Fuzzy Systems and Data Mining VII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210173

Do Reviews Influence Real Estate Marketing: The Experience Combing with Natural Language Processing

Ting WU[a], Guang YU[a, 1], Tong LI[a], Dan SHANG[b] and Wuwu YAN[c]
[a] *School of Management, Harbin Institute of Technology, P.R.China*
[b] *Heilongjiang Eldath Educational Information Consulting Co., Ltd, P.R.China*
[c] *Pang An property & casualty insurance company of China. LTD, P.R.China*

Abstract. In the Internet era, since online user reviews play an important role in various fields, various industries including real estate industry attach great importance to that. However, according to the existing literatures, there is no clear conclusion that whether online user reviews have an impact on real estate marketing. In order to figure out this problem, this paper will combine traditional real estate theories and machine learning technology to mine data on Chinese real estate online user reviews. We use natural language processing technology and panel data regression analysis method to explore whether the emotional tendency of online user reviews have an impact on the price of second-hand housing in real estate companies, and research more deeply about its impact on marketing of real estate companies. Our research provides a reference for real estate companies to make effective marketing strategies.

Keywords. Estate marketing, Reviews, Sentiment analysis, Word2vec, Natural Language Processing

1. Introduction

In the Internet era, online user reviews play a pivotal role in product marketing. Especially for enterprises, online user reviews can not only influence on the judgment of customers but also change their desire of purchasing. At the same time, they greatly influence on corporate brand image and provide companies with opportunities for better business activities. Therefore, online user reviews are widely valued by companies.

There are many existing researches on online user reviews in recent years. In the field of marketing, online user reviews can boost sales of fashion shopping, electronics and video games [1; 2], and help companies identify consumer needs and adjust marketing strategies to increase their competitiveness [3; 4]. In addition, digital marketing strategies affect both the quantity and value of online reviews and indirectly impact on hotel performance [5]. Online user reviews can also help customers make reasonable shopping decisions and reduce the risk of shopping [6]. On the electronic word-of-mouth, researching online user reviews can help companies identify potential electronic word-of-mouth customers to improve product sales [7]. On the other hand,

[1] Corresponding author, Guang YU, School of Management, Harbin Institute of Technology, 92 Xidazhi Street, Nangang District, Harbin, P.R.China; E-mail: yug@hit.edu.cn.

reveal the mechanism that affects companies' reputation and provide references for companies to change their marketing strategies[8; 9]. In forecasting and recommendation system, Tingting Song, Siering M, Bakhshi S and others[10-12] have shown that online user reviews can assistant to improve the performance of the movie box, airline and catering industry's forecasting and recommendation systems. These studies have fully demonstrated that user reviews play different roles in various fields.

Due to this, online user reviews are also valued by real estate marketers. In order to promote real estate marketing by that, major real estate companies set up homepages, community forums or user reviews windows on different platforms (as shown in Figure1.), and spend human labour and material resources to maintain and control them. However, there are few related literatures available about whether online user reviews affect real estate sales. Therefore, there is great uncertainty for real estate marketing strategies by online user reviews.

In order to solve the above problems, this paper intends to analyze the impact of online user reviews on real estate marketing. We add other online user reviews factors based on traditional real estate marketing theory[13-15]. The research data are from three Real estate related websites. By web crawler technology, we collect and figure out sales data, online user evaluation data and reviews about 12 real estate companies in around 10 months. And then we start our research by combining with natural language processing technology and panel data regression analysis method.

The research uses natural language processing technology to calculate the text data. Initially, through web crawler technology, relevant data is accessed online. Based on PyCharm development environment, the structure of the unstructured raw data obtained by web crawler is set up. And real estate companies' related data and user reviews data are extracted and stored in a table format. Secondly, we use Python language for data preprocessing, which include data cleaning, using the jieba lexicon to perform word segmentation processing on online Chinese text-data user reviews, then the user reviews dataset is formed. Thirdly, based on the skip-gram model from word2vec, a word vector calculation tool launched by Google in 2013[16; 17], the user reviews dataset is vectorized and the user review dictionary is constructed.

