ICIC Express Letters
Volume **10**, Number **7**, July **2016**

 ISSN 1881-803X
pp. 1591–1597

STUDY ON SUBJECTIVE STATEMENT SCREENING OF PRODUCT REVIEWS BASED ON THE FEATURE COMBINATION METHOD

Tieli Sun[1,2], Weiqiao Guan[3], Fengqin Yang[1,2,*], Hongguang Sun[1,2,*]
Chuande Ji[1], Meijia Song[1] and Yanwen Li[1,2]

[1]School of Computer Science and Information Technology
Northeast Normal University
[2]Key Laboratory of Intelligent Information Processing of Jilin Universities
No. 2555, Jingyue Ave., Changchun 130117, P. R. China
*Corresponding authors: {yangfq147; sunhg889}@nenu.edu.cn

[3]Flight Theory Department
Aviation University of Air Force
Changchun 130022, P. R. China

Received December 2015; accepted March 2016

Abstract. *Subjective statement screening is a fundamental part of sentiment analysis of product reviews. Its aim is to extract the subjective texts from product reviews and filter objective texts that do not contain any sentimental information. The existing screening methods are mainly based on only one feature. However, single feature is insufficient to mark all subjective or objective statements. We propose a feature combination method based on the N-pos feature, the word feature and the syntactic dependency feature for subjective statement screening. The experiments on the real datasets show that the proposed feature combination method significantly improves the accuracy of subjective statement screening and is really feasible and effective.*

Keywords: Product reviews, Feature combination, Sentiment analysis, Subjective statement screening

1. **Introduction.** With the rapid development of electronic commerce, many product reviews emerge on the Internet. For customers and manufacturers, product reviews can provide a lot of useful information [1]. Sentiment analysis of product reviews has become a hot topic in the field of information processing. Product reviews include the subjective statements and the objective statements. Subjective statements express the views and feelings of users and are data source of sentiment analysis. To avoid the interference of the objective statements, we need to screen the subjective statements from product reviews and filter the objective statements. Therefore, subjective statement screening is a fundamental and significant task for sentiment analysis of product reviews [2].

The subjective statement screening methods can be divided into the semantic method, the statistical method and the method based on graph. Yao and Peng [3] selected the stable features including personal pronouns and interjections, and verified the effectiveness of the semantic method. Ye et al. [4] proposed the method that computes the subjective degree of sentences according to the part of speech (POS) of continuous double words. Zhang et al. [5] proposed three syntactic structure templates for the subjective expressions. The statements that satisfy one of the proposed templates are regarded as the subjective statements.

Due to the flexibility and complexity of natural language, the subjective statement screening has become a challengeable task. We cannot obtain ideal results only using one feature. To capture more subjective information, we adopt the word feature, the POS feature and the syntactic structure feature simultaneously to extract subjective statements. Firstly, we construct the subjective and objective word features. Secondly, we use the

N-pos model to express the texts and extend N to 5. Thirdly, we extract the syntactic dependency feature on the basis of syntactic parsing and use the labels of the syntactic structure as statistical features. Finally, the above three features are combined to screen subjective statements. Experimental results on the real datasets show that our feature combination method improves the accuracy of subjective statement screening significantly and it is feasible and effective.

2. **Related Work.** Wiebe et al. [6] chose the POS feature, punctuation and the sentence position as the features to express subjective emotions of English texts. The average accuracy of subjective statement selection is 72.17%.

In N-pos model, the words in sentences are classified according to POS and n successive POS from an N-pos tuple. The appearance of a word in the N-pos model depends on the POS of $N-1$ word ahead [7]. The text can be represented by a vector consisting of all N-pos tuples in the text.

