Advances in Social Science, Education and Humanities Research, volume 329
4th International Conference on Contemporary Education, Social Sciences and Humanities (ICCESSH 2019)

Classification Model of Online Review Credibility on E-commerce Platform

Zimo Li
School of Economics and Management
Beijing Jiaotong University
Beijing, China

***Abstract*—With the increasing popularity of e-commerce, the phenomenon of network ghostwriters brushing praise, merchants deliberately denigrating competitors, and users using comment templates to cope with the comment task is emerging in an endless stream. Starting from the credibility of online reviews, this paper studies the data characteristics of fake reviews, and constructs a logistic regression model from the four dimensions of the length of reviews, the similarity of reviews, the difference between the text emotional tendency and actual score, and the number of days between the confirming receipt and the release of comments, so as to classify the online reviews' credibility. The validity of the method is verified by using the review set of iPhone on Jingdong platform.**

Keywords—online review; credibility; logistic regression; sentiment analysis

I. INTRODUCTION

The rapid development of the Internet has not only changed people's lifestyle, social mode and thinking patterns, but also reconstructed the industrial form, social economy and even spawned emerging industries such as social media, e-commerce, Internet +, and Internet of things. With the popularity of the Internet, online shopping attracts many consumers and netizens because it is not limited by time, place, type, flexible payment method, no need to go to stores, low price and other advantages. According to the survey, as of June 2017, the number of Chinese Internet users had reached 751 million, of which 480 million users of shopping. Online shopping has become a shopping model that is increasingly favored by consumers today. Customers can not only browse and select products on the e-commerce platform, but also comment on the platform. As e-commerce capabilities continue to expand, online customer reviews are an important information resource for enterprises, consumers and researchers. For enterprises, mining customer reviews can not only draw user portraits provide personalized recommendations to different levels of customers, but also attract new customers, improve customer retention and customer loyalty. For consumers, they can refer to the comments of other users before purchase to further understand the characteristics of the product. After purchase, they can release their own feelings of use and summarize their own shopping experience to urge merchants to improve their shortcomings, and provide reference for other potential buyers. For researchers, mining and analyzing customer reviews is an important means to study the purchase decision-making process and influencing factors of consumers' purchasing decisions.

As merchants gradually realize the importance of users' online comments, a variety of fake comments began to flood. It is not uncommon for businesses to manipulate comments, hire Internet mercenaries to praise, and deliberately slander their competitors. At the same time, motivated by benefits, buyers use comment templates to cope with evaluations, publish comments unrelated to the products to improve user activity level, and over-praise the products to get "praise cash back" rewards from merchants are also numerous. These behaviors reduce the credibility and quality of online comment on e-commerce platforms. In the era of big data, the quantity of information increases exponentially, and the number of comments on popular commodities can reach hundreds of thousands, and the quality of online comments is uneven, resulting in consumers need to spend a lot of time and energy searching for comments and measuring credibility. In the long run, it will not only interfere with consumers' purchase decisions and reduce the shopping experience, but also lead to consumers' declining trust in online reviews on e-commerce platforms. The existence of fake comments is not conducive to the healthy development of e-commerce platforms. Therefore, it is of great significance for enterprises to classify the credibility of online reviews of e-commerce platforms and take corresponding measures to avoid the proliferation and spread of fake comments, so as to formulate marketing strategies and promote the healthy development of e-commerce platforms.

II. LITERATURE REVIEW

Credibility is a concept in the field of mass communication, also known as credibility perception. Comment credibility generally refers to the degree to which the comment information published by the commentator is recognized by others [1] or the degree to which the comment recipient unconditionally trusts the content expressed by the source comment [2].

