Online Comment Text Analysis with Improved Feature Weight

Chaoju Hu*, Xiaojie Yang

Department of computer, North China Electric Power University, Baoding 071000, China
**Corresponding author e-mail: y2296665134@163.com*

ABSTRACT. *In the product reviews of online shopping platforms, the star rating and text comments given by the same user often appear. These data are processed by the unreasonable scoring system to give the merchants a false rating and mislead consumers. In order to improve the scoring system, an improved feature weighting method combining text review content and star rating is proposed. Firstly, the weighting rules were defined. Secondly, according to the part-of-speech evaluation function in the short text, the feature word selection was combined with the star rating and the text comment content. The CBOW model was used to train the word vector. Finally, the text classification method SVM was used to obtain the final result. This model was applied to two text datasets and compared with the traditional TF-IDF feature selection. The results show that the algorithm can effectively improve the accuracy and F1 value of text classification.*

KEYWORDS: *online user review, SVM, short text classification, sentiment analysis*

1. Introduction

With the advancement of technology, online shopping has gradually become a part of people's lives. In the process of selecting goods, users will not only consider the value of the goods themselves, but also the favorable comment rate of the products will be included in the evaluation scope as the basis for purchasing or not [1]. At present, the favorable comment rate of goods is often determined by the user's star rating. For those who give five-star reviews but make bad reviews on the service or product quality in the text comments, they will not lower the favorable rate. Such a scoring system ignores the influence of short text comments on the favorable comment rate of goods and stores, and the judgment results are one-sided and misleading, and there are certain defects [2].

In order to improve the online shopping scoring system and provide consumers with accurate value judgment basis, some scholars have conducted a tendency analysis on the content of text comments, and based on the results of the analysis of the content of the comments, combined with the evaluation of star ratings, the favorable comment rate [3-4]. Some scholars have introduced methods of embedding words to improve the similarity of texts to analyze the emotional orientation of texts. Although trend analysis can improve the text evaluation system, the flaws in the method of extracting feature values cannot be ignored. In the process of classifying short texts, extracting feature values is an important data preprocessing process, mainly using feature words and feature weights to represent text data. Its shortcomings are also obvious. It ignores the contribution of part of speech and meaning in semantics, which leads to problems such as feature redundancy, feature sparsity and dimensional disaster. In order to solve this problem, Yuan et al. [5] studied the semantic relevance of texts and extracted the high-order textual and semantic features of the comments, thus improving the text classification effect; Zhang Zhenhao et al. [6] studied the similarity of text keywords. And propose a text classification framework; Mohamed et al. [7] proposed a kernel function between unstructured text descriptions based on distributed semantics, which can effectively predict the classification data set. However, these models ignore the contribution of part of speech to text, especially the analysis of emotions such as degree adverbs.

Therefore, for the sentiment analysis combined with the commentary short text, this paper proposes an improved feature weight short text classification method based on the characteristics of short text, trying to define the multi-factor weight rule, and combine the part of speech in the text with the star rating in the customer review. After that, the CBOW (Continuous Bag-of-Words) model is used to train the feature word vectors, and finally the classification and sentiment orientation analysis are performed. The experimental results show that the method is feasible and has good classification results.

2. Methodology

2.1 Update mechanism

After users purchase goods on major shopping websites, they will let users carry out star-class Evaluation. At the same time, users can publish their own shopping experience. Often, the favorable comment rate of products on the website is determined by star rating, which is quite different from user comments[3]. Therefore, when evaluating shops and services, it is not possible to use only star ratings and ignore the specific content of text comments. In addition, The focus of each user's attention will vary with the length of the short text, and the expressed emotion will also be manifested by the degree of adverbs. Secondly, after using the user for a period of time, the user can also evaluate the product again, that is, additional evaluation. This part is also an indispensable part of analyzing the quality of the product. According to the above analysis, the feature weight calculation method proposed in the literature [8] and the user comment time feature influencing factor model proposed in the literature [9] are modified. A keyword extraction method is proposed, which combines the multi-factor weighting strategy of integrating star rating, part of speech information and additional evaluation:

$$\mathrm{Weight(w)} = \alpha\mathrm{Weight}_{eva}(\mathrm{w}) + \beta\mathrm{Weight}_{ten}(\mathrm{w}) + \gamma\mathrm{Weight}_{add}(\mathrm{w}) \qquad (1)$$