Finally, based on GitHub open source of Random Forest algorithm[18; 19], which uses the decision tree to vote together to determine the advantages of classification results, we analyze online user reviews data of sentiment orientation, and obtain the quantitative representation of user reviews data.

The research combines with real estate marketing theory[13-15] and panel data regression analysis method. First of all, we establish a regression model about the impact of online user reviews on real estate sales. We define the explained variables and the explanatory variables based on real estate marketing theory. Next, performing unit root test and cointegration test ensure the stability of the variables and avoid the occurrence of pseudo-regression. Finally, by using the individual fixed-effects model for analysis, this research establishes the models, which are the models for analysis in terms of both two explained variables, according to adding user reviews or not.

The purpose of this research is to reveal the relationship between online user reviews and real estate marketing.

The structure of this paper is as follows: the second part describes the methodology, including data collection and preprocessing, user reviews sentiment analysis and exploratory analysis; the third part shows the results of exploratory analysis; and the fourth part summarizes the research briefly.

Figure 1. Online user reviews windows

2. Methodology

Real estate marketing is mainly product-oriented because it has the characteristics of fixedness, single-piece, large amount of value and high unit price. For example, real estate marketing based on 4P theory mainly focuses on product, price, place and promotion[15]. With the changes in the marketing environment, real estate marketing also considers the needs of consumers. But real estate prices which include second-hand housing prices are still an important factor in marketing. The factors affecting real estate prices are mainly divided into three parts, that are general factors, regional factors and individual factors. Among them, the general factors mainly include economic factors, social factors, administrative factors, and so on; regional factors mainly include the prosperity of business services, traffic convenience, urban facilities, environmental conditions, etc.; and individual factors mainly include land and buildings. Among the above three parts of factors, regional factors are more concerned by consumers, so the impact on the price of housing, especially the price of second-hand housing is more obvious.

Based on the above theory, this paper selects second-hand housing as the explained variables for considering that the new housing price is relatively stable, and the second-hand housing price fluctuation range is large. Specifically, that are the difference between the price of second-hand housing and the price of the district, and the difference between the price of second-hand housing and the price of real estate. In addition, when setting explanatory variables, this paper considers the regional factors of housing prices, and adds the impact of consumer reviews, which would help to explore the relationship between user reviews and real estate marketing. Specifically, the explanatory variables include the online user evaluation dimension and online user reviews, that are price, district, transportation, support, environment and online reviews.

2.1. Data collection and preprocessing

2.1.1. Data collection

The data in this paper comes from Chongqing Online Real Estate, National Real Estate Market Data Center and Fangtianxia platform. Chongqing Online Real Estate has official real estate authority data. National Real Estate Market Data Center has a real estate

market database covering the whole China. Fangtianxia platform is the leading professional network platform for real estate industry and provides services such as owners' forums and community websites for all real estate projects.

Web crawler is used to collect online user related information, user evaluation data and user review data of real estates from January 1, 2018 to October 31, 2018 from the above network platforms, which includes real estate name, reviews time, user ratings of various factors (including price, district, transportation, support and environment) and user reviews. And then collect second-hand housing prices from real estates from Chongqing online real estate. A total of 12 real estates and 973 data were collected.

2.1.2. Data preprocessing

Data preprocessing flowchart is shown in Figure 2.

First of all, as the original data obtained by web crawler is unstructured data, which contains irrelevant information for the research, the information related to the research needs to be extracted from the original data and stored for the form data available. Based on Pycharm development environment, we use Python programming language to extract real estates' name, user name, user reviews time, user evaluation data (including price, district, transportation, support and environment), user reviews and real estate second-hand housing prices from the original data. After being extracted, these data are saved as a tabular form.

