An Integrated Method for Micro-blog Subjective Sentence Identification Based on Three-Way Decisions and Naive Bayes

Yanhui Zhu, Hailong Tian, Jin Ma, Jing Liu, and Tao Liang

School of Computer and Communication, Hunan University of Technology, Zhuzhou 412008, Hunan, China
{swayhzhu,tianhailongbmg}@163.com

Abstract. Microblog's subjective sentence recognition is the basis of it's public opinion analysis further research .Therefore, its recognition accuracy is crucial for future research work. Owing to the imprecision or incomplete of information, the precision of traditional SVM, NB and other machine learning algorithms that for microblog's subjective sentence recognition is not ideal. Presents a method based on the integrated of three-way decision and Bayesian algorithms to distinguish microblog's subjective sentence. Compared with traditional Bayesian algorithms, Experimental results show that the proposed integrated approach can significantly improve the accuracy of subjective sentence's recognition .

Keywords: Three-way decision, Bayes, micro-blog, subjective sentence.

1 Introduction

With the rapid development of Internet, for people to obtain information ， micro-blog has become an important channel. Statistical Report ,which was on Internet Development 33rd China Internet Network [1] shows that by the end of December 2013,China micro-blog users reached 281 million, and in netizen the Micro-blog unilization ratio was 45.5%. In recent years，many scholars launched the study of micro-blog identificational subjective sentences，Yang Wu et al [2] by using Bayes classifier, researched the classification of subjective and objective to micro-blog statement. Firstly, they analyzed the main differences between micro-blog text and other texts, and extract some features of subjective and objective clues for the characteristics of expression about micro-blog text. Then, they researched to 2-POS mode on the best select way, finally, as semantic features by feature words and the objective and subjective clues, as grammatical features by 2-POS mode, to study their impact on the classification results using Naïve Bayes classifier. Experimental results show that the method of taking into account semantic features and the structural characteristics of grammatical is better than the mehtod of only considering a feature better. Its classification precision was 81.9%, the recall 80.5%, F-Measure 81.2%.

D. Miao et al. (Eds.): RSKT 2014, LNAI 8818, pp. 844–855, 2014.
DOI: 10.1007/978-3-319-11740-9_77

Alexander Pak et al [3] to select N-Gram and speech tagging of micro-blog as feature, identification research on subjective sentences of micro-blog using Naïve Bayes classifier, compare with the two kinds of classifiers that SVM(support vector machine) and CRFs, expermental results show that, subject sentence identification of micro-blog based on Naïve Bayes was the best.

It's a binary classification problem about identification of subjective sentence. Since micro-blog has characteristics of short text, colloquial, randomness, etc, the subjective information is often imprecise, incomple, and the classification accuracy rate is not high enough .This paper presents the method of subjective sentence recognition that micro-blog based on Three-Way Decision and machine learning algorithms integrated，the method using decision-rough sets and Three-Way Decision as theoretical basis，achieve identification that subjective sentence of micro-blog through integration with Bayes classifier that many previous studies and the best effect. The experiments show that the method can significantly impove the accuracy of subjective sentence recongnition.

2 Three-Way Decision Theory

Three-Way Decision theory [4-6] is proposed in rough sets [7-8] and decision-rough sets [9-13] by Yao, based on this model, Yao study on the semantic of positive region, negative region, boundary region in the rough sets theory and proposed to explain the rough sets rules extraction problem from the perspective of three-way decisions.

Decision-rough set is the expand of Pawlak algebra rough sets and 0.5-probability rough sets[10].the core of rough sets is the definition of the upper and lower approximation.

