Journal of Computers Vol. 29 No. 1, 2018, pp. 86-98
doi:10.3966/199115992018012901008

Combining Features to Meet User Satisfaction: Mining Helpful Chinese Reviews

Lizhen Liu, Shiwei Zhang, Wei Song, Hanshi Wang[*]

Information and Engineering College, Capital Normal University, Beijing 100048, China
liz_liu@126.com, 2141002047@cnu.edu.cn, wsong@cnu.edu.cn , necrostone@sina.com

Received 24 September 2016; Revised 3 March 2017; Accepted 4 May 2017

Abstract. Product reviews have become the important recourse in the online environment of Internet. However, the quality of the reviews is spotty and this influences the accuracy and the reliability for data mining. This paper focuses on how to excavate the helpful product reviews buried under the mass of data. The proposed method is as follows: filter words using the best first-search strategy, use latent Dirichlet allocation (LDA) to get the topic distribution, use the Kullback-Leibler (KL) divergence to calculate the similarity, extract the popular opinion of reviews, observe the difference between the popular opinion and the review, perform emotion detailing by getting the specific value of each attribute, consider the credibility as well as the metadata of the reviews, and finally train the weights of feature vectors according to the support vector machine (SVM). Experimental results demonstrate the ability of the proposed method to significantly improve the classification accuracy.

Keywords: KL divergence, LDA, machine learning, popular opinion, the best first-search strategy

1 Introduction

With the development of the Web 2.0, online reviews provide a new data resource for users of the internet. These reviews are written by users detailing their experience about a specific product. User-generated reviews provide useful information to other customers and have become an important factor when making a purchase decision. However, e because of the openness of internet, anyone can post a review and this creates a flood of massive data. Users do not have much time to read the reviews one by one. It is hard for them to distinguish which ones are helpful and which ones are not. How to classify reviews based on helpfulness and uselessness has become a very important issue.

What is the meaning of helpfulness? Helpfulness or “practicability” means that the review contains enough information, and it has a higher supporting role for the customer’s perception and judgment of the commodity. On the contrary, useless is “junk”, which means the review has no reference value for customers. Furthermore, the futile reviews as act as “noise” which can negatively affect the customer’s reasonable perception and judgment of the product. In view of this phenomenon, Amazon provides consumers with a quick way to browse reviews. Customers are allowed to vote whether the review is helpful or useless, and Amazon ranks reviews according to the ratio of effective voting. Other e-commerce sites will extract sentiment orientation label from reviews. However, the biggest problem is that the effective voting needs time to get votes. A new review cannot be compared with earlier reviews, which causes it to drown in the flood of data. Even though the new review has enough information, it will not be presented to customers. So we need a mechanism to automatically analyze the review’s helpfulness, select the useful ones from the large amount of data and help customers make right purchase decision quickly.

Now, the vast majority of studies take the helpful information as features, and then the reviews are ranked from high to low based on the feature. Some studies consider the text format of the review,

[*] Corresponding Author

particularly the length, writing style, the votes received and the timeliness. Others consider the content, such as semantic analysis, emotional polarity analysis, the architectural feature of reviews and syntactic analysis.

In order to train a good classification model and improve the result of ranking, this paper adopts hybrid features to build the classified model to classify online reviews. First of all, we get the topic distribution of each review and all reviews, then to calculate the similarity of topic distribution using the Kullback-Leibler divergence. Before using LDA to get the topics of reviews, we filter the words using the best first-search strategy. Then, we extract the popular opinion of reviews, and observe the difference between the popular opinion and each review. We perform the emotion detailing and get the specific value of each attribute. We also consider the credibility of the review as well as its metadata. Finally, we train the weights of feature vectors according to SVM.

The rest of this paper is organized as follows: Section 2 introduces the related works. Section 3 introduces the way to filter words based on the best first-search strategy. In Section 4, we discuss the classification of emotional tendency based on fine granularity. In Section 5, we present the hybrid features for training. Section 6 discusses the results of the experiment. Finally, Section 7 presents our conclusion and the future work.

