Int. J. Mach. Learn. & Cyber. (2018) 9:75–84
https://doi.org/10.1007/s13042-015-0347-4

ORIGINAL ARTICLE

Sentimental feature selection for sentiment analysis of Chinese online reviews

Lijuan Zheng[1,2] · Hongwei Wang[2] · Song Gao[2]

Received: 18 September 2014 / Accepted: 5 March 2015 / Published online: 19 March 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract With the growing availability and popularity of online reviews, the sentiment analysis arises in response to the requirement of organizing useful information in speed. Feature selection directly affects the representation of online reviews and brings a lot of challenges to the domain of sentiment analysis. However, little attention has been paid to feature selection of Chinese online reviews so far. Therefore, we are motivated to explore the effects of feature selection on sentiment analysis of Chinese online reviews. Firstly, N-char-grams and N-POS-grams are selected as the potential sentimental features. Then, the improved Document Frequency method is used to select feature subsets, and the Boolean Weighting method is adopted to calculate feature weight. At last, experiments based on online reviews of mobile phone are conducted, and Chi-square test is carried out to test the significance of experimental results. The results suggest that sentiment analysis of Chinese online reviews obtains higher accuracy when taking 4-POS-grams as features. Besides that, low order N-char-grams can achieve a better performance than high order N-char-grams when taking N-char-grams as features. Furthermore, the improved document frequency achieves significant improvement in sentiment analysis of Chinese online reviews.

Keywords Online reviews · Sentiment · Feature selection · Statistical machine learning

1 Introduction

With the boost of online reviews, a large quantity of users' opinions on certain products or services are generated and spread over the Internet, thus techniques of sentiment analysis rise in response to the requirement of retrieving valuable information in speed. A growing body of research suggests the promising applications of online reviews in e-commerce, financial risk, health care and so on [1–3]. Although online reviews serve numerous functions, sentiment analysis of online reviews remains a challenging problem.

Generally, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual sentiment polarity of a document. There are two kinds of methods for sentiment analysis: semantic orientation approach and statistical machine learning approach. The semantic approach determines the sentiment of a document based on extracted sentimental words and phrases [4]. The statistical approach determines the sentiment of a document based on extracted sentimental features and the machine learning approach [5].

The statistical approach outperforms the semantic approach with respect of the accuracy of sentiment analysis [6]. When we adopt the statistical approach for sentiment analysis, every online review should be expressed as a vector containing selected sentimental features. Features selection directly affects the representation of online reviews and determines the performance of sentiment analysis. However, little attention has been paid to feature selection of Chinese online reviews so far. Therefore, this paper is motivated and focused on the effects of feature selection on sentiment analysis of Chinese online reviews.

✉ Hongwei Wang
hwwang@tongji.edu.cn

[1] School of Business, Liaocheng University, Liaocheng 252000, China

[2] School of Economics and Management, Tongji University, Shanghai 200092, China

2 Related works

Feature selection is a widely employed technique for reducing dimensionality among practitioners. It aims to choose a small subset of the relevant features from the original ones according to certain relevance evaluation criterion, which usually leads to better learning performance (e.g., higher learning accuracy for classification), lower computational cost, and better model interpretability. Feature selection is applicable to a wide range of areas from data mining [7], pattern recognition [8], regression learning [9], to sentiment analysis [10]. Sentimental features are expected to not only reflect the sentimental information but distinguish different online reviews. The effective feature extraction algorithms will facilitate sentiment analysis of online reviews. The common-used sentimental features for sentiment analysis are n-grams. We thus develop a review of previous studies based on n-gram features and feature extraction.

2.1 N-gram features for sentiment analysis

To select sentimental features is of importance for determining the performance of sentiment analysis. Various types of n-gram features have emerged for capturing sentiment cues in English reviews, including character, word, and part-of-speech (POS). Table 1 provides a summary of n-gram features used for sentiment analysis of English online reviews.

The first step of processing n-gram features in online reviews is Chinese word segmentation (CWS). Recent research focus in Chinese word segmentation has been placed on the development of statistical algorithms, such as maximum match algorithm [11], conditional random field (CRF) algorithm [12, 13], Evolutionary algorithm [14] and so on.

Character n-grams are letter sequences. For example, the word “hate” can be represented with the following letter sequences “ha, at, te, hat, ate”. While character n-grams were used mostly for style analysis previously, they have been shown to be useful in related sentiment analysis recently [15].

