Predicting Reader's Emotion on Chinese Web News Articles

Shuotian Bai[1], Yue Ning[1], Sha Yuan[2], and Tingshao Zhu[1,⋆]

[1] Institute of Psychology,
University of Chinese Academy of Sciences, CAS
Beijing 100101, China
tszhu@psych.ac.cn
[2] Institute of Acoustics, CAS,
Beijing 100190, China
{baishutian10,ningyue09,yuansha10}@mails.gucas.ac.cn

Abstract. Currently, more and more information are spreading on the web. These large amounts of information might influence web users' emotion quite a lot, for example, make people angry. Thus, it is important to analyze web textual content from the aspect of emotion. Although much former researches have been done, most of them focus on the emotion of authors but not readers. In this paper, we propose a novel method to predict readers' emotion based on content analysis. We develop an emotion dictionary with a selected weighting coefficient to build text vectors in Vector Space Model, and train Support Vector Machine and Naive Bayesian model for prediction. The experimental results indicate that our approach performs much better on precision, recall and F-value.

Keywords: Reader Emotion, Emotion Classification, Emotion Dictionary.

1 Introduction

More and more information spread on the web with the rapid development of Internet. The amount of online information increases dramatically, including news, blogs, etc.. On one hand, these large amounts of information can meet people's information need with the help of information retrieval techniques; On the other hand, people would like to express their emotion or meet their emotional needs on the web. For example, a severely depressive fan of rock and roll needs heartwarming stories more than rock music. Traditional search engines focus on meeting informational needs to retrieve the music immediately but not heartwarming ones. Therefore, they can not meet the user's emotional needs sometimes.[8]

To identify people emotional needs, we need to identify their emotion after accessing web content, that is, their emotional preference. Emotion plays an important role in human intelligence which helps people adapt to environment,

⋆ Corresponding author.

Q. Zu, B. Hu, and A. Elçi (Eds.): ICPCA-SWS 2012, LNCS 7719, pp. 16–27, 2013.

arouse individual motivation, organize mental activities and smooth the process of interpersonal communication. Contents on the web may trigger different moods of readers, which will influence their life at the end. It is very useful to detect the users affective state, thus to improve the performance and user interfaces of various web applications. Nowadays, affective information is pervasive on the web, especially online news.[23] Although much research have been done on text emotion classification, they focus on the sentiment of content, instead of readers' emotion triggered by the content. It is a challenging task to predict the reader's emotion on web information [12]. In this paper, we propose to build a reader emotion classification system based on emotion dictionary and machine learning, and evaluate its performance on simplified Chinese News.

The rest of this paper is organized as follows. Section 2 will describe some related work by other researchers. We elaborate our experiment method and system in section 3, and Section 4 shows the experiment results and analyzes of different algorithms. At last, Section 5 provides the conclusion of our work and the future work.

2 Related Work

Text emotion classification is a emerging topic on Data Mining (DM)[4] and Information Retrieval (IR)[3]. Much research has been done on this direction.

Yue Ning et al. [20] introduced a Chinese text emotion classifier with five sentiment categories. They added an emotion lexicon in feature extraction process which would increase the weight of emotion tokens and decrease the weight of non-emotion tokens. However, a part of tokens will be ignored in there system in the process of feature extraction which makes the predictor function badly on short text dataset.

Kevin et al. [21] built a classification system on reader emotion[8] on news articles of Yahoo. They classified news articles into different emotion categories using various combinations of feature sets. There feature sets contains the Chinese character bigrams and metadata. But reader emotion of a news article didn't have a strong correlation with metadata such as publishing time or event location. Therefore, the accuracy of the implemented eight-category SVM classifier was even lower than 60% for some categories.

Hu et al. [15] implemented a Naive Bayesian text classification model. Their results show that the Naive Bayesian classification can achieve a good performance on pure text classification. However, this classifier just categorizes texts into two classes, positive and negative. Emotion of a reader contains a lot such as happy, angry, sad, moving. A more detailed prediction of emotion is needed.