Definition 2.1. [8]: *$G = \{g_1, g_2, \ldots, g_n\}$ is a set of POS, $g(w)$ is the POS of the word w, $g\left(w_{i_1}^{i_2}\right)$ denotes $g_{w_{i_1}}, g_{w_{i_1+1}}, \ldots, g_{w_{i_2}}$, and c represents the context of the word w. Supposing a sentence s consists of the word sequence $w_1, w_2, \ldots, w_n$, the occurrence probability of the word w_i is defined as Formula (1) in N-pos model.*

$$P(w_i = w|c) = P\left(w_i = w|g\left(w_{i-N+1}^{i-1}\right)\right) \tag{1}$$

[4] used the Bi-pos model. For example, in the sentence "*Its price is very affordable, my friends like it!*", all Bi-pos for the above sentence are "*r-n*", "*n-v*", "*v-adv*", "*adv-adj*", "*adj-pun*", "*pun-pron*", "*pron-n*", "*n-v*", "*v-pron*" and "*pron-pun*". Each item is a Bi-pos and reflects the subjective emotion. So they are called subjective Bi-pos patterns. [9] extended the Bi-pos model and proposed Tri-pos model. The Tri-pos model can expand the context of the words.

[10] used the pre-defined syntactic structure templates to classify the text sentiment and obtained high accuracy. However, this method needs to analyze a large corpus and summarize the rules manually.

[3] shows that product reviews usually contain the words as follows. (1) Emotional words. They usually refer to adjectives that can express opinions and feelings. (2) Degree adverbs. In order to emphasize feelings or opinions, users usually use emotional degree adverbs before adjectives. (3) Indicative verbs. People use words such as "*advocate*", "*consider*" and "*feel*" after personal pronouns to express opinions. (4) Modal words. They are commonly used at the end of the sentence to express a kind of tone. These words are also important to extract the subjective statements.

In this paper, we extend the N-pos feature with the word feature and syntactic dependency feature to describe the language model more accurately.

3. **Our Feature Combination Method for Subjective Statement Screening.** Usually, product reviews contain two parts, which are the subjective and objective statements. Subjective statements reflect opinions and emotional expression. Objective statements are the description of facts. To analyze the polarity of product reviews, it is essential to screen subjective statements. In this paper, we use the word, POS and the syntactic structure as subjective statement features of product reviews.

3.1. **The word feature of the subjective and objective statements.**

(1) The subjective word feature

We select the word feature of subjective statements by using the word feature in [3] and HowNet [11]. We use 6 word lists in HowNet. They are the positive sentiment word list, the negative sentiment word list, the positive evaluation word list, the negative evaluation word list, the degree adverb list and the indicative verb list.

(2) The objective word feature

Through the analysis of a large number of product reviews, we find that the objective statements in product reviews usually contain the words as follows. (1) The third person pronouns. In the objective statements, the third person often appears, such as "*father*" and "*colleagues*". (2) Verbs. Verbs often appear in objective product reviews, such as "*buy for someone*" and "*take for someone*", where the "*buy*" and "*take*" are verbs. The real user is not the buyer. (3) Status words. For example, in the sentence "*The TV is in use and I'm not sure it is good or not.*", "*in use*" is a status word. Users tend to know the product quality and function after a period of use. In this paper, we manually sort out above words as word features of the objective statements.

3.2. **The *N*-pos feature.** The N-pos model is the simplification of the N-gram model. In the N-pos model, the parameter N is the order of the model and its value determines the accuracy and complexity of the model. As the N's value increases, we can describe the dependency among words more accurately and the model becomes more complicated at the same time. So the appropriate N's value is a compromise between the accuracy and complexity of the model [12]. In this paper, we extend the N-pos pattern to different value of N. In this way, we can expand the context of words and describe the language model more detailedly and accurately. N-pos pattern extraction mainly includes 4 steps as follows.

Step 1. We preprocess and parse the product reviews using ICTCLAS *(http://ictclas.nlpir.org/)*.

Step 2. We construct the N-pos ($N = 2, 3, 4, 5$) statistical language model, i.e., the Bi-pos model at $N = 2$, the Tri-pos model at $N = 3$, the Tetra-pos model at $N = 4$ and the Penta-pos model at $N = 5$.

Step 3. We calculate the CHI square statistics [13] of subjective and objective N-pos ($N = 2, 3, 4, 5$) patterns and sort them.

Step 4. We extract a certain number of N-pos patterns as features according to the value of CHI square statistics for each type. The first 5 Bi-pos and Tri-pos patterns are shown in Table 1 and Table 2, respectively.