In the above formula, Weight(w) represents the weight of the word w in the text d. $\mathrm{Weight}_{eva}(\mathrm{w})$, $\mathrm{Weight}_{ten}(\mathrm{w})$, and $\mathrm{Weight}_{add}(\mathrm{w})$ respectively represent the star rating, the part of speech information, and the weight of the additional comment in the text d. α,β,γ are weighting coefficients, and their sum is 1. The calculation formula of Weight_{ten} is referenced [10], and the weight calculation formulas of Weight_{eva} and Weight_{add} are:

$$\mathrm{Weight}_{eva}(\mathrm{W}_i) = \frac{tf(w_i,d)\times log(\frac{N}{n_w}+0.01)\times eva_{w_i}}{\sqrt{\sum_{w\epsilon d}\left[tf(w_i,d)\times log(\frac{N}{n_w}+0.01)\times eva_{w_i}\right]^2}} \qquad (2)$$

Where $tf(w_i,d)$ represents the word frequency of the word w_i in the text d, N is the total number of texts, and add_{w_i} represents the evaluation level weighting value of the word, and its specific value needs to be defined according to the research content of the literature [11]. In this literature, degree adverbs are classified into six categories, which are used as weights for part of speech. The weight values corresponding to degree adverbs are as follows:

Table 1 Weight value of degree adverb

Degree level	Code	Examples	Weights
Super	δ	exceed, more than	0.29
extremely	ε	most., very, very much	0.24
very	η	quite a little, really	0.18
More	τ	such, enough	0.14
a little	ω	quite	0.09
Lack	φ	relatively, slight, not very	0.06
No			0

In the preprocessing stage, when the data set contains degree adverbs, it can be replaced according to the code in the above table. If a text contains multiple adverbs of degree, you can add them and calculate the weights comprehensively. If the same degree adverb appears multiple times, you need to set the word frequency within a reasonable threshold. This paper sets 5; For short texts that do not contain degree adverbs, it is recorded as 0 in the calculation formula of degree adverb weight, which means $\beta\mathrm{Weight}_{ten}(\mathrm{w}) = 0$. Similarly, if the short text does not contain additional comments, the feature weight should also be 0, which means $\gamma\mathrm{Weight}_{add}(\mathrm{w}) = 0$.

2.2 Training feature word vector

This paper uses the Hierarchical Softmax-based CBOW model proposed by Mikolov [12]. This language model generally has three layers, an input layer (word vector), a hidden layer and an output layer (softmax layer). The model is shown below:

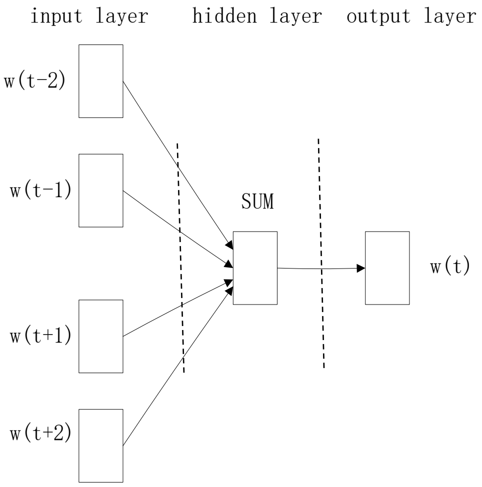

Figure. 1 CBOW model

Enter the one-hot encoded context word $\{x_1,\dots,x_C\}$, where the window size is C, the data dictionary is V, and the input vector is connected to the hidden layer through an V × N-dimensional weight matrix W,

$$\mathrm{h} = \frac{1}{C} W \cdot \left(\sum_{i=1}^{C} x_i\right). \tag{3}$$

Obtaining the output result of the hidden layer is calculating the weighted average of the input vector, and then the N-dimensional vector of the hidden layer is connected to the output layer through a N × V-dimensional weight matrix W, and the input of each node of the output layer is calculated.

