Sentiment Recognition of Online Chinese Micro Movie Reviews Using Multiple Probabilistic Reasoning Model

Wei Xu, Zhi Liu, Tai Wang, Sanya Liu
National Engineering Research Center for E-Learning, Huazhong Normal University, Wuhan, China
Email: xuwei-ccnu@163.com, liuzhi8673@gmail.com

***Abstract*—Online review websites provide an important channel for people to share their opinions. In this paper, we research the sentiment recognition technology on online Chinese micro movie reviews. As the sentiment expressed in Chinese is subtle and the feature space is very sparse, we adopt n-grams to develop sentiment feature space, and then propose an ensemble learning algorithm based on random feature space division method, namely Multiple Probabilistic Reasoning Model (M-PRM), for supervised document level sentiment classification. This algorithm captures discriminative sentiment features and makes full use of them. Comparing with this algorithm, we apply other four machine learning methods: Multinomial NaiveBayes (MNB), Probabilistic Reasoning Model (PRM), Sentiment-word method (SWM) and SVM on two micro movie review datasets. Results show that M-PRM achieves better classification performance than other methods.**

***Index Terms*—sentiment recognition, Chinese movie review, M-PRM, ensemble learning**

I. INTRODUCTION

Recent years have seen rapid growth in movie websites, many people exchange their positive or negative attitudes towards movies online with a few sentences or some words, the short movie reviews are called micro movie reviews, and they are useful for moviegoers when they are choosing which movie they want to watch. Due to the numerous micro reviews available online, it is worthwhile to discuss the problems that how to classify micro movie reviews into positive or negative categories automatically and calculate the ratio of reviews which contain positive sentiment polarity. In this paper, we discuss these problems on online Chinese micro movie reviews.

In order to recognize sentiment polarity in reviews, Turney [1] introduces an unsupervised learning method which can determine semantic polarity based on a training set, the training set contains one hundred billion non-overlapping words. Yu and Hatzivassiloglou [2] present an unsupervised statistical technique to determine semantic polarity at sentence level. Maite Taboada extracts sentiment polarity based on lexicon [3]. These researches focus on lexicon-based sentiment recognition, sentiment dictionary can be used in different domains, so that this method is versatile. However, plot is an important part of a movie, in some online micro movie reviews, reviewers use sentiment words to summarize what sentiment contained in the plot, such as tragedy, repression, and it's an objective sentiment in the movie, not the subjective opinions from reviewers after they watched the movie. Moreover, the expression of emotion in Chinese is subtle, reviewers often express their opinions without using obvious sentiment words, such as "我看得快要睡着了" (I almost fell asleep when I was watching the movie) conveys a negative attitude towards the movie without sentiment words. In addition, micro movie reviews belong to short texts, the distinctive features of short texts are sparse; a review may consist of few words, even one word. Thus it is clear that lexicon-based sentiment recognition method is not applicable in the domain of micro movie review. Bo Pang [4] presents a traditional text classification method that compares the number of positive sentiment words with the number of negative sentiment words in a text. In the experiment part, we apply this method for comparing performance of different algorithms.

In this paper, we mainly focus on the reviews written in Chinese. Chinese grammatical structure is complex and there aren't any space signs between two words like English, it's easy to lose important discriminative information when we divide Chinese characters into words. In order to improve classification performance, n-grams method can be utilized. N-grams method applies contiguous text characters sequence to represent text information and writeprint [5]; it is statistically stable and suitable for represent online short text information. Unfortunately, n-grams method increases dimensions of feature considerably [6][7]. In this case, traditional single classifier model and the method which is comparing the number of positive sentiment words with negative sentiment words are noise sensitive. Thus, the issue that how to make full use of the whole features' discriminations in feature space is worthy of discussion.