Plaban et al. [22] presented a method for classifying news sentences into multiple emotion categories. The corpus contained 1000 news sentences and the emotion labels are considered as anger, disgust, fear, happiness, sadness and surprise. They compared the performance between machine classification and human classification of emotion. In both the cases, it is observed that some ambiguous emotion (emotion which combining anger and disgust) is hard to predict

in human classification. They used words present in the sentences and the polarity of the subject, object and verb as features. In this experiment, the best average precision was computed to be 79.5% and the average class wise micro F1 is found to be 59.52% when anger and disgust classes are combined and surprise class is removed.

Some other research on Chinese text classification[16][2] use the similar method. Most of their works are based on traditional text classification. Some researchers conducted experiment on the massive web blog information[9][11] as corpus. They analyze both author or reader emotion[6][7], but with low precision.

All the above workings indicate that building an emotion classification system is quite important for sentiment analysis. The emotion classification is a fundamental work which can be used in other system. But the accuracy of them can not satisfy further application.

3 Methods

In this paper, we propose to predict the reader's emotion on Chinese news articles using Support Vector Machine(SVM)[14] and Naive Bayesian(NB)[13] based on emotion dictionary. The system flow chart is shown in Fig. 1.

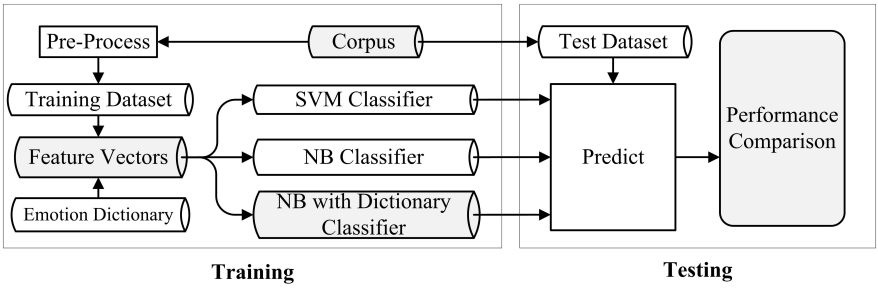

Fig. 1. System flow chart

We download the online society news articles as the corpus and divide it into training data set and testing data set. Before training the classifier, we run preprocessing on the training data set, including HTML tags removing, Chinese word segmentation and stop word removing. Using the emotion dictionary, we can construct the text vector for each news article. After training with different algorithms, we get different classifiers. In the testing part, we use the classifiers on the prediction of reader emotion on testing data set. Finally, we compare the performance of the different classifiers.

3.1 Dataset

Our corpus comes from Sina society news articles[10]. Sina society news web page supports an emotion voting function. After reading each news article, the reader

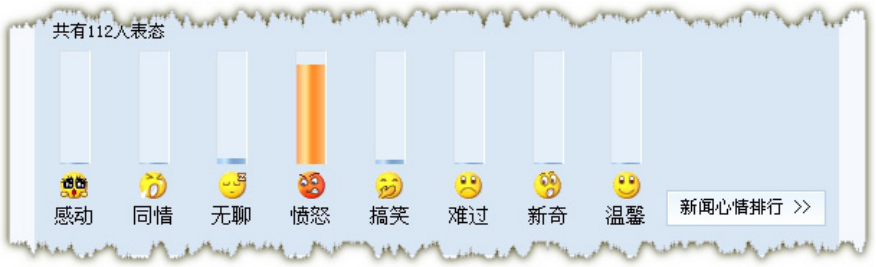

Fig. 2. Eight Moods of Sina Society News

can vote an emotion label: *Moving, Pity, Sad, Boring, Funny, Heartwarming, Surprised or Angry* that best describes his/her feeling shown in Fig.2. Therefore, we implement a crawler to download the corpus from Sina society news.

However, this corpus contains some noisy samples which would be useless for model training. In order to get a high-quality dataset, we need to make a sampling on the corpus. Our sampling strategies are shown below:

- Some ambiguous news which makes no sense or makes reader get multi-emotions should be neglected.
- Some out-of-fashion news which has vote count lower than a threshold (500) should be neglected.
- News articles with too many words or too few words should be neglected.
- Some categories should be neglected if the article number of this category is small.

With the four strategies, we can get a high-quality dataset with less noises. We download 14000 social news articles from Sina.com networking services between April, 2007 to November, 2009. Then we manully filter out some samples according to strategies 1 to 3. Finally, the article number distribution table of each category is listed in Tab. 1.

