A TARGET-DEPENDENT SENTIMENT ANALYSIS METHOD FOR CHINESE MICRO-BLOG

[1]YONGHENG WANG, [2]HANGWEI

[1,2]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
E-mail: wyh@hnu.edu.cn, weihang830314@hnu.edu.cn

Abstract- Recently with a large number of people publish Micro-blog to share their opinion, Micro-blog becomes an important data resource for opinion mining and sentiment analysis. The traditional methods usually get poor accuracy since they ignore structured semantic information and the target of sentimental expressions. In this paper we propose a novel target-dependent sentiment analysis method. This method obtains the structured information from syntax tree using the convolution tree kernel of Support Vector Machine. The syntax tree is then pruned according to target with the help of domain ontology and syntactic paths library. Using this method we can eliminate the effect of irrelevant appraisal expressions. Experimental results on two corpuses with different targets show that the performance of our method is much better than the traditional methods.

Keywords- Pruning Strategy, Sentiment Analysis, Syntax Tree, Tree Kernel

I. INTRODUCTION

Micro-blog is one of the most popular applications in the age of Web 2.0 which allows users to publish posts of up to 140 characters in length and sharing their opinions with their friends. Since its emergence, micro-blog has attracted a lot of people to record their lives, discuss the hot topics, express and share their opinions on it.It becomes an important resource of mining people's opinions and sentiments. Micro-blog provides effective data to support customer satisfaction survey, public sentiment detection, sociological research, etc. The information provided by micro-blog is various and growing rapidly. It is hard to gather and process them only relying on manual scanning. Sentiment analysis technique can be used to classify the texts automatically according to the sentiments expressed by the authors and it can solve the problem of information chaos to a great extent and help to locate the required information fast and accurately. Comparing to the traditional textual data, micro-blog posts have some new characters such as limited size, often containing spelling and grammatical mistakes, using variety of emoticons and colloquial expressions, etc. These characters make it more difficult to determine the sentiment polarity of micro-blog.

Micro-blog posts often contain diverse evaluation objects, some of which are irrelevant to the given targets. However, the state-of-the-art sentiment analysis methods basically work in a target-independent way that all the sentiment expressions appearing in the posts will be regarded as being opinionated about the target, no matter it is right or not, which will lead to some mistakes. To solve this problem, we propose a target-dependent sentiment analysis method. Frequently-appearing nouns are used to build the domain ontology. We detect the irrelevant objects in the posts and then prune the irrelevant appraisal expressions from syntax tree according to the syntactic paths library. At last, we utilize the pruned syntax trees and unigrams as classification features. Experimental results show that our method is much more effective than the traditional target-independent methods.

The following sections are arranged as follows. In section II, we introduce some related works. A brief introduction of convolution tree kernel and composite kernel is proposed in section III. The target-dependent pruning strategy is described in section IV. In section V, we introduce the experimental results. Then a discussion and conclusion is presented in section VI.

II. RELATED WORKS

A. Target-independent Sentiment Analysis

Target-independent sentiment analysis methods assign sentiment orientations to the posts but do not concern the target of the sentiments. Current methods, mainly include lexicon based and machine learning based methods, often work on a target-independent way.

The lexicon based methods perform sentiment analysis based on a function of opinion words in a sentence and use the dominant polarity of the opinion words to determine its polarity. Fiaidhi et al. used three scoring scheme to score the tweets, including subtracting positives from negatives, TF-IDF weighted scheme and Latent Dirichlet allocation (LDA).Yang et al. utilized How Net and current opinion lexicon to constructa more comprehensive lexicon for internet textual data. However, sentiment analysis for micro-blog usually get poor performance since they contain many kinds of special emotional words and expressions that do not exist in the general opinion lexicons.

The machine learning based methods treat sentiment

classification simply as a special case of text classification. These methods need large sets of labeled data to train sentiment classifiers. Pang et al. first applied machine learning based method on text classification. They selected unigram, bigram, part-of-speech(POS) tag, and the positions of opinion words as features. Naïve Bayesian(NB), Maximum Entropy(ME), and Support Vector Machine(SVM) are used to build the classifiers. Davidov et al. utilized Twitter tags and smileys as sentiment labels which avoided the need of manual annotation and allowed identification and classification of diverse sentiment types of tweets. In order to exploit the social context of tweets, Aisoposet al. introduced a method that relied on content-based and context-based features. Tan et al. and Zhanget al. combined the lexicon based and learning based method to avoid labor-intensive and time-consuming manual labeling. The performance of classifiers is also improved with their method.