Table 1. The top five Bi-pos pattern

No.	subjective pattern	objective pattern
1	pron-v	v-pron
2	aux-pun	v-aux
3	int-pun	aux-pun
4	adv-adj	num-quan
5	prep-adj	quan-quan

Table 2. The top five Tri-pos pattern

No.	subjective pattern	objective pattern
1	adv-adj-adj	quan-v-v
2	n-adj-num	pun-conj-pron
3	pron-quan-aux	n-adj-n
4	adv-adj-int	pron-quan-v
5	v-quan-adv	nums-noun of locality-nums

3.3. **Syntactic dependency feature.** [14] used syntactic structure templates summarized from training corpus manually to extract opinion sentences. However, this method can only obtain some special syntactic structures. On the contrary, we use syntactic structure labels as the statistical features to select subjective sentences. Thus, subjective sentences selected are not restricted by syntactic structure templates.

The extraction algorithm of the syntactic dependency feature is as follows.

Step 1. We preprocess the product reviews and tag POS on them using ICTCLAS.

Step 2. We do syntactic analysis on the product reviews by Stanford parser *(http://nlp.stanford.edu-/software/lex-parser.shtml*).

Step 3. We remove the words and POS tags, and retain the syntactic dependency labels which can be used as the classification features.

3.4. **Our feature combination method.** Through the above processes, we have obtained the word feature, the N-pos feature and the syntactic dependency feature, respectively. By directly concatenating the above three kinds of feature vectors, we get a combined feature vector. Then we use Chi-square to select a certain number of features as effective features from the combined feature vector. Finally, we apply TF-IDF [15] to weighting the selected effective features and obtain a combined feature vector for subjective statement screening of product reviews.

4. **Experiments and Analysis.** In the experiments, the experimental dataset includes 2700 TV and 4335 iPad air product reviews crawled from Jingdong Mall (*www.360buy.com*). Four students with different professional backgrounds classify the product reviews into subjective and objective statements according to the unified annotation requirements. Finally, we select 2000 TV and 3143 iPad air product reviews as subjective statements, and 700 TV and 1192 iPad air reviews as objective statements. For each category, we extract the 2/3 statements as the training set and the rest as test set.

The experimental procedures are shown as follows.

(1) Preprocessing data. Segmenting words and tagging POS of the words with ICTCLAS (*http://ictclas.nlpir.org/*).

(2) Extracting the features of subjective and objective word, the N-pos pattern and syntactic dependency, respectively.

(3) Combining the three features effectively. By directly concatenating each feature vector together, we get a high dimensional feature vector.

(4) Applying Chi-square to selecting 150 features as effective features from the high dimensional feature vector.

(5) Applying TF-IDF to weighting the selected effective features and obtaining a combined feature vector.

(6) Using Naïve Bayesian classifier in the Weka platform (*http://weka.wikispaces.com/*) to extract subjective statements.

The evaluation standards of classification are mainly the precision (P), recall (R) and F value. They are defined as Formulas (2)-(4).

$$P = \frac{A}{A+B} \tag{2}$$

$$R = \frac{A}{A+C} \tag{3}$$

$$F = \frac{2 * P * R}{P+R} \tag{4}$$

where, A is the number of statements assigned correctly to the current class, B is the number of statements that do not belong to the current class but are assigned to the class and C is the number of statements that actually belong to the current class but are not assigned to the class.

We conduct the experiments using different combined features. In the end, we select the optimal experimental results for each combined feature. The experimental results on TV and iPad air product reviews are shown in Table 3 and Table 4, respectively. Sub_P, Sub_R and Sub_F denote the performance of the subjective statement screening, and Ob_P, Ob_R and Ob_F denote the performance of the objective statement screening.

