Send Orders for Reprints to reprints@benthamscience.ae

Open Access

Combining Semantic Comprehension and Machine Learning for Chinese Sentiment Classification

Jianfeng Xu[1,2] ,Yuan Xu[1], Yuanjian Zhang[1,2,*] and Yu Li[1]

[1]*Software College of Nanchang University, Nanchang 330047, China;* [2]*School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*

Abstract: Semantic comprehension-based and machine learning based are two major methods for the classification of Chinese sentiment. The advantage of semantic comprehension-based method is that it can classify text among domains and achieve satisfied portability. However, the accuracy of classification is limited. Although the accuracy derived from supervised machine learning method is much better, the portability is rather poor due to randomly selection of samples and subjective labeling of semantic orientation. In this paper, a hybrid framework combining the advantages of the two methods was proposed. The text features were extracted preliminary based on semantic comprehension and were optimized by a novel information gain method. The features expressed in vector space model were integrated with traditional machine learning algorithm. Experiments show that support vector machine has the best discriminative power compared to other machine learning algorithms. Additionally, this framework improves portability and accuracy as compared to both semantic comprehension-based methods and machine learning based methods.

Keywords: Chinese information processing, sentiment analysis, semantic comprehension, machine learning.

1. INTRODUCTION

Currently, text sentiment classification method is roughly divided into two categories: semantic comprehension-based and supervised machine learning-based. In the literature [1], the merit and weakness of the two major methods are critically analyzed. The advantages of semantic comprehension-based method are unsupervised, no training corpus are required and normally with stable portability because of well-constructed sentiment lexicon, however the accuracy is not promising. Supervised machine learning method has better accuracy, but it needs to tag corpus manually before classification, and its portability is rather poor.

The first solution of semantic comprehension-based is phrase template-based. In the literature [2], the relation between orientation and linguistics property of conjunctions is measured. The orientation of those words are obtained through clustering method. In the literature [3], Pointwise Mutual Information is adopted to evaluate the similarity between phrases and paradigm words. It is proved that the result is better if the various dependencies between words are considered. In the literature [4], the linguistic nature of conjunctions is considered during the similarity measurement between words and selected paradigm words. The synonym structure in WordNet is designed to decrease the possibility of misclassification.

The second solution of semantic comprehension-based is semantic pattern-based. In the literature [5], a table of with certain orientation and a pattern library of words with confirmed orientation is constructed for the matching with predefined rules. In the literature [6] a hybrid framework which combines a context-aware sentiment lexicon with standard HowNet is proposed. It is proved to be more rational since both general idea and domain view is included. In the literature [7], Fuzzy Domain Sentiment Ontology is proposed to analyze the polarity on different components of product, which is also a combination of sentiment dictionary HowNet and synonym lexicon.

Method based on machine learning mainly depended on traditional text classification techniques. In the literature [8], it is proved that SVM method is the most efficient compared to Maximum Entropy and Naive Bayes. In the literature [9], terms with strong sentiment degree are used as input for the construction of a SVM-based classifier. In the literature [10], multiple probabilistic reasoning model (M-PRM) is proposed for the sentiment recognition of online micro-movie review. The M-PRM can be more efficient if more iteration is performed.

In allusion to the problems mentioned above, this paper proposed a hybrid framework for Chinese sentiment classification, which uses the idea of semi-supervised learning by combining the advantages of semantic comprehension with machine learning. The semi-supervised learning mechanism can keep the accuracy and require limited effort to label the text meanwhile and it is proved to be efficient on opinion mining on products [11]. After describing the essence of the proposed framework in Section 2, in Section 3 a series of

