International Journal of Database Theory and Application
Vol.7, No.6 (2014), pp.191-202
http://dx.doi.org/10.14257/ijdta.2014.7.6.17

A Hybrid Strategy for Fine-Grained Sentiment of Microblog

Ouyang Chunping, Luo Lingyun, Zhang Shuqing and Yang Xiaohua

School of Computer Science and Technology, University of South China, Hunan Hengyang, 421001, China
ouyangcp@gmail.com

Abstract

Currently, most sentiment analysis of microblog has been focused on coarse-grained sentiment analysis, but fine-grained sentiment is better for reflecting the opinion of the public when they are facing the social focus. Therefore, a hybrid strategy which is a combination of Naïve Bayesian and two-layer CRFs is put forward, which has been applied to the fine-grained sentiment analysis of Chinese microblog. First, microblog is classified into two types: sentiment and non-sentiment by using Naïve Bayesian classification algorithm. And then the first-layer CRFs model is built for the topic emotional sentence. Finally CRFs algorithm is used again to do multi-classification to assign a specific sentiment category. Experimental results show that a good result in sentiment identification based on the combination of Naïve Bayesian and CRFs, and also show the advantage of the combination of Naïve Bayesian and CRFs interrelated with emotional sentence extraction based on CRFs.

Keywords: *microblog; fine-grained sentiment analysis; Naïve Bayesian; Conditional Random Fields*

1. Introduction

Along with the development of Web 2.0, social media platform is becoming a very important tool for sharing public information. As a state-of-art social media platform, microblog has attracted very large amount of users because it is quick, abbreviate and efficient in information spreading [1]. Take Sina microblog for example, currently, the number of its registered users has become 300 million, and the number of blogs they posted daily has broken through 100 million [2]. Information users posted on microblog is usually their reactions and opinions to recent events. Since sentiment expresses people's inner feeling, it is an important source for recognizing people's opinion to events. Therefore, sentiment recognition and analysis for microblog plays an important role in monitoring public opinions through large amount of microblog data.

Sentiment analysis can be divided into two kinds: coarse-grained and fine-grained. Coarse-grained sentiment analysis only uses two categories, namely, "positive" and "negative", to recognize the sentiment in microblog. However, fine-grained sentiment can usually better demonstrate users' opinions, which divides coarse-grained sentiment further into people's real sentiments such as "joy", "happy", "sorrow", "disgust", etc. As for now, research on microblog sentiment analysis in the literature mostly focuses on coarse-grained sentiment analysis, which decides if the micropost expresses positive or negative feeling. Two methods are usually used: machine-learning or semantic-calculation.

The process for machine-learning based sentiment analysis goes as follows: At first, select and define the features of microblog, and train the classifying model using algorithms such as Bayesian, Maximum Entropy, K-Nearest Neighbor, SVM, to finally achieve the goal of classifying microblogs' sentiments into positive or negative[6, 7, 8, 9]. Different from the machine learning method, semantic-calculation based sentiment analysis firstly define a sentiment thesaurus, then calculate the weight of each feature word using mapping rules to complete the classification [10, 11]. The critical part in the machine-learning based method is

ISSN: 2005-4270 IJDTA

the selection of features, while that for the semantic-calculation based method is the completeness of the thesaurus. They are both critical to the final classification results. As a result, some researchers proposed a way to combine the two methods. For example, Jiang [13] and DmitryDavidiv [14] used sentiment thesaurus to help features selection, along with using the machine learning method to recognize the positive or negative sentiments is effective on English microblogs. However, because of the complexity of Chinese words, the sentiment thesaurus for Chinese is less complete than that for English, which costs efficiency in later classification.

To deal with difficulties in Chinese microblogs such as oral, non-normalized words and signals, native researchers tried to improve the representation of text features. Xie [15] combined 4 features together and used SVM method to analyze sentiments in Sina microblogs. Results showed that the highest accuracy is 66.467% and 67.283%, using subject-related and non-subject-related features separately. Xu and Lin [16] took words and structures in sentences into account, selected 9 semantic features which affect the sentiment of the whole sentence, and combined manual and automatic ways together to construct the sentiment ontology. Their method was an initial attempt in semantic based sentiment analysis. Li and Cao [17] studied from the linguistics perspective and leveraged the "sentiment tendency definition" calculation method based on weight-first to get the feature word which best represents the sentiment tendency among all the words in the phrase. Then they analyzed the sentiment tendency and its strength of the phrase according to the way the characteristic words were combined. The method is meaningful in fine-grained text sentiment analysis.

