

# Chinese Microblogs Sentiment Classification using Maximum Entropy

Dashu Ye, Peijie Huang<sup>\*</sup>, Kaiduo Hong, Zhuoying Tang, WeijianXie, Guilong Zhou

College of Mathematics and Informatics, South China Agricultural University,

Guangzhou 510642, Guangdong, China

dashuye4@stu.scau.edu.cn, pjhuang@scau.edu.cn, kd\_hong@163.com,  
tangzhy@yeah.net, tsewkviko@gmail.com, zglong2016@163.com

## Abstract

This paper presents our Chinese microblog sentiment classification (CMSC) system in the Topic-Based Chinese Message Polarity Classification task of SIGHAN-8 Bake-Off. Given a message from Chinese Weibo platform and a topic, our system is designed to classify whether the message is of positive, negative, or neutral sentiment towards the given topic. Due to the difficulties like the out-of-vocabulary Internet words and emoticons, polarity classification of Chinese microblogs is still an open problem today. In our system, Maximum Entropy (MaxEnt) is employed, which is a discriminative model that directly models the class posteriors, allowing them to incorporate a rich set of features. Moreover, oversampling approach is used to hand the unbalance problem. Evaluation results demonstrate the utility of our system, showing an accuracy of 66.4% for restricted resource and 66.6% for unrestricted resource.

## 1 Introduction

Recent years have witnessed the tremendous growth of the online social media. In China, Weibo, a Twitter-like microblog service, attracted millions of users. Unlike traditional blogs, microblogs are comparatively short (140 words max at a time), instantaneous, and fast-spreading, which means, when some events happen, people's attitude towards them can be found on the Weibo platform (Such as Sina, Tencent, NetEase etc.) immediately. And connected by online social ties, their comments

are likely to affect other users who read them or even the subsequent development of the event.

Since the enormous amount of users and its great effect, people find it necessary to take an insight look at this new form of message. Researches on microblogs fall into multiple areas, such as extraction of messages (Liu et al., 2012), extraction of opinion sentence (Ding, Liu, 2008; Liu et al., 2013), and determination of sentiment orientation (Ding, Liu, 2008; Go et al., 2009; Zhang et al., 2014). Generally speaking, researchers want to find out what people think through what they post on Weibo platform.

Weibo users share their different ideas towards a same topic, and these messages they post may be of positive, negative or neutral sentiment. By classifying the polarity of a piece of microblog, we can find out an overall attitude towards the very topic of the user who posts it. Therefore, sentiment polarity classification is undoubtedly a hotspot of microblog-based research. Nowadays, sentiment polarity of microblogs has been used in many fields, such as predicting book sales (Gruhl et al., 2005), predicting movie sales (Mishne et al., 2006), predicting future product sales (Liu et al., 2007) and investigations of the relations between breaking financial news and stock price changes (Schumaker et al., 2009). Moreover, the indirect assessment of public mood or sentiment from the results of soccer games (Edmans et al., 2007) and from weather conditions (Hirshleifer et al., 2003) have been proposed.

It is common scenery using machine learning approach such as Naïve Bayes and SVM to modeling the sentiment polarity by vectorizing the message under the technology of bag-of-words. But models can be easily suffered for the sparsity of a data matrix, especially when it comes to modeling a short text. We use Maximum Entropy (MxEnt for short) to perform

---

<sup>\*</sup> Corresponding author

satisfactory results. As a discriminative model, MaxEnt directly model the class posteriors, allowing them to incorporate a rich set of features without worrying about their dependencies on one another, which gives us a rather flexible way to construct appropriate features to cope with problem. In addition, oversampling approach is used to handling the unbalance problem. Evaluation results demonstrate the utility of the proposed method.

The rest of this paper is structured as follows. Section 2 describes the background of the task and related work. In Section 3, we briefly present the proposed CMSC system. Section 4 elaborates on the constructing of our system. Section 5 describes the experimental evaluation and the results analysis. Finally, the last section summarizes this paper and describes our future work.

## 2 Background and Related Work

In recent years, sentiment analysis (SA) has made a hit in the NLP research community (Jiang et al., 2011). Lots of areas are being researched, such as emotion tagging, emotional element extraction, polarity classification and so on. As to the sentiment classification area, two kinds of methods have been used for text-based sentiment classifications.