$$u_j = {v'}_{wj}^{T} \cdot \mathrm{h} \tag{4}$$

Where ${v'}_{wj}^{T}$ is the jth column of the output matrix W', and finally the output word y after one-hot encoding is calculated.

$$y_{c,j} = p\left(w_{y,j} | w_1,\dots,w_c\right) = \frac{exp(u_j)}{\sum_{j'=1}^{V} exp(u'j)} \tag{5}$$

2.3 Classification method

The Support Vector Machine (SVM) classification algorithm is based on statistical learning theory and structural risk minimization theory. It has a strict theoretical basis and can solve small samples, high dimensions and nonlinear problems. And this algorithm is more suitable for the two-category problem [13]. The traditional SVM algorithm is often applied to supervised learning such as pattern recognition, regression analysis and classification. At the same time, when the feature extraction of data, the SVM algorithm proves that it has certain robustness to data noise [14]. In this paper, the SVM model is mainly used for text classification to classify comments.

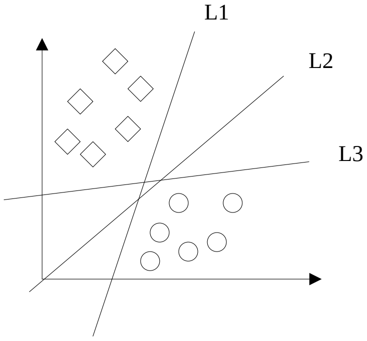

Figure. 2 Two-dimensional plane classification

The SVM model is formed based on the improvement of the two-dimensional plane classification. For the classification of the two-dimensional plane (Fig. 2), the classification effect of the line L_2 is better than that of L_1 and L_3 in order to separate the circle and the square in the figure. Similar to L_2, the two types of samples are separated and the probability that the sample points fall on both sides of the boundary is as high as possible. The plane separating the sample points in space is called hyperplane in high-dimensional space, and the sample point closest to this hyperplane in high-dimensional space is the support vector. Therefore, the support vector is also the only criterion for determining hyperplane. Assuming that $y = (x_1, x_2 \ldots x_n)$ is a point in the sample, where x_i is represented as the i th characteristic variable, the model that divides the hyperplane in the feature space can be expressed as $f(x) = w^t x + \mathrm{b}$, then the distance d from the point to the hyperplane can be calculated by the following formula :

$$d = \frac{|w_1 * x_1 + w_2 * x_n + \ldots w_n * x_n + b|}{\sqrt{\mathrm{w}_1{}^2 + \mathrm{w}_2{}^2 + \ldots + \mathrm{w}_n{}^2}} = \frac{|W^T * X + b|}{\|W\|} \tag{6}$$

Similar to finding the optimal interval for a low-dimensional hyperplane, the mapped data is converted into a dual problem:

$$max\ L(\alpha) = \sum_{i=1}^{n} \alpha_i \ - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\ \alpha_j y_i y_j {x_i}^T x_j \tag{7}$$

$$s.t.\ \ \sum_{i=1}^{n} \alpha_i\, y_i = 0, \alpha_i \geq 0, i = 1,2,3 \ldots, n \tag{8}$$

Text-classified data is not completely linearly separable, and a kernel function is introduced in order to separate data that is inseparable in two-dimensional space. Support vector machine algorithms are also available in multiple versions, mainly determined by different kernel functions [15]. Common kernel functions include linear kernel, polynomial kernel, Gaussian kernel, exponential kernel, Laplacian kernel and radial basis kernel function. This paper uses a radial basis kernel function.

$$\mathrm{k}(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \tag{9}$$

3. Experiment analysis

3.1 Experimental data

In order to verify the validity of the feature method proposed in this paper, this paper uses the user replies data crawled from Jingdong Mall by the network crawler to constitute the experimental corpus.

The experimental corpus is mainly two categories of mobile phones and laptops. The content of the crawl includes the first comment content, additional comment content, and star rating. It includes 5,730 comment data, including 3,216 mobile phone data and 2,514 laptop data. At the same time, using Python to obtain the data needs to be normalized, deleted or filled with missing values and other pre-processing work, and then manually label the text. The data standard format is shown in Table 2. Statistics on the original data favorable comment rate are shown in Table 3. The corpus participle uses the open source HanLP word segmentation tool.