For some applications based on microblog (such as public opinion supervision), fine-grained sentiment analysis explores the blogger's opinion more accurately, and detects users' sentiment changes more effectively. So, we propose a hybrid approach of native Bayesian and two-layer CRF to achieve fine-grained microblog sentiment analysis. Applying two-layer CRF method on the classification results of the native Bayesian model can reduce chaos, thus improve the accuracy of the result, and two-layer CRFs model is adopted to shorten the training time and improve results.

The paper is organized as follows: Chapter 2 introduces the whole framework; Chapter 3 explains how to discriminate between sentiment and non-sentiment using the native Bayesian classification algorithm and Chapter 4 illustrates the design of two-layer CRF model for fine-grained sentiment analysis; Chapter 5 includes analysis and discussion of the results; Conclusion and further work are made in Chapter 5.

2. Hybrid Strategy for Fine-Grained Sentiment Analysis

The aim of fine-grained sentiment analysis of microblog is to classify each blog's sentiment into one of the followings: anger, disgust, fear, happiness, like, sadness and surprise. Since microblog sentiment analysis is based on the fact that the microblog does have sentiment, we should first decide whether the microblog has sentiment or not, through which the fine-grained sentiment analysis accuracy can be improved. We propose an integrated method in this paper, comprised of three steps: First, we use the native Bayesian classification algorithm to decide if the microblog has sentiment or not; Secondly, we extract emotional sentence of microblog using the CRFs algorithm for binary classification; And then CRFs algorithm is used again to do multi-classification to assign a specific sentiment category. In Figure 1 our pipeline for analyzing fine-grained sentiments in microblog is shown.

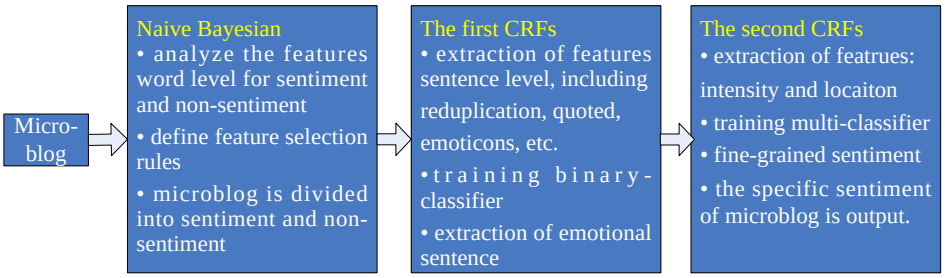


Figure 1. Analysis Pipeline for Fine-grained Sentiment in Microblog

Whichever algorithms we execute, the main task is feature selection.

Feature selection for Naïve Bayesian classifier: selecting the feature word that can be used to draw a clear distinction between sentiment and non-sentiment.

Feature selection for the first CRFs: several features are extracted for emotional sentence extraction, which include word unigram, POS features, syntactic features, sentiment features, emoticons, and reduplication.

Feature selection for the second CRFs: we analyze the influential factor of identifying the specific sentiment of microblog on two major grounds, which are intensity of sentiment and location of sentiment sentence.

3. Naïve Bayesian Classifier Design

3.1. Feature Selection

Feature selection is a data pre-processing procedure in text classifying algorithms. Words which obviously represent the category of the text are selected, while others are deleted. For such purpose, we need to do characteristic selection after parsing to differ feature words with sentiment from those without sentiment. In this paper, we first parse the microblogs in our training set, and then calculate the occurrences of each word under two conditions: sentiment and non-sentiment. At last, according to formula (1), we select those words whose occurrences changed dramatically under the two conditions as the characteristic words.

$$\frac{wordfreq_{1-i}}{wordfreq_i} > k \,\&\&\, wordfreq_{1-i} > n \,(i = 0,1) \qquad (1)$$

When i=0, it shows that the ratio of two frequency is greater than K, which the denominator is the frequency of the word examined in non-sentiment, and the numerator is the frequency of the word examined in sentiment. Moreover the frequency of the word examined in sentiment is greater than n. When i=1, the formula operate the opposite way around.