Definition of the upper and lower approximation of Pawlak algebra rough sets:

$$\begin{aligned} \overline{apr}(X) &= \{x \in U \mid [x] \cap X \neq \Phi\}, \\ \underline{apr}(X) &= \{x \in U \mid [x] \subseteq X\}. \end{aligned} \tag{1}$$

Definition of positive region $POS(X)$, negative region $NEG(X)$, boundary region $BND(X)$ based on the upper and lower approximation:

$$\begin{aligned} POS(X) &= \underline{apr}(X) = \{x \in U \mid [x] \subseteq X\}, \\ NEG(X) &= U - \overline{apr}(X) = \{x \in U \mid [x] \cap X = \Phi\}, \\ BND(X) &= \overline{apr}(X) - \underline{apr}(X) = \{x \in U \mid [x] \cap X \neq \Phi \wedge \neg([x] \subseteq X)\}. \end{aligned} \tag{2}$$

For the set of states $\Omega = \{X, \neg X\}$, there is the probability of conditions ,following:

$$p(X \mid [x]) = \frac{|X \cap [x]|}{|[x]|},$$
$$p(\neg X \mid [x]) = \frac{|\neg X \cap [x]|}{|[x]|} = 1 - p(X \mid [x]). \tag{3}$$

Which $|\cdot|$ represents a potential of collection, namely the number of collection, $[x]$ said equivalence class. The three domains of Pawlak algebra rough sets is described by using the probability as follows:

$$\begin{aligned} POS(X) &= \{x \in U \mid P(X \mid [x]) \geq 1\}, \\ BND(X) &= \{x \in U \mid 0 < P(X \mid [x]) < 1\}, \\ NEG(X) &= \{x \in U \mid P(X \mid [x]) \leq 0\}. \end{aligned} \tag{4}$$

The rough sets mode only using two extreme values of probability, it lack of fault tolerance when applied to classification decisions. Based on this, Yao et al proposed a decision-rough set model .Suppose $0 \leq \beta < \alpha \leq 1$, as a pair of thresholds, and define three domain as follows:

$$\begin{aligned} POS_{(\alpha,\beta)}(X) &= \{x \in U \mid P(X \mid [x]) \geq \alpha\}, \\ BND_{(\alpha,\beta)}(X) &= \{x \in U \mid \beta < P(X \mid [x]) < \alpha\}, \\ NEG_{(\alpha,\beta)}(X) &= \{x \in U \mid P(X \mid [x]) \leq \beta\}. \end{aligned} \tag{5}$$

When object x belongs X, orders λ_{pp}, λ_{np}, λ_{bp} as the loss function of x divided into $POS(X)$, $NEG(X)$, $BND(X)$, when x belongs to $\neg X$, appropriate orders λ_{pn}, λ_{nn}, λ_{bn} as the loss function of divided into the same three domains. Shown in the following table:

Table 1. Table of Loss Function

	$POS(X)$	$BND(X)$	$NEG(X)$
X	λ_{pp}	λ_{bp}	λ_{np}
$\neg X$	λ_{pn}	λ_{bn}	λ_{nn}

The risk of equivalence class divided into three domains is defined as:

$$\begin{aligned} R(POS(X) \mid [x]) &= \lambda_{pp} p(X \mid [x]) + \lambda_{pn} p(\neg X \mid [x]), \\ R(BND(X) \mid [x]) &= \lambda_{bp} p(X \mid [x]) + \lambda_{bn} p(\neg X \mid [x]), \\ R(NEG(X) \mid [x]) &= \lambda_{np} p(X \mid [x]) + \lambda_{nn} p(\neg X \mid [x]). \end{aligned} \tag{6}$$

Given the minimum rish decision rule by Bayes decision theory:

If $R(PO(S_X) \mid [x]) \leq R(BND(D_X) \mid [x])$ and $R(PO(S_X) \mid [x]) \leq R(NEG(G_X) \mid [x])$ then $x \in PO(S_X)$;

If $R(BND(D_X) \mid [x]) \leq R(PO(S_X) \mid [x])$ and $R(BND(D_X) \mid [x]) \leq R(NEG(G_X) \mid [x])$, then $x \in BND(D_X)$;

If $R(NEG(G_X) \mid [x]) \leq R(PO(S_X) \mid [x])$ and $R(NEG(G_X) \mid [x]) \leq R(BND(D_X) \mid [x])$ then $x \in NEG(G_X)$.