A. Credibility Classification Model

Previous studies have usually transformed the classification of credibility into a dichotomous problem, i.e. dividing reviews into two categories: real reviews and fake reviews. Commonly used classifiers include decision tree, support vector machine, Bayesian, k-means and other models. Ren Yafeng [3] et al. optimized the selection of comments by genetic algorithm, and then used unsupervised hard and soft clustering to identify fake comments, and finally achieved a classification accuracy of 84.5%. Miao Yuqing et al. [4] solved the problem of unbalanced number of positive and negative samples based on SMOTE oversampling technology, and then identified fake comments by constructing random forest. Wang Fei [5] et al. used the Bayesian network model with hidden variables to study the dynamic behavior of customers on MovieLens, a film review website. Taking credibility as the criterion, Li Jing [1] et al. conducted an in-depth study on the characteristics of fake comments, and constructed a logistic regression model by integrating the characteristics of users, merchants and reviews. Experiments show that the AUC value of the classification model is 89.3%, with good accuracy.

Through comparison, it can be found that although the clustering method shows a good effect in identification, the setting of the threshold in the experiment is an important factor affecting the experimental results, and when the number of clusters is too large, it is difficult to reflect the comparison between fake comments and real comments, making the results not intuitive enough. Although the random forests model did better on the results, the number of trees in the forest is one of the important parameters affecting the performance of the model. When the number of trees is small, the error of the experimental results may be larger, on the contrary, when the number of trees is too large, the time cost of the model will also increase, and a lot of time and energy will be spent in the process of selecting the appropriate number of trees. Overall, the logistic regression model does not need to repeatedly debug the parameters, and its fitting results are more intuitive. It can also be seen that the influence of different indicators on classification results is strong or weak. Therefore, this paper will develop a credibility classification of online reviews based on logistic regression models.

B. Influencing Factors of Credibility

Through research, we can find that there are differences between fake reviews and real reviews in the form of expression. Fake reviews usually show the following characteristics: a) the product features and usage feelings covered by the comments are rare; b) the comments are irrelevant to the product; c) the length of the review is longer or shorter than normal; d) the emotional tendency of the comment content do not match the actual score; e) the comments are consistent or similar to others' comments.

Based on the above characteristics of fake reviews, the indicators used by researchers in the credibility classification are mainly divided into three perspectives: the reviewers, the merchants and the content. From the perspective of the commentator, Guo Guoqing [6] believes that the degree of disclosure of the individual user information of the commentator reflects the level of the commentator's identity, which in turn affects the credibility of the comment. Walczuch R [7] et al. believed that the age, cognition and cultural literacy of the commentator are the main influencing factors of the credibility of reviews. The time factor of comment release is one of the important indicators to measure credibility. Chen and Lurie [8] pointed out that the time interval of consumer comments is largely affected by experience; and the longer the time interval is, the more credibility it shows. In addition, in some studies, the time factor is defined as the time interval between now and publications to reflect the timeliness of the comments, but Chen and Tseng [9] pointed out that the timeliness of the comment is not an effective indicator of credibility classification. Therefore, the metrics selected in the reviewer's perspective are the time interval between receipt and publishing.

In summary, this paper focuses on the difference between the length of the comment and the overall mean, the difference between the comment sentiment and the actual score, and the similarity of the review text, combined with the intervals days in the commentator's perspective. The review content information is merged with the commentator information to construct a logistic regression model, and the effectiveness of the method is verified by experimental data sets.

III. ONLINE REVIEW CREDIBILITY CLASSIFICATION MODEL

The credibility classification method proposed in this paper preprocesses the online comment data crawled by the web crawler, and then scores the text content based on the sentiment dictionary and the user-defined dictionary, and calculates the difference between the text and the user's actual score. Secondly, the text similarity and text length of each comment are calculated. Finally, the classification model is constructed to realize the classification of the credibility of the comments, and the data is used to verify the validity of the method. The technical route of this paper is shown in "Fig. 1".

A. Data Preprocessing

In this paper, Python's Chinese word segmentation toolkit, Jieba were used to segment and deactivate online comments. It provides users with three word segmentation modes: full mode, accurate mode and search engine mode. In view of the research direction of this paper, we adopt the accurate model suitable for text analysis. Additionally, the Jieba segmentation system enables users to load user-defined dictionaries, and customize some colloquial expressions or nascent words that are common but difficult to recognize correctly, so as to improve the algorithm's recognition rate.