Word n-grams include bag-of-word n-grams (e.g., unigrams) and higher order word n-grams (e.g., bigrams, trigrams) [16, 17]. In Pang’s experiment, Unigrams, Bigrams and Unigrams + Bigrams were selected as features, and Boolean value is used as weights. The experiment showed that Unigrams are superior in the performance of sentiment analysis, but Bigrams couldnot achieve the expected accuracy [18]. Cui et al. pointed out that the corpus in Pang’s experiment was too small to take advantage of n-grams ($n \geq 3$). Then let n be 1, 2, 3, 4, 5 and 6 respectively, and the experimental results showed that high order n-grams improved accuracy of analysis [19]. Opposing to the Cui’s conclusion, Ng et al. [20] discovered that combining Unigrams and Bigrams with Trigrams improved the performance of SVM; but when using Unigrams, Bigrams and Trigrams separately, accuracy of analysis declined with the increase of order.

Part-of-speech n-grams are very useful for sentiment analysis given the pervasiveness of adjectives and adverbs. Therefore, some studies have employed word plus POS n-grams as sentimental features. Turney proposed five combinations of words containing adjectives, adverbs or so to recognize language sentiment [21]. Mullen, et al. regarded the phrases extracted from Turney’s five-sentiment-combination model as the value phrases, and applied WorldNet to calculate all the adjectives’ EVA (evaluative), POT (potency) and ACT (activity) values. The three values above together with the SO (semantic orientation) formed the features, and the experimental result indicated that this analysis method was better than the previous ones [22].

2.2 Feature extraction for sentiment analysis

Noise and redundancy in the feature space prevent many quality features from being incorporated due to computational limitations. Therefore, larger n-gram feature sets require the use of feature extraction algorithms to extract quality feature subsets. Most feature extraction algorithms such as document frequency (DF), information gain (IG), Chi-square statistic (CHI), and mutual information (MI) have been used for sentiment analysis [23].

Document frequency algorithm is to sum up the number of documents which contain the feature. IG algorithm is an entropy-based feature extraction algorithm. The larger the entropy of a review is, the greater the uncertainty of classifying the document will be. In the CHI algorithm, features are filtered by measuring the association degree between itself and a definite category. The basic principle of MI is similar to the CHI algorithm, measuring the co-occurrence between features and categories. The features

Table 1 N-gram features used for sentiment analysis

N-gram category	Examples	Prior studies
Character n-grams	C, h, ch, ca, ar, cha, ara	[15]
Word n-grams	Character, character of, the character of	[16–20]
POS n-grams	The- det, character- noun, of- prep	[21, 22]

selected by MI method have the characteristics of high frequency of occurrence in a certain category, but low frequency in the others.

Ng et al. [24] compared DF, MI, IG, and the CHI methods, and the results showed that DF, IG and CHI were better than MI. Liu et al. [25] compared DF, IG, and CHI methods, and the experimental results indicated that DF was better than CHI and IG. Yao et al. [26] used DF, IG, CHI and MI to reduce dimensionality, and the experimental result showed that DF was the best one and MI was not suitable for the sentiment analysis of Chinese online reviews. Because DF method has been shown to work well for sentiment analysis, it is used to select feature subsets in this paper.

2.3 Research gaps and questions

Sentiment analysis research requires large quantities of sentimental features. Therefore, it is a key to select features properly characterizing the online reviews. Based on the literatures, we have identified appropriate gaps and questions.

Most studies have used limited sets of n-gram features, typically employing one category in sentiment analysis research in English. Due to the difference of language structures between English and Chinese, Chinese has its unique way of sentimental expression so that the n-gram features characterizing English online reviews are unable to be directly applied to Chinese ones. Moreover, few studies have attempted to integrate these heterogeneous n-gram categories into combined feature sets for the sentiment analysis of online reviews.

Noise and redundancy in the feature space prevent many quality features from being incorporated due to computational limitations. Therefore, feature extraction algorithms have been used for sentiment analysis, and DF method has been demonstrated to work well for extracting of feature subsets. However, few studies have attempted to discuss the factors influencing the performance of DF extraction algorithm.

3 Proposed approach

In this paper, statistical machine learning techniques are incorporated into the domain of Chinese online comments to automatically classify user reviews as positive or negative. The basic process is as follows: (1) divide the text into some segments; (2) select a sequence of sentimental features according to the training sample set $T = T (t_1, t_2,\ldots, t_n)$; (3) evaluate every text from the training sample set and test sample set to generate vector $D = D (t_1, w_1, t_2, w_2, \ldots, t_n, w_n)$, abbreviated to $D = D (w_1, w_2, \ldots, w_n)$, where w_k is the weight of feature t_k. (4) transform unstructured online reviews into structured data.