For $R(X \mid [x]) + R(\neg X \mid [x]) = 1$, orders

$$\begin{gathered}
\lambda_{pp} \leq \lambda_{bp} < \lambda_{np}, \\
\lambda_{nn} \leq \lambda_{bn} < \lambda_{pn}, \\
(\lambda_{pn} - \lambda_{bn})(\lambda_{np} - \lambda_{bp}) > (\lambda_{bp} - \lambda_{pp})(\lambda_{bn} - \lambda_{nn}), \\
\alpha = \frac{(\lambda_{pn} - \lambda_{bn})}{(\lambda_{pn} - \lambda_{bn}) + (\lambda_{bp} - \lambda_{pp})}, \\
\gamma = \frac{(\lambda_{pn} - \lambda_{nn})}{(\lambda_{np} - \lambda_{pp}) + (\lambda_{pn} - \lambda_{nn})}, \\
\beta = \frac{(\lambda_{bn} - \lambda_{nn})}{(\lambda_{bn} - \lambda_{nn}) + (\lambda_{np} - \lambda_{bp})}.
\end{gathered} \tag{7}$$

The above rules has the following equivalent description:

If $p(X \mid [x]) \geq \alpha$ then $x \in PO(S_X)$;

If $\beta < p(X \mid [x]) < \alpha$ then $x \in BND(D_X)$;

If $p(X \mid [x]) \leq \beta$ then $x \in NEG(G_X)$.

3 Three-Way Decision Approach to Microblog Subjective Sentence Recognition

To the positive , negative and boundary region of decision-theoretic rough set,Yao proposed positive rules, negative rules, boundary rules of three-way decisions .Sets said as follows:

$$\text{positive rules}: R(X \mid [x]) \geq \alpha, [x] \subseteq POS_{(\alpha,\beta)}(X)$$

$$\text{negative rules}: R(X \mid [x]) \leq \beta, [x] \subseteq NEG_{(\alpha,\beta)}(X)$$

$$\text{boundary rules}: \beta < R(X \mid [x]) < \alpha, [x] \subseteq BND_{(\alpha,\beta)}(X)$$

That is: $\forall x \in U$, the rate of $P(X \mid [x])$ is greater than or equal to the α .Then divided $[x]$ into positive region of X,At the same time make positive region decision, Other formula explained in the same way. And U is a finite set of objects.

Use $p(L \mid x)$ represent the rate that x is belong to subjective sentences category, It microblog whether belongs to the subjective sentences three-way decisions representation are as follows:

$p(L \mid x) \geq \alpha$, Decide x is subjective sentences;
$p(L \mid x) \leq \beta$, Decide x is not subjective sentences
$\beta < p(L \mid x) < \alpha$, It microblog whether belongs to the subjective sentences is uncertain, made by artificial processing decisions.

Three-way decisions model divided microblog into subjective sentences classes, not subjective sentences classes and boundary classes and compared with two classification model, not simply added a category, transformed the two classification problems into the multi-classification problems, But building on the basis of the decision-theoretic rough set for the target concept collection of positive ,negative and boundary region depiction.

4 Evaluation Standard

For positive class (category of subjective sentences), and negative class (category of nor subjective sentences) of the three-way decisions, this paper uses four evaluation indexes which are namely ,precision rate (P), recalling rate (R), F and macro-average. The table as follows:

Table 2. Contingency Table of Categories

	Actually belong to positive classes documents	Actually belong to negative classes documents
Judged belong to positive classes documents	a_{pp}	a_{pn}
Judged belong to negative classes documents	a_{np}	a_{nn}
judged belong to boundary classes documents	a_{bp}	a_{bn}

Then, there are evaluation indexes of positive classes and negative classes, as follows:

$$P_p = \frac{a_{pp}}{a_{pp} + a_{pn}} \qquad R_p = \frac{a_{pp}}{a_{pp} + a_{np} + a_{bp}},$$
$$P_n = \frac{a_{nn}}{a_{nn} + a_{np}} \qquad R_n = \frac{a_{nn}}{a_{nn} + a_{pn} + a_{bn}},$$
$$F_p = \frac{2 \times P_p \times R_p}{P_p + R_p} \qquad F_n = \frac{2 \times P_n \times R_n}{P_n + R_n},$$
$$P_{avg} = \frac{P_p + P_n}{2} \qquad R_{avg} = \frac{R_p + R_n}{2} \qquad F_{avg} = \frac{F_p + F_n}{2}. \tag{8}$$

In the formula, P_p, P_n and P_{avg} denote the precision for the positive classes ,negative class and macro-average, Similarly, other the same subscript do the corresponding explanation.

