Value perception impact and countermeasures analysis of new energy vehicle purchase behavior based on consumer level user review big data mining

Yihang Lv[1,*], and *Qin* Liu[2]

[1]School of computer, Wuhan University of Technology, China
[2]Wuhan University of Technology, China

Abstract. The development of new energy vehicles is inseparable from the drive of consumers. Therefore, to explore the influencing factors of purchase behavior from the consumer's personal level is helpful for businesses to adopt corresponding sales strategies and the government to adopt relevant policies. Based on the individual level of consumers, this paper constructs a new energy vehicle purchase behavior prediction model from the review text, and explores the predictive effect of consumer personal factors on the purchase behavior of new energy vehicles. First of all, this paper proposes a quantitative method of consumer individual level factors, which combines word-of-mouth reviews with statistics. In this method, word2vec is used to train word vectors in word-of-mouth corpus to mine initial keywords, and core keywords are selected through statistical correlation analysis. Secondly, based on the core keywords of consumers' personal level, the gbdt model is constructed to predict the purchase behavior of new energy vehicles. The results show that the probability of correctly predicting consumers' purchase behavior is more than 72%.

1 Introduction

As the new energy vehicles are in the initial stage, the market is in the state of being developed. How to promote the new energy vehicles and depict the personal portraits of consumers is a key problem. The lack of understanding of consumers is an unavoidable problem to be solved in the further development of new energy vehicles.

The prediction of consumers' purchase behavior is a hot topic for many scholars. Han Minqi studied the behavior of users in the decision-making process by mining the implicit feedback data of users, and found that the implicit feedback data can better reflect the shopping tendency of consumers [1]. Hu Dongbo used decision tree to mine the questionnaire of e-commerce consumers [2]. Liu et al. Used a large number of user behavior data such as browsing, clicking, purchasing, and so on, through support vector machine (SVM) to predict the future purchase of online consumers, and obtained satisfactory results[3]. Based on the historical shopping behavior data and demographic

* Corresponding author: isunnygirl@qq.com

data, silahtaroglu et al. Used decision tree and neural network to predict whether customers would buy the goods in their shopping cart[4]. The traditional prediction of consumer purchasing behavior mainly starts from the historical data, but consumers play an important role in the process of purchasing. This paper puts forward a prediction model of purchasing behavior based on the processing of comment information from the perspective of consumers.

With the rapid development of the Internet and the rise of the network, online reviews have become the main channel to feed back consumers' demand for products and pain points. How to quickly and effectively mine the review data and predict the purchase behavior of consumers has become a major research direction. In terms of online reviews, some scholars construct emotional dictionaries and rich corpora to conduct statistical analysis on sales keywords. From the perspective of negative comments, some scholars extract the dissatisfaction degree of consumers to improve the direction of decision-making. In terms of consumer purchasing behavior, some scholars believe that technical factors predict the purchase behavior from the aspects of vehicle type and displacement, and some scholars consider it from the social level, and influence the purchase through Baidu search index and consumers' attention. However, more scholars will predict the purchase behavior of consumers from the perspective of appearance, interior, space, power, comfort and so on, ignoring the important part of consumers' purchase behavior.

In this paper, a new energy vehicle purchase behavior prediction model based on new energy consumers is proposed. At the level of keyword structure, the paper extracts the key words from the perspective of economic category, personal risk and values, and finally processes the keywords to get the prediction model of purchasing behavior.

2 Comment information measurement method based on consumer's personal level

2.1 Data preprocessing

Data acquisition. In this paper, through the establishment of Python crawler code, we can crawl consumers' comments on new energy vehicles from the new energy vehicles section of eckar automobile network, a total of 14218 items, which are saved in TXT file format.

Data preprocessing. The main purpose of data preprocessing is to remove the useless information in the text, and make the data clearer and easier to analyze. It is mainly divided into three aspects: de duplication, de-noising and deletion of short sentences. De duplication is to delete the repeated comments in the comment information to reduce the interference of redundant information. The amount of data after de duplication was 8723.

Denoising is to delete some special characters such as expression and website address in the comment information, which is lack of practical significance and research value, and will affect the following word segmentation work.