3.1 Sentimental features for sentiment analysis

We choose the ICTCLAS system developed by the Institute of Computing Technology of Chinese Academy (http://ictclas.org/) to do the word segmentation and POS tagging. After word segmentation, some n-grams are selected as initial sentimental features. The n-grams with no contribution to sentiment analysis like stop words are dropped. We incorporate a rich set of n-gram features, consisted of all the categories discussed in the literature reviews. The feature sets are shown in Table 2.

1. N-char-grams

N-char-grams take characters (letters, spaces, symbols, etc.) as the basic units. N-char-grams of size 1 is referred to as a “1-char-grams”; size 2 is a “2-char-grams”; size 3 is a “3-char-grams”; and size 4 or more is simply called an “n-char-grams”. We combined adjacent characters into N-char-grams to discern sentiment of the document, for example, “不” followed by “错” becomes “不错” in the 2-char-grams.

The advantages of N-char-grams are as follows: Linguistics processing such as POS tagging of the document is unnecessary; It has good fault tolerance in spelling mistakes and does not rely on any prior knowledge; Dictionaries and regulations are unnecessary. An example of N-char-grams is shown in Table 3.

Table 2 The feature sets used for sentiment analysis

Label			Description
N-char-grams	1-Char-grams		Character sequences
	2-Char-grams		
	3-Char-grams		
N-POS-grams	1-POS-grams	“Noun-grams”/“adjective-grams”/"verb-grams”/“adverb-grams”	Combination of POS n-grams and word n-grams
	2-POS-grams	“Adjective-grams” + “adverb-grams”	
	3-POS-grams	“Verb-grams” + “adjective-grams” + “adverb-grams”	
	4-POS-grams	“Noun-grams” + “verb-grams” + “adjective-grams” + “adverb-grams”	

Table 3 Examples of feature selection

File	1-Char-grams	2-Char-grams	3-Char-grams
性价比不错。(the performance-price ratio is pretty good)	性 价 比 不 错	性价 价比 比不 不错	性价比 价比不 比不错

Table 4 Examples of N-POS-grams

File	Noun-grams	Adjective-grams	Verb-grams	Adverb-grams
按键后总感觉顿一下才有反应,电池不耐用,两天充一次。(it is slow to response after pushing buttons. The battery is not durable and needs to get charged every two days)	键 一下 电池 两天	不耐用	按 感觉 顿 充	总 才
看电影很方便,听歌感觉不错。(it is easy to see a movie and feels good to listen to the music)	电影 歌	方便 不错	看 听 感觉	很

2. N-POS-grams

N-POS-grams employ word n-grams plus POS n-grams as sentimental features. We choose the ICTCLAS system developed by the Institute of Computing Technology of Chinese Academy (http://ictclas.org/) to do the word segmentation and POS tagging, and then select the N-POS-grams according to the following rules. Four kinds of N-POS-grams (adjectives, adverbs, verbs and nouns) are selected as potential features from the training set.

To consider the importance of the negative words, such as “不” “不是” (“no” “not”), these negative words will be treated as a whole with their next contiguous adjectives, adverbs or verbs. For example, “我不喜欢手机的外形, 看上去不舒服。” (“I dislike the appearance of the mobile phone because it looks uncomfortable”). The features should be “不喜欢” (“dislike”) and “不舒服” (“uncomfortable”), rather than “喜欢” (“like”) and “舒服” (“comfortable”).

The POS of a word often varies with different context of a sentence. The same word with different POS should be treated as different features. For example, “很出色” (“perfect”) appearing in “显示屏效果很出色” (“the performance of the display screen is perfect”) is an adjective; while “很出色” (“perfectly”) in “功能很出色地满足了日常需要” (“this function satisfies daily needs perfectly”) is an adverb. “很出色” (“perfect”) being adjective and adverb are discerned as two different features.

The words following an adjective or adverb like “的”, “得” and “地” will be deleted when selecting feature words. For example, if “好看的” appears in a sentence, we only select “好看” as the adjective. Two examples of N-POS-grams are shown in Table 4.

3.2 Feature extraction for sentiment analysis

Large feature space span hundreds of thousands of features, feature extraction can provide greater insight into important class attributes, extract quality feature subsets, and potentially improve accuracy of sentiment analysis. DF algorithm has been shown to work well for extracting of feature subsets. However, few studies have attempted to discuss the factors influencing the performance of DF extraction algorithm. DF extraction algorithm is to sum up the number of documents which contain certain features. Therefore the DF value indicates the frequency of a feature t_i in all reviews. The calculation equation is as follows.

$$DF(t_i) = \sum_{j=1}^{M} N\left(C_j, t_i\right) \tag{1}$$

where $M = 2$, C_j denotes class j, which is “Positive” (C_1) or “Native” (C_2), and $N(C_j, t_i)$ denotes the number of reviews with feature t_i that belongs to C_j.