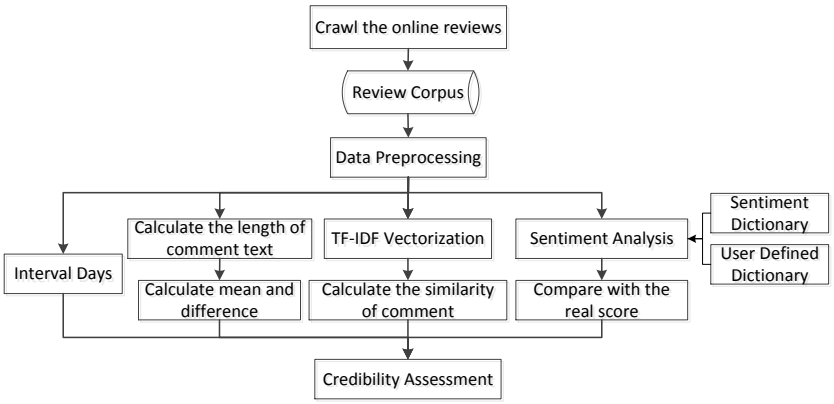

Fig. 1. Technical route of credibility classification model.

In daily expression, in order to improve the smoothness of sentences and accurately express the relationship between words and sentence, conjunctions such as "of", "and" are often used, as well as punctuation marks such as commas, periods, exclamation marks and question marks. However, the use of stop words not only occupies the space of storage, but also reduces the efficiency of text processing. Therefore, it is necessary to remove the stop word, further purify the comment content, and only retain the words that reflect the product characteristics and customer evaluation. After this step is completed, the comment corpus is obtained.

B. The Length of the Reviews

Generally speaking, the longer the comment is, the more product information it covers and the more comprehensive the user experience is. However, according to the research, in order to highlight the good characteristics or deliberately emphasize the disadvantages of the product, users tend to boast excessively or depreciate the product, so the evaluation of the product or service tends to exceed the standard level of ordinary comment length. The credibility of these comments is low. In addition, for the purpose of coping or getting rewards, some users only use short words to summarize the experience and usage feelings, and cannot provide opinions and suggestions for other users' purchase decisions. Therefore, the credibility of extremely short comments is also low. On the whole, when the length of comment text is within a certain range, the credibility of comment is higher. In this paper, the length of the comment after word processing is taken as one of the dimensions to measure the credibility of the comment. The expression of the character length of the commentary text in Article I is shown in (1).

$$\mathrm{Len}_i = \mathrm{review}_i.\mathrm{length} \qquad (1)$$

After obtaining the character length of each comment, the average text length of the entire comment corpus is calculated using (2). The difference between the comment and the total mean is calculated and recorded as Dis_i, as shown in (3).

$$\overline{\mathrm{Len}} = \sum_{i=1}^{n} \mathrm{Len}_i / n \qquad (2)$$

$$\mathrm{Dis}_i = \left| \mathrm{Len}_i - \overline{\mathrm{Len}} \right| \qquad (3)$$

C. Similarity of Comment Text

In this paper, texts are vectorized using TF-IDF before calculating text similarity. TF-IDF is a common weighted statistical method in text mining. TF (Term Frequency) refers to the frequency of a specific term in a document or text. The more times it appears, the greater the TF value for that term. In (4), TF_t refers to the word frequency of the term t, W_t refers to the number of times the term t appears in a document, and W refers to the number of all terms in the document. IDF (Inverse Document Frequency) is the inverse text frequency, referring to the frequency of a certain term in the entire corpus or all documents. In (5), IDF_t refers to the inverse text frequency of the term t, N refers to the total number of documents in the corpus, and N_t refers to the number of documents containing the term t. The denominator is added 1 to avoid it being 0. IDF is a criterion to measure the universal importance of a term, that is, the more times a term appears in all documents, the smaller its IDF value and the lower its importance. For example, conjunctions and prepositions such as “and”, “in”, and “is” may appear more frequently in a comment than the product feature, but if they appear in each comment, their IDF value will be lower, which proves that their importance is lower than the product feature. TF-IDF in (6) are weighted by word frequency and inverse text frequency to extract keywords and weigh the importance of keywords.