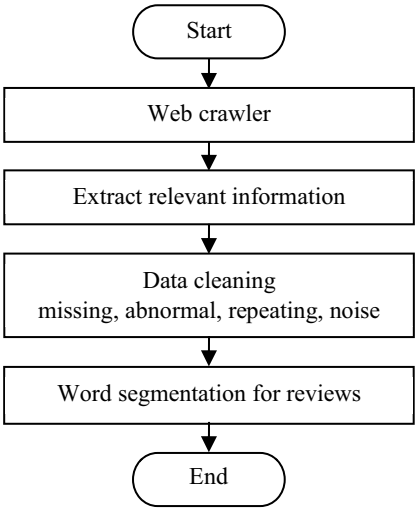

Figure 2. Data preprocessing flowchart

Next, we clean the data in tabular form by using python programming language, which includes missing values, outliers, duplicate values, and noise data. The first step is to delete or fill the missing values and the outliers; the next is to combine the repeated values by sorting and data similarity calculation; and the last step is to use the regression method to smooth the noise data. After cleaning, there are 263 data for 6 real estate companies available.

At the end of data processing, the text of user reviews is processed by using jieba lexicon of the GitHub open-source community, which is a third-party library of Chinese word segmentation with excellent performance, and uses a Chinese vocabulary to determine the probability of association between Chinese characters.

Through jieba lexicon, Chinese characters with high probability would form a phrase and build up the result of word segmentation[20]. Jieba lexicon participle supports three modes of word segmentation, namely precise mode, full mode and search engine mode. In precise mode, text can be accurately separated and there are no redundant words, in which it is suitable for text analysis; while in full mode, all possible words in the text can be scanned with excellent computing performance, but there is redundant and it cannot solve the ambiguity problem. Based on precise mode, search engine mode split the long words again to improve the recall rate, which is suitable for search engine segmentation. According to the characteristics of each mode and the applicable scenarios, our research selects precise mode for user reviews segmentation processing and obtains a list of user reviews data.

2.2. Comment sentiment analysis

2.2.1. Comment data dictionary construction based on word2vec

Before conducting sentiment analysis of the reviews data, the reviews text needs to be converted into vectors by using skip-gram model in word2vec. Word2vec is a natural language processing tool launched by Google in 2013, and skip-gram model is an effective method to learn high-quality distributed word vector representation[16; 17], which is able to capture a large number of precise syntactic and semantic word relationships. The specific equation is expressed as follows: for a given set of training words, w_1, w_2, …, w_T, the purpose of skip-gram model is to maximize the average log probability, which is shown in Eq. (1).

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c<j\le c,\, j\neq 0} logp(w_{t+j} \mid w_t) \tag{1}$$

Where c is the size of the training context. While the larger c is, the larger the training set, which means higher training accuracy. And $p(w_{t+j} \mid w_t)$ is defined by a softmax function shown in Eq. (2).

$$p(w_O \mid w_I) = \frac{exp(v'^{T}_{w_O} v_{w_I})}{\sum_{w=1}^{W} exp(v'^{T}_{w} v_{w_I})} \tag{2}$$

Where v_w and v'_w respectively represents the input and output vector representations of, and W is the number of words in the vocabulary.

Due to the high computational complexity of $\nabla \log p\,(w_O|w_I)$, it is really inefficient to train the model directly using the basic skip-gram formulation. Therefore, word2vec also provides two efficient training methods (including negative sampling and hierarchical softmax) to improve training speed and quality of skip-gram model. By using the computationally efficient skip-gram model here, high quality word and phrase representations would be learned in this section. In order to prepare for subsequent sentiment analysis of reviews, this section uses word2vec from GitHu b's open source to perform a vector representation for reviews dataset.

2.2.2. Sentiment analysis

In this section, Random Forest algorithm is used to classify the sentimental tendency of user reviews data. Random Forest algorithm is a supervised learning algorithm based on Bagging integration that is built by the based learning device of decision tree, and further introduces random attribute selection in the decision tree training process[18; 19]. Different from traditional decision tree which selects an optimal attribute in the attribute set of the current node when selecting the partition attribute, Random Forest algorithm selects the randomly selected feature to construct the optimal segmentation, that is, Random Forest algorithm only considers the random subset used to segment the nodes, and then selects an optimal attribute from this subset for partitioning. The schematic diagram of Random Forest algorithm structure is shown in Figure 3.