After selecting some feature words, we do the part of speech (POS) analysis on the remaining words. We discover that POS of some word occurrences differ a lot when expressed subjectively or objectively. For example, in the training set, [/face] signals appeared 972 times in blogs with sentiments and 138 times in microblogs without sentiments. As a result, we choose 19 POS of word whose occurrences changed dramatically from microblogs with sentiment to those without and add them to the feature word set. They are: [/t]、[/m]、[face]、[/o-y-e-ry]、[/rz]、[/z]、[/b]、[/ne]、[/n]、[/w]、[/rr]、[/r]、[/a]、[/v]、[/d]、[/p]、[/c]、[/u], and [/s].

3.2. Naïve Bayesian Classification Model Design

Naïve Bayesian classification algorithm (NB) is a hypothesis-based prior probability learning algorithm based on the Bayesian theorem. Intuitively, we assume the class set is pre-defined. To decide which class the new term belongs to, we calculate the probability of each class' appearance under the condition that the new term appears. The class with the biggest probability is the one that the term belongs to.

Definition 1. Let $A=\{a_1,a_2,a_3,...,a_m\}$ where A is a term to be classified, and a_i is the characteristics for A. Let $B=\{b_1,b_2,b_3,...,b_n\}$ where B is the class set and b_i represents each class. Calculate $P(b_1|A)$, $P(b_2|A)$, $P(b_3|A)$, … , and $P(b_n|A)$ respectively. If there exists $P(b_k|A)=\max\{P(b_1|A),P(b_2|A),...,P(b_n|A)\}$, then $A\in b_k$. This process is called naïve Bayesian classification.

From Definition 1, we can see that calculating the conditional probabilities is the most critical step. To calculate the value of $P(b_i|A)$, we need to conduct supervised learning. The annotated training set is used to collect the conditional probabilities of each characteristic property under different classes, i.e., $P(a_1|b_1)$, $P(a_2|b_1)$, …, $P(a_m|b_1)$; $P(a_1|b_2)$, $P(a_2|b_2)$, …, $P(a_m|b_2)$; $P(a_1|b_n)$, $P(a_2|b_n)$, … , $P(a_m|b_n)$. Consider the fact that each characteristic property is independent, we get the formula of calculating conditional probabilities of classes based on the Bayesian theorem as follows:

$$P(b_i|A)=P(A|b_i)P(b_i)/P(A)=P(a_1|b_i)P(a_1|b_i)...P(a_m|b_i)P(b_i)=P(b_i)\prod_{j=1}^{m}P(a_j|b_i) \qquad (2)$$

$P(b_i)$ is the prior probability for the class b_i, $P(a_j|b_i)$ is the posterior probability of a_j under class b_i. In our paper, we only have two classes b_0 and b_1, representing the class with sentiment and the class without sentiment, separately. The formulas for calculating the prior and posterior probabilities are shown in formula (3) and (4) separately.

$$p(b_i)=\frac{\sum(a|b_i)}{A}(i=0,1) \qquad (3)$$

$$p(a_j|b_i)=\frac{\sum F(a_j)|b_i+1}{\sum(a|b_i)+\sum c} \qquad (4)$$

$\sum(a|b_i)$ denotes that the total number of feature words belong to category b_i; A denotes the total number of feature words in training set; $\sum(a|b_i)$ denotes the frequency of a certain word in category b_i; $\sum c$ denotes the total number of non-repeated feature words. According to the above formulas, the process of classifying microblogs can be divided into three steps:

Step 1: Pre-processing. The purpose is to prepare for naïve Bayesian classification. We first construct the feature vectors on the training set after parsing, and then choose the proper calculation model. The polynomial model and the multi-variate bernoulli model are two commonly used Bayesian models for text classification. The polynomial model works on words, the representative variable for each word is the times it occurs in the file. The multi-variate bernoulli model works on files, the representative variable for each word is Boolean. Consider the facts that the amount of words in microblogs is small and word is a very important unit in microblogs, we choose the polynomial Bayesian classification model in our experiment.

Step 2: Classifier training. The aim of this procedure is to generate the classifier. Mainly, we calculate the probabilities that the two classes (with or without sentiment) appear in the

training set, denoted as $P(b_i)$ (i=1 represents the one with sentiment, vice versa). We also calculate the conditional probability of every feature word under sentiment and non-sentiment respectively. In this process, the input data includes the featrue sets and the training samples. The output is a classifier.