1. The factors influencing the performance of DF extraction algorithm

In the DF extraction algorithm, the key statistics are the number of pitive/negative corpora including the feature t_i. However, these statistics affected invisibly by the following two factors:

The different number of positive corpora and negative corpora is an important factor influencing the DF extraction algorithm. If there are more negative corpora than positive ones, the values of negative features would be inevitably exaggerated, thus it would increase the probability of extracting negative features.

The length of corpora is another important factor influencing the DF extraction algorithm. If the length of negative corpora is longer than that of positive one, there are more features in negative corpora than in positive one, thus the negative features would have more probability to be extracted and occupy the chance of positive features.

2. Improved DF extraction algorithm

The calculation of $N\left(C_j, t_i\right)$ is adjusted in order to eliminate the influence of the two factors above. $NFT(C_j, t_i)$ instead of $N\left(C_j, t_i\right)$ is used in the DF algorithm. The calculation equation is as follows.

$$FT\left(C_j\right) = \sum_{i=1}^{N_j} N\left(C_j, t_i\right) \tag{2}$$

$$NFT\left(C_j, t_i\right) = \frac{FT\left(C_j\right)}{NC_j} \times N\left(C_j, t_i\right) \tag{3}$$

$$IDF\left(t_i\right) = \sum_{j=1}^{M} NFT\left(C_j, t_i\right) \tag{4}$$

where N_j and NC_j respectively denotes the number of features in the corpora belonging to class C_j and the number of the corpora belonging to class C_j. FT (C_j) indicates the total number of features selected from each class (positive or negative). $NFT\left(C_j, t_i\right)$ indicates the relationship between the average length and the number of corpora belonging to each class. $IDF(t_i)$ represents the improved DF extraction algorithm.

3.3 Feature weigh calculation for sentiment analysis

Feature weighting methods include Boolean weights, term frequency (TF), term frequency–inverse document frequency (TF–IDF), and so on. Pang adopted Boolean weighting method to do experiment and the accuracy of sentiment analysis reached 82.9 %, which was better than the other weighting methods [18]. The reason is that positive or negative inclination depends on whether the words appear in the language or not, rather than their frequency of occurrences. Therefore, after feature selection, we use Boolean weight method to set weights.

3.4 SVM operation for sentiment analysis

Statistical machine learning can determine the sentiment of a document based on sentimental features and the machine learning approach. The commonly used approaches for text classification are support vector machines (SVM), maximum entropy (ME), naïve bayes (NB), recurrent neural networks (FNN) and so on [27–32].

Past studies show that SVM outperforms other classifiers in term of the classification performance, especially in the case of limited training sample. For this reason, we prefer SVM for sentiment analysis. We select the LIBSVM, a SVM classifier developed by Taiwan University (www.csientu.edu.tw/~cjlin/libsvm/) to conduct the experiment. Specific steps are illustrated in Fig. 1.

4 Experimental and evaluations

4.1 Experiment procedures

We conduct sentiment analysis experiments on mobile phone reviews. We create a corpus of mobile phone reviews, which are taken from a famous electronic commerce website JingDong (www.360buy.com). A crawler is developed by Java to randomly download 1500 positive reviews and 1500 negative reviews. The corpus is divided into four sets: the positive reviews as a training corpus, the negative reviews as a training corpus, the positive reviews as a test corpus, and the negative reviews as a test corpus. For example,

Positive review: 外观漂亮,屏幕够大! (the appearance is good, and the screen is large enough!)

Negative review: 电池不好,触摸屏不够灵敏。 (the battery is not good, and the touch screen is not sensitive enough).

The basic process of the sentiment analysis experiments is illustrated in Fig. 2. In this paper, we first select N-char-grams and N-POS-grams as the potential features of sentiment text, then adopt DF extraction algorithm and IDF extraction algorithm for feature subset selection, and at last, obtain various vectors for experiments under different thresholds.