Table 3. The experimental results of different combined features on TV product reviews

Combination	Sub_P	Sub_R	Sub_F	Ob_P	Ob_R	Ob_F
Bi-pos	**0.947**	**0.692**	**0.800**	**0.618**	**0.927**	**0.741**
Tri-pos	0.959	0.670	0.789	0.613	0.948	0.745
Tetra-pos	0.953	0.608	0.743	0.559	0.943	0.702
Penta-pos	0.710	0.996	0.829	0.911	0.091	0.166
Bi-Tri-pos	0.923	0.726	0.813	0.635	0.887	0.740
Bi-Tri-Tetra-pos	**0.964**	**0.704**	**0.814**	**0.633**	**0.951**	**0.760**
Bi-Tri-Tetra-Penta-pos	0.941	0.673	0.785	0.602	0.921	0.728
WF (Word Feature)	0.958	0.728	0.827	0.650	0.941	0.768
WF-Bi-Tri-pos	**0.970**	**0.792**	**0.872**	**0.712**	**0.954**	**0.815**
WF-Bi-Tri-Tetra-pos	0.957	0.780	0.860	0.707	0.938	0.807
WF-Bi-Tri-Tetra-Penta-pos	0.963	0.745	0.840	0.666	0.947	0.782
SDF (Syntactic Dependency Feature)	0.909	0.555	0.689	0.520	0.896	0.658
SDF-Bi-Tri-pos	0.911	0.783	0.842	0.801	0.919	0.856
SDF-Bi-Tri-Tetra-pos	0.929	0.807	0.864	0.821	0.935	0.875
SDF-Bi-Tri-Tetra-Penta-pos	**0.939**	**0.819**	**0.875**	**0.832**	**0.944**	**0.884**
WF-SDF-Bi-Tri-pos	0.930	0.844	0.885	0.850	0.933	0.890
WF-SDF-Bi-Tri-Tetra-pos	0.948	0.860	0.902	0.866	0.951	0.906
WF-SDF-Bi-Tri-Tetra-Penta-pos	**0.949**	**0.870**	**0.908**	**0.874**	**0.951**	**0.911**

Table 4. The experimental results of different combined features on iPad air product reviews

Combination	Sub_P	Sub_R	Sub_F	Ob_P	Ob_R	Ob_F
Bi-pos	**0.957**	**0.798**	**0.870**	**0.630**	**0.906**	**0.743**
Tri-pos	0.954	0.683	0.796	0.509	0.910	0.653
Tetra-pos	0.983	0.528	0.687	0.398	0.971	0.564
Penta-pos	0.797	1.000	0.887	1.000	0.003	0.006
Bi-Tri-pos	**0.962**	**0.770**	**0.855**	**0.602**	**0.919**	**0.728**
Bi-Tri-Tetra-pos	0.964	0.760	0.850	0.594	0.926	0.724
Bi-Tri-Tetra-Penta-pos	0.965	0.758	0.849	0.592	0.927	0.722
WF (Word Feature)	0.984	0.712	0.826	0.561	0.969	0.711
WF-Bi-Tri-pos	**0.978**	**0.769**	**0.861**	**0.611**	**0.954**	**0.745**
WF-Bi-Tri-Tetra-pos	0.979	0.752	0.850	0.594	0.957	0.733
WF-Bi-Tri-Tetra-Penta-pos	0.978	0.748	0.848	0.590	0.956	0.730
SDF (Syntactic Dependency Feature)	0.929	0.554	0.694	0.430	0.888	0.580
SDF-Bi-Tri-pos	0.955	0.789	0.864	0.618	0.901	0.733
SDF-Bi-Tri-Tetra-pos	**0.965**	**0.803**	**0.877**	**0.640**	**0.923**	**0.756**
SDF-Bi-Tri-Tetra-Penta-pos	0.960	0.768	0.853	0.599	0.915	0.724
WF-SDF-Bi-Tri-pos	0.978	0.794	0.877	0.637	0.954	0.764
WF-SDF-Bi-Tri-Tetra-pos	**0.983**	**0.812**	**0.890**	**0.661**	**0.963**	**0.784**
WF-SDF-Bi-Tri-Tetra-Penta-pos	0.980	0.772	0.864	0.615	0.959	0.749

From Table 3 and Table 4, we can find that the Bi-pos feature has obtained the best results among all single N-pos features because the Bi-pos feature is a kind of stable feature in natural language, such as the adverb + adjective pattern. However, the Bi-pos pattern is too short to model all pattern information in the review texts. So the POS features with different lengths compensate each other and the multiple combined N-pos features improve the performance of any single N-pos feature generally. Due to the sparsity of the Tetra-pos feature in reviews, the Tetra-pos feature does not improve the results. The word feature is as important as the POS feature and the syntactic dependency feature also plays a role that cannot be ignored. The word feature and syntactic dependency feature also improve the performance of the multiple combined N-pos features. The feature combination method based on the multiple N-pos features, the word feature and the syntactic dependency feature gets the best results. The experimental results also indicate that the word and the syntactic dependency feature are essential to the subjective statement screening of product reviews. Obviously, our proposed method is effective to screen subjective statements.