1874-4443/15 2015 Bentham Open

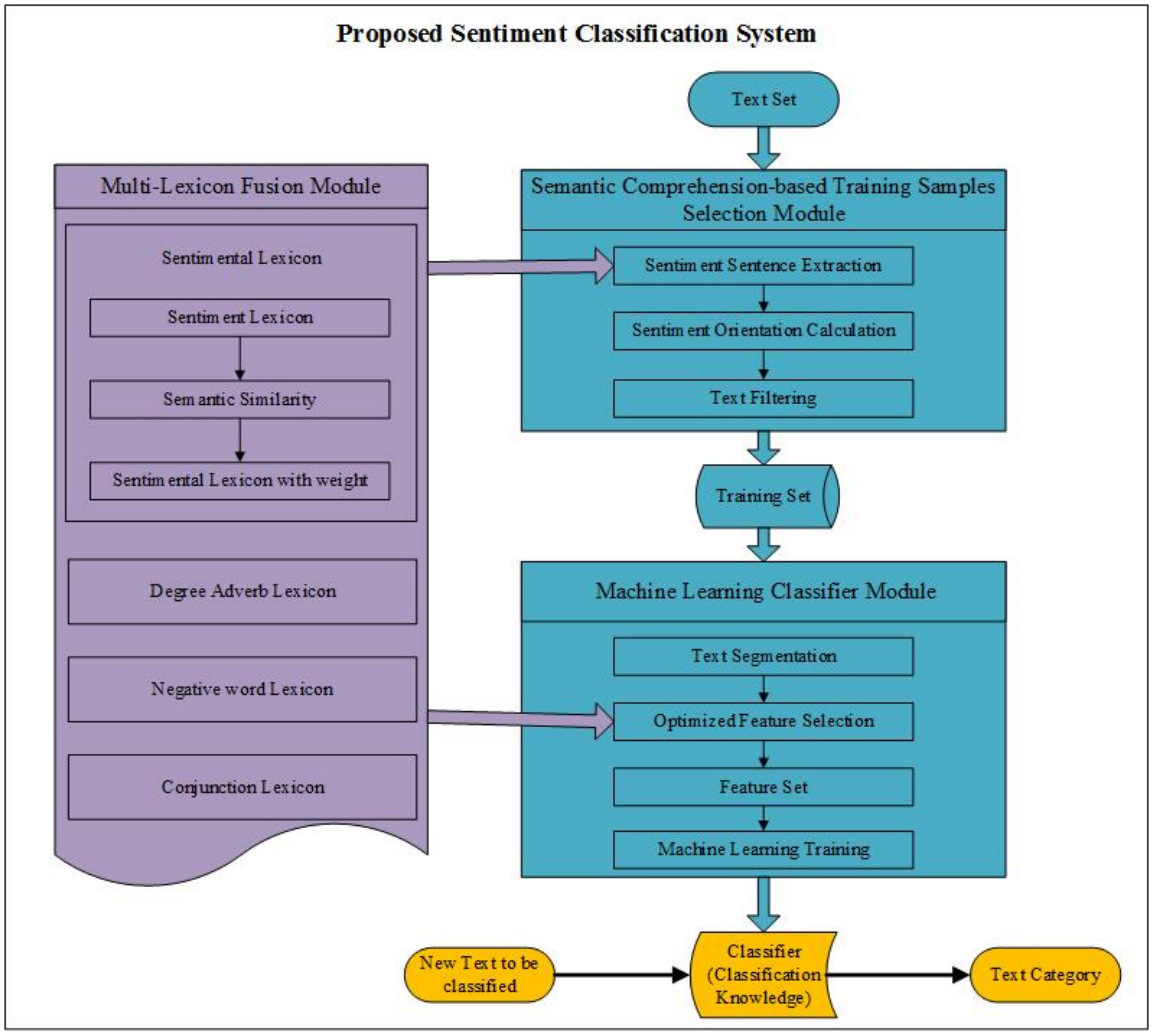

Fig. (1). Hybrid Algorithm Framework.

experiments are conducted to validate its performance in terms of classification and portability. Finally, the conclusion is listed in Section 4.

2. HYBRID ALGORITHM FRAMEWORK

Based on the limitation of the semantic comprehension-based and machine learning-based for the identification of text sentiment, a hybrid classification algorithm is proposed to combine the advantages of the two methodologies, as is shown in Fig. (**1**). The classifier is trained after two main module named as semantic comprehension-based training samples selection module and machine learning classifier module respectively.

2.1. Main Functions of Multi-Lexicon Fusion

Sentimental Lexicon: Words are selected from HowNet and words not commonly used are removed manuanlly. These selected words are summarized into initial sentimental lexicon.