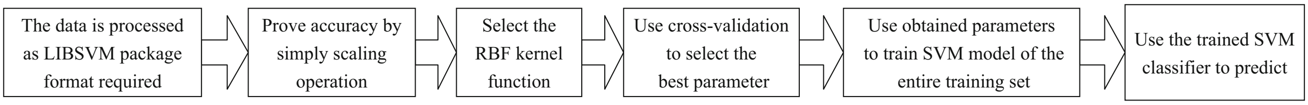
The data is processed as LIBSVM package format required
Prove accuracy by simply scaling operation
Select the RBF kernel function
Use cross-validation to select the best parameter
Use obtained parameters to train SVM model of the entire training set
Use the trained SVM classifier to predict

Fig. 1 SVM operation process

Fig. 2 The basic process of experiment

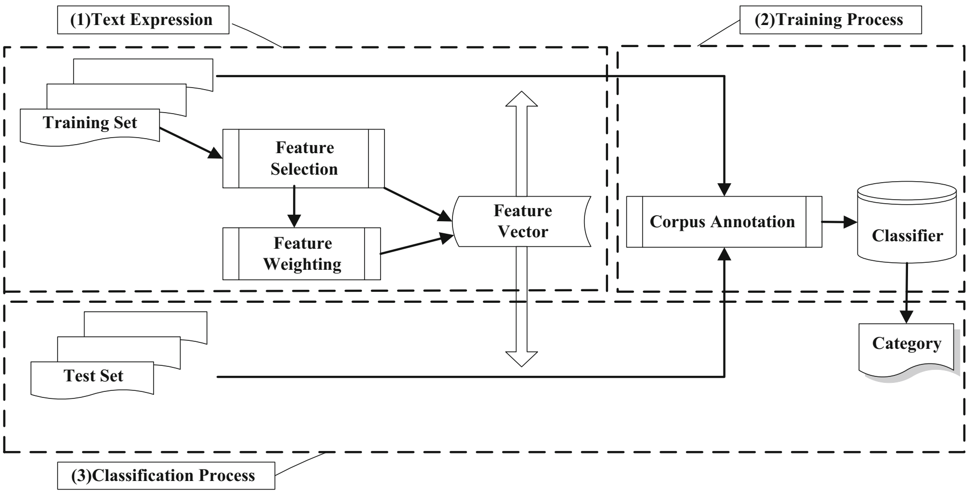

Table 5 The performance of 1-POS-grams and 4-POS-grams

POS	Accuracy (%)				
	124–3	123–4	234–1	341–2	Average
Nouns-grams	75.67	70.00	73.33	78.67	74.42
Verbs-grams	74.33	80.67	79.67	77.00	77.92
Adjective-grams	89.67	93.33	93.67	92.67	92.33
Adverb-grams	82.00	91.67	90.00	90.00	88.42
4-POS-grams (combining noun, verb, adjective and adverb)	89.33	94.67	96.32	92.33	93.16
χ^2					69.34
P					0.00

4.2 Comparative experiments on different part of speech

This paper chooses the ICTCLAS System developed by the Institute of Computing Technology of Chinese Academy (http://ictclas.org/) to do the word segmentation and POS tagging. And four types of POS are used to select features: noun, verb, adjective and adverb. Each type supposedly has different contribution to sentiment analysis, thus it is necessary to analyze the contribution of different POS to sentiment analysis.

To remove the influence of the randomness of single-validation on experimental results, the experiment adopts cross-validation. 1200 reviews are divided into 4 groups randomly, each containing 300 reviews (half are positive). Then three groups of them are used as training set in turns and the remaining group as test set. Average result of the four groups is the final accuracy. In Table 5, for example, "124–3" means "group 1", "group 2" and "group 3" as training set, and "group 4" as testing set. Nouns, adjectives, adverbs, and verbs are selected respectively as features, and DF algorithm is adopted to select top 150 features. The results are shown in Table 5 and Fig. 3.

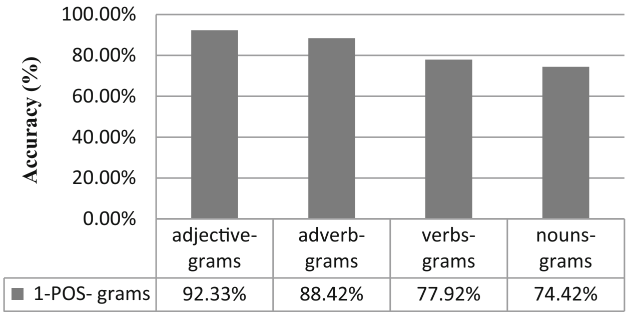

Fig. 3 The performance of 1-POS-grams

The table on the right of the figure shows the average percentage accuracy, and Chi-square test points to significant differences in accuracy of sentiment analysis when using 1-POS- grams and 4-POS-grams as features respectively.

The average values of Table 5 indicate that the performance is best when adopting 4-POS-grams to select features. But it takes much time to select all the nouns, verbs, adjectives and adverbs from massive reviews.