Step 3: Differ sentiment microblog from non-sentiment microblog. The aim of this procedure is to divide microblogs into the two classes: the class with sentiment and the class without sentiment. At first, we construct the microblog's feature vector. Then we calculate the product of the probabilities of each characteristic word under the two conditions (sentiment and non-sentiment) separately, which are formula (5) and (6):

$$P1 = P(sentiment)\prod_{j=1}^{m} P(a_j \mid sentiment) \quad (5)$$

$$P2 = P(non-sentiment)\prod_{j=1}^{m} P(a_j \mid non-sentiment) \quad (6)$$

After comparing the values of P1 to P2, we choose the bigger one as the sentiment class the microblog belongs to. In this process, the input data includes the classifier and the characteristic vector of the blog to be classified. The output is a mapping between microblogs and the two sentiment classes.

4. Sentiment Analysis based on Conditional Random Fields

Using the naïve Bayesian classification model, microblogs can be divided into the sentiment and non-sentiment. The objective of fine-grained sentiment analysis is to refine the emotional microblog. According to sentiment thesaurus constructed by Dalian University of Technology, fine-grained sentiment is divided into seven kinds, including anger, disgust, fear, happiness, like, sadness, surprise. The one of the seven sentiments is outputted as the final result. In order to accomplish the above tasks, two-layer Conditional Random Fields (CRFs) model is proposed. Once used in the emotional sentence extraction, another is used in the fine-grained sentiment analysis.

4.1. Conditional Random Fields Model

Conditional Random Fields is one notable variant of a Markov random field. Conditional Random Fields are a type of discriminative undirected probabilistic graphical model. The nodes in the graph represent random variables; the edges between the nodes represent dependencies between random variables.

X is given the observed values, and Y is a set of random variables. Conditional Random Fields are used to calculate the conditional probability of random variables Y on observations X. CRFs define as follows:

Definition [19]: Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$, so that Y is indexed by the vertices of *G*. Then (X, Y) is a conditional random field in case, when conditioned on *X*, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v \mid X, Y_w, w \neq v) = p(Y_v \mid X, Y_w, w \sim v)$, where $w \sim v$ means that *w* and *v* are neighbors in *G*.

From the view of the fundamental theorem of Markov random field, Conditional Random Fields is composed of an undirected graph and a set of potential functions cliques. The conditional probability distribution represented by a Conditional Random Field is given by

$$P(Y \mid X) = \frac{1}{Z_x} \prod_{c \in C} \Phi_c(x_c, y_c) \quad (7)$$

where Z_x, known as the partition function(Normalization factor), is given by

$$Z_x = \sum y \bullet \prod_{c \in C} \Phi_c(x_c, y_c) \quad (8)$$

$\Phi_c(x_c, y_c)$ is the potential function for the state of the variables that appear in *c* clique, *c* represents *a* clique in all cliques *C* , x_c and y_c represent the variables participated in clique *c* . Conditional Random Fields are often conveniently represented as log-linear combination. Each clique potential is replaced by an exponentiate weighted sum of features of the state:

$$\Phi_c(x_c, y_c) = \exp(\sum_k \lambda_{kc} f_{kc}(x_c, y_c)) \qquad (9)$$

Conditional random fields commonly used in the labeling or analysis of sequence data, such as Natural Language Processing, e.g. [20]. In principle, the graph model layout of conditional random fields can be arbitrarily given, but popular layout is a linear chain. It is because that there are efficient algorithms on the training, inference, or decoding of the linear chain CRFs.

Use of sentiment-related words is not the sole means of expressing emotion. Often a sentence, which otherwise may not have a sentiment word, may become emotion bearing depending on the context or underlying semantic meaning. Sometimes, the emotional expressions contain direct sentiment words, reduplication, colloquial words and foreign words. On the other hand, the emotional expressions contain indirect sentiment words and emoticons. As the microblog corpus has sentence level sentiment annotations, sentiment words can not be used as the sole standard for the sentiment expression of microblog. For example, two sentences in a microblog, "sunny" occurs in a subjective sentence, "boring" occurs in a objective sentence, the sentiment of microblog will expressed by the sentiment annotation of subjective sentence. Thus extracting emotion sentence can clear away the obstacles and improve the precise. To address the problem, we proposed a two-layer CRFs model:

- First layer CRFs: input is a sentence of the microblog, output is "S-Yes" or "S-No". "S-Yes" denotes the sentence is an emotions sentence of the microblog, "S-No" is on the contrary. In this layer, feature selection focus words, including direct sentiment words, indirect sentiment words and emoticons, and so on.
- Second layer CRFs: input is the result of the first step, and output is a one of the seven sentiments including anger, disgust, fear, happiness, like, sadness, surprise. The intensity of sentiments and location information of sentiment sentence are the features in this layer.