Semantic Similarity: Paradigm words lexicon is related to three aspects named high word frequency, strong sentiment orientation and certain scale number respectively. The weights of sentiment words are calculated by using semantic similarity calculation formula of HowNet. In this paper, we adopt the weighted formular to balance the proportion of commendatory and derogatory terms in paradigm words. Therefore, adjustable parameter α and β are introduced, as is shown in (1)

$$O(w)=\alpha\times\frac{1}{m}\sum_{i=1}^{m}Sim(w,POS_i)-\beta\times\frac{1}{n}\sum_{j=1}^{n}Sim(w,NEG_j) \qquad (1)$$

Where *POS* is commendatory set and *NEG* is derogatory set, $POS_i \in POS$, $NEG_j \in NEG$, m is the number of commendatory terms and n is the number of derogatory terms, $Sim(w,POS_i)$ and $Sim(w,NEG_j)$ are similarity calculation formula of HowNet, α and β are adjustable parameters of formula with adjustment. Let threshold be 0, if $O(w)>0$ then it is commendatory, if $O(w)<0$ then it is derogatory, otherwise it is neutral.

Table 1. Notations used in Text Sentiment Orientation Algorithm.

Notation	Meaning
$w_i(1 \le i \le n)$	word sets after segmentation
$S=\{w_1,w_2,w_3,...,w_n\}$	a sentiment sentence
$O(w_i)$	the sentiment value of w_i
$E=\{w_1,w_2,w_3,...,w_z\}$	sentimental lexicon
$O(w_j)'(1 \le j \le z)$	the weight of w_j
$N=\{n_1,n_2,n_3,...,n_m\}$	negative word lexicon
$D=\{d_1,d_2,d_3,...,d_p\}$	degree adverb lexicon
$L=\{l_1,l_2,l_3,...,l_y\}$	conjunction lexicon
$O(w_i)'$	weight after HowNet similarity calculation
D+W	degree adverb+sentiment word
N+W	negative word+sentiment word
N+D+W	negative word+degree adverb+sentiment word
D+N+W	degree adverb+negative word+sentiment word

Multi-lexicon Fusion: Error or low-degree words in sentimental lexicon are removed and a multi-lexicon module is developed by fusing degree adverb lexicon, negative word lexicon and conjunctions lexicon. It is acknowledged that the polarity of some words may be opposite if they are used to modify different objects. Therefore, after normalize the $O(w)$ of the rest sentiment words to [-1,+1], the dynamism of polarity is adjusted according to the formular (2).

$$O(w)'' = O(w)' \times c \qquad (2)$$

Where $O(w)'$ is the normalized polarity weight of dynamic words and $O(w)''$ is polarity weight of dynamic words with modifiers. If the word polarity is different from the origin, let c be -1, otherwise c equals to 1.

For degree adverb lexicon, different sentiment value is assigned to terms so that the $O(w)$ of a sentence can be adjusted. For conjunction lexicon, words are categorized into three groups named coordinating, progressive and adversative respectively with three corresponding value assigned to represent corresponding sentiment level.

2.2. Main Functions of Semantic Comprehension-Based Training Samples Selection

Sentiment Sentence Extraction: Sentiment sentences are extracted based on common collocation patterns and sentimental lexicon. There are nine commonly used patterns, includes n+adj; n+v+adj; adj+of+v+n; adj+of+v; v+of+adj; v+n+adj; v+adj and adv+adj.

Sentiment Orientation Calculation: With the combination of sentence pattern collocations and multi-lexicon module, this value is calculated by sentiment orientation calculation algorithm. The notations and corresponding alogirthm are shown in Table **1**:

The following is the description of sentiment orientation calculation algorithm.

Algorithm *:TSOC(S,E,L,N,O(wj)',D)*

Input: $S=\{w_1,w_2,w_3,...,w_n\}$, $E=\{w_1,w_2,w_3,...,w_z\}$, $L=\{l_1,l_2,l_3,...,l_y\}$, $N=\{n_1,n_2,n_3,...,n_m\}$, $O(w_j)'(1 \le j \le z)$, $D=\{d_1,d_2,d_3,...,d_p\}$

Output: $\sum O(S)$

Begin

(1) Traverse the sentiment sentence set and all words of each sentence, if $w_i \in E$ and it does not belong to dynamic word then $O(w_i)=O(w_i)'$, or if $w_i \in E$ and it belongs to dynamic word then get $O(w_i)$ according to (2); if $w_i \notin E$ and $w_i \in (adj,v,n)$, then normalize $O(w_i)$. When the value is greater than 0.05, let it be mined potential sentiment word and give it a weight. Do not assignment w_i in other conditions.

(2) In sentiment sentence, if $\exists w_i \in D$ and $w_i \notin N$. The calculation formula is as following:

If $O(d_p)<0, O(w_i)>0$,

then $D+W=(1+d_p)\times O(w_i)$.