This paper applies n-grams method for feature set extraction, multiple probabilistic reasoning model (M-PRM) algorithm is adopted to recognize sentiment

Project number: 2011BAK08B03，2011BAK08B05, NCET-11-0654 and CCNU09A02006.
Corresponding author: Tai Wang, wangtai@mail.ccnu.edu.cn.

doi:10.4304/jcp.8.8.1906-1911

polarity of reviews, this algorithm applies dynamic random allocation methods to divide sentiment feature subspaces, exploits discriminative features in different subspaces, so that it has higher level of robustness. This research has the following characteristics:

1. Experiments in this paper are based on real-life dataset, according to the habit of reviewing, reviewers usually browse the reviews which have been posted online as a guideline, and in other words, the sentiment polarity contained in previous reviews may have influence on the latter ones. We choose dataset within a period of time, the training set consists of the reviews collected in the first part of this period of time and the test set is collected in the rest part.
2. Due to the high cost of hand-label, we use the score given by the reviewer in each review to divide the review into either positive or negative categories automatically.
3. The proposed algorithm in this paper, selecting feature subsets from feature space dynamically, can capture more finer-grained sentiment information.
4. The ensemble mechanism has better recognition robustness and stronger distinguishing ability than single classifier. To some degree, different recognition results are the reflection of the popularity of different movies.

The organization of this paper is as follows. Firstly, we have a detailed introduction about proposed algorithm in Section II.Then in Section III, several experimental results by different algorithms on the same real-life dataset are shown and analyzed. Conclusion is given in Section IV.

II. SENTIMENT RECOGNITION USING MULTIPLE PROBABILISTIC REASONING MODEL

A. Framework of sentiment recognition

Online micro movie reviews, as a main channel of Web communication about movies, are important sources for recognizing the sentiment polarity of online short texts. Due to the special characteristics of online short texts, traditional text classification methods are no longer suitable for micro movie reviews classification. Some complicated methods need to be introduced. The main procedure of sentiment recognition is described as follows:

- Extract character n-grams features from training samples and normalize these features.
- Employ PCA [8] to reduce the dimensionality of feature space.
- Apply M-PRM method to build the sentiment recognition model.
- Fuse all predictive results to achieve a final decision for the sample with unknown class.

The framework of sentiment recognition of online micro movie reviews is shown as Figure 1.

B. Probabilistic Reasoning Algorithm

Probabilistic Reasoning Model (PRM) [9] combines PCA (Principal component analysis) with the Bayes classification rule. This model leverages PCA for compact the original feature space to obtain discriminative principal components. The key point of this algorithm is that covariance matrix of each class is estimated by calculating the within-class scatter under the

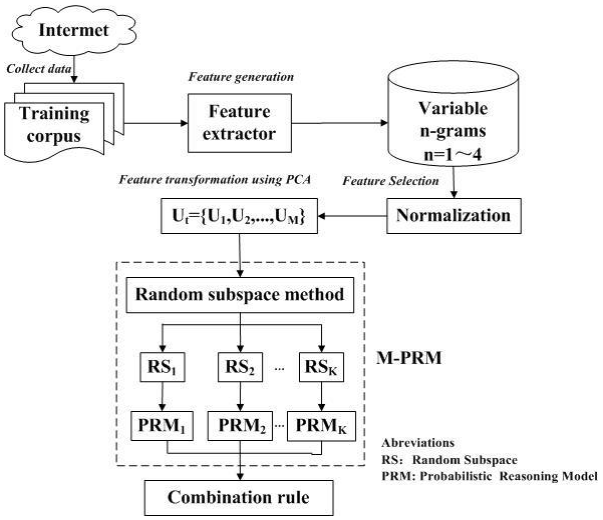

Figure 1. A framework of sentiment recognition for online micro movie reviews.

Gaussian distribution. All the within-class covariance matrices are assumed to be diagonal matrices; each factor on the diagonal is obtained by calculating variances in the one-dimensional (1-D) PCA space.

$$\delta_i^2 = \frac{1}{T}\sum_{k=1}^{T}\left\{\frac{1}{N_k - 1}\sum_{j=1}^{N_k}(x_{ji}^{(k)} - m_{ki})^2\right\} \quad (1)$$

Where T is the total number of sentiment classes, N_k is the number of samples belonging to the class ω_k, $x_{ji}^{(k)}$ is the i th component of the sample, m_{ki} is the mean value of the i th component in k-th sample, and $M_k = E(X \mid \omega_k)$.