Text short sentence deletion refers to the deletion of data with too few words in the comment information. From the perspective of analysis, the less words mean that the review information contains less information value, which is likely to be generated randomly by consumers and has no feedback significance and research value, such as "very satisfied". In this paper, the comment data within 10 words is defined as useless data, so the comments within 10 words are deleted to improve the quality of comment information. After the deletion of short sentences, the amount of data is 8529.

Text segmentation. The segmented text data is large and difficult to analyze and process. Extracting reasonable and effective text keywords is an important operation in text processing. In this paper, Python's open source Jieba library is used for accurate word

segmentation of the comment set. The algorithm is simple to use and efficient, and the basic vocabulary is up to 350000. In order to reduce the amount of data, save the storage space of the text and improve the efficiency of word segmentation, we introduce the out of use vocabulary, in order to improve the credibility of the data and increase the professionalism of the vocabulary, build a user-defined vocabulary of new energy vehicle consumers. The comment text is segmented to generate candidate key word database.

2.2 Initial keyword definition

Text segmentation. The segmented text data is large and difficult to analyze and process. Extracting reasonable and effective text keywords is an important operation in text processing. In this paper, Python's open source Jieba library is used for accurate word segmentation of the comment set. The algorithm is simple to use and efficient, and the basic vocabulary is up to 350000. In order to reduce the amount of data, save the storage space of the text and improve the efficiency of word segmentation, we introduce the out of use vocabulary, in order to improve the credibility of the data and increase the professionalism of the vocabulary, build a user-defined vocabulary of new energy vehicle consumers. The comment text is segmented to generate candidate key word database.

Table1. Initial keywords (Part).

Level	Subdivision type	Initialization keywords
Economic category	Related policies	Car subsidy, car registration
	Automobile consumption	Automobile oil price, automobile price and purchase tax
	Economic capacity	Cost performance, saving money, maintenance price
Personal risk	Family composition	Child seats, family members, space
	Vehicle environment	Charging, after sales service, road conditions
	Conditions of use	Mileage, safety, endurance
Values	environmental awareness	Green, environmental protection, energy saving, low emission
	Purpose of car purchase	Shopping, work, travel
	Appearance and personality	Beautiful, comfortable and control

2.3 Keyword mining based on word2vec model

Car purchase review is an important content sharing channel for consumers to feedback on car purchase information. Due to the rapid development of the network, a large number of consumers will increase more information to express the reasons for making purchase decisions after purchasing goods, and upload the reasons for consumption decisions after using goods to assist other consumers to purchase. Reviews contain different dimensions of product information, but also contain a large number of subjective factors that affect the purchase behavior based on the consumer level. In this paper, we use the comments of Eka automobile network as corpus to mine keywords that can represent the impact of consumers' personal level on purchase behavior.

Based on the above review word segmentation data, word2vec algorithm is used to train the word vector model, and each word is mapped to a certain vector dimension. By

calculating the semantic similarity between the candidate keyword and the primary keyword, the keyword breadth is expanded.

In this paper, the skip gram model of word2vec algorithm is used. The dimension of the word vector is set to 100 and the training window is set to 5. The word vector model is trained by using the comment data after word segmentation. The cosine formula is used to define the semantic similarity between candidate words and primary keywords. The larger the cosine value is, the closer the semantic similarity is. The top 10 items with the largest semantic similarity are selected and expanded to 240 keywords to establish the initial key word database.

Table 2. Correlation analysis of primary keywords and candidate keywords(Part).

Primary keywords	Candidate keywords	Correlation coefficient
Car subsidy	Government	0.9508
	Car price	0.9235
	Cost effective	0.9103
	Not expensive	0.9103
Car registration	Lottery	0.9730
	Change	0.9616
	Big city	0.9601
Automobile oil price	Rise	0.9708
	Spending money	0.9384
Price	Being close to the people	0.8182
	Discount	0.8181
Child seat	Pick up	0.9473
	Go to school	0.8881
	Self-driving travel	0.8511
Green	Contribution	0.8132
	Clean	0.7870
Energy conservation	Low cost	0.8639
Manipulation	Easy to use	0.8476
	Convenient	0.8309

2.4 Improving the validity of word frequency and sentiment analysis

The effectiveness of word frequency was improved. Due to the development of big data, data can more and more reflect the social needs and trends. In this paper, the statistical measurement method based on big data is used to eliminate the initial keywords and measure the emotion. According to the number of initial key words, the number of keywords involved in the initial key words is filtered, and the keywords that are less involved in the comments and the keywords that are relatively few and have no obvious difference are deleted.