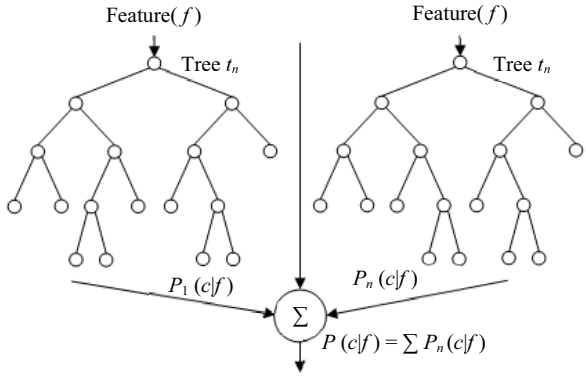

Figure 3. Schematic diagram of Random Forest structure

Random Forest consist of multiple CART (Classification and Regression Trees) and each decision tree is independent of each other. When a sample is input, each decision tree will get a result. The final result will be voted, and the one with the most votes will be the output result. Random forest has two random sampling processes, namely row and column sampling:

(1) In row sampling, the number of random sampling samples is the same as the number of input samples, and sampling with replacement is used. In this way, the input samples of each tree are not all samples, which can prevent the occurrence of over-fitting.

(2) Column sampling is randomly sampled without replacement according to a certain proportion. Assuming that there are M features and the number of samples is m, where $m << M$, the ratio can be $\sqrt{M}$, $1/2*\sqrt{M}$ *or* $2*\sqrt{M}$.

The decision tree is built from the sampled data. A certain node of the decision tree either cannot continue to split, or all its samples point to one category. Row sampling and column sampling together ensure randomness. When the number of layers is low, even without pruning, over-fitting will not occur.

The training process of Random Forest is as follows:

(1) Given training set S, test set T and feature dimension F. Determine the parameters, that is, the number of CARTs used t, the depth d of each tree, the number of features f used by each node, and the termination conditions (including the minimum number of samples on the node s, and the minimum information gain on the node m).

For the 1-t tree, i=1-t,

(2) From *S*, there is a training set *S(i)* with the same size as *S*, randomly selected as the sample of the root node, and training is started from the root node.

(3) If the termination condition is reached on the current node, set the current node as a leaf node.

1) If it is a classification problem, the predicted output of the leaf node is the largest type *c*(*j*) in the current node sample set, and the probability p is the proportion of *c*(*j*) in the current sample set.

2) If it is a regression problem, the prediction output is the average value of each sample value of the current node sample set. Then continue to train other nodes.

If the current node does not meet the termination condition, f-dimensional features are randomly selected from the F-dimensional features without replacement. Use this f-dimensional feature to find the one-dimensional feature *k* with the best classification effect and its threshold *th*. The samples with the *k-th* dimension of the sample on the current node less than th are divided into the left node, and the rest are divided into the right node. Continue to train other nodes.

(4) Repeat (2) and (3) until all nodes are trained or marked as leaf nodes.

(5) Repeat (2), (3), (4) until all CARTs have been trained.

The prediction process of random forest is as follows:

For the 1-*t* tree, *i*=1-*t*,

(1) Starting from the root node of the current tree, judge whether to enter the left node (<*th*) or the right node (>=*th*) according to the threshold *th* of the current node, until reaching a certain leaf node, and output the predicted value.

(2) Repeat (1) until all t trees have output predicted values.

1) For a classification problem, the result is the class with the largest sum of predicted probabilities in all trees, that is, the *p* of each *c*(*j*) is accumulated.

2) For a regression problem, the result is the average of the outputs of all trees.