5 Experimental Design

5.1 Microblog Feature Extraction

This paper uses the method on the base of combination of dictionary and statistical analysis to select microblog's candidate subjective features, and information gain (IG) method to extract features. All steps are as follows:

Step1：Structuring basic domain of view dictionary. Using the positive and negative emotion words of HowNet [14]. For positive and negative evaluation words, there are 8746 words after removing duplicate words.

Step2：Using multiple word segmentation system [15] to build custom thesaurus for the recognition of the domain of view word. To get 5820 custom thesaurus words .

Step3：Constructing view words in candidate domain. Combined the word in step1 and step2, so as to remove duplicate words, and then there will be 13827 words of the candidate domain of view words.

Step4：Expanding the domain of view words by the conjunctions word dictionary. To get 14064 words of the candidate domain of view words.

Step5：Statistics corpus question mark and exclamation point are occupying a larger proportion. Besides that there is differences of its proportion from its subjective sentence and non-subjective sentence. So the question mark and exclamation point can be candidate subjective features.

Step6：Building the candidate subjective features. Emerged step4 and step5, and see this combination as custom thesaurus to separate Corpus's words and added into segmentation system of ICTCLAS2014[16]. Extracted the same words from segmentation corpus and custom thesaurus, and see it as candidate subjective features. Finally get 6232 candidate subjective features.

Step7：Using IG to extract candidate subjective features, the extracting number is determined by experiment .

5.2 Threshold—An Explanation

Make the following assumptions:

$$\begin{aligned} \lambda_{pp} &= \lambda_{nn} = 0, \\ \lambda_{np} &= \eta\lambda_{bp}, \\ \lambda_{pn} &= \eta\lambda_{bn}, \\ \lambda_{bn} &= 2\lambda_{bp}. \end{aligned} \tag{9}$$

Because in the massive micro-blog text environment, subjective sentence recognition is a consideration of sensitive issues. The cost of subjective micro-blog into non-subjective micro-blog is smaller than non- subjective micro-blog divided into subjective micro-blog. There are:

$$\begin{aligned} \alpha &= \frac{(\lambda_{pn} - \lambda_{bn})}{(\lambda_{pn} - \lambda_{bn}) + (\lambda_{bp} - \lambda_{pp})} = \frac{2\eta - 2}{2\eta - 1} = 1 - \frac{1}{2\eta - 1}, \\ \beta &= \frac{(\lambda_{bn} - \lambda_{nn})}{(\lambda_{bn} - \lambda_{nn}) + (\lambda_{np} - \lambda_{bp})} = \frac{2}{\eta + 1}, \\ \gamma &= \frac{(\lambda_{pn} - \lambda_{nn})}{(\lambda_{np} - \lambda_{pp}) + (\lambda_{pn} - \lambda_{nn})} = \frac{2}{3}. \end{aligned} \tag{10}$$

$\eta > 2$, as $\alpha > \gamma > \beta$. The final value for η is determined by experiment.

5.3 The Three-Way Decision Classifier Design Based on NB

Using the prior probability and class-conditional estimate probability $p(L \mid x)$ [17]:

$$p(L \mid x) = \frac{p(L) \times p(x \mid L)}{p(x)}. \tag{11}$$

(1) NB probability estimation model uses the Bernoulli model :

$p(L)$ = The total number of documents subjective sentence / The total number of documents in the training text .

$p(x_i \mid L)$ =(The number of contain L of subjective sentences and x_i of features +1) / (The total number of documents L of subjective sentence +2) .