2 Related literature

Due to the increasing number of online reviews, they have more practical significance. More and more people are beginning to analyze these online reviews. Some of them analyze the emotion in the reviews; others analyze the construction. There are people who study the extraction of product attribute words. Jo and Oh [1] tackled the problem of automatically discovering what aspects are evaluated in reviews and how sentiments for different aspects are expressed. They put forward sentence latent Dirichlet allocation SLDA, a probabilistic generative model that assumes all words in a single sentence are generated from one aspect, then extend SLDA to Aspect and Sentiment Unification Model (ASUM) and applied them to reviews of electronic devices and restaurants. In some domestic and foreign researches, the utility analysis is transformed into a classification or ranking task, and extracted the features of multiple dimensions that are relevant with reviews and build the prediction or ranking model by making use of machine learning. In Wang et al. [2], novel method of opinion mining was proposed and evaluated by a collection of real online product reviews. A hierarchical fuzzy domain sentiment ontology (FDSO) has been introduced by this approach, which defined a space of product features and corresponding opinions, thus making it possible for a product to be classified and scored by commonly accepted features. Therefore, it enhances the user experience to search a product and compare it with other products feature by feature. Yu et al. [3] designed the acquisition algorithm based on the dictionary, rule and LDA to formalize each review. And also selected reviews by greedy algorithm, which maximized the coverage of reviews' product attributes after selecting high-quality review from product review set. Hong et al. [4] made use of the user preferences with language characteristics to train two categories of classifier and a ranking system of support vector machine (SVM). Lee and Choeh [5] adopted the multilayer perceptron neural networks to predict the level of reviews. Liu et al. [6] thought of three factors which are helpful for potential customers to make purchase decisions are reviewers experience, writing style of reviews and the review timeliness. The study by extracting relative attributes from the IMDB reviews data to fit the three factors by building non-linear regression model, then predict the helpfulness of reviews. Zhiyu [7] study the prediction method from four dimensionalities, that is, evaluation objective, feature selection, evaluation methods and evaluation target. To satisfy the users' individual needs, Miao et al. [8] put forward a view search system, which considered the topic relevance and the time dimension of information quality when ranking the reviews. Krishnamoorthy [9] dealt with the reviews using four aspects: language characteristics, the metadata of reviews, reviews' legibility and the reviews' subject. Cheung et al. [10] analyzed the impact of the degree of consumer involvement degree and the profession of customers on shopping decision. Hong [4] put forth three user preferences as features to classify and rank reviews: information needs, reviews credibility and popular opinion of reviews. In the aspect of ranking, Krestel and Dokoohaki [11] combined the star and LDA to build a model and rank reviews by summary-focused ranking, sentiment-focused ranking, topic-focused ranking. Ramkumar et al. [12] proposed a complete system which started from reviews collection through the analysis of reviews to get a score calculation of a product. The novel methods which were used in this system are the calculation of spam level of each review and the calculation of scores for each feature of a product. Fuzzy logic was used to calculate the spam level scores and the product ratings. Scaffidi et al. [13] presented a new search system called Red Opal that enables users to locate products rapidly based on features. Their fully automatic system examines prior customer reviews, identifies product features, and scores each product on each feature. Kannan [14] provided an apples-to-apples comparison of features for review utility prediction. They surveyed and tested existing methods for usefulness prediction on a corpus of Amazon product reviews with an SVM classier and regression. The motivation was two-fold: to provide a standardized system to measure individual feature usefulness, and to explore possible synergies between different features for utility classification and regression. Also, there were many learners to assess reviews by verifying the assumption they put forward. Lee and Kao [15] verified the impact of positive and negative tendencies on review prediction.

This paper uses the best first-search strategy to filter words, select the high quality words to get the topic by LDA, then compare the topic of one review with the total reviews to get the similarity matrix. Secondly, we extract the popular opinions and analyze the degree of matching between each review and the popular opinion. We refine the emotional tendency of each review feature, and build a model to get the popular emotion tendency of each feature. We then combine the reliability of reviews and the metadata to train the classifier.

3 Filter Words Based on the Best First-search Strategy

As we all know, the input to LDA is a bag of words (BOW). The basic thought is to ignore the word order, grammar, syntax and only look at the text as a collection of words, with each word independent of the other. Let’s say there are two documents:

(1) Bob likes to play basketball; Jim likes to play, too.

(2) Doc2: Bob also likes to play football games.

Based on the two text documents to construct a dictionary:

{1:”Bob”, 2. “like”, 3. “to”, 4. “play”, 5. “basketball”, 6. “also”, 7. “football”， 8. “games”, 9. “Jim”, 10. “too”};

The two documents can be represented as the vector:

1: [1, 2, 1, 1, 1, 0, 0, 0, 1, 1];

2: [1, 1, 1, 1 ,0, 1, 1, 1, 0, 0];

However, the reality is that every word has relation more or less. This assumption has a big effect in training the model of the LDA. Therefore, in this paper, we use the best first-search strategy to select the collection of words. Filtrate words to form the best collection and ensure the efficiency of the LDA.

We make the words as the predictors and calculate the correlation between two words and the relevance between the word and the target. We need the relevance of two words as small as possible; as for their correlation, the higher the better.

In this paper, mutual information is used to calculate the relevance between two words.

$$I(w_i,w_j)=\log\frac{p(w_i,w_j)}{p(w_i)p(w_j)}\approx\log\frac{A*N}{(A+C)(A+B)} \quad \textbf{(1)}$$

$$\overline{r_{w_i,w_j}}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}I(w_i,w_j)}{n*(n-1)} \quad \textbf{(2)}$$

A is the number of co-occurrence of . B is the time that w_i,w_j appears in the document but the w_j is not in the document. C is the times that the w_j is in the document but the w_i is not in the document. N is the total number of reviews.

Information gain mainly assesses how much information entropy that the item can bring. This paper uses the information gain to calculate the correlation between the predictor and the target.

$$IG(W) = H(C) - H(C \mid W) = H(C) - P(W)H(C \mid W) + P(\overline{W})H(C \mid \overline{W}) \quad \textbf{(3)}$$

$$IG(w_i) = -\sum_{i=1}^{k} p(c_i)\log p(c_i) + p(w_i)\sum_{i=1}^{k} p(c_i \mid w_i)\log p(c_i \mid w_i) + p(w_i)\sum_{i=1}^{k} p(c_i \mid \overline{w_i})\log(c_i \mid \overline{w_i}) \quad \textbf{(4)}$$

$$\overline{r_{cw}} = \sum_{j=1}^{n} \frac{IG(w_i)}{n} \quad \textbf{(5)}$$

The classification of reviews is a problem of binary-class, so the collection of classification is $\{c_1, c_2\}$ so the k is 2.

$$Merit_s = \frac{n * \overline{r_{cw}}}{\sqrt{n + n*(n-1)*\overline{r_{w_i w_j}}}} \quad \textbf{(6)}$$

where Merits is the heuristic "merit" of a feature subset S, $\overline{r_{cw}}$ is the average word-class correlation, and $\overline{r_{w_i,w_j}}$ is the average word-word intercorrelation. The correlation-based feature selection (CFS) aims to reduce dimension [16]. However, we use this method here to filter words, meaning, we make the words as the features for the moment.