This process produces a wide range of versatility and enables better performance models. Random Forest algorithm exhibit powerful performance in many real-world tasks because of the advantages of simple, easy to implement, lower computational cost and good generalization performance. We classify user reviews data through Random Forest algorithm from GitHub's open source. The labels of sentiment analysis were independently annotated by three researchers, and finally they were collated with each other. Data that cannot be agreed upon will be eliminated, so we can guarantee the accuracy of the proofreading data. We use the labeled data as training data, put it into the model for training, and make predictions on the unlabeled data. In the end, we used the 10-fold cross-validation to valid, and the accuracy of Random Forest algorithm is 63.7, so we finish scoring the emotional tendencies of user reviews data. The accuracy of the prediction result is the problem, and finally the error is introduced into the regression model. Through analysis, it is found that the correct examples of predictions have sentence sentiment consistency and sentiment clarity, while wrong examples are manifested as emotional inconsistencies, such as inconsistent evaluations of different aspects in a sentence, or even large differences.

2.3. Exploratory analysis

Further, we use panel data regression analysis to research the impact of online user reviews on real estate sales[21-23]. We define two explained variables as the difference between the price of second-hand housing and the price of the district, and the difference between the price of second-hand housing and the price of real estate. And the explanatory variables are the evaluation dimension and online reviews of the online user for the real estate that includes price, district, transportation, support, environment and online reviews.

Before starting panel data regression analysis, the unit root test and the cointegration test are needed to ensure the stability of the variables, avoid the occurrence of pseudo-regression and ensure the validity of the subsequent analysis results. The unit root test results of panel data regression analysis show that the variables are stable in all cases and are single-order in the same order. Since the purpose of the cointegration test is to investigate whether a linear combination of a set of non-stationary sequences has a cointegration relationship, it's unnecessary to perform the cointegration test on each variable.

This research will use individual fixed-effect models to analyze. There're four models to research the impact of online user reviews on real estate sales. For two explained variables (the difference between the price of second-hand housing and the price of the district, and the difference between the price of second-hand housing and the price of real estate), considering whether the factor of user reviews is taken into account, we establish regression models as Eqs. (3) and (4):

$$PD_1 = \alpha_{01} + \alpha_{11}COM_1 + \alpha_{21}PRI_1 + \alpha_{31}DIS_1 + \alpha_{41}TRA_1 + \alpha_{51}SUP_1 + \alpha_{61}ENV_1 + \varepsilon_1 \quad (3)$$

$$PD_2 = \alpha_{02} + \alpha_{12}COM_2 + \alpha_{22}PRI_2 + \alpha_{32}DIS_2 + \alpha_{42}TRA_2 + \alpha_{52}SUP_2 + \alpha_{62}ENV_2 + \varepsilon_2 \quad (4)$$

Where PD_1 is the difference between the price of second-hand housing and the price of the district, PD_2 is the price of second-hand housing and difference between real estate prices, and *COM*, *PRI*, *DIS*, *TRA*, *SUP*, and *ENV* respectively represents user reviews, user evaluation about price, district, transportation, support and environment.

3. Results

For the difference between the price of second-hand housing and the price of the district, Model 1 is a model in which user reviews isn't taken into account while Model 2 is; and for the difference between the price of second-hand housing and the price of real estate, Model 3 is a model in which user reviews isn't taken into account while Model 4 is.

Table. 1 shows the results of panel data regression analysis. Among them, the units of *PD1* and *PD2* are *RMB* yuan, and the other variables are scores (1-5 points). The coefficient of regression analysis is determined by R-squared in panel analysis. Tabel.1 shows that the R^2 of each model is greater than 70%, indicating that each model has a good fitting effect on the sample data. This shows the effectiveness of the model.

Table 1. Results of panel data regression analysis

	Model 1 (PD_1, no COM)	Model 2 (PD_1, contained COM)	Model 3 (PD_2, no COM)	Model 4 (PD_2, contained COM)
COM	-	281.4906	-	-1094.152
	-	(0.124410)	-	(-0.345290)
PRI	-5017.377***	-4962.971***	-7937.725***	-8149.199***
	(-3.431980)	(-3.215353)	(-3.871720)	(-3.769799)
DIS	4217.333**	4148.147**	7462.432***	7731.358***
	(2.359367)	(2.190837)	(2.976994)	(2.915602)
TRA	127.6533	88.05184	493.6364	647.5670
	(0.104681)	(0.069054)	(0.288657)	(0.362620)
SUP	-1048.875	-1086.942	-324.6592	-176.6957
	(-0.717813)	(-0.719430)	(-0.158436)	(-0.083507)
ENV	-815.5719	-800.0388	234.4921	174.1153
	(-0.523779)	(-0.505857)	(0.107388)	(0.078609)
R-squared	0.868042	0.868094	0.703902	0.704804
Adjusted R-squared	0.805362	0.800450	0.563255	0.553422
F-statistic	13.84875	12.83327	5.004758	4.655790