The within-class densities are calculated with Gaussian distribution as follows:

$$P(X \mid w_i) = \frac{1}{(2\pi)^{\frac{m}{2}}\prod_{j=1}^{m}\delta_j}\exp\left\{-\frac{1}{2}\sum_{j=1}^{m}\frac{(x_i - m_{ij})^2}{\delta_j^2}\right\} \quad (2)$$

Where δ_j and m_{ij} denote the covariance matrix and mean of class ω_i respectively.

MAP rule is one of the most classic fusion rules which involves the comparison of conditional probability densities for each class. The rule can achieve the minimum error rate in classification. The PRM applies the rule to fuse all posterior probabilities from all predictors. In our experiment, the prior probabilities are set to be equal. Thus, the classification function is represented as following.

$$\sum_{j=1}^{m}\frac{(x_j-m_{ij})^2}{\delta_j^2}=\min_k\left\{\sum_{j=1}^{m}\frac{(x_j-m_{kj})^2}{\delta_j^2}\right\}\Rightarrow X\in\omega_i \quad (3)$$

C. Multiple-Probabilistic Reasoning Model

To decrease the redundancy and exploit all discriminative information in the feature space. Random subspace method (RSM) [10] is employed to divide the original feature space to a set of n-grams with sentiment information. M-PRM (Multiple-Probabilistic Reasoning Model) applies dynamic random allocation methods to divide several subspaces, and then generates corresponding PRM-based classifiers. The ensemble of classifiers has the better generalization ability with a high accuracy. By using the ensemble method, different base classifiers have diverse complementary ability. Finally, the ensemble system can obtain the higher robustness than a single classifier.

Even after feature selection, the dimension of feature space is still high. By dividing the feature space to some subsets and submitting each one to a base classifier (BC), the number of BCs in the ensemble and the dimension of each random subspace should be predefined. Here, for the sake of simplicity, we suppose each feature subset includes the same number of the feature, and then calculate the feature's number N_{fs} is $floor(\alpha * N_f)$ in each feature subset by the specified selection rate α, where N_f is the total feature number.

In this paper, the proposed the feature sampling method is performed based on the random subspace and the weight distribution DD of features. The major procedure of feature space partition mechanism is described as follows.

1) Form a random number $nRand$ on (0,1).
2) Construct N_f intervals on the interval (0,1) according to the probabilistic distribution DD_i of features.
3) Locate the upper boundary of $nRand$. If $nRand$ is located in the i th interval, get the i th feature from the original feature set.
4) Else $i \leftarrow i+1$ and return the step 3).
5) Repeat the steps (1) ~ (4) N_{fs} times.

Description of M-PRM Algorithm.

Input:

1) Let $TS=\{(x_1,y_1),\ldots,(x_N,y_N)\}$, where $y_i\in Y$ with class label $x_i\in X=\{1,\ldots,N\}$, and N is the total number of classes, Y is sample space.
2) The number of BCs: M

Output: The optimum ensemble H

Steps:

(a) Initialize the distribution of feature weights DD_i.
(b) According to the sampling method mentioned above, divide the sample space into several feature subsets, the original sample set is reduced to $\tilde{S}=(TS,r)$, where r is the subspace dimension.
(c) Form a sample subset on $\tilde{S}$ as an input for the learning algorithm, and train a basic classifier h_t.
(d) Repeat the steps (b) ~ (c) for M times to build ensemble classifiers $H=\{h_1,\ldots,h_M\}$.