At every step, the best attribute is chosen and appended to the current best subset. If adding the new word does not cause improvement, then the search goes back to the next unexplored subset and starts from there. Finally, a common stopping criterion is controlling the level of backtracking allowed [17].

4 The Classification of Emotional Tendency Based on Fine Granularity

In past studies, only the sentiment polarity was considered. For example, a sentence is either positive or negative. However, it is enough to analyze the polarity; we also need to quantize the value of the emotional intensity, which can meticulously indicate the effectiveness of the reviews about the product. Because the effectiveness when compared to the popular opinion can affect people's purchase decision, the distance of emotional intensity between the review and the popular opinion can get tricky. The comparison is not only between both the polarity, which the result is 0 (the same polarity by subtracting) or ± 2 (subtraction of the different polarity). The polarity of emotional words is easy to determine, but the intensity of the emotional words is multifaceted.

The sentiment words are usually adjectives, and the adverbs reflect the degree of magnitude of adjective; therefore, adverbs are an important indicator of emotional intensity. Although the adjective itself can represent intensity, the intensity that the broader range of adjectives are similar. What's more, adverbs are divided into some levels. According to the HowNet, the adverbs are divided into six levels: extreme, very, more, -ish, insufficiently, over. Each level has a different effect on the sentiment words, so we assign a value for each. The degree from low to high in order is 0.1, 0.2, 0.4, 0.6, 0.9, 1.3.

However, the sentiment words in the HowNet are only divided into positive and negative, so we use the ontology emotional words in Chinese of Dalian University of Technology. Lexical ontology joins the emotional category "good" for a more detailed division of good feelings. In their document, there is the polarity as well as the intensity of the word. The negative polarity in there is represented as 2; we changed it to -1.

When reading the reviews, it is not difficult to observe that there are 6 kinds of circumstances:
Adverb + emotional words/ emotional words +adverb. The emotional tendency is the multiplication of both, the tendency is:

$$v = t_{adv} * t_{sw} \quad \textbf{(7)}$$

tsw is the value of the emotional tendency for the sentiment word.

t is the value of the adverb degree.

v is the value of the emotional intensity about the attribute word in this review

Negative word + positive emotional words. If the tendency of the emotional word is positive, adding the negative word represents the emotion that is negative, and the tendency is:

$$v = -1 * t_{sw} \tag{8}$$

For example, "not comfortable" has an emotional tendency of:

v(not comfortable)=-1*t(comfortable)

Negative word + negative emotional words. The tendency of the emotional word is negative and adding the negative word which does not represent the opposite polarity. The negative word to the negative polarity only weakens the intensity of the negative emotional words. We found that if the intensity of the sentiment word is greater than 0.5, its polarity is still negative after adding the negative word. On the contrary, if the intensity of sentiment word is less than 0.5, it will become the opposite polarity after adding the negative word. So, the way to calculate its tendency is:

$$v = -1 * (0.5 + t_{sw}) \tag{9}$$

Adverb + neutral words/ neutral word + adverb. In the sentence, if it only has the neutral word, the attitude of the reviewer usually is objective. However, if the negative word is added before the neutral word, the attitude is changed. We can feel the feeling of disappointment to the product, for example, "very general". In this circumstance, the tendency is:

$$v = 0.5 - t_{adv} \tag{10}$$

Negative word + adverb + emotional words. The negative word has weakened the intensity of adverb, and the tendency is:

$$v = (2 - t_{adv}) * t_{sw} \tag{11}$$

For example: "not very comfortable", the tendency is: v(not very comfortable)=(2-t(very))*t(comfortable). 2 is the sum of extreme value range.

Adverb + negative word + emotional words. In this circumstance, the tendency is:

$$v = t_{adv} * (-1) * t_{sw} \tag{12}$$

5 Model Features

Let $R = \{r_1, r_2, \cdots, r_n\}$ be a set of reviews in the dataset. Each can be represented as a tuple containing four elements [ZF, PF, RF, MF] where ZF describe the topic relevance between each reviews and all reviews; PF is the difference between each review's opinion and the popular opinions; RF is the review reliability; and MF is the review metadata, such as review data, pictures, stars, readability and so on. Our aim is to divide all reviews into two categories: helpfulness and useless. Therefore, we use C to represent the result, that is:

$$C_I = \begin{cases} 1; & if\ Y_i > \tau \\ 0; & else \end{cases} \tag{13}$$

C is the result of classification, Y is the score of calculating review's helpfulness, τ is the threshold value. =1 is the mean that the review is judged to be helpful after assessing the review.

F as the feature matrix, which is an n*m dimension, where n is the number of reviews and m is the number of features.