The first one is relied on rules and lexicon containing a specific sentiment (Ding, Liu, 2008). It simply accumulates the number of lexicon expressing the same emotion for a given text, independently. And output the corresponding emotion with the highest frequencies. However, the shortcoming is it relies too much on the quality of sentiment lexicon and thus hard to cover the network language arose spontaneously.

The other one is mainly based on the machine learning approach. These method have been employ in text classification and continue to be used in short-text like microblog sentiment classification. Classical model like Naïve Bayes and SVM can be found among the text mining. Turney (2002) applied unsupervised learning method on review classification. Similar work in Movie-Review domain using supervised machine learning technique is researched by Pang et al.(2002) and Go et al. (2009) who use the emoticon in twitter and build the model using MaxEnt, NB and SVM. In Chinese microblogs, abundant emoticon may be more useful in classification. Tang et al. (2014) employ deep learning (DL) method for twitter sentiment

classification. Also there are some other methods being used now, e.g. KNN, RNN. However, most of the existing approaches use the bag-of-words technology. As it is known to all, microblog with a limitation of no more than 140 Chinese characters, bag-of-words technology may bring in a challenge of feature sparsity.

In our method, we utilize the flexibility of the feature function in Maximum Entropy approach to incorporate these two kinds of methods mentioned above. We use the segmentation result instead of using the word vector directly which increase dimension of the feature space obviously. In addition, we add rule-based feature which is a supplement of the feature.

## 3 System Overview

The flowchart of the proposed Chinese microblog sentiment classification (CMSC) system is shown in Figure 1.

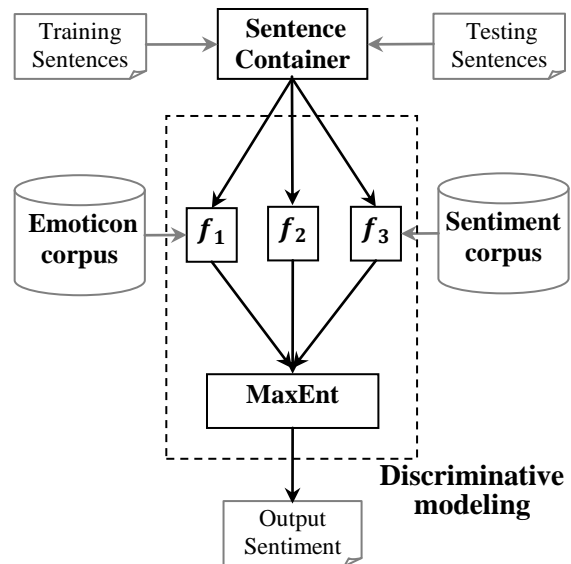


Figure 1. Flowchart of the CMSC system.

The system can be separated mainly into four parts: sentence container, language model, emoticon corpus and sentiment corpus. It performs microblogs sentiment classification as the following step:

Step 1: A given sentence is required to put into the sentence container, and only the Chinese characters and some specific notation remain in the sentence after this phase. We found that phrase containing digital number like ‘2014年8月15日’ and punctuations like ‘,’ or ‘。’ do not make any sense when it comes to the predicting phase, as a result they are suggested to

be deleted from the original sentence before coming to the next part.

Step 2: Extract structured feature using the feature functions. 3 feature functions were employ into the language model in current system. Feature function  $f_1$  compares the number of emoticon expressing the similar sentiment,  $f_2$  makes original mircroblog message as input and yields word segmentation, feature function  $f_3$  just like  $f_1$ , but replaces the emoticons by sentiment words.

Step 3: In this training phase, Maximum Entropy (MaxEnt) regarded as discriminative model yields a satisfactory performance. MaxEnt, one of the most power approach in linguistics modeling, directly models the class posteriors, allowing them to incorporate a rich set of features no need for considering their independence.

Step 4: As the same as training phase, when it comes to the predicting phase, we wanted to extract the structured features among the given testing sentence, and apply them into trained MaxEnt model, to get the corresponding output sentiment.