III. EXPERIMENTS

A. Dataset

We collect micro movie reviews from a website which is a main website offering movie reviews in China (http://www.mtime.com/). We collect two datasets from two movies, one is "The Flowers of War" (represented by M1), and the other is "Flying Swords of Dragon Gate" (represented by M2). The two datasets are collected from December 16, 2011 to December 24, 2011. We only select the reviews which contain scores (stars). For each dataset, the training set consists of the reviews collected in the first part of this period of time (from December 16, 2011 to December 18, 2011), and the test set consists of the reviews collected in the rest part of this period of time (from December 19, 2011 to December 24, 2011). The sentiment polarity of each review can be recognized into three classes: positive, negative, neutral. In this paper, we only focus on discriminating between positive and negative sentiment polarity, hence. We consider that the reviews with scores between 1 and 6 belong to negative class (NEG), and the reviews with scores between 8 and 10 belong to positive class (POS). The data distributions of the two movies are shown as Table I.

TABLE I.
DATASET

		Num of POS	Num of NEG	POS/NEG	Avg characters of per text
M1	Training set	808	104	7.77	19.4309
	Test set	679	102	6.66	17.9206
M2	Training set	712	144	4.95	21.6578
	Test set	587	171	3.43	20.9169

Firstly, according to the characteristics of Chinese micro movie reviews, we construct a stop word list for them, and eliminate the stop words from reviews in the data pre-processing step. Secondly, there are many four-character idioms in Chinese expression, we set the max grams to four while applying n-grams method to extract features, hence large amounts of redundant features are produced, then we only retain the features whose frequencies are more than six, sort the retained features descending and select the first 2000 features in the descending sequence. Thirdly, PCA can reduce dimensions of feature space, further extract feature vectors and reduce the effect of noise on the classifier. After PAC change, M1 retains 897 features and M2 retains 850 features, then according to the accuracy testing with Support Vector Machine (SVM) based on different number of features retained after the process of

PCA, the accuracy of M1 reaches maximum with 500 features and the maximum accuracy of M2 is with 700 features, so we use these number of features for M-PRM. Fourthly, the maximum number of iterations of base classifiers is 50; it means that there are 50 base classifiers at most for constructing ensemble classifier. Some of the stop words and retained words are shown in Table II.

TABLE II.
STOP WORDS AND RETAINED WORDS

Stop words	的, 地, 了 (they are specific particles in Chinese)
Retained words	好(good), 效果(performance), 赞(excellent), 沉重(heavy)

B. Experimental Settings

In our experiments, we apply four machine learning methods to compare with M-PRM; they are Multinomial NaiveBayes (MNB), PRM, Sentiment-word method (SWM) and SVM. For SWM, we adopt HowNet semantic lexicon, within which there are 4566 words in positive part (e.g., "赞"(excellent), "漂亮"(beautiful)) and 4370 words in negative part (e.g., "烂"(bad), "悲伤"(sad)). SWM simply decides sentiment polarity of a review by counting the number of the positive and the negative words in the review.

To evaluate the performance of M-PRM, we adopt five indices: Recall, Accuracy, POSRATE, Kappa and Training Time. POSRATE means the proportion of predicted positive reviews in the whole test set, it represents the popularity of movies, POSRATE is a measure of validity for M-PRM, evaluating whether the POSRATE relates directly to the movie's total score, which is extracted from "http://www.mtime.com/". The indices can be calculated according to Table III.

TABLE III.
EVALUATION

	Predicted positive reviews	Predicted negative reviews
Actual positive reviews	true positive (tp)	false negative (fn)
Actual negative reviews	false positive (fp)	true negative (tn)

$$Recall = \frac{tp}{tp + fn} \tag{4}$$

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn} \tag{5}$$

$$Posrate = \frac{tp + fp}{tp + fn + fp + tn} \tag{6}$$

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{7}$$

Where, Recall is the recall ratio for actual positive reviews. In (7), p_o is the probability of the actual outcome and p_e is the probability of the expected outcome as predicted by chance [11].

C. Experimental Results and Analysis

Based on previous work, classification accuracy reaches a maximum at 0.5 sampling rate for the same number of base classifiers iterations [10]. Figure 2 shows the different accuracies of M-PRM with different number of iterations at 0.5 sampling rate. The classification accuracy of M1 rises from 70.13% to 84.34%, and M2 rises from 63.61% to 81.56%. With the increasing of the number of iterations, that is to say, with the increasing of the number of base classifiers, classification accuracy is increasing and classification ability of ensemble classifier is strengthened. It proves that important sentiment classification information is contained in each feature subspace; hence each feature subspace has discrimination. In addition, there are mutual complementary classification capabilities among these base classifiers, a classification error in one subspace may be corrected by the right classification in the other feature subspace.