5.1 Topic Relevance

In section 3, we have filtered some words, now we use the remainder words as the input to get the topic of the reviews. The remainder words make up a set of vocabulary the number of which is m. In the flow of LDA, each document d in the document set D is a sequence of words {w1,w2,…,wm}. LDA hope training out of the results of the two vectors according to the document set D as input. The two vectors include topic and words in the topic. In this paper, the document set D is R, each document is the review r.

The core of the LDA formula is as follows:

$$p(w|r)=p(w|z)*p(z|r) \tag{14}$$

where $p(w|r)$ is the probability of the word in the review r; is the probability of word in the topic; is the probability of topic in the review.

For each review, the formula is:

$$p(w|r)=\sum_{j=1}^{|z|}p(w_i|z_j)*p(z_j|r) \tag{15}$$

For all reviews, the formula is:

$$p(w_i|R)=\sum_{j=1}^{|z|}p(w_i|z_j)*p(z_j|R) \tag{16}$$

Here, we need to calculate the topic relevance between each review and all reviews using the Kullback-Leibler divergence. The KL-divergence is always used to measure the distance between two probability distributions.

The distance of review ri from all reviews is:

$$D_{KL}(p(z|r_i)\|p(z_j|R)=\sum_{j=1}^{|z|}p(z_j|r_i)\log(\frac{p(z_j|r_i)}{p(z_j|R_i)}) \tag{17}$$

$p(z|r)$ is the topic distribution for review r. D_{KL} is bigger, the relevance is higher. Using this method to calculate the distance of every review and all reviews. The reason for making the value as a feature to classify reviews is that the topic of all reviews represents some points of concern for the users. If the topic relevance is higher, the review gets peoples' attention more than whether it is helpful or not.

In the experiment, we set 10 themes for each product. Fig. 1 and Fig. 2 show the probabilities of the 10 themes for each product.

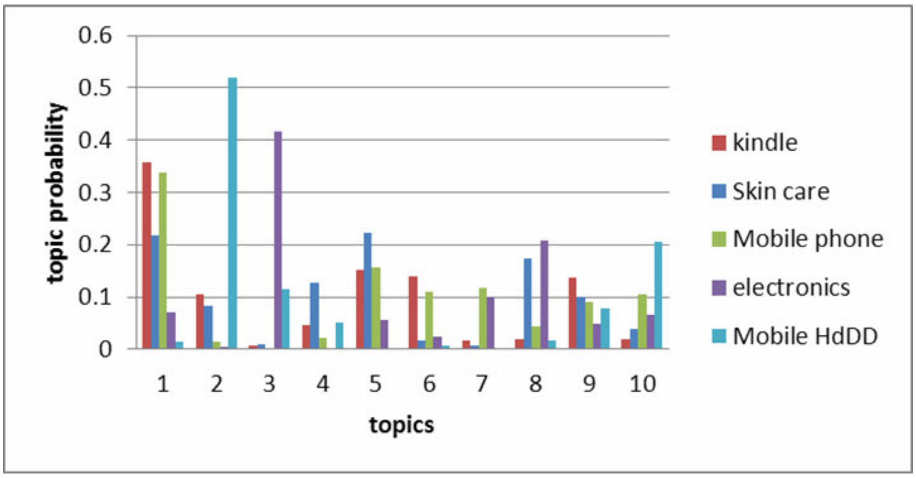

Fig. 1. The results of ZF about five products

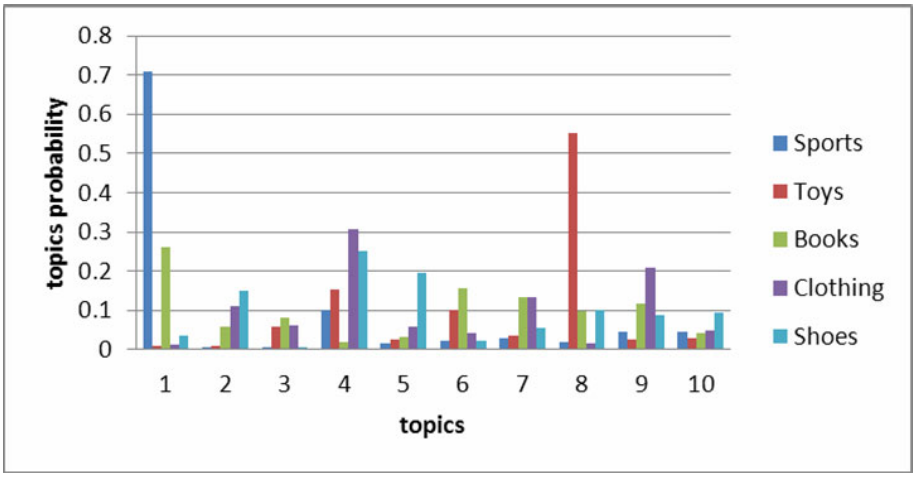

Fig. 2. The results of ZF about other products

5.2 The Difference of Opinions

In section 4, we introduced a method to calculate the degree of emotional tendency for product attribute based on the fine grit. Now, we use the graph model on the second floor to chart the tendency of popular opinion, and then build a simple template which includes the product attributes and the value of attribute's emotional tendency.