$$\mathrm{TF}_t = W_t / W \qquad (4)$$

$$\mathrm{IDF}_t = \log\left(\frac{N}{N_t + 1} \right) \qquad (5)$$

$$\text{TF-IDF}_t = \text{TF}_t * \text{IDF}_t \quad (6)$$

Firstly, all the words in the corpus are exported and duplicated to generate a comment dictionary. Then, all the words in the dictionary are sequentially compared with the words in each document, and all the documents are traversed to calculate the value of TF-IDF. Thus, all words in each document can be represented by their TF-IDF value, i.e. each comment can be mapped into vector space. Cosine similarity is the cosine value of the angle between two plane vectors, which reflects the similarity between the two vectors. The smaller the angle is, the greater the cosine will be, and the greater the similarity will be. The formula for calculating the cosine similarity is shown in (7).

$$\text{Similarity}(u,v) = \cos\left(\vec{u},\vec{v}\right) = \frac{\vec{u}\cdot\vec{v}}{\left\|\vec{u}\right\| \times \left\|\vec{v}\right\|} = \frac{\sum_{i=1}^{n}\left(x_i \times y_i\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(y_i\right)^2}} \quad (7)$$

D. Difference Between Emotional Tendency and Actual Score

Emotional polarity analysis is to study the emotions, subjective feelings and emotional tendencies expressed by users in comments. The research on emotional polarity divides emotional tendency into positive, neutral and negative. In this paper, the sentiment analysis method based on sentiment dictionary is adopted to match the emotional words in the comments with the dictionary, and the emotional score of each comment is calculated. "Table I" shows examples of different parts of speech and weights in the sentiment dictionary.

The obtained emotional score was divided into five grades and compared with the actual score given by the user. The difference between the emotional score of the text and the actual score of the user was calculated by (8), and recorded as E_i. Among them, $real\_score_i$ represents the actual score of the user of the Article I, and $text\_score_i$ denotes the emotional polarity score of the Article I.

$$E_i = \left|real\_score_i - text\_score_i\right| \quad (8)$$

TABLE I. EXAMPLES OF DIFFERENT PART OF SPEECH AND WEIGHTS OF THE SENTIMENT DICTIONARY

Part of Speech	Weights	Examples
adverb of degree	2	absolutely, exhaustively, extremely, thorough, too
	1.5	very, exceptional, quite, extraordinarily
	1.25	more, so, farther, especially
	0.5	more or less, slightly, a little, inevitable, some
inverse words	-1	no, not
exclamation mark	1	!
positive words	1	not bad, good, praise, worth, affordable, cheap, satisfied, like, love, nice, cost-effective
negative words	-2	poor, bad, defective, slow, angry, insufficient, garbage, general

E. Interval Days of the Review

Due to the difference in online purchasing behavior, users will give evaluation in different periods after confirming receipt. Upon confirmation, the user experience may only be subjective feelings of product appearance and supporting services. With the extension of use time, users may gradually find the advantages and disadvantages of products and services, such as standby time, operating speed and after-sales service. Therefore, it is speculated in this paper that the interval days between the date when the user confirms receipt of goods and gives evaluation may have some potential correlation with the credibility of the comment. In order to study the potential relationship, this paper also takes the interval days as one of the dimensions in the credibility model. The number of days in the dataset varies from 0 to 63 days, as shown in "Fig. 2". As can be seen from the figure, more than 60% of the comments were published within 0-7 days, and the number of comments decreases as the number of the intervals increases.

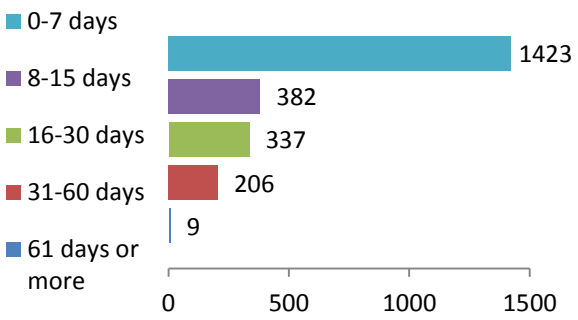

Fig. 1. Distribution of interval days.