4.2. Different Features Definition

An important characteristic of Conditional Random Fields model is that various features can be flexibly defined and additional independence assumptions or internal constraints do not need to be considered. On the other hand, it also shows that feature selection has a great influence on the final results. According to the research result, the combination of multiple features in comparison with a single feature generally shows a reasonable enhancement of any classification system [21]. Consequently, through manually reviewing the microblog corpus and their language specific characteristics, the feature sets for emotional sentence extraction and fine-grained sentiment analysis are defined, as shown in the Table 1 and Table 2.

Table 1. Feature Set for Emotional Sentence Extraction

Feature	Feature symbol	Description
Emoticons	Em	Judge the emoticons whether to be included in the sentence. For example,  often occurs in the emotional sentence.
POS information	POS	Chinese word is tagged "noun,verb,adj,adv, nw,idiom,prep", and statistics for each part of

		speech in the sentence frequency.
Reduplication	Dup	Use "Dup-0" and "Dup-1" to represent indicate whether the repeated words. For example, Chinese sentence "he is 太(very)太(very)太(very) good"; The repeated word "very" is used to express praise, so the sentence is considered as an emotional sentence.
Special punctuation symbols	SPS	"?" or "!" is included in the sentence, which is used as an intensive word in emotional sentence.
Quoted sentence	Qu	Judge double quotation marks whether to be included in the sentence. Generally, quoted sentence do not contain author's emotion.
Context feature	CF	Parallelism sentence is usually used to make statements, so which is not considered as an emotional sentence.

Emotional sentence extraction is binary classification based on CRFs, and then fine-grained sentiment analysis is multi-classification based on CRFs, the feature set is shown in Table 2.

Table 2. Feature Set for Multi-classification CRFs

Feature	Feature symbol	Description
ntensity	AN, DI, FE, HA, LI, SA, SU	According to sentiment thesaurus, each sentiment word has intensity of sentiment. For example, "AN-8" denotes that the intensity of sentiment "anger" is eight in this sentence. "DI-0" denotes that sentiment "disgust" is not included in this sentence sentiment.
Location	Loc-F,Loc-M,Loc-L	Loc-F: emotional sentence at the beginning of a mciroblog text. Loc-F: emotional sentence at the middle of a mciroblog text. Loc-F: emotional sentence at the end of a mciroblog text.

4.3. Model Training

We train our CRFs model using CRF++, a highly efficient general purpose CRFs toolkit written in C++ [22]. CRF++ allows the definition of both unigram and bigram features, where unigram features are related to the prediction of a single observation in a sequence (first order Markov) and bigram features are related to the prediction of pairs of observations (second order Markov). Unigram features generate a total of L×N distinct features, where L is the number of output classes and N is the number of unique features. Bigram features generate L×L×N distinct features. In our task, the first layer CRFs employs unigram template, bigram template is adopted by the second layer CRFs, and the window size of feature template both are 3.

When running the train command "%crf_learn template_file train_file model_file>> info.txt", the parameter "-f NUM" is set to "-c 5", which meant that only the features appears not less than 5 times will be calculated when CRF++ is training. Additionally, in the case of the CRFs trained on independent observations, we remove all but node features, so as to avoid an artificial decrease in performance. At last, command "%crf_test –m model_file test_file >>result.txt" is running, and then the accuracy of model is achieved by assessing the file "result.txt".

5. Experimental Results

5.1. Second-order Headings

Experimental data is from NLP&CC 2013 microblog sentiment analysis corpus. The corpus is mainly from Sina microblog, a total of 4000 annotated sample data and 10000 annotated testing data are included. In this corpus, not only the sentiment of microblog is annotated, but also the sentiment of each sentence is annotated. A hybrid strategy is presented in experiment scheme, different machine learning algorithms are used to solve different problem. Firstly, microblog is divided to sentiment and non-sentiment by using Naïve Bayesian. In this part, extract 3500 microblogs from 4000 microblogs annotated manually as the training corpus. Then CRFs is adopted to analyze the fine-grained sentiment of microblog. In order to shorten the training time and improve results, we have adopted a two-layer CRFs model. In the first layer, we carried out the binary classification, determine whether a sentence is an emotional sentence, then do multi-classification to assign a specific sentiment category. In this part, 2171 emotional microblogs annotated manually from 4000 microblogs is extracted as the training corpus.