If $O(d_p)>0, O(w_i)>0$,

then $D+W=O(w_i)+(1-O(w_i))\times O(d_p)$.

If $O(d_p)>0, O(w_i)<0$,

then $D+W=O(w_i)-(1+O(w_i))\times O(d_p)$.

(3) In sentiment sentence, if $\exists w_i \in N and w_i \notin D$. The calculation formula is as following:

$$N+W=-\frac{2}{3}\times O(w_i).$$

(4) In sentiment sentence, if $\exists w_i \in D and w_i \in N$. The calculation formula is as following:

$$N+D+W=-0.5\times O(w_i).$$

If $O(d_p)<0$,

then $D+N+W=(-1+O(d_p))\times O(w_i)$.

If $O(d_p)<0, O(w_i)>0$,

then $D+N+W=-O(w_i)-(1-O(w_i))\times O(d_p)$.

If $O(d_p)>0, O(w_i)<0$,

Table 2. Three Groups Random Dataset.

	Total Number of Text	Commendatory Text Number	Derogatory Text Number
Group 1	300	150	150
Group 2	300	150	150
Group 3	300	150	150

then $D+N+W=-O(w_i)+(1+O(w_i))\times O(d_p)$.

(5) Summarizing $O(w_i)$ as this sentence weight. Looping until all sentences have been traversed.

(6) For sentences with conjunction, adjust weights according to its part-of-speech and summarize these weights as text sentiment orientation.

End

Text Filtering: Training samples are sorted based on sentiment orientation and texts with higher value are selected as training samples.

2.3. Main Functions of Machine Learning Classifier

Text Segmentation: text segmentation and word tagging are conducted by using Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS, *http://ictclas.nlpir.org*). An initial feature set is formed after deletion of stop words.

Feature Selection:, most representative feature words are selected from initial feature set by using optimized feature selection method and feature subsets are formed. Word frequency and word sentiment degree are the most representative features of a given text, however, the value of them should be normalized to make the result more comprehensive. The features then are expressed by vector space model (VSM) [12] initially followed by using the widely used information gain method [9] .

Feature Weight Calculation: Each feature weight is calculated by using term frequency- inverse document frequency (TF-IDF [13]) method and expressed as feature vector.

Machine Learning Training: After input data normalization and machine learning training, a sentiment classifier model is built. Given that SVM has superior discrimination for text classification, we use it as the machine learning methods.

Sentiment Prediction: For the unknown test data samples, a classifier and classification results are obtained after previous four steps. The feasibility of this classifier is measured based on classification criteria.

3. EXPERIMENT AND ANALYSIS

The purpose of the experiment is to validate the proposed hybrid framework is superior to the traditional methods. This paper mainly evaluates the index of classification accuracy and the portability performance.

3.1. Data Set and Evaluation Metrics

Date Set: The selected Chinese sentiment text corpus is from COAE (*http://www.searchforum.org.cn/tansongbo /senti_corpus.jsp*). We select hotel review corpus for this experiment. 300 texts (half is commendatory texts, half is derogatory texts) are selected from 4000 tagged hotel review texts and randomly select three times to constitute different experimental datasets, as is shown in Table **2**:

Evaluation Metrics: Classification results are evaluated by Precision, Recall and F-measure. These three evaluation metrics calculation are defined in formula (3),(4) and (5).

$$Precision:\ P=\frac{P_c+P_d}{2}\left(where P_c=\frac{a}{a+b}, P_d=\frac{d}{d+b}\right) \quad (3)$$

$$Recall:\ R=\frac{R_c+R_d}{2}\left(where R_c=\frac{a}{a+c}, R_d=\frac{d}{d+c}\right) \quad (4)$$

$$F\text{-}measure:\ F=\frac{2\times P\times R}{P+R} \quad (5)$$

With reference to a confusion matrix in Table **3**, *a* and *d* refer to the number of correctly classified commendatory or derogatory text; *b* and *c* refer to the number of wrongly classified commendatory or derogatory text.

3.2. Performance Analysis of Hybrid Framework Framework Method on Text Sentiment Classification

Hybrid framework that combines optimized semantic comprehension and traditional machine learning is used for this experiment (Method IG_II). Experimental results compare with method proposed in [14] (Method DSC) , traditional machine learning methods (Method DF_I, Method X^2_I and Method IG_I) to validate hybrid framework availability. Detailed results are shown in Table **4**:

From Table **4**, it is observed that Method IG-II has greatly improve than Method DSC and traditional machine learning methods only (Method DF_I, Method X2_I and Method IG_I). Average Precision increases 9.7% and 2.8%, average Recall increases 12.5% and 2.77% and average F-measure increases 11.1% and 2.8% respectively. It is thus proved the effectiveness of hybrid framework.