The average values of Fig. 3 indicate that when selecting the adjective-grams as the features, a better performance will be obtained. The result is consistent with our common

Table 6 The performance of N-POS-grams

POS	Accuracy (%)				
	124–3	123–4	234–1	341–2	Average
4-POS-grams (combining noun, verb, adjective and adverb)	89.33	94.67	96.32	92.33	93.16
3-POS-grams (excluding nouns)	93.67	94.33	96.33	93.33	94.42
2-POS-grams (excluding nouns and verbs)	90.67	92.00	94.00	94.33	92.75
1-POS-grams (adjective-grams)	89.67	93.33	93.67	92.67	92.35
χ^2					1.07
P					0.78

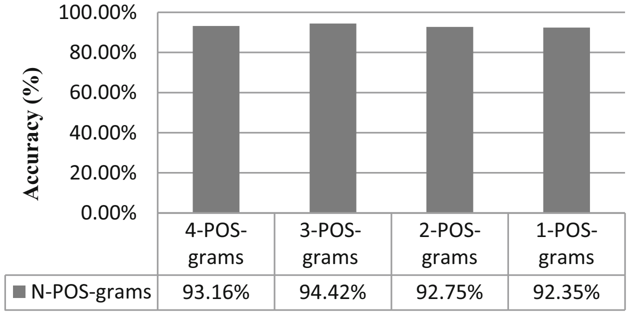

Fig. 4 The performance of N-POS-grams

sense that adjectives contain more sentiment information and makes more contribution to sentiment analysis than other POS.

4.3 Comparative experiments on different N-POS-grams

The average values of Table 5 and Fig. 3 indicate that the accuracy when selecting nouns or verbs as features is low. The low accuracy means that the contributions of nouns and verbs to analysis are limited, and they even have negative influence on overall accuracy. So, further experiments are conducted after excluding verbs and nouns. The changes of accuracy are shown in Table 6 and Fig. 4.

Chi-square test shows that there is no significant difference in accuracy when using 1-POS-grams, 2-POS-grams, 3-POS-grams and 4-POS-grams as features respectively. From the average results of Table 6, we can know,

1. After excluding nouns, the performance of 3-POS-grams is satisfied. The possible reason is as follows: most nouns selected from the reviews, such as “功能” (“function”) and “电池” (“battery”), are neutral. These words appear frequently but have little contribution to analysis.
2. After excluding nouns and verbs, accuracy of 2-POS-grams is also higher. The possible reason is that the combination of adjectives and adverbs is more sentimental. For example, “非常漂亮” (“very beautiful”), the combination can reflect the feeling of reviewer better than a single adverb.

4.4 Comparative experiments on different N-char-grams

N-char-grams are selected as the sentimental features, and DF algorithm is adopted to select top 150 features. The analysis results are shown in Table 7 and Fig. 5.

From the Table 7 and Fig. 5, we can know:

1. There are significant differences in accuracy between 4-POS-grams and N-char-grams, and the performance of sentiment analysis is best when adopting 4-POS-grams as sentimental features.
2. Two-word phrases are usually thought to be used most frequently in Chinese, so accuracy of 2-char-grams should be the best. But the experiment shows an opposite result when selecting N-char-grams as features, accuracy declines with the increase of the order. The accuracy of 1-char-grams is better than that of 2-char-grams and close to that of 4-POS-grams. Possible reasons are as follows:

If 4-POS-grams are selected as features, there will be many individual characters, such as “快” (“fast”), “紧” (“tight”) and “难” (“hard”). These characters are obvious in sentiment polarity and have greater influence on sentiment analysis. 1-char-grams choose frequent individual characters as features, so the features selected by 1-char-grams and by 4-POS-grams have much in common.

If 2-char-grams are selected as features, the importance of features would be underestimated. For example, “大方” (“generous”) and “大气” (“liberality”), both expressing the same sentiment polarity, appear ten times and 15 times respectively. When DF algorithm is used to the features ranking, they rank the 150th and the 100th respectively. But if we select 1-char-grams as the features, “大” (“big”) will appear 25 times with the 80th ranking. The 80th ranking reflects the importance of selected features more accurately to some extent.