Table 2 Original data format

Category	Original comment	Additional comments	Star rating
Mobile phone	The phone has good sound quality and beautiful colors, but the memory is too large. Facial unlocking is very useful	Null	3
Mobile phone	Mobile phone is good, just without headphones	Null	5
computer	The computer is easy to use and the customer service attitude is good. The fan sound is a bit loud, the customer service says that the game sound is big.	I bought it for two days, the fan sound can't stand it.	5
computer	I just bought it today, it will drop 200 directly.	Successfully applied for a refund, thank you	1

Table 3 Evaluation statistics

Comment data	Mobile phone	Proportion	Calculation	Proportion
all	3526		2529	
favorable	3432	0.9733	2378	0.9401
neutral	15	0.0043	7	0.0028
bad	26	0.0074	14	0.0055
additional	53	0.015	130	0.0514

3.2 Evaluation Index

In order to make the experimental results more accurate and more contrast, this paper uses the commonly used evaluation criteria for text classification, such as recall ratio and accuracy and comprehensive evaluation index F metric [16].

Accuracy rate:

$$p = \frac{\mathrm{tp}}{tp+fp} \tag{10}$$

Recall rate:

$$R = \frac{\mathrm{tp}}{tp+fn} \tag{11}$$

F metric:

$$F_{\alpha}(\mathrm{P},\mathrm{R}) = \frac{(\alpha^2+1)PR}{\alpha^2(P+R)} \tag{12}$$

When α=1, it becomes the F1 metric, $\mathrm{F1} = \frac{2PR}{P+R}$.

Where tp represents the number that actually belongs to this class and has been correctly retrieved, fp represents the number that is not actually in the class but has also been retrieved, and (tp+fp) represents all the numbers actually retrieved, fn represents the number that actually belongs to the class but has not been retrieved, and (tp+fn) represents all the numbers that should be retrieved. This paper mainly uses the F1 metric value, combined with the results of the accuracy rate P and the recall rate R. When F1 is higher, the experimental method is more reasonable.

3.3 Results and analysis

In this paper, the SVM model, KNN (k-Nearest Neighbor) model and Naive Bayesian Model (NBM) are used to verify the validity and feasibility of the proposed feature weight method. Because this article mainly studies the favorable comment rate, so the text label is set to both praise (set to 1) and non-praise (set to 0), while using the traditional TF-IDF feature weighting and a feature method combining star rating and text commenting proposed by this paper. Using the F1 metric as the evaluation criteria, the results as shown in Table 4 were obtained, where N-KNN and N-NBM are the results of the evaluation using the method herein.

According to the tabular data, the SVM evaluation results are generally better than the KNN model and the naive Bayesian model, because the support vector machine SVM has a strong theoretical basis, which can ensure that the found extreme solution is the global optimal solution rather than the local minimum, SVM has a good generalization ability for unknown samples [17].

The classification evaluation effect with improved feature weight is generally better than the classification effect based on common feature extraction. It shows that there are a lot of star rating evaluations and inconsistencies in the commentary short text, and there may be a large number of adverbs in the text to express the user's personal shopping sentiment. The model SVM-1 is the experimental result without considering user comments and additional comments. SVM-2 considers the evaluation results of three aspects comprehensively. Comparing the above two methods, it is found that the method proposed in this paper is better than the comparison method, which proves that the model uses text comment data and degree adverb weight to be more consistent with the customer's real emotion, which will also improve the effect of text classification.