We have been applying "NB+SVM" and "NB+KNN" into the sentiment analysis of Chinese microblog, and have taken part in the NLPCC 2013 evaluation. The effectiveness of the above methods are proved by the good ranking of the subimitted results [23]. To verify the effectiveness of the two-layer CRFs, some comparison experiments are carried out, data selection in detail is shown in Table 3.

Table 3. Feature Set for Multi-classification CRFs

Model		Training Set	Test Set
SVM		3500 sample data	500 sample data
KNN		3500 sample data	500 sample data
NB+SVM	NB	4000 sample data	10000 testing data
	SVM	2172 emotional sample data annotated manually	6593 emotional testing data calculated automatically using NB
NB+KNN	NB	4000 sample data	10000 testing data
	KNN	2172 emotional sample data annotated manually	6593 emotional testing data calculated automatically using NB
NB+CRFs	NB	4000 sample data	10000 testing data
	First layer	2172 emotional sample data annotated manually	6593 emotional testing data calculated automatically using NB
	Second layer	The number of microblog remains unchanged, but the number of sentence decreased significantly.	

5.2. Results and Observations

Wherever Times New Roman is specified, Times Roman, or Times may be used. If neither is available on your word processor, please use the font closest in appearance to Times New Roman that you have access to. Please avoid using bit-mapped fonts if possible. True-Type 1 fonts are preferred.

In our experiments, ICTCLAS released by CAS is employed to segment words and to tag POS and CRF++0.58 is employed to train model. In additional, the classification and

intensity of sentiment word is based on sentiment thesaurus constructed by Dalian University of Technology, which is enriched by some network vocabulary and emoticons.

Three evaluation metrics, that is, Precision (P), Recall (R), and F1-measure (F) are used for performance evaluation for binary classification of sentiment. And three other metrics, Macro Precision (MP), Macro Recall (MR), Macro F1-measure (MF) are used for evaluating the efficiency of fine-grained sentiment analysis. The formulas of evaluation standards are listed as follows.

$$\text{Precision} = \frac{\#system\_correct(emotion = Y)}{\#system\_proposed(emotion = Y)} \tag{10}$$

$$\text{Recall} = \frac{\#system\_correct(emotion = Y)}{\#annotated\ (emotion = Y)} \tag{11}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

$$Macro\_\text{Precision} = \frac{1}{6}\sum_i \frac{\#system\_correct(emotion = i)}{\#system\_proposed(emotion = i)} \tag{13}$$

$$Macro\_\text{Precision} = \frac{1}{6}\sum_i \frac{\#system\_correct(emotion = i)}{\#annotated(emotion = i)} \tag{14}$$

$$Macro\_F1 = \frac{2 \times Macro\_\text{Precision} \times Macro\_\text{Recall}}{Macro\_\text{Precision} + Macro\_\text{Recall}} \tag{15}$$

In above formulas, #system_correct denotes the number of the submitted results which match with the manually annotated results, #system_proposed denotes all the number of the submitted results, #annotated denotes the number of the manually annotated results, and i denotes a specific sentiment belongs to the seven sentiment.

For creating the results, we carried out five classification methods, including SVM, KNN, NB+SVM, NB+KNN, NB+CRFs, the last three of which are hybrid method. In Table 4 the comparison results for sentiment identification are provided, and the results for fine-grained sentiment analysis is shown in Table 5. Taken as a whole, it is that a satisfactory result can be acquired adopting hybrid methods, and the efficiency of sentiment identification is greater than fine-grained sentiment classification. In Table 4, the results show the hybrid methods that provide the higher performance. But "NB+SVM" and "NB+CRFs" give identical results, because they only hire Naïve Bayesian to identify the sentiment of microblogs. Moreover, for "NB+KNN" method, although classification for sentiment and non-sentiment using Naïve Bayesian, the final results will be modified by KNN model, hence "NB+KNN" has high performance than two other hybrid methods. From the results of Table 5, "NB+CRFs" has the highest accuracy in MP and MF. Thus it can be seen that two-layer CRFs can improve the precision of fine-grained sentiment analysis. The results also show that the emotional sentence extraction based on CRFs has contributed to the final results.

