真空包装 (Vacuum packed), 外观 (Appearance), 味道 (Flavor), 产地 (Origin), 四溢 (Overflowing), 产品 (Product), 颗粒 (Grain), 出来 (Come out), 香味 (Aroma), 产品包装 (Product packaging), 香气 (Aroma), 速度 (Speed), 米饭 (Rice), 口感 (Taste), 风味 (Flavor), 下单 (Place an order), 好吃 (Yummy), 牌子 (Brand), 回购 (Repurchase), 物流 (Logistics), 包装 (Package), 不错 (Not bad), 服务态度 (Service attitude), 看到 (See), 继续 (Continue), 价格 (Price), 购买 (Purchase), 感觉 (Feeling), 发货 (Shipping), 值得 (Worth it), 小哥 (Delivery personnel), 时候 (Time), 基本 (Basic), 满意 (Satisfied), 质量 (Quality), 收到 (Received), 品牌 (Brand), 开始 (Begin), 一般 (General), 信赖 (Trust), 放心 (At ease/Social trust), 包装袋 (Packaging bag), 好评 (Good reviews), 推荐 (Recommended), 购物 (Shopping), 公司 (Company), 配送 (Delivery), 看起来 (seem that), 便宜 (Cheap), 觉得 (Feel like), 超市 (Supermarket), 划算 (Good value), 朋友 (Friends), 需要 (Need), 追评 (Catch-up review), 挺好吃 (Quite tasty).

The more often a word appears in the reviews, the bigger its circle on the graph indicates. Words that occur more frequently at the same time or in more comparable semantics (or contexts) are grouped together and shown in the same color.