3.3. Experiment and Analysis of Classifier Portability

SVM-based method usually has higher accuracy, but the portability is not stable since it requires vast manpower and material resources to get plenty of tagged training samples for each field. Classification accuracy of semantic comprehension-based method is not high, but it can still be used in a

Table 3. Confusion Matrix.

		Observed	
		Commendatory	**Derogatory**
Predicted	**Commendatory**	a	b
	Derogatory	c	d

Table 4. Result of different methods in terms of sentiment classification (%)

		Group 1	Group 2	Group 3	Mean
DSC	Precision	76.58	70.34	75.87	74.26
	Recall	74.21	68.7	72.45	71.79
	F-measure	75.38	69.51	75.65	73.51
DF_I	Precision	80.75	80.53	81.65	80.98
	Recall	80.5	80.5	81.5	80.83
	F-measure	80.62	80.51	81.57	80.9
X^2_I	Precision	79.53	80.5	80.65	80.23
	Recall	79.5	80.5	80.5	80.17
	F-measure	79.51	80.5	80.57	80.19
IG_I	Precision	81.75	81.25	82.32	81.77
	Recall	80.82	80.35	82	81.06
	F-measure	81.28	80.80	82.16	81.41
IG_II	Precision	83.5	83	87.96	84.82
	Recall	83	82.6	87.5	84.37
	F-measure	83.25	82.8	87.73	84.59

field without tagged training samples. We proposed a hybrid method based on optimized semantic comprehension and SVM, and this experiment validates its portability. We choose laptop reviews corpus collected by Songbo Tan Team as corpus set with 2000 commendatory texts and 2000 derogatory texts included. 1000 pairs of texts are selected as corpus for this experiment and its steps are as follows:

Step 1) 1,000 pairs of tagged sentiment corpus are disrupted as unknown corpus and classified by optimized semantic comprehension method. Top *K* pairs with strong degree are selected as training samples of SVM. Classifier 1 is obtained after training.

Step 2) K pairs from corpus are selected as training samples of SVM. Classifier 2 is obtained after training.

*Step 3)*100 pairs of texts (100 commendatory and 100 derogatory) from other 1000 pairs of corpus are selected as test corpus. The performance of proposed method is measured by F-measure with different *K*. The results are shown in Figure.2:

From Fig. (**2**), it is observed that the proposed algorithm has stable portability. When training samples are un-tagged, we tag them by using optimized semantic comprehension method and train these tagged samples with SVM. When the number of samples is small, the accuracy is getting higher. However, the accuracy is reducing along with the increment of samples while the number reaches to a threshold. The reason is that the accuracy of semantic comprehension-based method is not high and the error-tagged sample become more with the increment of samples, thus reduce classifier performance. For ordinary SVM classifier, with the increment of samples, the accuracy is gradually increased due to the less influence from noise samples.