F. Credibility Classification Model Construction

The credibility classification of online reviews can be regarded as a dichotomous problem, and logistic regression model is one of the commonly used methods in classification problems. In addition, it is one-sided to judge the credibility of online reviews only from the perspective of users or merchants. Therefore, the credibility classification model of this paper incorporates the perspective of commentator and comment content, which integrates the length of the review, the difference between the emotional score and the actual score, the similarity of comment text and the interval days between the confirmation and the evaluation, and uses the logistic regression model for fitting. By substituting the three variables into the Sigmoid function of the logistic regression model, the expression of the logistic regression model in this paper can be obtained, as shown in (9).

$$p(x)=\frac{1}{1+\exp\{-(c_0+c_1Dis_i+c_2Sim_i+c_3M_i+c_4Day_i)\}} \quad (9)$$

Dis_i refers to the difference between the text length of Article I comments and the average length of all comments. Sim_i refers to the text similarity between Article I and the corpus. M_i is the difference between the actual score and the emotional score of Article I. Day_i indicates the interval days of Article I.

IV. Experimental Results and Analysis

A. Data Set and Evaluation Indicators

In this paper, it crawled the reviews of iPhone XS from Jingdong mall by python's web crawler, and got 2358 pieces after deduplication. In addition to the textual content of the comments, each message also includes the user's publishing time, the interval days, and the overall rating. The rating ranges from 1 star to 5 stars, all of which are integers. Among the online reviews crawled, 2186 were highly credible comments and 172 were low. The experiment extracts 80% of the reviews as training set, and the remaining 20% as test set. The performance of the method is evaluated by recall, precision and F-score.

B. Contrast Experiments

According to the ratio of 8:2, the processed 2358 data were divided into training set and test set. The accuracy, recall rate and F-score were calculated, and the ROC curve was drawn. The ROC curve of the credibility classification model is shown in "Fig. 3". As can be seen from "Fig. 3", the model presents better results and accuracy on the experimental data set, with an accuracy rate of 94.33%, a recall rate of 99.08% and an F-score of 97.17%.

In order to analyze the influence of the four dimensions selected by the model on the experimental results, this paper removed one dimension in turn and compared the new model constituted by other dimensions with the original model. "Fig.4" compares the accuracy rate, recall rate and F-score of the new model and the original one.

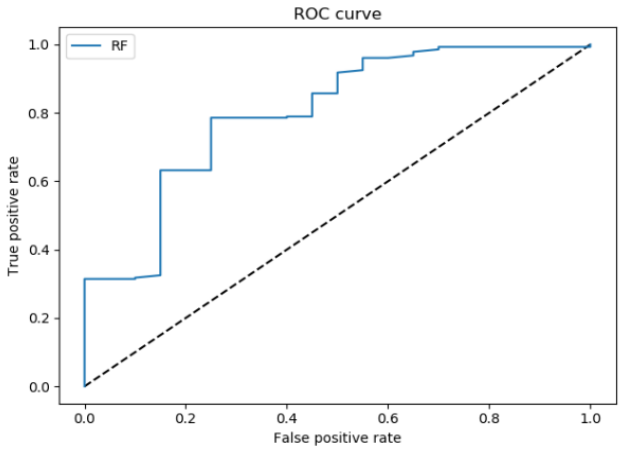

Fig. 2. Credibility classification model ROC curve.

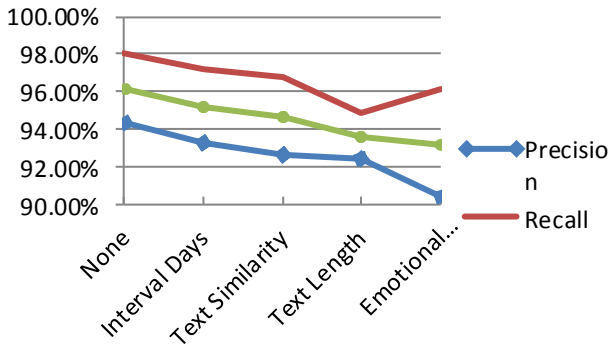

Fig. 3. Comparison of experiment results.