(2) Assume that each feature is independent to each other .

(3) Total Probability : $p(x) = \sum_{L_i} p(x \mid L_i) p(L_i)$.

5.4 Experimental Result and Analysis

Experimental corpus as "Chinese micro-blog sentiment analysis evaluation data set" [18] , Choose one of 3415 in the corpus , After feature selection, feature extraction

and feature weighting, using the vector space model for text representation. With constructing micro-blog decision table, in the experiment shows that the number of features being the 3000, based on a subjective sentence micro-blog NB best recognition performance. With this micro-blog decision table, Threshold test, Experimental set the interval [2,22], In steps of 0.5 threshold value of the experiment. Experimental results shown in Figure 1, Figure 2, Figure 3.

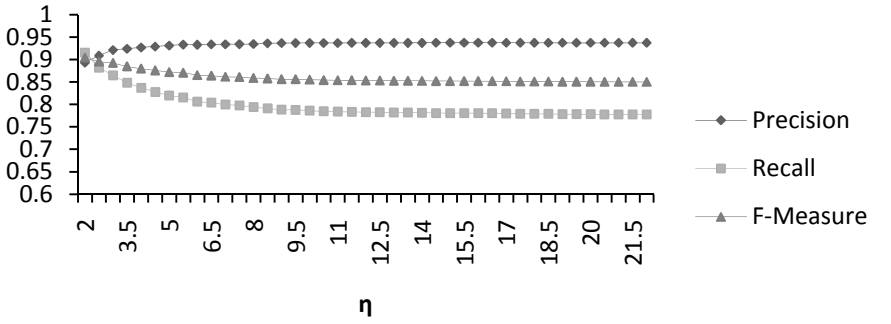

Fig. 1. Parameter η experimental result for the positive class

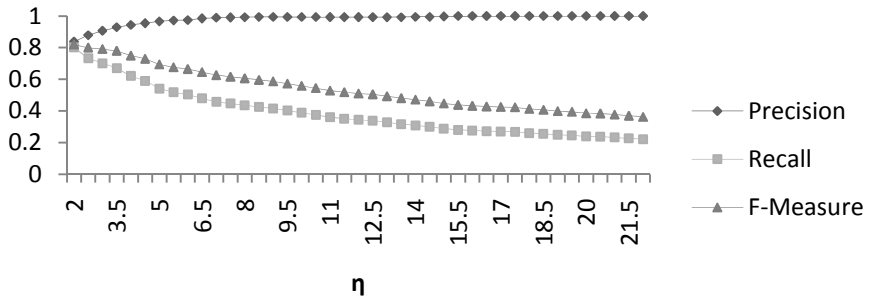

Fig. 2. Parameter η experimental result for the negative class

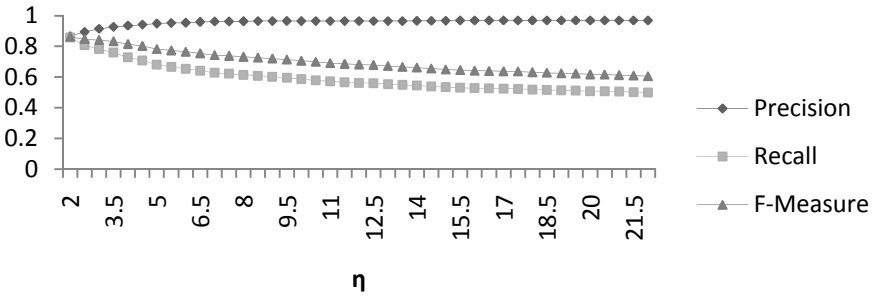

Fig. 3. Parameter η experimental result for the macro-average

By comparison chart shows the experimental results, The overall trend with increasing values of η, precision rate on the rise, The recall and F-measure show a downward trend. Precision rate of the positive class is highest at $\eta = 14.5$, Precision rate of the negative class is 1 at $\eta = 15.5$, Precision rate of the macro-average is highest at $\eta = 15.5$. The results shown in Table 3, Table 4, Table 5:

Table 3. Three-way decisions improve precision rate effectively of the positive class

The positive class $\eta = 14.5$		
Precision	Recall	F-Measure
0.937398	0.780598	0.8518

Table 4. Three-way decisions improve precision rate effectively of the negative class

The negative class $\eta = 15.5$		
Precision	Recall	F-Measure
1	0.28039	0.437984

Table 5. Three-way decisions improve precision rate effectively of the macro-average

The macro-average $\eta = 15.5$		
Precision	Recall	F-Measure
0.968699	0.53049	0.645

Considering the precision, accuracy, recall and F-measure, $\eta = 3$ shows the best results of the experiment and results were compared with the NB classification. The results are shown in Table 6, Table 7, Table 8, Figure 4, Figure 5, Figure 6:

Table 6. Comparison of experimental results of the positive class at $\eta = 3$

	The positive class		
	Precision	Recall	F-Measure
Three-way	0.921294	0.864914	0.8923
NB	0.823259	0.975521	0.8929

Table 7. Comparison of experimental results of the negative class at $\eta = 3$

	The negative class		
	Precision	Recall	F-Measure
Three-way	0.906852	0.70058	0.79048

NB	0.932584	0.61786	0.74328

Table 8. Comparison of experimental results of the macro-average at $\eta = 3$

	The macro-average		
	Precision	Recall	F-Measure
Three-way	0.914073	0.78275	0.842
NB	0.877922	0.79669	0.818

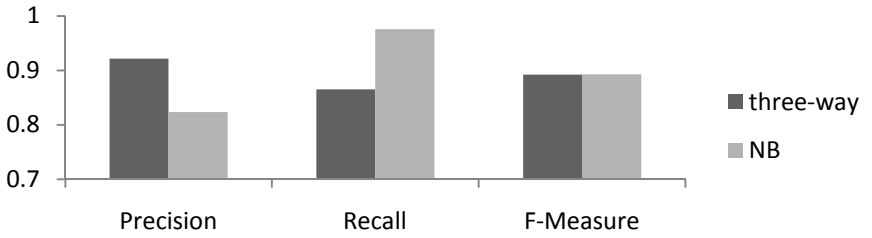

Fig. 4. Comparison of the positive class results

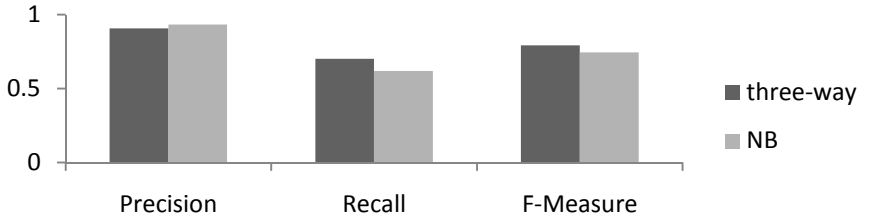

Fig. 5. Comparison of the negative class results

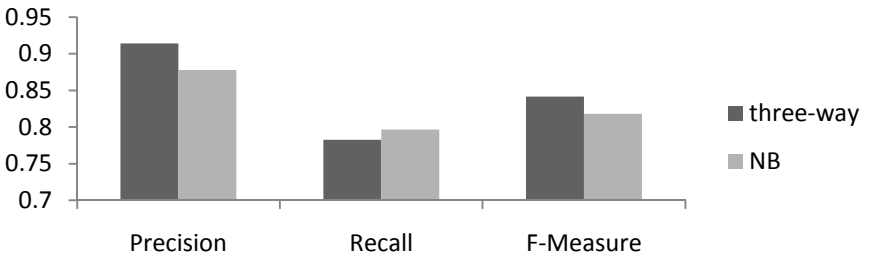

Fig. 6. Comparison of the macro-average results

By comparing the experimental results, compared with the NB method, Precision is improved by nearly 10%, F-Measure basically unchanged, Recall rate decreased slightly according to Three-Way Decision, which means the macro-average experimental result was significantly higher than NB. It described the Three-Way Decision method ,meanwhile maintaining the overall identification performance of subjective sentence, can significantly improve the accuracy of classification. There are a lot of subjective micro-blog statements in the mass micro-blog text environment. It plays a very important role for subsequent emotional research and improves the accuracy of the analysis of public opinion that improving the recognition accuracy of subjective sentence.