B. Target-dependent Sentiment Analysis

People may mention multiple targets in one micro-blog post. Target-dependent methods classify the posts according to the sentiments expressed towards the given target. HuandLiu detected the nouns and noun phrases as product's features, and for each feature, the nearby adjective was regarded as its effective opinion. Ding and Liu modified the method in by aggregating all the opinion words in one sentence according to the distance between them and the target. Nasukawa and Yi applied the semantic analysis with syntactic parser and sentiment lexicon to identify the semantic relationships between the sentiment expressions and subjects. The polarity of sentiment expressions are determined according to the rules made manually. This kind of methods cannot analyze complicate sentences and have low recall because it is hard to make a comprehensive rule set.Jianget al. followed the learning based methods and they introduced seven kinds of target-dependent features, combining with content features and sentiment lexicon features. Their method needs to make the feature-extracting rules manually and these rules only applied to English. Lixinget al. filtered the sentences irrelevant to the target, but the filtering strategy was not so accurate and their method could not tackle the situation that one sentence might contain multiple opinion targets.

III. TREE KERNEL-BASED METHOD

Most of the learning based methods represent a text as a vector by extracting flat features from diverse resource. These flat features usually include n-gram, POS tag, negations, and so on. However, they ignore word order and syntactic relations between words in a sentence which are extremely important and useful for sentiment classification, especially in micro-blog classification because of posts' limited size and lacking of enough information for classification. To solve this problem, we extract structured features from syntax trees of micro-blog sentences, and utilize the convolution kernel of SVM to obtain structured information.

A. Convolution Tree Kernel of SVM

Convolution tree kernel(CTK) counts the number of common subtrees as the syntactic similarity between two parse trees. The similarity of two syntax trees T_1 and T_2 can be calculated by:

$$K_C(T_1,T_2) = \sum_{n_1 \in N_1} \sum_{n_2 \in N_2} \Delta(n_1,n_2) \quad (1)$$

where N_1 and N_2 are sets of nodes in tree T_1 and T_2, and $\Delta(n_1,n_2)$ is the number of the common subtrees rooted at n_1 and n_2.$\Delta(n_1,n_2)$ can be calculated in polynomial time by applying the following recursive rules:

1) If the productions at n_1 and n_2 are different, then $\Delta(n_1,n_2)=0$;

2) Else if both n_1 and n_2 are pre-terminals, $\Delta(n_1,n_2)=1\times\lambda$;

3) Else

$$\Delta(n_1,n_2) = \lambda \times \prod_{j=1}^{nc(n_1)} (1 + \Delta(ch(n_1,j), ch(n_2,j))) \quad (2)$$

where$nc(n_1)$ is the number of children of n_1, ch(n,j) is the j^{th} child of node n, and λ is a decay factor introduced to make the kernel values less variable with respect to the subtree's size.

B. Composite Kernel of SVM

Beside the structured features, the flat features are also very important for sentiment classification. The new SVM tool provides the support of composite kernel consists of convolution tree kernel and basic kernels (such as linear kernel and polynomial kernel).Therefore we can obtain information from both structured and flat features. The composite kernel can be represented as $\lambda K_1+\tau K_2$, in which K_1 is the convolution tree kernel and K_2 is the basic kernel. The parameters λ and τ indicate the contribution of these two kernels.

Panget al. had discovered that using unigram as classification feature could obtain the best performance. We also utilize the unigram as flat features. In order to remove noise and reduce dimensions, we utilize the CHI square test to filter unigrams that have low contribution to classification. We calculate the χ^2 scores for each unigram and rank them in descending order, so that we can select top n unigrams as features and discard the rest.

Given term t and class c_i, let A be the number of times t and c_i co-occur, B be the number of times t occurs without c_i, C be the number of times c_i occurs without t, D be the number of times neither t nor c_i occurs, and N be the number of documents, then χ^2 can be calculated by:

$$\chi^2(t, c_i) = \frac{N(A \times D - B \times C)^2}{(A + C)(B + D)(A + B)(C + D)} \quad (3)$$

IV. TARGET-DEPENDENT PRUNING STRATEGY

Syntax tree can provide rich information for classification, but the completed syntax tree include too much noise and the cost for classification is high. Zhanget al. provided two kinds of syntax tree pruning strategies: adjectives-based and sentiment words-based. Liet al. pruned the dependency tree according to the frequently-appearing dependence relationships. Focusing on target-dependent sentiment analysis, we propose a pruning algorithm to prune the target-irrelevant appraisal expressions from syntax trees.