## 4 Maximum Entropy based CMSC System

### 4.1 Maximum Entropy Modeling

Now we give a brief introduction to Maximum Entropy (MxEnt for short) for our CMSC system. MaxEnt has a wide application in real word especially in statistical modeling and pattern recognition (Berger et al., 1996).

Given a set of training data  $\{(x_i, y_i)\}_{i=1}^N$ , where where  $x_i$  represents for the contextual information and  $y_i$  stands for the corresponding target output. MaxEnt is derived from the idea that we wanted to find a most uniform distribution under the given constraints:

$$C \equiv \{p \in P | p(f_i) = \tilde{p}(f_i) \text{ for } i \in \{1, 2, \dots\}\}, \quad (1)$$

where  $P$  is the whole hypothesis space,  $p(f_i)$  is the expected value with respect to  $p(y|x)$ , namely the entire conditional probability distribution given by the model, while  $\tilde{p}(f_i)$  is the expected value with respect to empirical function  $\tilde{p}(x, y)$ .  $f_i$  named as feature function or feature for short, describes the relation we interested in, between input  $x$  and the target output  $y$ . As usual, it can be presented in the form of:

$$f(x, y) = \begin{cases} 1 & \text{if } y \text{ and } x \text{ agree with some relation,} \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

which act key role when making a decision. We will discuss the construction of feature function in the coming section.

To find out the most uniform distribution  $p^*$ , a mathematical theory was used to measure the uniformity of conditional distribution  $p(y|x)$ :

$$H(p) \equiv - \sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x) \quad (3)$$

With this definition in hand, we derived

$$p^* = \underset{p \in C}{\operatorname{argmax}} H(p) \quad (4)$$

Naturally, the learning model is equivalent to optimize the following function with constrains:

$$\begin{aligned} \max_{p \in C} H(p) &= - \sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x). \quad (5) \\ \text{s. t. } \sum_{x,y} \tilde{p}(x) p(y|x) f(x, y) &= \sum_{x,y} \tilde{p}(x, y) f(x, y), \\ \sum_y p(y|x) &= 1. \end{aligned}$$

We can transform this constrain problem to unrestraint one using the Lagrange multipliers from the theory of constrained optimization. And for short, we finally get the parametric form of maximum entropy principle:

$$p_\lambda(y|x) = \frac{1}{Z_\lambda(x)} \exp \left( \sum_i \lambda_i f_i(x, y) \right). \quad (6)$$

Here  $Z_\lambda(x) = \sum_y \exp(\sum_{i=1}^n \lambda_i f_i(x, y))$  is called the normalizing constant, just for meeting the constrain of  $\sum_y p(y|x) = 1$ .

### 4.2 Feature Function

In this subsection, we introduce the feature construction of our CMSC system. Feature function regarded as central to the performance of MaxEnt, gives us a flexible way to express interesting evidence.

Generally speaking, feature function can broadly split into observation feature and statistical one.

For example, sentence containing positive words may convey positive sentiment. And we can roughly construct a feature function

$$f(x, y) = \begin{cases} 1 & y = 1 \text{ and } \text{contain\_posi\_element}(x), \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

Feature function of this kind directly finds out the interested evidence in the given sentence without any calculation.

Another kind of feature function utilizes statistic approach to dig out the latent information which may get a magical performance. For example, it can be sometimes, as easily as

$$f(x, y) = \begin{cases} 1 & y = -1 \text{ and } \text{CEM}(x) = 3, \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

where  $\text{CEM}(x)$  counts appearances of exclamation mark among the given sentence.