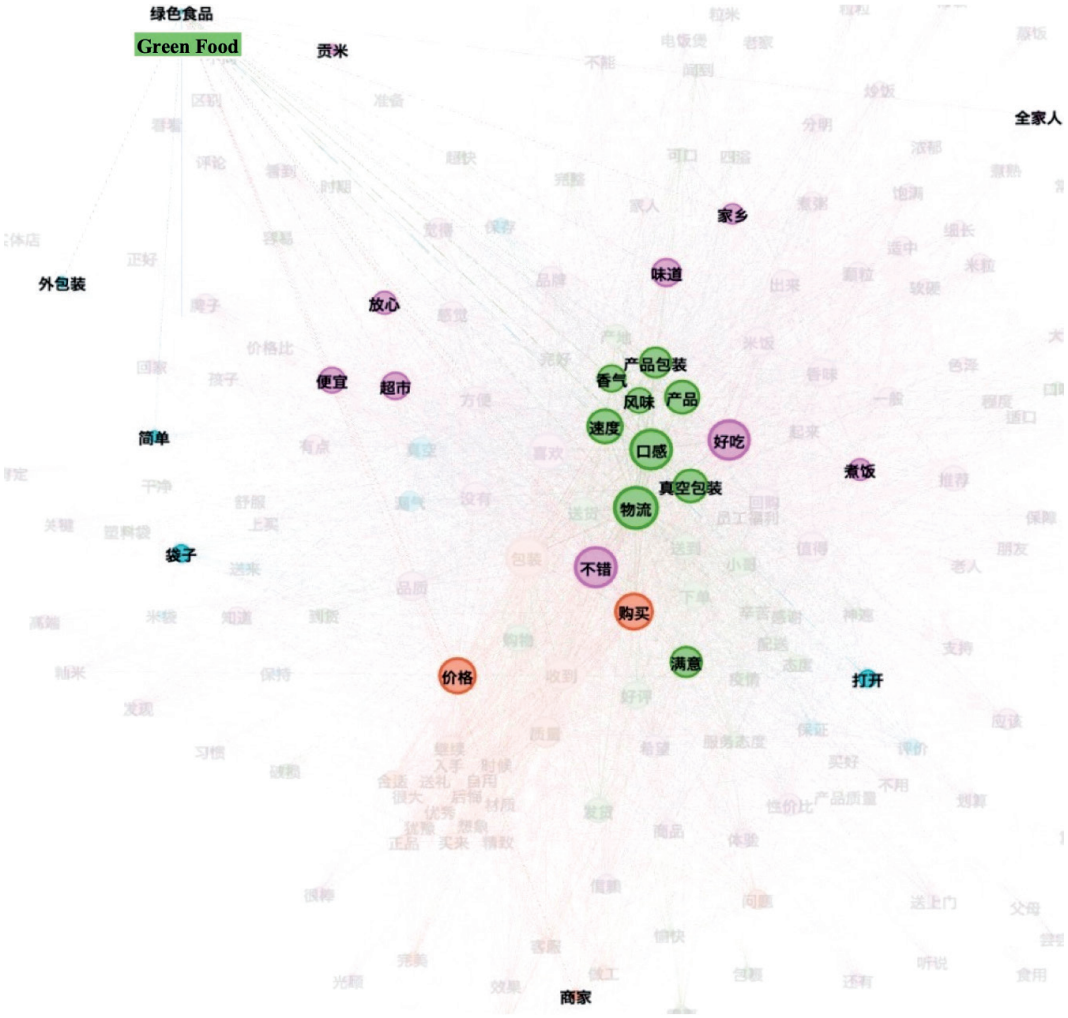

Fig. 6. Co-occurrence network graph connected with variable "Green food" (Product G)

Notes: The keyword for this co-occurrence network is 绿色食品 (Green food), and the words that co-occurred with 绿色食品 (Green food) are as follows: 贡米 (Special rice (Once reserved for the nobility)), 全家人 (Whole family members), 家乡 (Hometown), 外包装 (Outer package), 味道 (Flavor/Taste), 放心 (At ease/Social trust), 便宜 (Cheap), 产品包装 (Product packaging), 超市 (Supermarket), 香气 (Aroma), 产品 (Products), 风味 (Flavor), 简单 (Easy), 速度 (Speed), 好吃 (Yummy), 口感 (Taste), 煮饭 (Cooking), 真空包装 (Vacuum packed), 物流 (Logistics), 袋子 (Bag), 不错 (Not bad), 购买 (Purchase), 价格 (Price), 满意 (Satisfied), 打开 (Open), 商家 (Merchants).
The more often a word appears in the reviews, the bigger its circle on the graph indicates. Words that occur more frequently at the same time or in more comparable semantics (or contexts) are grouped together and shown in the same color.

References

Gentzkow, M. et al. (2019) Text as data. *J. Economic Literature*, **57**, 535-574.

Hastie, T. et al. (2009) *The Elements of statistical learning: Data mining, inference, and prediction, Second Edition*. Springer New York, NY.

Huang, H. et al. (2022) Exploring public attention about green consumption on Sina Weibo: Using text mining and deep learning. *Sustainable Production and Consumption*, **30**, 674-685.

Sanders, R. (2006) A market road to sustainable agriculture? ecological agriculture, green food and organic agriculture in China. *Development and Change*, **37**, 201-226.

Storm, H.et al. (2020) Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*, **47**, 849-892.

Xu, H. et al. (2023) Green-labelled rice versus conventional rice: Perception and emotion of Chinese consumers based on review mining. *Foods*, **12**, 87.

Yu, X. et al. (2014) Willingness to pay for the "green food" in China. *Food Policy*, **45**, 80-87.

Zhao, H. (2009) The development and present situation of "three kinds of food" certification system in China: non-polluted food, green food and organic food. *J. Food System Research*, **16**, 14-28 [In Japanese].

Zhou, J. et al. (2017) Habit spillovers or induced awareness: Willingness to pay for eco-labels of rice in China. *Food Policy*, **71**, 62-73.

Zhu, Q. et al. (2013) Green food consumption intention, behaviors and influencing factors among Chinese consumers. *Food Quality and Preference*, **28**, 279-286.