Table 4 Comprehensive model evaluation results (%)

combination	Mobile phone			computer		
	P	R	F1	P	R	F1
SVM	98.26	95.74	96.98	98.41	94.74	96.54
KNN	96.44	90.18	93.21	95.37	91.65	93.47
NBM	95.72	92.35	94.01	96.84	93.22	95.00
N-KNN	97.57	96.01	96.78	97.23	96.44	96.83
N-NBM	97.11	92.84	94.93	96.84	95.01	95.92
SVM-1	98.01	96.24	97.12	96.59	97.02	96.80
SVM-2	99.23	96.84	98.02	98.67	98.00	98.33

4. Conclusion

weighted method of the fusion of comments proposed in this paper effectively improves the classification accuracy of short texts such as customer reviews, which also improved the method of using only star rating to determine the favorable comment rate in the past. In the process of commenting, users often have comments that do not match the star rating. After a period of use, they will add comments to the product and make the latest feedback. The method fully considers the user's shopping and evaluation habits, and comprehensive text comments ensure that the data of favorable comment rate is objective. However, with the deepening of the research, it was found that some product reviews were affected by some click farming. Using fake ways to improve online store rankings, get sales and praise will have a great impact on text analysis. Therefore, in view of this phenomenon, it is still of great significance to study how to eliminate the influence of inaccurate comments such as network brushing. Therefore, in view of this phenomenon, it is significance to study how to eliminate the influence of inaccurate comments such as click farming.

References

[1] L.L. Xu, J.S. Fu and C.H. Jiang (2015). A Product Ranking Algorithm Based on Wilson Interval of Users' Positive Ratings. Computer Technology and Development, vol.25, no.5, p.168-171.

[2] Y.J. Li, Y.L. Li (2015). The Influence of Seller's Manipulation of Online Reviews. Soft Science, vol.29, no.12, p.135-139.

[3] H.X. Yuan (2018). Research on Tendency Analysis of OnlineShopping Comment Based on SVM. Chongqing Normal University.

[4] C.R. Li (2018) Sentiment Analysis And Visualization Research of Online News Users' Comments. Harbin Institute of Technology

[5] X. Yuan, M. Sun, Z. Chen, et al, Semantic Clustering-Based deep hypergraph model for online reviews semantic classification in cyber-physical-social Systems [J]. IEEE Access, 2018 6 (1): 17942-17951.

[6] Z.H. Zhang, Y. Guo and M.Q. Han (2015). Research on short text classification based on keyword similarity. Application Research of Computers, 1-6 [2019-03-14]. https: //doi.org/10.19734/j.issn.1001-3695.2018.04.0440.

[7] M. Elhoseiny, A. Elgammal and B. Saleh, Write a classifier: predicting visual classifiers from unstructured text [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017 39 (12): 2539-2553.

[8] C.Ma, R.F. Guo and C.Gao (2018). Short Text Clustering Algorithm with Improved Feature Weight. Computer Systems & Applications, vol.27, no.9, p.210-214.

[9] Y.F. Zhang, L.H. Peng and C.Hong (2019). An Empirical Study on Time-series Correlation Characteristics of Online Users'Follow-up Review Behaviors: Taking the Mobile Phone Review Data on Jingdong Mall as an Example. Information studies: Theory & Application, 1-11 [2019-03-06]. http://kns. cnki. net/kcms/detail/11.1762. G3.20181030.0921.004.html.

[10] T.C. Li, Y.Y. Xi and B.Wang (2015). Improved Short Text Hierarchical Clustering Algorithm. Journal of Information Engineering University, vol.16, no.6, p.743-748, 752.

[11] Z.W. Zhou (2017). Data Analysis on Customer Satisfaction Based on Commodity Comment: Taking the Reviews of Jingdong Mobile Platform As An Example. Zhejiang Gongshang University.

[12] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. Computer Science, 2013, 3 (1): 1-12.

[13] S.Z. Tu, J.Yang. and L.Zhao (2019). Filtering Chinese microblog topics noise algorithm based on a semisupervised model. Journal of Tsinghua University (Science and Technology), vol.59, no.3, p.178-185.

[14] VAPNIK V. Statistical learning theory.1998 [M]. Wiley, New York, 1998: 1.

[15] Q.S. You, J.X.Wang. and X.Y. Zhang (2019). SVM-Based Analysis on Food Safety Sampling and Inspection Data. Software Engineering, vol.22, no.2, p.29-31.

[16] H. Li. Statistical learning method (2012). Tsinghua University Press

[17] H.Y. Wang, J.H. Li. and F.L. Yang (2014). Overview of support vector machine analysis and algorithm. Application Research of Computers, vol.31, no.5, p.1281-1286.