In CMSC system for about 3 feature functions were used to encoding the evidence:

$$f_1(x, y) = \begin{cases} 1 & \text{CPE}(x) > \text{CNE}(x), \\ -1 & \text{CPE}(x) < \text{CNE}(x), \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

where  $\text{CPE}(x)$  calculates the number of positive emoticons, while  $\text{CNE}(x)$  calculates the number of negative emoticons. Emoticon is one of the most expressive elements among Chinese microblog. Not only can users type character to compose emoticon like ‘:p’, but can also utilize the system build-in emoticon distinguished by square brackets ‘[微笑]’, which will shows in a more lively way ‘😊’. For the diversification and the irregularity of character-composed emoticon, we just consider the build-in one in our CMSC system.

$$f_2(x, y) = \text{WS}(x), \quad (10)$$

where  $\text{WS}(x)$  returns a vector of words derived from the word segmentation of given sentence  $x$ . Unlike English and other language, Chinese sentences compose of single characters. As a result, word segmentation technology can split sentence into words without losing its original semantic in some way. Knowing about the shortcoming of bag-of-words technology that introduces a vast scale of zeros, we just directly use the word vector as part of input feature shrinking the feature space from ten thousands

down to tens and without bringing in redundant zeros. Jieba word-segmentation tool<sup>1</sup> was used to enhance the performance. Jieba segmentation tool provides 3 patterns of word segmentation, including default mode, full mode and search engine mode.

$$f_3(x, y) = \begin{cases} 1 & \text{CPW}(x) > \text{CNW}(x), \\ -1 & \text{CPW}(x) < \text{CNW}(x), \\ 0 & \text{otherwise} \end{cases}, \quad (11)$$

where  $\text{CPW}(x)$  returns the appearances of positive word of a given sentence and  $\text{CNW}(x)$  counts the negative one. According to our intuition, sentences that contain much more positive word are more likely to convey positive emotion. Although sentence may contain no words conveying sentiment, it’s reasonable to categorize this sentence to the group of neutral sentiment.

### 4.3 Sample Weighting and Validation

In this subsection we talk about parameters decision and other tricky way to enhance the system performance. From the training data we learn that the total count of sentences with a specific sentiment is 900. This may result in unbalance problem. The class imbalance problem typically occurs when, in a classification problem, there are many more instances of some classes than others. In such cases, standard classifiers can be suffered by the large classes and ignore the small ones (Chawla et al., 2004). Existing method dealing with unbalance data as sampling methods which utilize sampling techniques to balance the data set can make a satisfactory performance. Under-sampling approach randomly picks up similar size of samples from the majority one, in order to generate a relatively balanced data set. Over-sampling balances the data set by reuse minority one.

Taking our situations, which we totally get for about 5000 training message including 400 messages of positive sentiment and other 500 messages conveying negative emotion, we weight every class in the minority side, to prevent from under fitting.

Here arises a question about how much weighting is suitable for mitigating the impact bringing by the unbalance problem. Validation set was used to figure out a proper weight adding to the minority class. Though experiments, it turn

<sup>1</sup> [github.com/fxsjy/jieba](https://github.com/fxsjy/jieba)

out that by using sample weighting method our system gets significant improvement with 10%-20% performance gain.

## 5 Experimental Evaluations

### 5.1 Task Description

In task of SIGHAN-8 Bake-Off, rules of Topic-Based Chinese Message Polarity Classification is as follows: Given a message from Chinese Weibo platform (Such as Sina, Tencent, NetEase etc.) and a topic, classify whether the message is of positive, negative, or neutral sentiment towards the given topic. For messages conveying both a positive and negative sentiment towards the topic, whichever is the stronger sentiment should be chosen.

Each participant is required to submit two kinds of results based on: (1) restricted resource for fair comparison, e.g. same sentiment lexicon, corpus, etc. that will be announced together with the test data; and (2) unrestricted resource. We believe that a freely available, annotated corpus that can be used as a common testbed is needed in order to promote research that will lead to a better understanding of how sentiment is conveyed in tweets and texts.

The evaluation metrics of both kinds of results, including precision rate, recall rate and F1-score, is provided by the Topic-Based Chinese Message Polarity Classification Task group. The confusion matrix shown in Figure 2 is to measure the related indicators.

Confusion Matrix		System Results	
		Positive	Negative
Gold Standard	Positive	TP	FN
	Negative	FP	TN

Figure 2. Confusion matrix.