Table 7 The performance of 4-POS-grams and N-char-grams

The number of features	Accuracy (%)				χ^2	P
	4-POS-grams	1-Char-grams	2-Char-grams	3-Char-grams		
50	87.29	89.67	81.33	62.33	86.99	0.00
100	93.31	93.00	85.67	65.67	113.60	0.00
150	96.32	95.00	85.67	68.00	127.10	0.00
200	94.65	94.00	86.33	70.33	95.81	0.00
250	95.65	94.00	86.33	71.33	95.03	0.00
χ^2	22.95	7.09	3.76	6.96		
P	0.00	0.13	0.44	0.14		

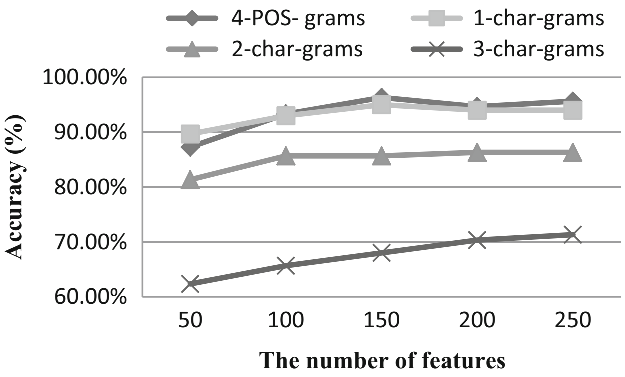

Fig. 5 The accuracy of 4-POS-grams and N-char-grams

If 3-char-grams are selected as features, the feature vectors corresponding to some of 3-char-gram documents (particularly the short ones) have many zeroes. This suggests that the main problem with the 3-char-gram model is likely to be data sparseness.

3. The number of selected features has a certain influence on accuracy. As Fig. 5 shows, at the beginning, the accuracy increases significantly with the increase of features, but the increase rate declines afterward, and the inflection occurs around 150 or 200. So the trade-off between efficiency and accuracy should be taken into consideration during the feature selection.

4.5 Comparative experiments on the DF algorithm and the IDF algorithm

In order to test and verify the IDF algorithm, comparative experiments are conducted based on mobile phone reviews. 1-char-grams are selected as the sentimental features, and DF/IDF algorithm is adopted to select top 200 features.

Table 8 demonstrates that there are significant differences in accuracy between DF algorithm and IDF algorithm. In other word, the improved DF achieves significant improvement in sentiment analysis of Chinese online reviews.

5 Conclusions

This paper takes Chinese online reviews as the research object, selects N-POS-grams and N-char-grams as features to characterize the sentiment text, adopts DF extraction algorithm or IDF extraction algorithm for feature subset selection, and studies the impact of feature selection on sentiment analysis through a series of experiments. The results demonstrate that:

1. If N-char-grams are selected as the features, precision would decline with the increase of order, i.e., 1-char-grams > 2-char-grams > 3-char-grams. And the accuracy of 1-char-grams is close to that of 4-POS-grams.
2. The best performance is achieved when combining noun, adjectives, adverbs, and verbs as the sentimental features. And adjective contains more sentiment information and makes greater contribution to sentiment analysis than other POS. Therefore, if N-POS-grams are selected as features, it is advisable to adopt adjectives as features to save time and improve efficiency.
3. The number of features has a certain influence on accuracy analysis, but not the more the better. So the trade-off between efficiency and accuracy should be taken into consideration during the feature selection.
4. The improved DF achieves significant improvement in sentiment analysis of Chinese online reviews.

Further research will focus on the following aspects:

1. Features based on grammatical structures may play an important role in the sentimental expression, such as “性价比高” (“performance-price ratio is high”) and “操作简单” (“operation is simple”), which are subject-predicate structures. Therefore, the effect of features

Table 8 Comparison of the DF algorithm and the IDF algorithm

Extraction algorithms	Accuracy (%)
DF	94.00 %
IDF	97.33 %
χ^2	4.02
P	0.02

based on grammatical structures on sentiment analysis is worth discussion.

2. The current study is based on the paragraph level. If both positive sentence and negative sentence appear in one paragraph, the analysis results will be influenced. Therefore, the study based on sentence level is also worth discussion.

Acknowledgments This work is partially supported by the NSFC Grant 70971099 and 71371144, the fundamental research funds for the Central Universities (1200219198), and Shanghai Philosophy and Social Science Planning Projects (2013BGL004).