5. **Conclusions.** This paper focuses on subjective statement screening of product reviews based on the feature combination. We combine the N-pos feature with the word feature and the syntactic dependency feature. The experimental results on the real datasets show the proposed method significantly improves the results obtained by any single feature and our proposed feature combination method is effective.

In the future work, we will summarize the difference between the subjective and objective statements of the product reviews from the aspects of network emotional symbols, and extract more detailed features.

Acknowledgments. This paper was sponsored by Jilin Provincial Science and Technology Department of China (Grant No. 20130206041GX), Changchun Science and Technology Bureau of China (Grant No. 14KP009), Jilin Province Development and Reform Commission of China (Grant No. 2013C036-5, [2013]779, 2014Y096), and the Doctoral Program of Higher Education of China (No. 20110043110011), respectively.

REFERENCES

[1] J. Zhong, S. Y. Yang and Q. G. Sun, Sentiment analysis for goods evaluation based on text classification, *Journal of Computer Applications*, vol.34, no.8, pp.2317-2321, 2014.

[2] J. S. Guerrero, J. A. Olivas, F. P. Romero and E. H. Viedma, Sentiment analysis: A review and comparative analysis of web services, *Information Sciences*, vol.311, pp.18-38, 2015.

[3] T. F. Yao and S. W. Peng, A study of the classification approach for Chinese subjective and objective texts, *The 3rd National Conference on Information Retrieval and Content Security*, pp.117-123, 2007.

[4] Q. Ye, Z. Q. Zhang and Z. X. Luo, Automatically measuring subjectivity of Chinese sentences for sentiment analysis to reviews on the Internet, *Journal of Information Systems*, vol.1, no.1, pp.79-91, 2007.

[5] B. Zhang, Y. Zhou and Y. Mao, Extracting opinion sentence by combination of SVM and syntactic templates, *International Conference on IEEE Natural Language Processing and Knowledge Engineering (NLP-KE)*, pp.1-7, 2010.

[6] J. M. Wiebe et al., Development and use of a gold standard dataset for subjectivity classifications, *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, pp.246-253, 1999.

[7] Z. Q. Zhang, *Patterns of Word Class Combination for Sentiment Analysis in Chinese*, Master Thesis, Harbin Institute of Technology, 2007.

[8] Y. Guan, K. Zhang and G. H. Fu, Computational language model based on statistics, *Application Research of Computers*, vol.6, pp.26-28, 1999.

[9] P. Zhao, Z. W. Zhao and X. Z. Tao, Analysis method of subjective and objective Chinese texts based on sematic TriPos model, *Application Research of Computers*, vol.29, no.9, pp.3285-3288, 2012.

[10] W. Y. Li and H. Xu, Text-based emotion classification using emotion cause extraction, *Expert Systems with Applications*, vol.41, no.4, pp.1742-1749, 2014.

[11] X. H. Fu, G. Liu and Y. Y. Gu, Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, *Knowledge-Based Systems*, vol.37, no.2, pp.186-195, 2013.

[12] Y. K. Xing and S. P. Ma, A survey on statistical language models, *Computer Science*, vol.30, no.9, pp.22-26, 2003.

[13] M. C. Wang, Z. Wang and K. Zhang, Rough set texts classification rule extraction based on CHI value, *Journal of Computer Applications*, pp.1026-1028, 2005.

[14] J. M. Chenlo and D. E. Losada, An empirical study of sentence features for subjectivity and polarity classification, *Information Sciences*, vol.280, pp.275-288, 2014.

[15] T. Joachims, A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization, *Proc. of the 14th International Conference on Machine Learning*, pp.143-151, 1997.