Note: ** indicates that the 0.05 level is significant; *** indicates that the 0.01 level is significant; the value in parentheses is the estimated T statistic; '-' indicates that the corresponding explanatory variable is not included in the model.

Explanatory variable of user's evaluation of price (*PRI*) is significant for the difference between the price of second-hand housing and the price of the district (PD_1), and the difference between the price of second-hand housing and the price of real estate (PD_2) at the 0.01 level. It shows that the impact of *PRI* on *PD1* and *PD2* is very significant and shows a negative correlation. When *PRI* decreases, *PD1* and *PD2* will increase tremendously. Explanatory variable of user's evaluation of district (*DIS*) is significant for the difference between the price of second-hand housing and the price of the district (PD_1) at the 0.05 level, while it's significant for the difference between the price of second-hand housing and the price of real estate (PD_2) at the 0.01 level. It shows that DIS has a significant impact on PD1 and PD2, and shows a positive correlation. That is, the larger the DIS, the larger the PD1 and PD2. But other explanatory variables, including online reviews (*COM*), online user's evaluation of transportation (*TRA*), support (*SUP*) and environment (*ENV*) aren't significant for both PD_1 and PD_2. That means *COM, TRA, SUP* and *ENV* have no effect on PD_1 and PD_2.

The results show that the emotional tendency of user reviews does not constitute an influencing factor of the difference between the price of second-hand housing and the price of the district (PD_1) and the price of second-hand housing and the price of real estate (PD_2).

4. Conclusion

Online user reviews have an increasing impact on various fields, especially in the field of marketing, which not only affect companies' reputation and marketing strategies, improve corporate recommendation and predicts system performance, but also help customers make consumption decisions and reduce consumption risks. In view of their huge influence, real estate marketers have been paying increasing attention to online user reviews and spending human labour and material resources to maintain and control them. However, according to the existing literatures, there is no clear conclusion that whether online user reviews affect real estate sales. And it brings great uncertainty to formulate real estate marketing strategies. Therefore, this paper explains the relationship between online user reviews and real estate sales, and provides a reference for real estate companies to make effective marketing strategies.

This research combines traditional real estate theories with natural language processing technology. Wetake online user reviews into account and analyze the impact of online user reviews on real estate marketing. The research firstly collects online user reviews data of real estate companies by using natural language processing technology, and then conducts sentiment analysis of online user reviews based on word2vec and Random Forest algorithm. Secondly, we adopt panel data regression analysis method to analyze the relationship between the emotional tendency of online user reviews and second-hand housing prices.

The results show that the emotional tendency of online user reviews does not affect the price of second-hand housing. It means online user reviews have no impact on the real estate marketing. Therefore, the real estate companies' marketing strategies based on online user reviews are worthless. Real estate companies can adjust marketing strategies based on the results of the research to improve marketing effectiveness.

When online user reviews are taken into account during analyzing real estate marketing strategies, traditional real estate marketing theory is improved. And that provides new theoretical ideas for subsequent research. At the same time, we find that natural language processing technology and real estate marketing strategy analysis can be combined. The research methods of real estate marketing strategy are enriched by using natural language processing technology to mine potential marketing influence factors.

While we use word2vec and Random Forest algorithm to conduct sentiment analysis of online reviews, there are two models, BERT and ERNIE, launched some time ago. BERT was launched by Google in 2018, and it achieved the best results than other models in different natural language processing tasks. And ERNIE, Baidu's NLP pre-training model, was released in 2019, and surpassed BERT in several Chinese tasks. Future research can expand the research dataset, and use BERT and ERNIE for sentiment analysis to explore the impact of online reviews on marketing. That is predicted to enrich the analysis methods of marketing strategies.