In Tab. 1, Boring, Sad, Pity and Heartwarming news articles make up a tiny proportion of the corpus. Based on strategy 4, we do not consider them and omit these 4 weak categories. Therefore, our further emotion predictions are all based on the remaining four categories, *angry, moving, funny and surprised.*

Until now, we get a better dataset, named imbalanced dataset with different sample number of 4 categories. In order to get a better performance of classification, a balanced dataset is necessary. Using strategies 1 to 3 again, we select 200 news articles from each category and build a balanced dataset.

3.2 Pre-process

The simplified Chinese documents need some pre-processes before training. Since the corpus comes from the webpage crawler, it is necessary to remove HTML tags first. Then we make a sampling according to strategies listed in the previous section.

Table 1. Emotion Distribution Samples

Category	Proportion	Number
Angry	75.26%	6255
Moving	13.58%	1129
Funny	6.94%	577
Surprised	3.36%	279
Boring	0.47%	39
Sad	0.24%	20
Pity	0.12%	10
Heartwarming	0.02%	2

Chinese articles need to be segment into tokens which is different from English articles. ICTCLAS[5], developed by Institute of Computing Technology(ICT) is one of the most popular tools of Chinese word segmentation. It performs quickly, and researcher can add customized dictionary based on the training purpose. The tool give each article an output of bag of tokens. These tokens are then filtered with a strop word list. This step removes the stop words in the simplified Chinese which appear frequently but have no actual meaning. After the above processes, each news article is represented as a bag of meaningful tokens.

3.3 Emotion Dictionary

The core idea of our method is to increase the weight of emotion tokens and decrease the weight of the other tokens. The emotion dictionary we use comes from four parts: the extended lexicon of TongYiCi CiLin[17], the original TongYiCi CiLin[19], HowNet[18] and corpus key tokens. CiLin was firstly published in 1996 by Shanghai CiShu, then extended by HIT-IRLab during 2006 to 2009. Emotion words in HowNet come from actual corpus corresponding with artificial screening. We also calculate χ^2 statistic for each token in corpus. The formula of χ^2 is shown as following:

$$\chi^2 = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

where A is the frequency that a term and a category occur together, B is the times that a term occurs while a category does not occur, C is the occurrences of a category without a term, D is the count that a term or a category neither occurs and N is the number of the total text set. χ^2 test is based on hypothesis test and measures the positive interdependency and negative interdependency between a word and a categorization. That is how irrelevant between a word and a categorization. We select the top 100 χ^2 statistic tokens.

In sum, we get an emotion dictionary with more than one thousand four hundred emotion words shown in Fig. 3.

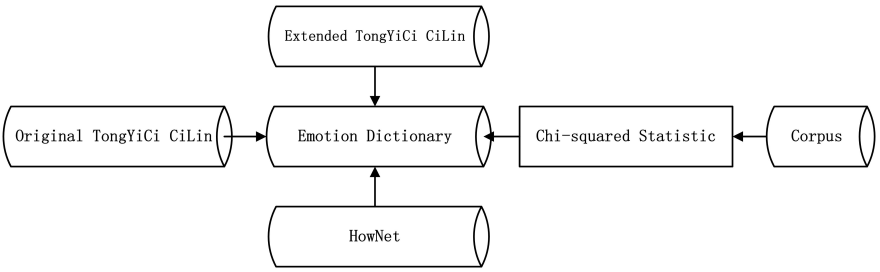

Fig. 3. System flow chart

Then we begin to build the text vector of each article in vector space model where articles are represented as

$$document = (token_1, weight_1; token_2, weight_2 ... token_n, weight_n); \quad (1)$$

An intuitive idea is to increase the weight of emotion words that appears in training dataset and decrease the weight of all the other non-emotion words. Therefore, we take the Bayesian conditional probability into consideration. For each token, the Bayesian conditional probability is

$$weight_{t,c} = \frac{T_{c,t} + 1}{\sum_t T_{c,t} + B} \quad (2)$$

where $T_{c,t}$ is the term frequency(the appearing time of the token) of token t in category c, B is the total number of tokens in all the categories. The "1" term is a smooth factor to avoid zero weight.