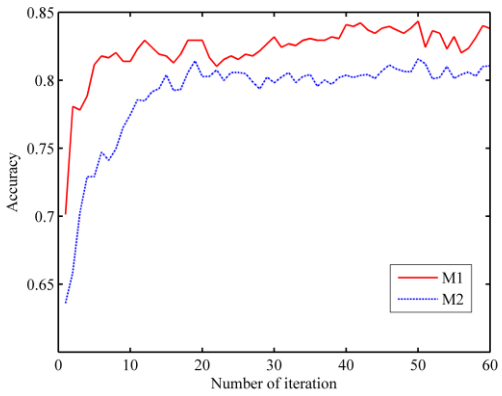

Figure 2. Results of accuracy by using different number of iterations

M-PRM can select subspaces randomly from the original feature space, covering most discriminative sentiment information. In this experiment, the maximum number of iterations is 50. Figure 2 shows that the classification accuracy tends towards stability when the number of iterations is more than 40, if we increase the number of iterations, instead of great improvement of classification accuracy, the system load will increase and the operating efficiency of ensemble system will decrease.

In this paper, we evaluate classification performance of M-PRM by comparing with MNB, PRM, SWM and SVM, the sampling rate of M-PRM is 0.5. Experimental results are shown in Table IV.

TABLE IV.
CLASSIFICATION PERFORMANCE OF SEVERAL ALGORITHMS

	M1					M2				
	Recall (%)	Accuracy (%)	POSRATE (%)	Kappa	Training Time(s)	Recall (%)	Accuracy (%)	POSRATE (%)	Kappa	Training Time(s)
MNB	88.65	82.21	84.68	0.2712	0.21	81.21	79.3	69.1	0.3781	0.19
PRM	70.1;	67.86	67.09	0.1401	0.13	73.08	70.58	65.17	0.2941	0.12
SWM	77.38	70.52	76.57	0.1711	--	79.71	74.54	82.32	0.3412	--
SVM	90.12	83.51	85.54	0.3015	0.69	84.52	80.48	70.8	0.3814	0.69
M-PRM	93.08	84.34	88.99	0.3445	0.83	86.2	81.56	77.18	0.4106	0.78

As it shown above, based on the dataset of M1, the classification performance of PRM is worst with 70.1% recall and 67.86% accuracy, but M-PRM obtains 93.08% recall and 84.34% accuracy. Because PRM is a single classifier, its ability of capturing discriminative sentiment information in feature space is weak, and it is noise sensitive. M-PRM is a multiple classifier, its ability of capturing discriminative sentiment information is strong, and it has higher robustness.

For SWM, its classification performance is between MNB and PRM. SWM is a lexicon-based algorithm for sentiment polarity recognition, the performance of this algorithm depends entirely on the quality of the sentiment lexicon. In this experiment, we adopt HowNet semantic lexicon, the lexicon has a large vocabulary (positive part has 4566 words, negative part has 4370 words), but there are two defects in the lexicon, one is that some words exist in both the positive and the negative part of the lexicon (e.g., “大” (big)), the other is that few popular words online contained in the lexicon (e.g., “顶” (support), “V5”). In addition, Chinese emotion expression is subtle and the features of micro movie reviews are sparse. Moreover, it needs to be noted that the evaluating indices of M1 are higher than that of M2 while using MNB, PRM, SVM [12][13] and M-PRM for testing; it shows that M1 is more popular than M2. But the values of these indices are reverse by SWM, because objective sentiment describing plots may contain in the reviews. M1 is a war movie, “剧情悲壮沉重, 很有震撼力”(The story is tragic and heavy, very shocking), in this positive review, “悲壮”(tragic) and “沉重”(heavy), which are negative words in lexicon, are used to describe the movie plot, it is a objective negative sentiment but considered to convey subjective negative sentiment. Thus it is clear that lexicon-based sentiment recognition method is not applicable in the domain of online micro movie review.