References

[1] Kawaf. F, Istanbulluoglu. D. Online fashion shopping paradox: The role of customer reviews and facebook marketing[J]. Journal of Retailing and Consumer Services, 2019, 48: 144-153.

[2] Cui. G, Lui. H.-K, Guo. X. The effect of online consumer reviews on new product sales[J]. International Journal of Electronic Commerce, 2012, 17(1): 39-58.

[3] Timoshenko. A, Hauser. J. R. Identifying customer needs from user-generated content[J]. Marketing Science, 2019, 38(1): 1-20.
[4] Cezar. A, güt. H. Analyzing conversion rates in online hotel booking: The role of customer reviews, recommendations and rank order in search listings[J]. International Journal of Contemporary Hospitality Management, 2016, 28(2): 286-304.
[5] De Pelsmacker. P, Van Tilburg. S, Holthof. C. Digital marketing strategies, online reviews and hotel performance[J]. International Journal of Hospitality Management, 2018, 72: 47-55.
[6] Hong. H, Xu. D, Wang. G. A, Fan. W. Understanding the determinants of online review helpfulness: A meta-analytic investigation[J]. Decision Support Systems, 2017, 102: 1-11.
[7] Zhao. P, Wu. J, Hua. Z, Fang. S. Finding ewom customers from customer reviews[J]. Industrial Management & Data Systems, 2019, 119(1): 129-147.
[8] Zhang. Z, Ye. Q, Law. R, Li. Y. The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews[J]. International Journal of Hospitality Management, 2010, 29(4): 694-700.
[9] Park. S, Nicolau. J. L. Asymmetric effects of online consumer reviews[J]. Annals of Tourism Research, 2015, 50: 67-83.
[10] Song. T, Huang. J, Tan. Y, Yu. Y. Using user-and marketer-generated content for box office revenue prediction: Differences between microblogging and third-party platforms[J]. Information Systems Research, 2019.
[11] Siering. M, Deokar. A. V, Janze. C. Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews[J]. Decision Support Systems, 2018, 107: 52-63.
[12] Bakhshi. S, Kanuparthy. P, Gilbert. E. Demographics, weather and online reviews: A study of restaurant recommendations[C]. Proceedings of the 23rd international conference on World wide web, 2014: 443-454.
[13] Wu. Z. H, Liu. C. B, Jin. H. Y. The analysis of the main influence factors of real estate price in shenzhen[M]. 2006: 1534-1538.
[14] Wu. J, Deng. Y, Liu. H. House price index construction in the nascent housing market: The case of china[J]. The Journal of Real Estate Finance and Economics, 2014, 48(3): 522-545.
[15] Chen. J, Han. W. The research on the marketing strategies theory and empirical based on the product value[C]. 2009 International Symposium on Marketing Management (ISMM 2009)-Marketing Innovations and Economic Development, 2009: 286-290.
[16] Mikolov. T, Sutskever. I, Chen. K, Corrado. G. S, Dean. J. Distributed representations of words and phrases and their compositionality[C]. Advances in neural information processing systems, 2013: 3111-3119.
[17] Mikolov. T, Chen. K, Corrado. G, Dean. J. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[18] Breiman. L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.
[19] Svetnik. V, Liaw. A, Tong. C, Culberson. J. C, Sheridan. R. P, Feuston. B. P. Random forest: A classification and regression tool for compound classification and qsar modeling[J]. Journal of chemical information and computer sciences, 2003, 43(6): 1947-1958.
[20] Sun. J. 'Jieba'chinese word segmentation tool. 2012.
[21] Hsiao. C. Analysis of panel data[M]. Cambridge university press, 2014.
[22] Canay. I. A. A simple approach to quantile regression for panel data[J]. The Econometrics Journal, 2011, 14(3): 368-386.
[23] Choi. I. Unit root tests for panel data[J]. Journal of international money and Finance, 2001, 20(2): 249-272.