Previous studies have classified emotional polarity into positive and negative, which is a rough division. There are levels of positive and negative effect on the emotion. Using the graph model on the second floor, we constructed an undirected graph. Fig. 3 shows a simple way to get the tendency of product attributes. On the left is the product attribute while on the right are the possible values of emotional tendency about the product attribute. The line between attribute with the degree of emotional tendency has weight and the weight is the number of lines.

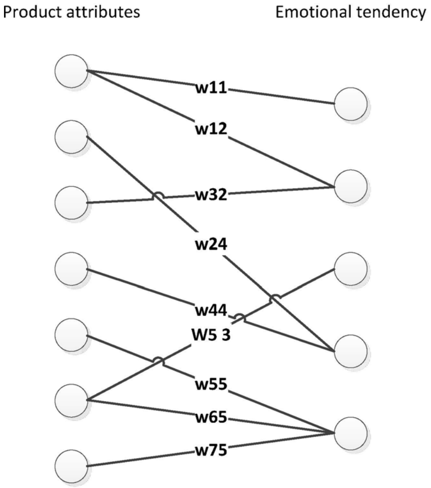

Fig. 3. The method of calculating the emotional tendency of each attribute word

The number of the product attribute is k, A is the set of product attributes, E is the set of emotional tendencies, $E = \{e_1, e_2, e_3, \cdots, e_N\}$. w_{ij} is the number of the tendency of product attribute a_i is e_j . After statistics the emotional tendency of attribute, we need to generate a simply template that include all product attribute and the value of tendency. For example, the product is a smartphone, the product attribute includes size, weight, sharpness, pixel, screen and so on. The final template is "(size,V1); (weight, V2); (sharpness, V3); (pixel, V4); (screen, V5) ". Similarly, we can use this way to represent each review, but the value of tendency is concrete instead of a fuzzy value like 1 or 0.5. The tendency of the product attribute in the template is the average weight, that is:

$$V_i = \frac{\sum_{j=1}^{N} w_{ij}}{N} \quad \textbf{(18)}$$

Here, we only provide the popular tendency of the extracted attribute words for the four products, as shown in Fig. 4, Fig. 5, Fig. 6 and Fig. 7.