Each index measure is as follows:

$$Precision (P) = TP / (TP + FP), \quad (12)$$

$$Recall (R) = TP / (TP + FN), \quad (13)$$

$$F1-score = 2 * P * R / (P + R). \quad (14)$$

Each indicator:

Precision (P): Precision rate of all sentiment Polarity Classification

Recall (R): Recall rate of all sentiment Polarity Classification

Precision+ (P+): Precision rate of positive sentiment Polarity Classification

Recall+ (R+): Recall rate of positive sentiment Polarity Classification

Precision- (P-): Precision rate of negative sentiment Polarity Classification

Recall- (R-): Recall rate of negative sentiment Polarity Classification.

### 5.2 Datasets

#### Corpus

Participants were asked to report results based on 2 kinds of resources, namely restricted resource and unrestricted resource.

The restricted resource is made up by two words set, the emotion ontology set provided by Dalian University of Technology and the sentiment words set provided by National Taiwan University (NTUSD). And we extract a vocabulary of 38553 words, 14039 words for positive sentiment, 19059 words for negative sentiment and 5376 words for neutral sentiment. We found that so many words in that file are useless since most of the people don't use that word in microblogs. We ignore such word and then construct a lexicon with the form shown as Figure 3.

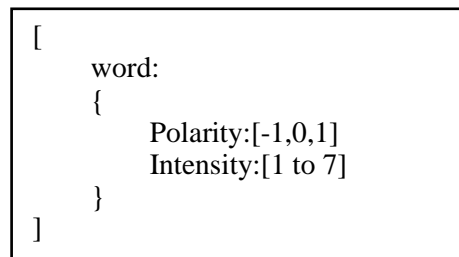


Figure 3. The storage form of lexicon.

It means that given a word, we bind it to its polarity and intensity.

As to unrestricted resource, we employ the sentiment analysis words set in "HowNet" (Li et al., 2002), which contains 91016 words in total. Besides, we collected 161 emoticons from training data as a emoticon lexicon, so as to solve the problem that the current corpus do not contain network language. And the construct the lexicon is as the same of the restricted ones.

#### Training Set and Validation Set

Original training resource and test resource of the task are messages, which fall into several given topics, extracted from the Sina Weibo platform. These messages are microblog posts on real-event topics from real users. We construct Training. By utilizing the feature function

introduced above, three type of final-run training set were constructed for both restricted and unrestricted, shown in Figure 4.

<p>Restricted Run:</p> <p>Run1 = <math>f_1</math></p> <p>Run2 = <math>f_1 + f_2 + f_3 + emVector</math></p> <p>Run3 = <math>f_1 + f_2 + emVector</math></p> <p>Unrestricted Run:</p> <p>Run1 = <math>f_1 + f_2 + f_3 + emVector</math></p> <p>Run2 = <math>f_1 + f_2 + f_3</math></p> <p>Run3 = <math>f_1 + f_2</math></p>
--

Figure 4. The combinations of features.

In Figure 4,  $f_1$  is the feature function comparing the number of emoticon with different sentiment polarity which has been discussed in formulation (9), feature function  $f_2$  takes word segmentation as input feature as show in formulation (10), feature function  $f_3$  constructed by formulation (11) simply compare the number of lexicon conveying different sentiments among the given sentence and  $emVector$  represented for a vector of emoticons which are extracted from the given sentence.

We randomly pick out 40% of microblogs from original training data for evaluating the performance by enumerating the combinations of constructed features. Through validation our model yields an acceptable result.

### 5.3 Experimental Setup

#### Word Segmentation

With a lot of word segmentation toolkits, we choose Jieba segmentation toolkit instead of ICTCLAS<sup>2</sup> on account of granularity and encoding problems.

The segmentation Jieba has tree mode, default mode, full mode and search mode. The differences are showed as Figure 5.

As shown in Figure 5, in “Default Mode” Jieba tries to separate every single Chinese character into a unique phrase on the semantic level, while in “Search Mode” long phrases generated base on “Default Mode” will be given a further segmentation. As showed in the example “不好过”, “Default Mode” results in “不好” and “不好过” when it comes to “Search Mode”. The shortcoming is it may bring extra noises into our system. Different from “Default

<p>Source: “@newcomer2009 最近油价太便宜了不好过啊!”</p> <p>Default Mode: “@/ newcomer2009/最近油价/太/便宜/了/不好过/啊/!”</p> <p>Search Mode: “@/ newcomer2009/最近/油价/太/便宜/了/不好/不好过/啊/!”</p> <p>Full Mode: “newcomer2009/最近/油价/太/便宜/了/不好/不好过/啊”</p>
--

Figure 5. Different modes of word-segmentation produced by Jieba.