Emotional analysis. Consumers' personal level has a high degree of dependence on emotional feedback, so it is necessary to conduct Emotional Analysis on keywords. Positive emotional words will greatly promote consumers to make purchase behavior, while negative words will hinder the purchase behavior. The frequency of unsatisfied words is compared with the frequency of satisfactory words. If the frequency of satisfactory comments is more than twice the frequency of unsatisfied comments, they are classified as positive word frequency and marked as 1. If the frequency of unsatisfied comments is not

less than 1 / 2 times of the number of satisfactory comments, it is classified as negative word frequency and marked as - 1. The weight definition of keywords.

Table 3. Improve the vocabulary (Part).

First level	Secondary level	Key word
Economic category	National policy	Subsidy, policy,purchase tax, unlimited number
	Car expenses	Oil price, cost performance
Personal risk	Vehicle environment	After sales service, license plate, speed
	Conditions of use	Safety, endurance, charging
Values	Environmental awareness	environment protection
	Personal preference	Control, space
	Purpose of car purchase	Pick up ,go to work

Positive word frequency weight= Frequency of satisfaction/(Satisfaction frequency + dissatisfaction frequency). (1)

Since the influence of negative word frequency will be greater than that of the same type of positive word frequency, the definition of negative word frequency should be strengthened appropriately.

Negative word frequency weight=2 * frequency of dissatisfaction / (frequency of satisfaction + frequency of dissatisfaction) (2)

Table 4. Word frequency weight table.

Key word	Tagging	Weight	Key word	Tagging	Weight
Subsidy	1	0.8571	After-sale service	-1	1.086
Policy	1	0.8208	Security	-1	1.4829
Purchase tax	1	1	Speed	-1	1.314
Unlimited number	1	1	Endurance	-1	1.9684
Oil price	1	1	Charge	-1	0.8757
Cost performance	1	0.9812	Space	-1	0.7049
Economical and practical	1	0.8602	Ride	-1	0.7885
Registration	1	0.9469	Back row	-1	1.3770
Pick up	1	0.8827	Suspension	-1	1.5736
Go to work	1	0.8721	Comfortable	-1	0.9736
Operation	1	0.7927	Environment protection	1	1

3 Consumer purchase behavior prediction based on gdbt algorithm

Gbdt belongs to boosting algorithm. It uses the value of negative gradient direction of loss function in the current model as the approximate value of residual, and then fits a cart regression tree. Because each iteration of gbdt needs to fit gradient value and continuous value, regression tree should be used. As a combination of decision tree and gradient boosting, gbdt algorithm has higher prediction accuracy, can better fit and classify nonlinear data, and can flexibly process various types of data, including continuous value and discrete value. Compared with SVM, the prediction preparation rate can also be higher in the case of relatively small parameter adjustment time. Using some robust loss functions, it is very robust to outliers. For example, Huber loss function and quantity loss function.

The prediction effect of weak learners is limited, that is to say, there is a certain gap between the predicted results and the real results (i.e., residual). Then we can train the

second weak learner, take the residual as the target to learn, and then add the prediction results of the two weak learners, then the prediction results will be better than that of one weak learner. But there is still a certain residual, we can repeat the appeal process and continue to learn the weak learner until the goal is achieved. This is the idea of boosting.

In this paper, we use gbdt algorithm based on keyword weight to predict consumer purchase behavior. The basic principle of gbdt algorithm is as follows:

3.1 Initializing weak learners

$$f_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c) \tag{3}$$

3.2 For M = 1,2,..., M has:

• Calculate I = 1,2,..., N for each sample, and calculate the negative gradient
• Take the residual obtained in the previous step as the new real value of the sample, and take the data $(x_i,\ R_i)$, i = 1, 2,... N as the training data of the next tree to get a new regression tree. The corresponding leaf node region of F(x) is R_{jm}, j = 1,2,...., J. Where j is the number of leaf nodes of regression tree t.
• The best fitting value was calculated for the leaf region J = 1,2,... J

$$\gamma_{jm=\underbrace{\arg\min}_{\gamma}} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma) \tag{4}$$