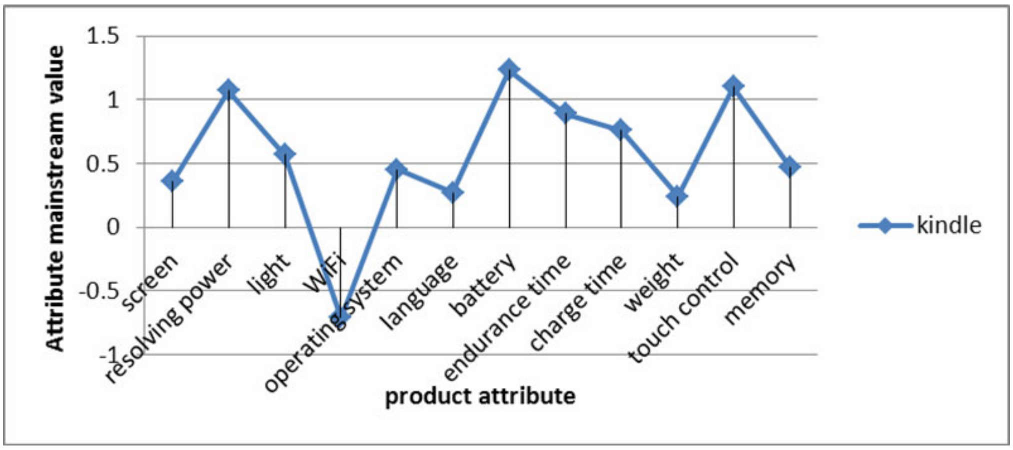

Fig. 4. The popular emotional tendency of kindle

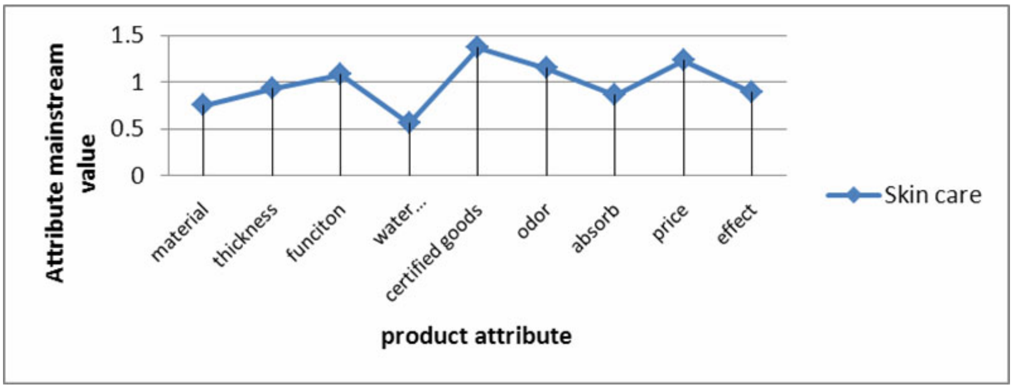

Fig. 5. The popular emotional tendency of skin care

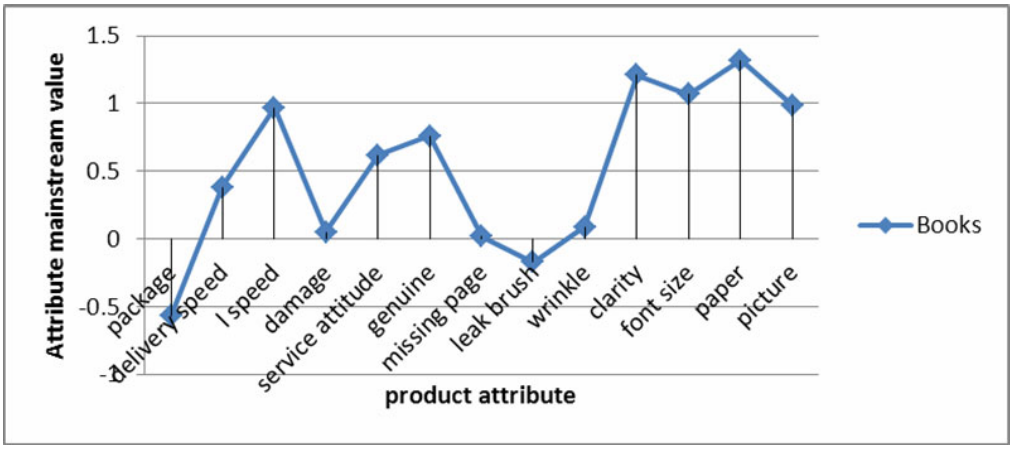

Fig. 6. The popular emotional tendency of books

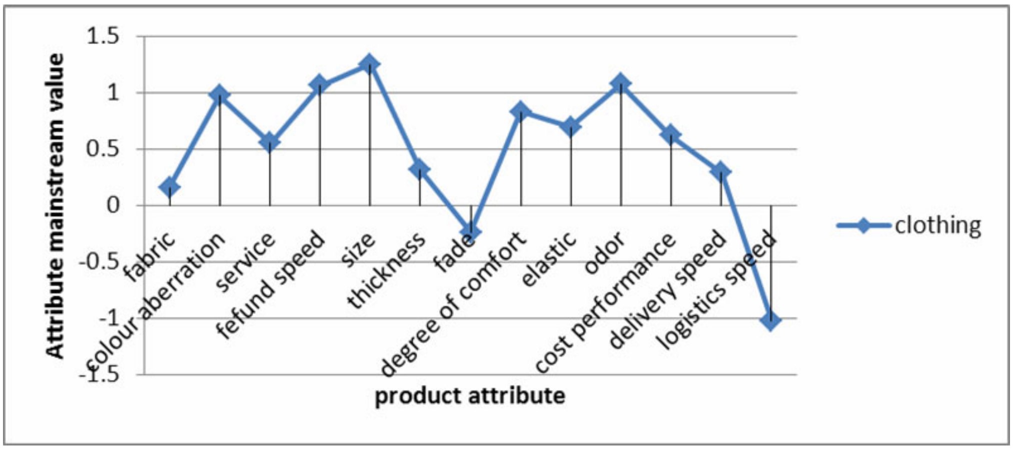

Fig. 7. The popular emotional tendency of clothing

After assigning the value of the product attribute, we consider whether the product attribute appears in the review. If the attribute word appears in the review, then we denote the word as 1 in the vector; otherwise, we denote the word as 0 in the vector. In the process of calculation, we also consider the weight of the attribute words. The reason is some words may be what the customers focus on while others are not. So the distance of the review from popular opinion is:

$$dis = \sum_{i=1}^{k} w_i * (v_i - V_i) * x_i \qquad \textbf{(19)}$$

Where w_i is the weight of attribute word, v_i is the accurate value about the attribute word's emotional tendency, x_i is 1 or 0, its mean the word whether appears in the review.

5.3 Review Reliability

Reviews are important for consumers when making a purchase decision. It not only affects the choice of the product, but also the choice of the shopping website. Unfortunately, some reviews are irresponsible and therefore, are not reliable. Therefore, we need to analyze the reviewers' credibility. In this paper, we summarized and classified the text characteristic dimension of credibility factors, constructed a text credibility calculation model, explored the main difficult points minutely and presented the terms of settlement. The paper does a research from the words and performs a source and reviewer's credibility analysis. This includes the identity of the reviewer, the reviewer's level and reviewer's number of posts. We use the model to evaluate the reliability of reviews.

5.4 Review Metadata

Reviews also contain other information, such as the title, the date it was published, the star rating, pictures, and so on. We cannot deny the effect of these data — people will know the general content through the title. On the other hand, customers can easily know the attitude of the reviewer about to the product through the star rating they gave. Also, some people prefer to read recent reviews. They ascribe greater reference value for more recent reviews because the product itself changes over time. Finally, the readability of a review is important. A fluent review is easy to read and can attract more customers to read. On the contrary, a review with sentence inversion or misused terms will make the customers lose their patience. We mainly considered the review metadata from three aspects:

Publish date. A review indicates the state of the product at a certain time. It is the difference between the time now and the time the review was released.

$$review\ time = \log(actual\ time - review\ data) \qquad \textbf{(20)}$$

where the review time is the difference between the time of customer browsed the product and the review publish date.