Mode” and “Search Mode”, “Full Mode” just extract legitimate phrase among the sentence without considering semantic validity, what’s more every punctuation like “@” and “!” were left out within this mode, which may lead to a failure when modeling a strong emotion . As a result “Default mode” is adopted into our modeling system.

#### Using the Tool of Maximum Entropy Model

Since the sparsity of feature, we try to use Maximum Entropy model to solve this problems. In addition, we adjusted the proportion of the sample to deal with the imbalance. In order to prevent from over fitting and get a good performance, we sample the training set holding out 40% of the data for tuning out the best weight for each class. We initialize the proportion of the positive, negative and neutral sample to be 10:7:1. With the sources including restricted sources and unrestricted sources introduced above and construct the tree kind of feature and then used the maximum entropy model to train. We used the maximum entropy model toolkit of Dr. Zhang Le<sup>3</sup>. Since the training set were small and thus training time was short, using the general iterate scale (GIS) algorithm produced robust results and we iterated it one thousand times.

### 5.4 Result and Discussion

#### Performance on Validation Set

By playing local evaluations on validation set with enumeration among different combinations of the features, top-3 performances can be seen

<sup>2</sup> [ictclas.nlpir.org/](http://ictclas.nlpir.org/)

<sup>3</sup> [homepages.inf.ed.ac.uk/lzhang10/maxent\\_toolkit.html](http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html)

from Table 1 running on restricted resources and Table 2 running on unrestricted resources. Performances in validation set best illustrate the validity of feature function we constructed. Under the combinations of Figure 4, as shown in Table 1, i.e. the restricted running performance, we learn that, a combination of different kinds of feature function is obviously superior to using a single one, and the best performance occurs when adopting the totally three types of feature functions (Run2) accompanying with an emoticon vector. We find the same when it comes to unrestricted round. By contrasting Run2 with Run3 in Table 2, 1% enhancement gains from adding feature function  $f_3$ .

### Evaluation Results

The Topic-Based Chinese Message Polarity Classification task of SIGHAN8 Bake-Off 2015

attracted 13 teams who submitted their testing results. Table 3 and Table 4 show the evaluation results of the task based on restricted and unrestricted resource respectively. The “Best” indicates the highest score of each metric achieved in the task. “Run” is the evaluation score of our system. And the “Average” represents the average score of all participants. As we can see from Table 3 and Table 4, we achieve a result close to the average level, but still have a long way from the best result, especially in F1+ and F1-.

Evaluation performances on the whole constructed feature space are not as good as that in the validation phase. It boils down to the reason that topics in given training resource are totally different from the testing ones. This may causes out-of-vocabulary problems. Focusing on

	Precision	Recall	F1	Precision+	Recall+	F1+	Precision-	Recall-	F1-
Run1	0.7792	0.7792	0.7792	0.3961	0.5223	0.4505	0.4157	0.4840	0.4473
Run2	<b>0.8113</b>	<b>0.8113</b>	<b>0.8113</b>	<b>0.5132</b>	<b>0.4968</b>	<b>0.5049</b>	<b>0.4531</b>	<b>0.5068</b>	<b>0.4784</b>
Run3	0.8047	0.8047	0.8047	0.4750	0.4841	0.4795	0.4408	0.4932	0.4655

Table 1. Score on validation set of restricted resources.

	Precision	Recall	F1	Precision+	Recall+	F1+	Precision-	Recall-	F1-
Run1	<b>0.8108</b>	<b>0.8108</b>	<b>0.8108</b>	<b>0.5099</b>	<b>0.4904</b>	<b>0.5000</b>	<b>0.4527</b>	<b>0.5023</b>	<b>0.4762</b>
Run2	0.7945	0.7945	0.7945	0.4402	0.5159	0.4751	0.4408	0.4932	0.4655
Run3	0.7843	0.7843	0.7843	0.4091	0.5159	0.4563	0.4245	0.4749	0.4483

Table 2. Score on validation set of unrestricted resources.