As can be seen from "Fig. 4", the new model constructed from the features of the remaining three dimensions is not as good as the original model in terms of accuracy, recall rate and F-score. When the emotional difference is removed, the accuracy decreases more than that of the other three characteristics, indicating that the emotional difference is the most important measure in the credibility classification. When the dimension of text length is removed, the recall and F-score fluctuate greatly, indicating that the length of comment text in the data set is uneven, and there is a large gap between ultra-long comment and short comment. When the interval days or text similarity dimensions is removed, the three evaluation indexes all decline to a certain extent, which shows that the interval days and text similarity are effective in judging the credibility of a comment. Through comparative experiments, it is proved that the credibility model proposed in this paper is effective, and the selected four dimensions can achieve better results than any other three dimensions.

C. Analysis

Through the credibility analysis of online comments identified by the model, it can be found that the comments with high credibility basically cover the description of product performance, experience and supporting services. For example, performance is reflected in the operating speed, battery, fingerprint identification, camera, screen proportion, etc. The experience is concentrated on the system, feel, display, sound quality and other aspects. Supporting services are logistics distribution, product packaging, after-sales service, etc. The less credible reviews are characterized by perfunctory tone and less descriptive content, e.g. product received, many gifts were given, praise. The users did not give a real product experience, so the review does not have high credibility.

"Fig. 5" shows the distribution of comment interval days under the two credibility types. As can be seen from "Fig.5", when the interval days are between 16-30 days, the credibility of the comment is the highest. When the interval days are too long or too short, its credibility will be reduced relatively, but not significant. This also confirms that the evaluation index fluctuated slightly but not significant when

the interval days dimension was removed in the comparison experiment. As shown in "Fig.6", when the credibility of the comment is low, the difference between the emotional polarity of the comment text and the actual score is greater than that of the comment with high credibility. This indicates that the more consistent the emotional tendency of the comment is with the actual score, the higher the credibility of the comment will be. In other words, the greater the difference between the emotional tendency reflected in the comment text and the actual score, the lower the credibility of the comment will be.

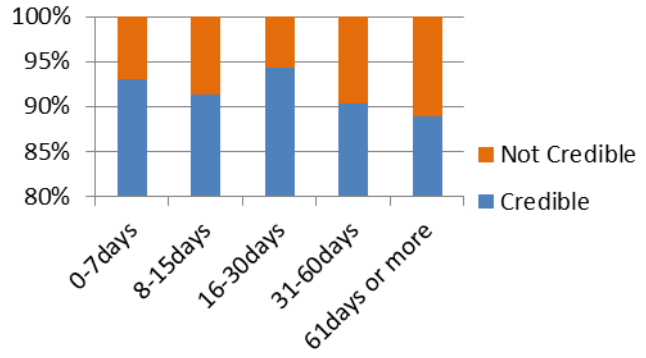

Fig. 4. The proportion of interval days under two types of credibility.

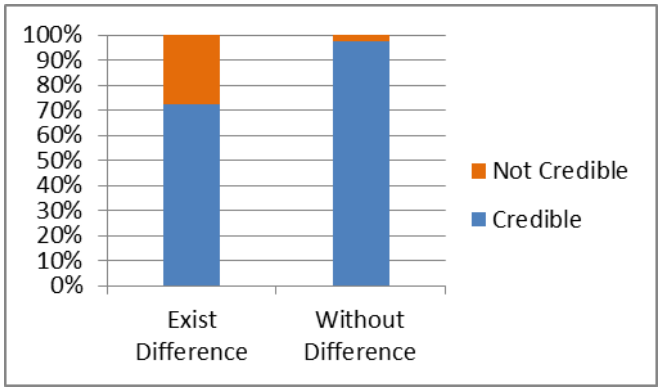

Fig. 5. Comparison of emotional difference under two types of credibility.