References

1. Li X, Xie H, Chen L, Wang J, Deng X (2014) News impact on stock price return via sentiment analysis. Knowl Based Syst 69:14–23
2. Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: the role of reviewer identity disclosure in electronic markets. Inf Syst Res 19 (3):291–313
3. Greaves F, Ramirez D, Millett C, Darzi A, Donaldson L (2013) Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ Qual Saf 22(3):251–255
4. Yang L, Xu LD, Shi ZZ (2012) An enhanced dynamic hash trie algorithm for lexicon search. Enterpr Inf Syst 6(4):419–432
5. Li HX, Xu LD, Wang JY, Mo ZW (2003) Feature space theory in data mining: transformations between extensions and intensions in knowledge representation. Expert Syst 20(2):60–71
6. Ye Q, Lin B, Li YJ (2005) Sentiment classification for chinese reviews: a comparison between SVM and semantic approaches. In: proceedings of the 4th international conference on machine learning and cybernetics. NY, USA: IEEE Press, pp 2341–2346
7. Xie ZX, Xu Y (2014) Sparse group LASSO based uncertain feature selection. Int J Mach Learn Cybern 5(2):201–210
8. Subrahmanya N, Shin YC (2013) A variational bayesian framework for group feature selection. Int J Mach Learn Cybern 4(6):609–619
9. Wei P, Ma PJ, Hu QH, Su XH (2014) Comparative analysis on margin based feature selection algorithms. Int J Mach Learn Cybern 5(3):339–367
10. Abbasi A, Chen H, Salem A (2008) Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans Inf Syst (TOIS) 26(3):12–21
11. Huang C (1997) Word segmentation issues in chinese information processing. Applied linguistics (in Chinese), p 1
12. Zhao H, Huang C, Li M (2006) An improved chinese word segmentation system with conditional random field. In: proceedings of the 5th SIGNAN workshop on Chinese language processing. Sydney, Australia, pp 162–165
13. Gao J, Li M, Wu A, Huang C (2005) Chinese word segmentation and named entity recognition: a pragmatic approach. Comput Linguist 31(4):531–574
14. Zhang D (2013) An evolutionary approach to automatic chinese text segmentation. In: ninth international conference on natural computation
15. Abbasi A, Chen H, Thoms S, Fu T (2008) Affect analysis of web forums and blogs using correlation ensembles. IEEE Trans Knowl Data Eng 20(9):1168–1180
16. Ghiassi M, Skinner J, Zimbra D (2013) Twitter brand sentiment analysis: a hybrid system using N-gram analysis and dynamic artificial neural network. Expert Syst Appl 40(16): 6266–6282
17. Remus R, Rill S (2013) Data-driven vs. dictionary-based word n-gram feature induction for sentiment analysis. In: 25th international conference of the German-Society-for -Computational-Linguistics-and-Language-Technology (GSCL). Darmstadt, Germany, pp 25–27
18. Pang B, Lee L, Vaithyanathan S (2002) Sentiment classification using machine learning techniques. In: proceedings of the conference on empirical methods in natural language processing, Philadelphia, US, pp 79–86
19. Cui H, Mittal V, Datar M (2006) Comparative experiments on sentiment classification for online product reviews. In: proceedings of the 21st national conference on artificial intelligence (AAAI-06), Boston, USA, pp 1265–1270
20. Ng V, Dasgupta S, Arifin N (2006) Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In: proceedings of the COLING/ACL main conference poster sessions, Association for Computational Linguistics, Morristown, NJ, USA, pp 611–618
21. Turney P (2002) Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of review. In: proceedings of the 40th annual meeting of the association for computational linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 417–424
22. Mullen T, Collier N (2004) Sentiment analysis using support vector machines with diverse information sources. In: proceedings of the 2004 conference on empirical methods in natural language processing, Barcelona, Spain, pp 412–418
23. Ng V, Dasgupta S, Arifin SMN (2006) Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews.In: proceedings conference computational linguistics, association for computational linguistics, pp 611–618
24. Ng HT, Goh WB, Low KL (1997) Feature selection, perceptron learning and a usability case study for text categorization. In: proceedings of the 20th annual Int'l ACM SIGIR conference on research and development in information retrieval, pp 67–73
25. Liu X (2011) Sentiment polarity classification on chinese reviews based on statistic natural language. Master's Degree Thesis, Tongji University
26. Wang HW, Yin P, Yao JN (2013) Text feature selection for sentiment classification of chinese online reviews. J Exp Theor Artif Intell 25(4):425–439
27. Rückstieß T, Osendorfer C, Smagt PVD (2013) Minimizing data consumption with sequential online feature selection. Int J Mach Learn Cybern 4(3):235–243
28. Xia HS, Peng LY (2009) SVM-based comments classification and mining of virtual community: for case of sentiment classification of hotel reviews. In: proceedings of the Int'l symposium on intelligent information systems and applications, pp 507–511

29. Phienthrakul T, Kijsirikul B, Takamura H, Okumura M (2009) Sentiment classification with support vector machines and multiple kernel functions. Lect Notes Computer Sci 58:583–592
30. Ye Q, Zhang ZQ, Law R (2009) Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst Appl 36(3):6527–6535
31. Moraes R, Valiati JF, Gaviao N, Wilson P (2013) Document-level sentiment classification: an empirical comparison between SVM and ANN. Expert Syst Appl 40(2):621–633
32. Wan X (2011) Bilingual co-training for sentiment classification of chinese product reviews. Comput Linguist 37(3):587–616