	Precision	Recall	F1	Precision+	Recall+	F1+	Precision-	Recall-	F1-
Run1	0.6435	0.6435	0.6435	0.1431	0.2734	0.1878	0.3480	0.3366	0.3422
Run2	0.6520	0.6520	0.6520	0.1441	0.2648	0.1866	0.3648	0.3355	0.3496
Run3	<b>0.6640</b>	<b>0.6640</b>	<b>0.6640</b>	<b>0.1631</b>	<b>0.2813</b>	<b>0.2065</b>	<b>0.3607</b>	<b>0.3174</b>	<b>0.3377</b>
Average	0.6866	0.6797	0.6829	0.2037	0.2293	0.1822	0.3970	0.2867	0.3108
Best	0.8357	0.8357	0.8357	0.6258	0.5139	0.5643	0.8232	0.6048	0.5961

Table 3. Evaluation score of restricted resources.

	Precision	Recall	F1	Precision+	Recall+	F1+	Precision-	Recall-	F1-
Run1	0.6577	0.6577	0.6577	0.1367	0.2734	0.1822	0.3489	0.3122	0.3295
Run2	<b>0.6664</b>	<b>0.6664</b>	<b>0.6664</b>	<b>0.1626</b>	<b>0.2899</b>	<b>0.2084</b>	<b>0.3784</b>	<b>0.3237</b>	<b>0.3489</b>
Run3	0.6435	0.6435	0.6435	0.1500	0.2908	0.1979	0.3407	0.3471	0.3439
Average	0.7013	0.6974	0.7007	0.2030	0.2285	0.1900	0.4456	0.3475	0.3660
Best	0.8536	0.8536	0.8536	0.5880	0.6207	0.6039	0.7918	0.6175	0.6938

Table 4. Evaluation score of unrestricted resource.

the F value among the microblogs containing a specific emotion, we got slightly superior to average level, gain from utilizing the oversampling method.

## 5.5 Error Analysis

As is shown in the two figures, we achieved a rather robust result. But on the other hand, it is also obvious that we still have a long way from the state-of-arts, and the potential of the Maximum Entropy model method is far from the state-of-arts, and the potential of the Maximum Entropy model method is far from fully exploited. The major weakness of our system fall down to the low recall rate, which might be the result of not applying enough feature functions. Figure 6 shows some typical error examples of our current system.

Case 1:

topic: 中国政府也门撤侨  
message ID: M00046783  
weibo: #我有话说# 【俄飞机赴也门撤侨遭阿拉伯联军阻挠 被迫折回】 这就是没有人缘的结果  
<http://t.cn/RAoi81S>  
standard polarity: 0  
system polarity: -1

Case 2:

topic: 孙楠退赛  
message ID: M00035167  
weibo: 【张靓颖谈孙楠退赛:做什么选择必有他的道理】张靓颖谈到早前孙楠在《我是歌手 3》总决赛时突然退赛,表示孙楠是简单直接的人,他做什么选择一定有自己的道理。并且举例孙楠帮忙照顾自己和其他歌手的一些琐事,大赞孙楠没那么复杂很干脆。<http://t.cn/RA67PA1>  
standard polarity: 1  
system polarity: 0

Case 3:

topic: 隆平高科超级稻  
message ID: M00048401  
weibo: 超级稻的所谓高产和美的空调一晚 1 度电有的一比。//【安徽万亩"隆平稻种"减产绝收】 <http://t.cn/RA67iNo>  
standard polarity: 0  
system polarity: 1

Figure 6. Error examples.

The first case is of neutral sentiment, but our system categorizes the text as the negative size. In Our system considers "没有人缘" the negative impact on weibo.

In the second case, our system judges the polarity as neutral sentiment because the message does not contain any sentimental words of our corpus. "没那么复杂" conveys a positive emotion by double negation. What more, It is still challenge for us to cope with long distance relation.

The third case, our system judges the polarity as positive sentiment because of the message contains the words "高产" and "美的", which affect the total sentiment of the weibo by the polarity of corpus. And microblog advertising of this kind do not make any contribution to modeling the sentiment polarity, but bring in unknown noise in some way.