V. CONCLUSION

Online review on shopping platform is a valuable resource in the information age. A large number of fake reviews will confuse consumers' judgment and affect the purchase decision-making process. Therefore, how to measure the credibility of reviews and identify fake reviews is crucial. The method proposed in this paper incorporates the length of comment text, the difference between comment emotional tendency and actual score, comment similarity and the number of days between receipt confirmation and comment release time into the logistic regression model for credibility classification. The empirical model achieved good results with an accuracy rate of 94.33%.

The credibility classification indicators selected in this paper are based on the characteristics presented by fake reviews, aiming to replace the complicated classification indicators with a few representative indicators, so as to avoid the absence of features. However, in practical application, the judgment of online review credibility may involve many factors and have a certain degree of subjectivity. In addition, there are fewer data sets available for review credibility studies, and the results may vary when the sample size is large.

REFERENCES

[1] Li Jing, Wu Guoshi, Xie Fei, Yao Xu, Qi Jiayin, and Sun Pengfei, "Research on fraud review detection model on O2O platform," ACTA ELECTRONICA SINICA, Beijing. China, vol. 44, 2016, pp. 2855-2860.

[2] Sun Shuying, "An empirical study of the influential factors for the information credibility of online consumers," Journal of Beijing institute of technology (social sciences edition), Beijing. China, vol. 10, December 2008, pp. 50-54.

[3] Ren Yafeng, Yin Lan, and Ji Donghong, "Deceptive reviews detection based on language structure and sentiment polarity," Journal of Frontiers of Computer Science and Technology, Wuhan. China, 2014, 8(3), pp. 313-320.

[4] Miao Yuqing, Qu Weijian, Liu Tonglai, Liu Shuiqing, and Wen Yimin, Detection of fake reviews based on sentiment polarity and over-sampling, Guangxi. China, vol. 35, July 2018, pp. 2042-2045.

[5] Wang Fei, Yue Kun, Sun Zhengbao, Wu Hao, and Feng Hui, "Analyzing rating data and modeling dynamic behaviors of users based on the Bayesian network," Journal of Computer Research and Development, Yunnan. China, vol. 54, 2017, pp. 1488-1499.

[6] Guo Guoqing, Chen Kai, and He Fei, "An empirical study on the influence of perceived credibility of online consumer reviews," Contemporary economy & management, Beijing.China, vol. 32,October 2010, pp. 17-23.

[7] Walczuch. R, and H. Lundgren, "Psychological antecedents of institution-based consumer trust in e-retailing," Information & Management, vol. 42, 2004, pp. 159-177.

[8] Chen. Z, and N. H. Lurie, "Temporal contiguity and negativity bias in the impact of online word of mouth," Journal of Marketing Research, vol. 50, 2013, pp. 463-476.

[9] C.C. Chen, and Y. Tseng, "Quality evaluation of product reviews using an information quality framework," Decision Support Systems, vol. 50, 2011, pp. 755-768.

[10] Gong Si-lan, Ding Shengchun, Zhou Xiawei, and Chao Naipeng, "An empirical research of online commodity reviews information credibility factors," Journal of intelligence, Nanjing. China, vol. 32, November 2013, pp. 202-207+180.

[11] Mudambi. S, & Schuff. D, "Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," Social Science Electronic Publishing, vol. 34, March 2010, pp. 185-200.

[12] Wang Zhuo, Li Zhun, Xu Ye, and Song Kai, "Detecting product review spammers based on review graphs," Computer Science, Shenyang. China, vol. 41, October 2014, pp. 295-299+305.

[13] Jensen. M. L. , Averbeck. J. M. , Zhang Z. , and Wright. K. B. , "Credibility of anonymous online product reviews: A language expectancy perspective," Journal of Management Information Systems, vol. 30, 2013, pp. 293-324.

[14] Li Chao, Xiang Jing, Xiang Jun, "Assessment method of credibility on online product reviews," Journal of Computer Applications, Hubei. China, vol. 39, 2019, pp. 181-185.