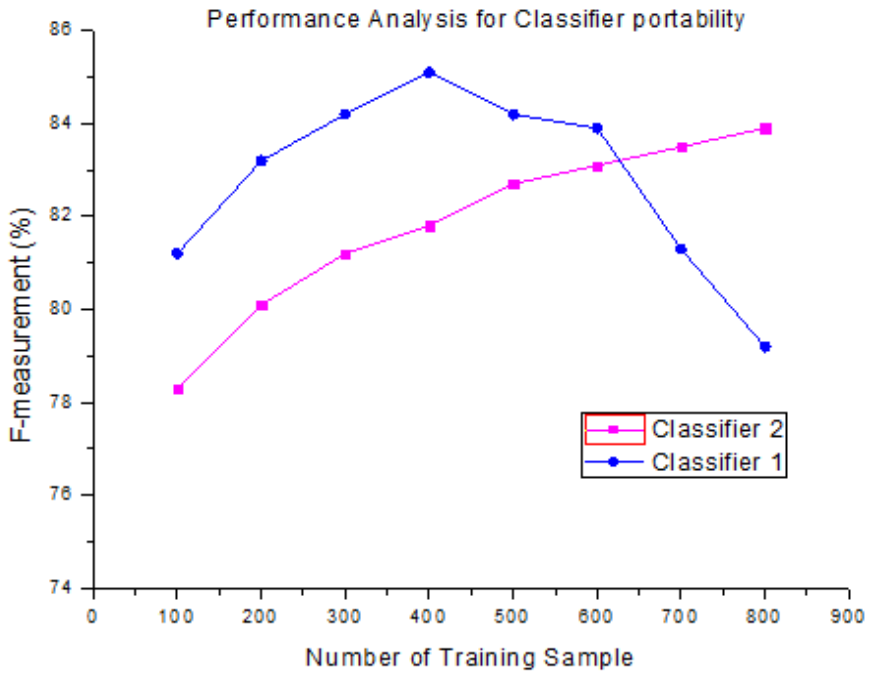

Fig. (2). Performance Analysis of Proposed Classifiers Portability.

CONCLUSION

The effectiveness of proposed method on training texts selection is validated. Obtained classifier has good classification performance with a small number but strong sentiment degree texts as samples. The reason is that we adopt the mechanism of hybrid on existing two major methodologies by implementing the proposed framework. During the design of algorithm, the word frequency with sentiment degree were integrated so that a relatively satisfied texts are filtered for machine learning procedure. It is thus more likely to obtain a result which is both time-saving and efficient.

There are several potential methods to increase the performance of this algorithm. Firstly, the specificity of different domains should be considered, which means the membership of words to domain should be evaluated. Secondly, it will be useful to response if it can quickly confirm the sentiment orientation towards some specific subjective. Last but not the least, the result of text sentiment will be more objective if it is voted by several machine learning methods. Currently, we are conducting these experiments and it will be completed in the near future.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This work is supported by the Chinese National Natural Science Fund (Grant No.61070139).

REFERENCES

[1] Y. Y. Zhao, B. Qin, and T. Liu, "Sentiment analysis," *Journal of Software*, vol. 21, no.8, pp. 1834-1848, 2010.

[2] V. Hatzivassiloglou, and K.R. McKeown, "Predicting the semantic orientation of adjectives," In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 1997.

[3] G. Recchia, and M.N. Jones, "More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis," *Behavior Research Methods*, vol. 41, no.3, pp. 647-656, 2009.

[4] X. Lin, W. Wang, and B. Wu, "A complementary method to determine semantic orientations of words based on WordNet," In: *Fuzzy Systems and Knowledge Discovery (FSKD), Eighth International Conference on*, 2011.

[5] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," In: Data Mining, ICDM *Third IEEE International Conference on*, 2003,

[6] L. Liu, M. Lei, and H. Wang, "Combining domain-specific sentiment lexicon with hownet for chinese sentiment analysis," *Journal of Computers*, vol. 8, no.4, pp. 878-883, 2013.

[7] H. Wang, X. Nie, L. Liu, and J. Lu, "A fuzzy domain sentiment ontology based opinion mining approach for chinese online product reviews," *Journal of Computers*, vol. 8, no.9, pp. 2225-2231, 2013.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Sentiment Classification using Machine Learning Techniques," In: The Conference on Empirical Methods, Natural Language Processing. pp. 79-86, 2002.

[9] L. H. XU, H. F. Lin, and Z.H. Yang, "Text Orientation Identification Based on Semantic Comprehension," *Journal of Chinese Information Processing*, vol.1,pp. 015, 2007.

[10] W. Xu, Z. Liu, T. Wang, and S. Liu, "Sentiment recognition of online Chinese micro movie reviews using multiple probabilistic reasoning model," *Journal of Computers*, vol. 8, no.8, pp. 1906-1911, 2013.

[11] L. Liu, Z. Lv, and H. Wang, "Extract Product Features in Chinese Web for Opinion Mining," *Journal of Software*, vol. 8, no.3, pp. 627-632, 2013.

[12] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM,* vol.18, no.11, pp. 613-620, 1975.

[13] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF* IDF, LSI and multi-words for text classification," Expert Systems With Applications, vol.38, no.3, pp. 2758-2765, 2011.

[14] L. Dang, and L. Zhang, "Method of Discriminant for Chinese Sentence Sentiment Orientation Based on HowNet," Application Research of Computers, 2010.

Received: June 16, 2015 Revised: August 10, 2015 Accepted: September 19, 2015

© Xu *et al.*; Licensee *Bentham Open.*

This is an open access article licensed under the terms of the (https://creativecommons.org/licenses/by/4.0/legalcode), which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.