• Update learner

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} \gamma_{jm} I(x \epsilon R_{jm}) \tag{5}$$

3.3 Get the final learner

$$f(x) = f_M(x) + \sum_{m=1}^{M} \sum_{j=1}^{J} \gamma_{jm} I(x \epsilon R_{jm}) \tag{6}$$

4 Experimental results and analysis

Data preparation. The experimental data of this paper are the satisfaction comments and dissatisfaction comments of Eka automobile network. The satisfaction comment data is defined as the consumer will make the purchase behavior decision, and the mark factor is 1; the unsatisfied comment data is defined as that the consumer will not make the purchase behavior decision, and the mark factor is 0. 75% of the review data are divided into training sets and 25% of the review data are divided into test sets. The review data are quantified according to the keyword weight in advance, and the lewnstein ratio of the keywords and the comment data is calculated respectively. The keyword vector is defined as the dimension multiplied by the weight multiplied by the lewnstein ratio. In order to increase the strength of the feature words, it is multiplied by 100 for each dimension. Then, the key words are summarized into seven dimensions: national policy, automobile expense, vehicle environment, use conditions, environmental awareness, personal preference and automobile use. The sum of keyword vectors of each dimension is represented as the feature. Then, the gdbt algorithm training is introduced to construct the prediction model. The number of decision trees is set to 3000, the maximum depth of each decision tree is set to 2, the minimum number of samples contained in each basic decision tree model leaf node is set to 0.1.

Keyword vector construction= Tagging * weight * 100 * Lewenstein ratio (7)

Analysis of experimental results. The accuracy of the model is as high as 0.8072. To a certain extent, the purchase behavior of new energy vehicles can be predicted from the perspective of consumers' personal level.

Table 5. Confusion matrix of consumer purchase behavior prediction.

The truth	Forecast results	
	purchase	Not purchased
purchase	TP:4344	FN:1656
Not purchased	FN:481	TN:2519

After digitizing the confusion matrix, the precision rate P, recall rate R and F1 values are obtained:

$$P = \frac{TP}{TP+FP} \tag{8}$$

$$R = \frac{TP}{TP+FN} \tag{9}$$

$$F1 = \frac{2*P*R}{P+R} \tag{10}$$

The accuracy of the model P is 0.724, the recall rate R is 0.9003, and F1 is 0.8025, which has high confidence.

5 Conclusion

Under the condition of big data, we can effectively grasp the pain points of consumers by analyzing the consumer's personal demand. The method of keyword extraction based on the combination of statistical theory and machine learning model makes keywords have not only data basis but also theoretical basis, which makes the confidence of keywords higher and more convincing. From the perspective of consumers, this paper constructs a prediction model of purchasing behavior, which can promote and guide the development of new energy vehicle enterprises. Therefore, according to the research, the following suggestions are put forward for enterprises and policies:

• Improving the comfort level, space design and endurance of new energy vehicles can greatly eliminate consumers' concerns and improve their purchase decisions.

• After sales service and a series of new energy vehicle postpartum problems should be further arranged in place to strengthen the after-sales problem solving of consumers and improve the probability of repeat customers.

• The state should solve the construction problem of basic energy storage equipment such as charging pile of new energy vehicles. Reasonable energy storage infrastructure will promote consumers to make purchase decisions

Acknowledgment

This work is supported by the National Social Science Foundation of China (No. 19BSH105), the Fundamental Research Funds for the Central Universities (WUT: 2020VI028), and National innovation and entrepreneurship training program for college students S202010497038 and 202010497018

References

1. Han Minqi. Mining selection *uncertainty in consumer behavior* [J]. Journal of Chongqing University of science and technology: Social Science Edition, 2017 (06): 39-43 + 51
2. Hu Dongbo, Xiao Xuan, Zhou Jin. *Analysis of mobile e-commerce user group characteristics based on data mining* [J]. Science and technology management research, 2013: 222-226
3. Liu X,Li J． Using *support vector machine for online purchase predication*［C］//2016 International Conference on Logistics,Informatics and Service Sciences(LISS) IEEE,2016: 1－6
4. Silahtarogu G, Donertasli H. *Analysis and Prediction of E- Customers'Behavior by Mining Clickstream Data*[C]. IEEE International Conference on Big Data. Santa Clara, CA, USA: IEEE, 2015.