Table 4. Experimental Result of Sentiment Identification

	Precision	**Recall**	**F-measure**
SVM	0.6212	0.7274	0.6701
KNN	0.6495	0.7220	0.6838
NB+SVM	0.6626	0.8012	0.7254
NB+KNN	0.6676	0.7982	0.7271
NB+CRFs	0.6626	0.8012	0.7254

Table 5. Macro Average of Fine-grained Sentiment Classification

	Macro Precision	Macro Recall	Macro F-measure
SVM	0.1922	0.1805	0.1862
KNN	0.2088	0.2308	0.2192
NB+SVM	0.2109	0.1996	0.2051
NB+KNN	0.2704	0.3064	0.2873
NB+CRFs	0.2899	0.2998	0.2948

Those results lead to the conclusion, that the method using "NB+CRFs" model is a suitable choice for the fine-grained sentiment analysis of microblog. With sentiment identification using Naïve Bayesian model results in a better performance.

6. Conclusions and Future Work

We presented a hybrid sentiment analysis approach for detecting seven sentiment classes on microblog. We contribute a mean for decreasing non-sentiment microblog and non-topic sentence interfere with the final results. We are able to identify two classes of sentiments with accuracy of 72.71% and senven classes with accuracy of 29.48%. Based on our approach, NB+CRF, it is possible to detect more microblog that are relevant for public opinion monitoring system than using traditional sentiment analysis methods.

Future work needs to be done to enhance the current machine learning models. In this case, first, a much larger labeled training set is needed. Second, feature selection for CRF should be further research. Study on how to combine the different feature and test the optimum combination. Furthermore, the hybrid strategy should be tested in different types of corpus to check its robustness.

Acknowledgements

This research work is supported by Hunan Provincial Natural Science Foundation of China (No.13JJ4076), the Scientific Research Fund of Hunan Provincial Education Department for excllent talents (No.13B101), Foundation of University of South China (No.2012XQD28), the Construct Program for the Key Discipline in University of South China (No.NHxk02), the Construct Program for Innovative Research Team in University of South China.

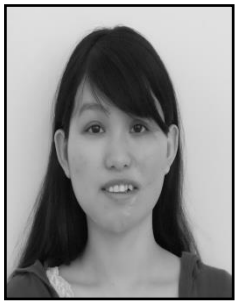
References

[1] Z. S. Chen, Q. W. Ting, S. Y. Zi, S. X. Zhi and S. Y. Chen,"Overview on Sentiment Analysis of Chinese Microblogging", Computer Applications and Software,vol. 30, no. 3, **(2013)**.
[2] C. Zhang, D. Zeng, J. Li, F. Y. Wang and W. Zuo,"Sentiment Analysis of Chinese Documents: From Sentence to Document level", Journal of the American Society for Information Science and Technology, vol.60, no.12, **(2009)**.
[3] C. R. Fink, D. S. Chou, J. J. Kopecky and A. J. Llorens,"Coarse- and Fine-Grained Sentiment Analysis of Social Media Text", Johns Hopkins APL Technical Digest, vol. 30, no. 1, **(2011)**.
[4] C. Zirn, M. Niepert, H. Stuckenschmidtand M. Strube, "Fine-Grained Sentiment Analysis with Structural Features", Proceedings of Fifth International Joint Conference on Natural Language Processing,**(2011)**; Chiang Mai, Thailand.
[5] B. Yuan, Y. Liu, H. Li, T. T. T. Phan, G. Kausar, C. N.Sing-Bik and W. Wahi, "Sentiment Classification in Chinese Microblogs: Lexicon-based and Learning-based Approaches", vol. 30, no. 3, **(2013)**.
[6] E. Kouloumpis, T. Wilson and J.Moore, "Twitter sentiment analysis: The good the bad and the omg!", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, **(2011)**.
[7] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data", Proceedings of COLING, **(2010)**; Beijing, China.
[8] L. J. Hao, Y. A. Min, Z. Y. Mei and C. Z. Jian, "Clssification of Microblog Sentiment Based on Naïve Bayesian", Computer Engineering & Science, vol. 34, no. 9, **(2012)**.