A. Construction of the Domain Ontology

An ontology is defined as an "explicit, machine-readable specification of a shared conceptualization", which can be used for modeling the terms in a domain of interest andthe relations among these terms. Formal Concept Analysis (FCA) is a mathematical data analysis theory, which is a user-driven step-by-step methodology for constructing domain ontology. The basic element of FCA is the concept, which is described via extension and intension. The former is a set of objects and the latter is a set of attributes.

Kontopoulos et al. adopted FCA to build domain ontology manually, and assigned a score to every attribute of one product. We utilize a semi-automatic FCA method to construct domain ontology:

1) Count the nouns co-occurring with the target, and rank them in ascending order of frequency;
2) Select objects and attributes from the nouns set from 1) and put them into the object set and attribute set;
3) Extract the concept according to the object set and regard the object set as the extension, then extract the common attributes from the attribute set to build the intension, furthermore, extract the unique attributes of objects and attach them with the corresponding object.

The steps 2) and 3) should be conducted manually. Fig.1 shows a domain ontology of smartphone. Notably, in the area of products or services, there are competitive relations among the objects in the extension of one concept, like Nokia and Samsung in Fig.1.

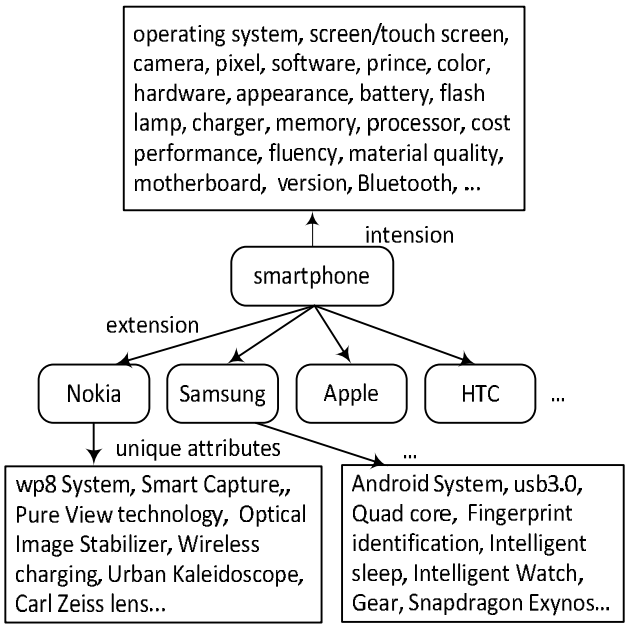


Fig. 1 Domain ontology of smartphone

B. Detection of the Appraisal Expressions

In order to prune useless part of syntax tree, we have to detect effective appraisal expressions from one sentence, which consist of the sentiment expression and its target. Zhaoet al. proposed a syntactic path based algorithm based on the assumption that, instead of disorganized, the syntactic paths between a sentiment expression and its target may occur with degree of regularity. We use similar strategy to detect appraisal expressions and provide a basis for pruning. The syntactic path between the sentiment expression and its target refers to the syntactic structure connects these two nodes in the syntax tree. Fig. 2 shows that the node "好 用 (easy-to-use)" is a sentiment expression and "相机 (camera)" is its target.The syntactic path between the two nodes is

NN→NP→NP→IP→VP→VP→VA. By counting the syntactic paths on large scale corpus, we can find that the effective paths occur much more frequently than the invalid ones. Generally, the targets are nouns which we can detect through POS tagging, and the sentiment expressions can be found by opinion lexicons. After extracting the syntactic paths from syntax tree, we need to carry out the generalization that makes paths with small difference aggregating to one path. This step aggregates the same adjacent elements in a path, for example, the path shown above can be generalized as NN→NP→IP→VP→VA. We select the most frequently-appearing paths to build the syntactic path library.