NaiveBayes, despite its conditional independence assumption doesn’t support in real-world situations, it still performs well in text categorization [14]. On the dataset of M1, MNB obtains 88.65% recall and 82.21% accuracy; SVM obtains 90.12% recall and 83.51% accuracy; M-PRM obtains best classification performance with 93.08% recall and 84.34% accuracy, because M-PRM is a multiple classifier system, the final classification result is determined by all base classifiers, some base classifiers may make wrong judgments, but these wrong judgments can be made up by right judgments from other classifiers, so that the final classification result is best.

Moreover, among the Kappa values of these five algorithms, M-PRM reaches the maximum, it shows that the classification abilities of the base classifiers are complementary, and M-PRM obtains the best consistency between the actual and the predicted class.

However, the training time of M-PRM is longest among these five algorithms, because M-PRM contains 50 base classifiers; in addition, it needs about 0.3s to divide sentiment feature subspaces, but the training time of M-PRM is in an acceptable range. SWM doesn’t have training time, because it is based on sentiment lexicon, doesn’t need to build train model.

For all the analyses above, we can draw the conclusion that the classification performance of M-PRM is best among the five algorithms.

We collect datasets within a period of time, the training sets are collected in the first part of this time and the test sets are collected in the rest part. From Table I we can see that the POS/NEG ratio on training set for M1 is 7.77, and it is 4.95 for M2, they are unbalanced datasets. For test sets, Figure 3 shows that all the recall values of M1 using M-PRM are above 90% with different sample rates, and the recall values are above 80% for M2, for one thing, the high recall values prove that the proposed algorithm has a good classification ability, for another thing, there are consistencies between the training sets and the test sets, the sentiment polarity contained in othe sentiment polarity of latter reviews (test sets), because reviewers usually browse some previous reviews before they give theirs, so that the previous reviews affect the latter reviews to some extent.

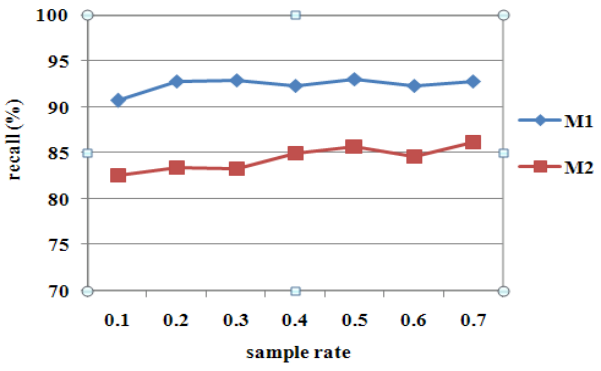

Figure 3. Recall curve of M-PRM with different sample rates.

Figure 4 shows that the POSRATE of M1 is higher than that of M2 by M-PRM, it means that M1 is more popular, it is in accordance with the fact that the total average score of M1 is 8.5 and M2 is 7.5 at "http://www.mtime.com/", and this positive correlation demonstrates the efficiency of the proposed algorithm.