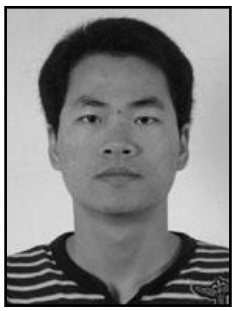
[9] W. Yang, J. Song and J. Q. Tang, "A Study on the Classification Approach for Chinese MicroBlog Subjective and Objective Sentences",Journal of Chongqing Institute of Technology, vol. 27, no. 1, **(2013)**.
[10] H. Saif, Y. He and H. Alani, "Semantic Sentiment Analysis of Twitter", Proceedings of the 11thInternational Semantic Web Conference, **(2012)**; Boston, USA.
[11] P. Lei, L. S. Shan, Z. G. Dong, "Sentiment Classification Method of Chinese Micro-blog Based on Emotional Knowledge", Computer Engineering, vol. 38, no. 13, **(2012)**.
[12] J. L. Zhang, S. T. Han, J. Wan, B. J. Zhu, L. Zhou, Y. J. Ren and W. Zhang,"IM-Dedup: an image management system based on deduplication applied in DWSNs", International Journal of Distributed Sensor Networks, vol. 2013, **(2013)**.
[13] L. Jiang, M. Yu, M. Zhou, X. Liu and T. Zhao, "Target-dependent Twitter Sentiment Classification", Proceedings of the 49 Annual Meeting of the Association for Computational Linguistics, **(2011)**; Oregon, USA.
[14] J. L. Zhang, W. Qian, X. H. Xu, J. Wan, Y. Y. Yin and Y. J. Ren, "WLBS: A W, "WLBS:AWeight-based Metadata Server Cluster Load Balance Strategy", International Journal of Advancement in Computing Technology, vol. 4, no. 1, **(2012)**.
[15] X. L. Xing, Z. Ming and S. M. Song, "Hierarchical Structure Based Hybrid Approach Sentiment Analysis of Chinese Microblog and Its Featrue Extraction", Journal of Chinese Information Processing, vol. 26, no. 1, **(2012)**.
[16] X. L. Hong, L. H. Fei, "Discourse Affective Computing Based on Semantic Features and Ontology", Journal of Computer Research and Development, vol. 44, no. 3, **(2007)**.
[17] D. X. Shuang, Z. Qibo and G. Yi, "A Survey on Sentiment Analysis Models", Scientific Journal of Psychology, vol. 1, no. 2, **(2012)**.
[18] X. L. Hong, L. H. Fei, P. Yu, R. Hui and C. Jianmei, "Constructing the Affective Lexicon Ontology", Journal of the China Society for Scientific and Technical Information, vol. 27, no. 2, **(2008)**.
[19] J. Lafferty, A. McCallum and F. Pereira, "Conditional random fields: Probabilisticmodels for segmenting and labeling sequence data", Proceeding of the 18th International Conference on Machine Learning,**(2001)**; Williamstown, MA, USA.
[20] Liu B, "Sentiment analysis and subjectivity", Handbook of natural language processing, vol. 2, **(2010)**.
[21] D. Das, S. Bandyopadhyay, "Emotion Tagging–A Comparative Study on Bengali and English Blogs", ICON, vol. 9, **(2009)**.
[22] T. Kudo, "CRF++: Yet another CRF toolkit", Software available at https://code.google.com/p/crfpp/downloads/list, **(2014)**.
[23] O. C. Ping, Y. X. Hua, L. L. Yan, X. Qiang, Y. Ying and L. Z. Ming, "Multi-strategy Approach for Fine-grained Sentiment Analysis of Chinese Microblog", ActaScientiarumNaturaliumUniversitatisPekinensis, vol. 50, no. 1, **(2014)**.

Authors

Ouyang Chunping received her Ph.D. degree from University of Science Technology Beijing, China, in 2011. Now, she is working in University of South China, where she is an associate professor and the dean of software engineering department. Her research interests are in Semantic Web, social computing and sentiment analysis of text.



Luo lingyun was born in 1981. She received his Ph.D. degree from CAS, China, in 2010. She is lecturer in the school of Computer and Technology, the University of South China. Her current research interests focus on Semantic Web and medical informatics.



Zhang Shuqing received his B.E. degree from Department of Computer Science of Huanghe Science and Technology College, Zhengzhou, in 2011. Now, he is working his M.E. Degree at University of South China. His research interests include natural language processing and data mining.

Yang Xiaohua received his Ph.D. degree from CAS, China, in 1999. From 2000 to 2001, he was a visiting scholar at Wollongong University. Yang is vice-president of University of South China, and professor of computer science, doctoral tutor. He has authored and coauthored over 50 research publications in peer-reviewed reputed journals. He has served as the program committee member of various international conferences and reviewer for various international journals. His research interests include natural language processing and information retrieval.