Therefore, we revised the formula of Naive Bayesian conditional probability with a weighting coefficient shown in the following:

$$weight_{t,c} = \begin{cases} \frac{(K_t+1)\times T_{c,t}+1}{\sum_t T_{c,t}+B+K_t\times N_c} & \text{if t is an emotion token;} \\ \frac{T_{c,t}+1}{\sum_t T_{c,t}+B} & \text{otherwise.} \end{cases} \quad (3)$$

where N_c is the total term frequency of emotion tokens in category c, K_t is a non-negative emotion word weighting coefficient. If token t is non-emotion word, then set $K_t = 0$. Otherwise, we set K_t a non-negative number. The classification will be back to traditional Naive Bayesian classification if K_t is always set zero shown in Fig. 4.

If K_t is positive infinity, the conditional probability is

$$weight_{t,c} = \begin{cases} \frac{T_{c,t}}{N_c} & \text{if t is an emotion token;} \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

This case with a positive infinity K_t will only take emotion tokens into account and omit all the other non-emotion tokens. Therefore, we need to build a balance between emotion tokens and non-emotion tokens and find the optimal value of K_t.

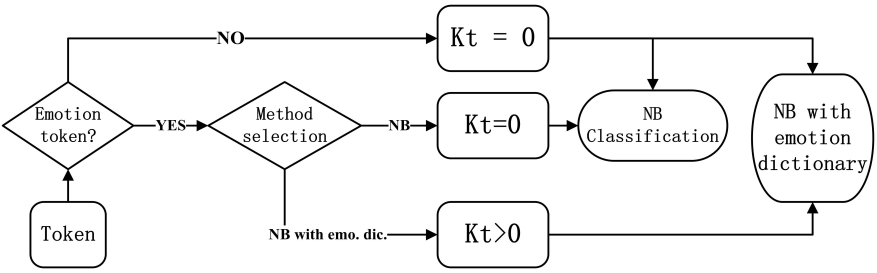

Fig. 4. Weight coefficient selection

Algorithm 1. TrainNBwithEmotionDictionary

Require: Class, C; Documents, D; Token set, V; Emotion dictionary, E;

$N_c \leftarrow CountEmotionTokenAppearing(D, E)$;
$B \leftarrow |V|$;
for $c \in C$ **do**
  $prior[c] \leftarrow \frac{N_c}{N}$;
  $text_c \leftarrow ConcateTextOfAllDocsInClass(D, C)$;
  for $t \in V$ **do**
    $T_{ct} \leftarrow CountTokensOfTerm(text_c, t)$;
  end for
  for $t \in V \&\& t \in E$ **do**
    $condprob[t][c] \leftarrow \frac{(K_t+1)\times T_{c,t}+1}{\sum_t T_{c,t}+B+K_t\times N_c}$;
  end for
  for $t \in V \&\& t \notin E$ **do**
    $condprob[t][c] \leftarrow \frac{T_{c,t}+1}{\sum_t T_{c,t}+B}$;
  end for
end for
return $V, prior, condprob$;

4 Experiments

In order to test the importance of emotion dictionary, we design two comparable experiments. Experiment 1 tests the classification without emotion dictionary. This experiment uses the classic algorithms(*SVM, NB*) in machine learning. The performance can be the baseline of our work. Experiment 2 works on the classification with emotion dictionary. This experiment uses our method when constructing text vectors. We also list the precision, recall and F-value of the two experiments.

4.1 Experiment 1

Experiment 1 focus the classification without emotion dictionary. While constructing the text vectors, we use formula (2) where K_t is always set zero for

any token. We firstly test the performance with Support Vector Machine. An SVM model is a representation of the examples as points in space. The core idea is to find a biggest clear gap that can divide examples of the separate categories. Test samples are then predicted to belong to a category based on which part they fall in. Since Support Vector Machine(SVM) is insensitive to data distribution, we choose it as training algorithm on imbalanced dataset. To simplify calculation, we use the method of chi-squared feature extraction. We select the top 500 keywords as features to do the training. We chose the C-SVC as the SVM type with $C = 40$. Using 10 fold cross-validation, the testing results are shown in Tab. 2.