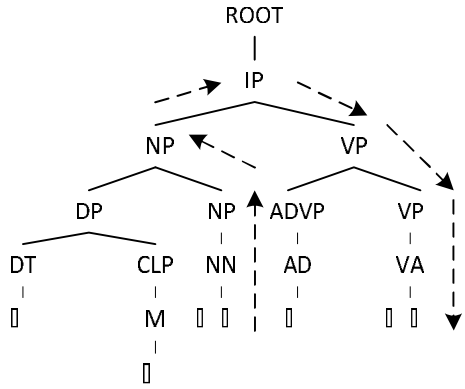


Fig.2 An example of syntax tree and syntactic path

C. Target-dependent Pruning Strategy

After the domain ontology and syntactic path library are built, we can prune the syntax tree. The basic idea is to prune the target-irrelevant parts from the syntax tree and the pruning strategy is shown as follows:

1) Obtain the noun set N and the sentiment expression set SEfrom a sentence by POS tagging and sentiment lexicon;

2) Put the objects in N but not in the ontology into the set composed by the objects to be pruned and denote it as IN;

3) Put objects in the ontology but have competitive relation with the target into set IN, if there are any attributes of these competitive objects appearing subsequently, put them into IN as well;

4) Get syntactic paths between every object inIN and every expression in SE, and find out effective paths by matching syntactic path library. Then we can obtain the appraisal expressions to be prunedand put them into set AE;

5) According to the appraisal expressions in AE, locate the corresponding nodes of sentiment expressions s and its target t, find out the common parent node p of them, then prune the subtrees of p that contains s or t. After that, if there is not any subtree rooted at p, prunep from the syntactic tree as well.

Fig. 3 illustrates the target-dependent pruning strategy. We assume that "诺基亚(Nokia)" is the target, and the syntax tree of sentence "诺基亚还不错，差评是给京东的(Nokia is good and the bad review is for JingDong)" is shown in Fig.3. We can find that the sentiment expression set is {不错(good), 差评(bad review)}, and the target set is {诺基亚(Nokia),京东(JingDong[1], a Chinese shopping website)}. With the help of domain ontology, we can find that "京东" is an object irrelevant to the target and should be pruned. The syntactic paths between (京东,不错) and (京东, 差评) are NN→NP→VP→IP→CP→IP→VP→VA and NN→NP→VP→IP→NP→NN. By matching the syntactic path library, we discover that the latter is effective but the former is not, so we can determine that "差评" is the effective sentiment expression of "京东". Weprune the corresponding structure from the syntax tree following the steps shown above and mark it by dashed line in Fig. 3.

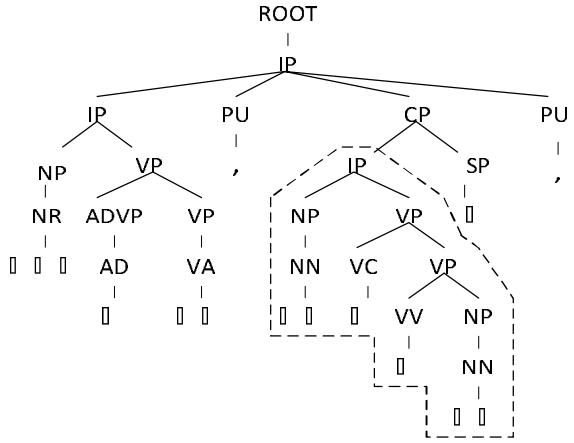


Fig.3 Pruning the syntax tree by target-dependent strategy

V. EXPERIMENTAL RESULTS AND ANALYSIS

Firstly, we need to build the syntactic path library. We utilize NLPIR Chinese segmentation system[2]for word segmentation and POS tagging, and Stanford Parser[3]for syntactic tree building.

The opinion lexicon we used is the NTUSD lexicon, which contains 2810 positive opinion words and 8276 negative ones.

Besides, we put 50 frequently-used positive internet words and 53 negative ones into the lexicon. The corpus we used to build the syntactic path library contains 36,042 micro-blog posts grabbed from Tencent[4] Micro-blog, which covers many fields like cars, smart phone, shopping websites, and so on.

Table.1 shows the five most frequently-appearing paths and their occurring-frequency.

[1]http://www.jd.com
[2]http://ictclas.nlpir.org
[3] http://nlp.stanford.edu/software/lex-parser.shtml
[4] http://t.qq.com

TABLE 1. FIVE MOST FREQUENTLY-APPEARING PATHS

Serial Number	Syntactic Path	Frequency
1	NN→NP→IP→VP→VV	17,267
2	NN→NP→NN	14,087
3	NN→NP→VP→VV	14,068
4	NN→NP→ADJP→JJ	11,056
5	NN→NP→IP→VP→VA	9,311

We select two keywords as the targets, i.e. "诺基亚 (Nokia)" and "唯品会(WeiPinHui[5], a Chinese shopping website)", and grab posts containing the targets from Tencent Micro-blog. We manually classify each post as positive or negative towards the target. The corpus with target "诺基亚" contains 304 positive posts and 340 negative posts, and corpus with target "唯品会" contains 1,212 positive posts and 1,063 negative posts.