6 Summary and Outlook

Using three-way decision, and set the reasonable Threshold of α, β ,while maintaining the overall recognition performance of subjective sentence, can effectively improve the recognition accuracy of subjective sentence. The next job of this paper is mainly on the extraction of α and β's threshold, while researching the consideration of microblog's three decisions. And probing a more effective evaluation criteria which based on three-way decision.

Acknowledgment. The research of this paper should thanks to the author Yanhui Zhu as a visiting scholar during the University of Regina in Canada in the guidance of Professor Yiyu Yao ,Here to express my deep gratitude to Professor Yiyu Yao.

This paper is sponsored by Project supported by the National Natural Science Foundation of China under Grant（No. 61170102）and the National Social Science Foundation of China under Grant（No. 12BYY045）.

References

1. China Internet Network Information Center. The 33th China Internet Development StatisticsReport [EB/OL], http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201301/P020140221376266085836.pdf
2. Wu, Y., Jingjing, S., Jiqiang, T.: A Study on the Classification Approach for Chinese MicroBlog Subjective and Objective Sentences. Journal of Chongqing University of Technology (Natural Science) 2013(01), 51–56 (2013)
3. Patrick, P.A.P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: Proceedings of International Conference on Language Resource and Evaluation. Lisbon: [s. n.], pp. 1320–1326 (2010)
4. Yao, Y.: An outline of a theory of three-way decisions. In: Yao, J., Yang, Y., Słowiński, R., Greco, S., Li, H., Mitra, S., Polkowski, L. (eds.) RSCTC 2012. LNCS, vol. 7413, pp. 1–17. Springer, Heidelberg (2012)
5. Yao, Y.Y.: Three-way decisions with probabilistic rough sets. Information Sciences 180, 341–353 (2010)

6. Yao, Y.Y.: The superiority of three-way decisions in probabilistic rough set models. Information Sciences 181, 1080–1096 (2011)
7. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11(5), 341–356 (1982)
8. Pawlak, Z.: Rough set:theoretical aspects of reasonsing about data. Kluwer Academic Publishers, Dordrecht (1991)
9. Yao, Y.Y., Wong, S.K.M., Lingras, P.: A decision –theoretic rough set model. In: The 5th International Symposium on Methodologies for Intelligent Systems (1990)
10. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximating concepts. International Journal of Man-Machine Studies 37, 793–809 (1992)
11. Yao, Y.Y.: Probabilistic approaches to rough sets. Expert System 20, 287–297 (2003)
12. Yao, Y.Y.: Probabilistic rough set approximations. International Journal of Approximate Reasoning 49, 255–271 (2008)
13. Wong, S.K.M., Ziarko, W.: A Probabilistic Model of Approximate Classification and Decision Rules with Uncertainty in Inductive Learning. Technical Report CS-85-23. Department of Computer Science. University of Regina (1985)
14. HowNet, http://www.keenage.com/html/c_index.html
15. Yanhui, Z., Yeqiang, X., Wenhua, W., et al.: Research on Opinion Extraction of Chinese Review. In: The Third Chinese Opinion Analysis Evaluation Proceedings, pp. 126–135. Conference Publishing, Shandong (2011)
16. ICTCLAS 2014 (2014), http://ictclas.nlpir.org/
17. Xiuyi, J., Lin, S., Xianzhong, Z., et al.: Threey-way decisions theory and its applications, pp. 61–79. Nanjing University Press, Nanjing (2012)
18. China Computer Federation. The Chinese Micro-Blog Emotional Analysis and Evaluation: Sample Data Sets [EB/OL] (July 01 2012), http://tcci.ccf.org.cn/conference/2012/pages/page04_eva.html