Star rating. Because the star rating of review shows the attitude of reviewer about the product, most customers prefer to read the reviews which include both advantages and disadvantages of the product. If the review only praised the product or only criticized the product, customers will think the review does not have contrast. Generallu, users want to read the reviews with three or four stars. For this reason, we give a bigger weight for four and three stars than others.

Pictures. Compared to the traditional word-of-mouth (WOM), electronic WOM can spread by the way of content and pictures. Text usually has a strict syntax, which can facilitate accurate explanation. On the other hand, the pictures can let consumers quickly get intuitive feelings about product. Some studies indicated that pictures can generate positive effect more easily than text. Therefore, we put the review's pictures into consideration.

6 Experiment

6.1 Dataset

We used Firefox to crawl data for 10 different product types listed in Table 1. The dataset is from the Chinese Amazon.com website. Before using the dataset, we removed duplicate reviews. Generally, if the matching degree of words exceeds 90% for two sentences, the two sentences are considered duplicates. We also used spam recognition algorithm to remove spam reviews. Generally, people use the rate of effective vote as the standard to compare the accuracy rate. We needed to set a threshold value for distinguishing whether the review is helpful or useless. Let's say Review 1 has 80 helpful votes and 20 dissenting votes and Review 2 has 8 helpful votes and 2 dissenting votes. We can see that the effective vote rate for both is 80%, but we think Review 1 is more helpful than Review 2. To solve this, we learned from Liu [6] that they only used reviews with at least ten total votes to ensure robustness of the results.

After cleaning the data, the number of product reviews per type is shown in Table 1.

Table 1. The type of data

Product type	Number of reviews	Product type	Number of reviews
Kindle	12530	Sports	1587
Skin care	6680	Toys	892
Mobile phone	6460	Books	2740
Electronics	2670	Clothing	5791
Mobile HDD	3350	Shoes	4953

Based on Kim et al. [18] and Ghose and Ipeirotis [19], the threshold value is 0.6. If the helpful voting rate is greater than 0.6, then we say the review is helpful. If the helpful voting rate is less than 0.6, the review is useless.

To build the classification model, we use SVM as a learning method. In the evaluation process, we ran a 10-fold cross-validation, using one fold as the test set and the others as training sets.

6.2 Result

Accuracy of each feature. A classification model is constructed in the experiment, which takes into account multiple features ZF, PF, RF and MF. The accuracy of each product is shown in Table 2.

It can be seen from the table that the feature PF has the highest accuracy. This is because the feature ZF adopts the LDA method. However, the Chinese reviews are usually very short; each user review is based on personal preference and thus, a comprehensive evaluation of reviews is lacking. Although LDA can reveal representative themes, the accuracy is still low due to the lack of words in the text. The second feature is designed to comprehensively compare the emotional tendency of each attribute and it is independent of the text size. As for the picture, few users posted pictures and thus, only a small number of reviews had this feature, making it auxiliary and not a decisive factor.

Table 2. The accuracy of each product type

Features	Kindle	Skin care	Cell phone	Electronics	mobile HDD
ZF	67.27	63.79	65.35	65.24	66.22
PF	69.37	72.81	73.26	73.16	69.56
RF	50.91	53.27	53.06	50.73	52.87
MF	51.86	52.29	54.35	52.97	53.55
ALL	83.26	84.35	85.14	82.68	81.96

Features	Sports	Toys	Books	Clothing	Shoes
ZF	63.28	65.25	64.67	68.24	66.10
PF	68.94	71.51	74.46	74.05	70.82
RF	49.89	52.28	51.46	53.46	52.53
MF	52.65	54.58	53.60	55.12	54.07
ALL	82.49	83.58	83.98	85.39	82.87

Accuracy of combined features. The experiment was also performed to explore the accuracy of the combined features: the combination of ZF and PF and the combination of ZF and MF. Because different products have different reviews, different expressions and different attractions, the product reviews with the highest and lowest accuracy were chosen for the experiment.

Table 3. Results of feature combinations for the product review with the highest accuracy

Feature	Accuracy (%)	Feature combinations	Accuracy (%)	Feature combinations	Accuracy (%)	Feature combinations	Accuracy (%)
ZF	68.24	ZF+PF	75.28	ZF+PF+RF	80.95	ZF+PF+RF+MF	85.39
PF	74.05	ZF+RF	69.32	ZF+RF+MF	76.34		
MF	55.12	PF+RF	75.13	ZF+PF+MF	82.56		
RF	53.46	PF+MF	76.11	PF+RF+MF	77.61		
		ZF+MF	70.15				
		RF+MF	59.43				

Table 4. Results of feature combinations for the product review with the lowest accuracy

Feature	Accuracy (%)	Feature combinations	Accuracy (%)	Feature combinations	Accuracy (%)	Feature combinations	Accuracy (%)
ZF	66.22	ZF+PF	71.35	ZF+PF+RF	77.86	ZF+PF+RF+M F	81.96
PF	69.56	ZF+RF	70.36	ZF+RF+MF	72.45		
MF	53.55	PF+RF	71.34	ZF+PF+MF	80.65		
RF	52.87	PF+MF	70..96	PF+RF+MF	76.24		
		ZF+MF	58.37				
		RF+MF	55.33				

Comparing the accuracy of individual features and the accuracy of feature combinations, we found that some feature combinations are not as accurate as other combinations with less accurate features. This is because the correlation degree of two very accurate features may be less than that of two inaccurate features. This means that an accurate feature may be very independent and thus, does not complement well other features.