In our experiments, we use SVM-light-TK[6] as the tree kernel-based classifier and the composite kernel-based classifier. We compare the performances of four feature selecting strategies: ①only use unigram as features and filter them by CHI square test, then use linear kernel as SVM kernel; ②only use convolution kernel and syntactic tree features; ③use unigram and syntactic tree as features, then use composite kernel as SVM kernel; ④use unigram and pruned syntactic tree as features. SVM-light-TK provides a parameter r to adjust the contribution of tree kernel and the composite kernel K can be denoted as:

$$K = Tree_Kernel \times r + Vector_Kernel \quad (4)$$

where the default value of r is 1, corresponding to the parameter λ in formula $\lambda K_1+\tau K_2$. Table.2 presents the Accuracy(A), Precision(P), Recall(R) and F-score(F) of four strategies. When it comes to composite kernel, we only present the best performance in Table.2 while r is changing.

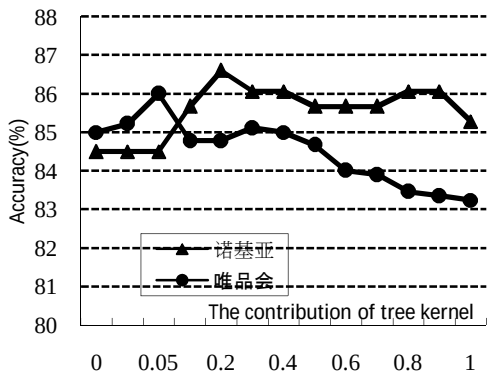


Fig.4 The relation between value of r and accuracy

[5] http://www.vip.com
[6] http://disi.unitn.it/moschitti/Tree-Kernel.htm

In our experiments, the changing of parameter r also affects the performance, and Fig.4 shows the relationship between accuracy and the value of r. From Fig.4 we can see that when target is "诺基亚" and the value of r is 0.2, our method can achieve best accuracy of 86.6%, and when target is "唯品会", we can get the best accuracy of 86.0% with the value of r is 0.05. Table 2 presents these two best performances, and we can set r to 0.05-0.2 in practice.

TABLE 2. THE PERFORMANCES OF FOUR FEATRUE SELECTING STRATEGIES(%)

Features and Kernels	诺基亚							唯品会						
	POSITIVE			NEGATIVE			A	POSITIVE			NEGETIVE			A
	P	R	F	P	R	F		P	R	F	P	R	F	
Unigram+CHI (Linear Kernel)	77.1	81.1	79.0	82.5	78.7	80.5	79.8	81.2	85.8	83.4	82.7	77.4	80.0	81.8
Syntactic tree (Convolution tree kernel)	80.7	51.2	62.5	67.3	89.0	76.6	71.2	78.4	67.8	72.6	68.0	78.1	72.6	72.6
Syntactic tree+unigram+CHI (Composite kernel)	87.1	78.6	82.5	82.7	89.7	86.0	84.5	85.8	84.7	85.2	82.8	84.0	83.4	84.4
Pruned syntactic tree+unigram+CHI (Composite kernel)	89.7	82.1	85.5	83.6	91.5	87.3	86.6	86.1	86.3	86.1	86.2	85.6	85.8	86.0

From Table 2, we can draw the following conclusions:①Our method achieves the best performance both in positive and negative post sets. The accuracy, precision, recall, and F-score of our method are all highest among all methods. This result shows that syntactic tree is a kind of useful structured feature which provides important syntactic information for micro-blog sentiment analysis. ②If we only use syntactic tree as classification features, the performance decreases sharply comparing tounigram features. The reason is that the tree kernel-based method cannot capture the flat features which are also very important for sentiment classification.

However, we can settle this problem by using composite kernel. The experimental result shows that composite kernel-based method can capture both structured and flat features and achieve better performance. ③The performance is further improved by using pruned syntactic tree. The reason is the target-dependent pruning method eliminates much noise and irrelevant structures in the syntactic tree. So we believe that the combined features representation and appreciate pruning strategy are both excellent selections for sentiment analysis of short texts, such as micro-blog posts and reviews.