## 6 Conclusion and Future work

This paper proposes the Chinese microblog sentiment classification (CMSC) system based on MaxEnt from team of South China Agricultural University (SCAU) that participated in the SIGHAN-8 Topic-Based Chinese Message Polarity Classification task. MaxEnt enables to incorporate a rich set of features, which gives us a rather flexible way to construct appropriate features to cope with sparsity problem caused by the characters limitation of microblog. In addition, oversampling approach is used to handling the unbalance problem.

It is our first attempt on Chinese grammatical error diagnosis, and our system achieves a result close to the average level. There are many possible and promising enhancements in the coming future. More appropriate features can be added to the system for a better modeling. Besides, existing sentiment corpuses and lexicons are filled with "book words" (literary, abstract and technical terms), while microblogs are usually in much less formal forms, with a significant amount of using of colloquial phrases, network language and even emoticons and pictures. Long distance relation and adverting detection are also a challenging research topic.

## Acknowledgments

This work was partially supported by National Natural Science Foundation of China under Grant No. 71472068, Science and Technology Planning Project of Guangdong Province, China under Grant No. 2013B020314013, and the



Innovation Training Project for College Students of Guangdong Province under Grant No. 201410564294.

## References

- Berger A, Pietra S.D, Pietra V.D. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, Vol. 22, No.1, pp. 5-9.
- Chawla, N.V., Japkowicz, N., and Kolcz, A., editors 2004. *SIGKDD Special Issue on Learning from Imbalanced Datasets*.
- Ding, X., Liu, B., and Yu, P.S. 2008. A holistic lexicon-based approach to opinion mining. *In Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08)*, pp. 231-239.
- Edmans, A, Garca, D, Norli Ø. 2007. Sports sentiment and stock returns, *The Journal of Finance*, Vol. 58, pp. 1967-1998
- Go A., Bhayani R., Huang L. 2009. *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford.
- Gruhl, D, Guha, R, Kumar, R, Novak, J, & Tomkins, A. 2005. The predictive power of online chatter. *In: Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 78-87. ACM, New York, NY, USA.
- Hirshleifer, D, Shumway, T. 2003. Good Day Sunshine: Stock Returns and the Weather. *The Journal of Finance*, Vol. 58, pp.1009-1032.
- Li S.J., Zhang J., Huang X., et al. 2002. Semantic computation in Chinese question-answering system. *Journal of Computer Science and Technology*, Vol. 17, No.6, pp. 933-939.
- Liu N., He Y. X., He F. Y., et al. 2013. The Method of Chinese Opinion Sentence Extraction and Polarity Identification Based on Sentimental Elements, *Advanced Materials Research*, Vols. 765-767, pp. 1406-1410
- Liu Y., Huang X., An A., et al. 2007. ARSA: a sentiment-aware model for predicting sales performance using blogs. *In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 607-614. ACM Press, New York
- Liu Z. Y., Chen X.X., Sun M.S. 2012. Mining the interests of Chinese microbloggers via keyword extraction. *Frontiers of Computer Science in China*, Vol. 6, pp. 76-87
- Jiang L., Yu M., Zhou M., et al. 2011. Target dependent Twitter Sentiment Classification. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pp. 151-160.
- Mishne G., Glance N. 2006. Predicting Movie Sales from Blogger Sentiment. *In Proceedings of the 21th National Conference on Artificial Intelligence (AAAI 2006)*. Menlo Park, California, pp.155-158.
- Pang B., Lee L., Vaithyanathan S. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp.79-86.
- Schumaker R. P, Chen, H. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, Vol. 27, No. 12, pp.1- 12.
- Tang D.Y., Wei F.R., Qin B., et al. 2014. A Deep Learning System for Twitter Sentiment Classification. *In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 208-212, Dublin, Ireland, August 23-24.
- Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pp. 417-424.
- Zhang J., Zhu B., Liang L.L., et al. 2014. Recognition and Classification of Emotions in the Chinese Microblog Based on Emotional Factor. *Acta Scientiarum Naturalium Universitatis Pekinensis*, Vol. 50, No.1, PP. 79-84 (In Chinese)