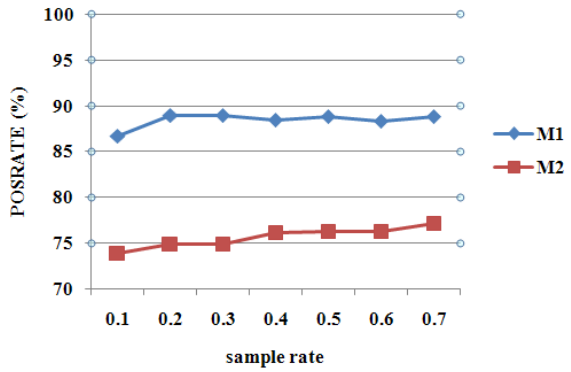

Figure 4. POSRATE curve of M-PRM with different sample rates

IV. CONCLUSION

In this paper, we have developed a random subspace approach based on sentiment features for recognizing sentiment polarity of short movie reviews. The ensemble mechanism at the feature-level captures the richest discriminative sentiment information. But the task is difficult because: 1) the high sparsity of sentiment features and 2) minor sentiment information hidden in short texts. Our approach has solved these problems. By using PCA to compact the whole feature space and dividing it into several subspaces, more finer-grained sentiment information can be exploited in multiple subspaces. Experimental results on the real-life dataset show that compared with MNB, PRM SVM and SWM, the proposed algorithm achieves better performance at the sentiment recognition of short texts with an acceptable accuracy rate, but its training time is longest, in the future, we need to try our best to reduce the training time of the proposed algorithm.

ACKNOWLEDGMENT

This work was supported by the National Key Technology R&D Program in the 12th Five-Year Plan (Grant No.2011BAK08B03, No.2011BAK08B05), Program for New Century Excellent Talents in University (NCET-11-0654) and self-determined research funds of CCNU from the colleges' basic research and operation of MOE (No.CCNU09A02006). We would like to thank Prof. Tai Wang and Prof. Sanya Liu who help us for this study, and reviewers for their comments which helped improve this paper.

REFERENCES

[1] P.D. Turney, M.L. Littman, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus", *Technical Report ERB-1094, National Research Council Canada, Institute for Information Technology*, 2002.

[2] Yu Hong, Vasileios Hatzivassiloglou, "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences", *Proceedings of the 2003 conference on Empirical methods in natural language processing* (EMNLP-2003), pp. 129-136, 2003.

[3] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede, "Lexicon-based methods for sentiment analysis", *Computational Linguistics*, vol. 37, no. 2, pp. 267-307, 2011.

[4] Pang Bo, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pp. 79-86, 2002.

[5] Sanya Liu, Zhi Liu, Jianwen Sun, Lin Liu, "Application of Synergetic Neural Network in Online Writeprint Identificat ion", *International Journal of Digital Content Technology and its Applications*, vol. 5, no. 3, pp. 126-135, 2011.

[6] Xuerui Wang, Andrew McCallum and Xing Wei, "Topical N-grams: Phrase and Topic Discovery, with an Application to Information Retrieval", *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)*, pp. 697-702, 2007.

[7] Fotis Aisopos, George Papadakis, Theodora Varvarigou, "Sentiment analysis of social media content using N-Gram graphs", *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, pp. 9-14, 2011.

[8] Qifa Qu, "Determination of Weights for the Ultimate Cross Efficiency: A Use of Principal Component Analysis Technique", *Journal of Software*, vol. 7, no. 10 (2012), pp. 2177-2181, Oct 2012.

[9] Liu C. and Wechsler H, "Robust Coding Schemes for Indexing and Retrieval from Large Face Databases", *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 132-137, 2000.

[10] Tin Kam Ho, "The Random Subspace Method for Constructing Decision Forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998.

[11] Andrew Rosenberg, Ed Binkowski, "Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points", *HLT-NAACL-Short '04 Proceedings of HLT-NAACL 2004*, pp. 77-80, 2004.

[12] Shanxiao Yang, Guangying Yang, "Emotion Recognition of EMG Based on Improved L-M BP Neural Network and SVM", *Journal of Software*, vol. 6, no. 8 (2011), pp. 1529-1536, Aug 2011.

[13] Hao Liu, Xiaoming Tao, Pengjun Xu, Guanxiong Qiu, "Classification of Bio-potential Surface Electrode based on FKCM and SVM", *Journal of Software*, vol. 6, no. 5 (2011), pp. 880-886, May 2011.

[14] David D.Lewis, "Naive(Bayes) at forty: The independence assumption in information retrieval", *European Conference on Machine Learning(ECM)*, pp. 4-15, 1998.

Wei Xu is a postgraduate at National Engineering Research Center for E-learning (NERCEL), Central China Normal University (CCNU), Wuhan, China. Her main research interest is text mining.