CONCLUSIONS AND DISCUSSIONS

Micro-blog data are important resource for sentiment analysis but it is much more difficult to determine their polarity due to their limited size and lacking of context information. Besides flat features, structured features provided by syntax tree are also very important for classification. Based on these observations we use composite kernel of SVM to capture both flat and structured information from posts. Especially, when multiple opinion targets appear in one sentence, we prune the target-irrelevant appraisal expressions from the syntax tree with the help of domain ontology and syntactic path library. Experimental results show that

our combined feature representation and target-dependent pruning strategy can greatly improve the performance of sentiment analysis for Chinese micro-blogs. Our appraisal expression detection method relies on opinion lexicon which means we cannot prune syntax trees that do not contains any opinion words. Besides, our pruning strategy cannot handle complicate sentences like comparative sentences and transitional complex sentences. So in the future we will consider how to discovery new opinion words in the internet and find more sophisticated methods for appraisal expression detection. More appropriate pruning strategy for complicate sentences.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (No.61371116) and the Hunan Provincial Natural Science Foundation (No.13JJ3046).

REFERENCES

[1] A. Pak, and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," In Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, 2010.

[2] F. Aisopos, G. Papadakis, K. Tserpes, and T. Varvarigou, "Content vs. context for sentiment analysis: Acomparative analysis over microblogs," In Proceedings of the 23rd ACM conference on Hypertext and social media, Milwaukee, USA, pp. 187-196, July 2012.

[3] X. Lixing, Z. Ming, and S. Maosong, "Hierarchical structure based hybrid approach to sentiment analysis of Chinese micro blog and its feature extraction," Journal of Chinese Information Processing, vol. 26, no. 1, pp.73-83, Jan. 2012(in Chinese).

[4] J.Fiaidhi, O. Mohammed,S. Mohammed, S. Fong, and T. Kim, "Opinion mining over twitterspace: Classifying tweets programmatically using the R approach, " In Proceedings of the 7th International Conference on Digital Information Management, Macau, China, Aug. 2012.

[5] C. Yang, S. Feng, D. Wang, N. Yang, and G. Yu, "Analysis on web public opinion orientation based on extending sentiment lexicon," Journal of Chinese Computer Systems, vol. 31, no. 4, pp.691-695, Apr. 2010.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," In Proceedings of Conference on Empirical Methods in Natural Language Processing, Philadelphia, Pennsylvania, USA, pp. 79-86, 2002.

[7] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitterhashtags and smileys," In Proceedings of the 23rdInternational Conference on Computational Linguistics, Beijing, China, pp. 241-249, Aug. 2010.

[8] S. Tan, Y. Wang, and X. Cheng, "Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples," Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development Information Retrieval, Singapore, pp.743-744,July 2008.

[9] L. Zhang, G. Riddhiman, D. Mohamed, H. Meichun, and B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis," HP Laboratories, Tech Rep: HPL-2011-89, 2011.

[10] M. Hu, and B. Liu, "Mining and summarizing customer reviews,"In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, pp.168-177, Aug.2004.

[11] X. Ding , and B. Liu, "The utility of linguistic rules in opinion mining," In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, pp. 811-812, July 2007.

[12] T. Nasukawa, and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," In Proceedings of the 2nd International Conference on Knowledge Capture, Sanibel Island, FL, USA, pp. 70-77, Oct. 2003.

[13] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, USA, pp.151-160, Jun. 2011.

[14] W. Zhang, P. Li, and Q. Zhu, "Sentiment classification based on syntax tree pruning and tree kernel," In Proceedings of the 7th Web Information Systems and Applications Conference, Inner Mongolia, China, pp.101-105, Aug.2010.

[15] P. Li, Q. Zhu, and W. Zhang, "A dependency tree based approach for sentence-level sentiment classification," In Proceedings of the 12th ACIS International Conference on Software Engineering, Artificial Intelligence, NetWorking and Parallel/Distributed Computing, Sydney, Australia, pp. 166-171, Dec.2011.

[16] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, "Ontology-based sentiment analysis of twitter posts," Expert System with Applications, vol. 40(10), pp.4065-4074, Aug.2013.

[17] Y. Zhao, B. Qin, W. Che, and T. Liu, "Appraisal expressions recognition based on syntactic path,"Joural of Software, vol.22, no.5, pp.887-898, May 2011(in Chinese).

★★★