Reviewer classification has become an increasingly important issue. In this paper, we studied the text features and classified reviews using SVM according to the features of text reliability, LDA theme, emotional tendency and picture posting. Experimental results on the 10 product types showed that among the four features, the accuracy has the highest level of 85.39% and the lowest level of 81.96%. It can also be seen from the Tables that the accuracy of the combination of the four features is higher than that of each feature alone and that of other combinations.

7 Conclusion

This paper delved into the effectiveness of online reviews and put forth a new method to classify them.

The new model used hybrid features to automatically assess review helpfulness.

Four features were defined. First, the theme model and the KL divergence were used to analyze the distribution difference between individual reviews and global reviews. Prior to this, we pre-process the text to filter the review words jointly using information entropy, correlation and the optimal priority selection algorithm. By doing this, the correlation between any pair of words is minimized while the correlation between the word and the target variable is maximized. Next, the emotional tendency of reviews is refined. Individual reviews are compared with the popular emotional tendency of reviews to clarify the emotional tendency and improve the classification accuracy. Afterwards, we examine the reliability and metadata of the reviews. These four hybrid features are used to construct a model for efficient and automatic classification of reviews. Experimental results demonstrate the ability of the proposed method to significantly improve the classification accuracy.

For this study, the focus was on the analysis of words of reviews, distribution of themes and the emotional tendency. In the future, we will perform in-depth research on the linguistic features of texts (syntax or grammar) and user preferences in order to construct a more effective classification model for more accurate and automatic evaluation of reviews.

Acknowledgments

This work was supported in part by National Science Foundation of China under Grants No. 61303105 and 61402304, the Humanity & Social Science general project of Ministry of Education under Grants No.14YJAZH046, the Beijing Natural Science Foundation under Grants No. 4154065, the Beijing Educational Committee Science and Technology Development Planned under Grants No. KM201610028015 and KM201410028017, and Beijing Key Disciplines of Computer Application Technology.

References

[1] Y. Jo, A.H. Oh, Aspect and sentiment unification model for online review analysis, in: Proc. the fourth ACM International Conference on Web Search and Data Mining, 2011.

[2] H. Wang, X. Nie, L. Liu, J. Lu, A fuzzy domain sentiment ontology based opinion mining approach for chinese online product reviews, Journal of Computers 8(9)(2013) 2225-2231.

[3] W. Yu, C. Sha, X. He, R. Zhang, Review selection considering opinion diversity, Journal of Computer Research & Development 52(5)(2015) 1050-1060.

[4] Y. Hong, J. Lu, J. Yao, Q. Zhu, G. Zhou, What reviews are satisfactory: novel features for automatic helpfulness voting, in: Proc. the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval, 2012.

[5] S. Lee, J.Y. Choeh, Predicting the helpfulness of online reviews using multilayer perceptron neural networks, Expert Systems with Applications 41(6)(2014) 3041-3046.

[6] Y. Liu, X. Huang, A. An, X. Yu, Modeling and predicting the helpfulness of online reviews, in: Proc. ICDM'08. Eighth IEEE International Conference on Data Mining, 2008.

[7] L. Zhiyu, Study on the reviews effectiveness sequencing model of online products, New Technology of Library and sinformation Service 4(2013) 62-68.

[8] Q. Miao, Q. Li, R. Dai, AMAZING: A sentiment mining and retrieval system, Expert Systems with Applications 36(3)(2009) 7192-7198.

[9] S. Krishnamoorthy, Linguistic features for review helpfulness prediction, Expert Systems with Applications 42(7)(2015) 3751-3759.

[10] C.M. Cheung, B.S. Xiao, I.L. Liu, Do actions speak louder than voices? The signaling role of social information cues in

influencing consumer purchase decisions, Decision Support Systems 65(2014) 50-58.

[11] R. Krestel, N. Dokoohaki, Diversifying customer review rankings, Neural Networks 66(2015) 36-45.

[12] V. Ramkumar, S. Rajasekar, S. Swamynathan, Scoring products from reviews through application of fuzzy techniques, Expert Systems with Applications 37(10)(2010) 6862-6867.

[13] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, C. Jin, Red Opal: product-feature scoring from reviews, in: Proc. the 8th ACM Conference on Electronic Commerce, 2007.

[14] A. Kannan, W. Zhou, H. Chen, Linguistic Feature Selection for Review Utility Prediction, 2014.

[15] K.T. Lee, D.M. Koo, Evaluating right versus just evaluating online consumer reviews, Computers in Human Behavior 45(2015) 316-327.

[16] M. Hall, Correlation based feature selection for discrete and numeric class machine learning, in: Proc. 17th Int'l. Conf. Machine Learning, 2000.

[17] T.L. Ngo-Ye, A.P. Sinha, The influence of reviewer engagement characteristics on online review helpfulness: A text regression model, Decision Support Systems 61(1)(2014) 47-58.

[18] S.-M. Kim, P. Pantel, T. Chklovski, M. Pennacchiotti, Automatically assessing review helpfulness, in: Proc. the 2006 Conference on Empirical Methods in Natural Language Processing, 2006.

[19] A. Ghose, P.G. Ipeirotis, Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics, Knowledge and Data Engineering 23(10)(2011) 1498-1512.