Table 2. Results of SVM on imbalanced dataset

Category	Recall	Precision	F-value
Angry	0.912	0.963	0.937
Moving	0.873	0.690	0.771
Funny	0.575	0.460	0.511
Surprised	0.818	0.720	0.766

We also test the performance with Naive Bayesian Classification. Since Naive Bayesian can get a good performance in balanced dataset, we use the balanced dataset. The main idea of Naive Bayesian Classification is the conditional probability which means the probability of a token appears in a category. We use both Support Vector Machine and Naive Bayesian with formula (2) methods on the balanced dataset and take a five-fold cross validation to get the average results in the following Tab. 3.

Table 3. Results of SVM and NB on balanced dataset

	SVM			NB		
Category	Recall	Precision	F-value	Recall	Precision	F-value
Angry	0.806	0.83	0.818	0.933	0.63	0.752
Moving	0.904	0.88	0.891	1	0.279	0.436
Funny	0.790	0.76	0.745	0.740	0.9	0.812
Surprised	0.832	0.83	0.831	0.535	0.93	0.679

From the results, the precision of category "Moving" is lower(0.279) compared to other categories(0.535 at least). The reason may be the contents in this class. Since the news articles in the moving class involve many different kinds of aspects of topics, it is an ambiguous emotion for an automatic classifier. This two systems can be the baseline of our work.

4.2 Experiment 2

In experiment 2, we add the emotion dictionary and compare the classification results with experiment 1.

To minimize the bias from other factors, experiment 2 tests on the balanced dataset. As explained in formula 3, we take emotion dictionary into consideration. The first question is how to set the weighting degree K_t of the dictionary. As discussed above, the system turns back to traditional text classification if K equals to zero (formula 2). On the other side, the system will be over-fit if K_t is equal to infinity(formula 4). Therefore, we give K_t several values and run a series of experiments on the examination of K_t. For each given value of K, we can calculate the average precision, average recall and average F-value.

We show three curves of weighting coefficient on precision, recall and F-value in Fig. 5. In Fig. 5, the horizontal axis is K_t and we find a local optimal value at K_t equaling to 1.2.

Table 4. Results of NB with Emotion Dictionary

Category	Recall	Precision	F-value
Angry	0.902	0.87	0.886
Moving	0.980	0.88	0.972
Funny	0.831	0.87	0.849
Surprised	0.877	0.94	0.907

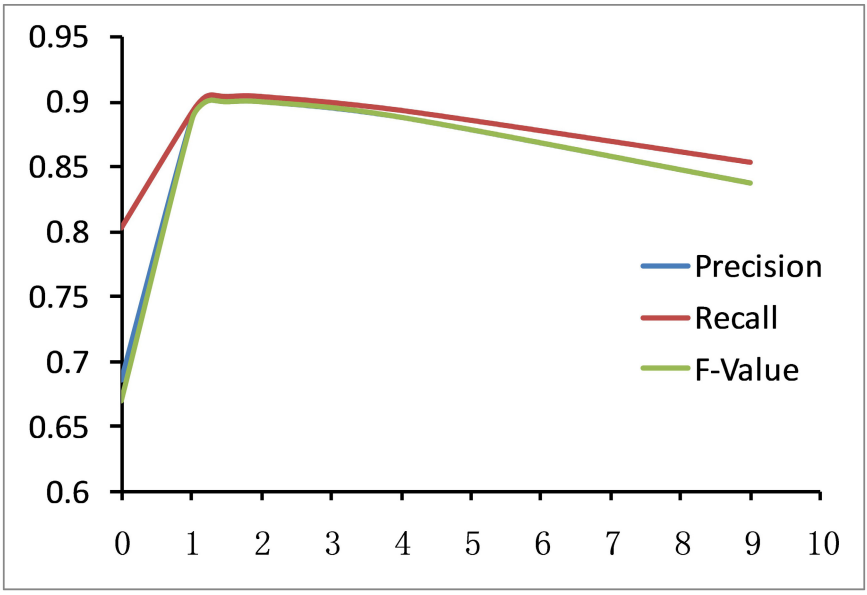

Fig. 5. Weighting coefficient experiment

From this result, we set $K_t = 1.2$ in the following experiments. The results with $K_t = 1.2$ are shown in Tab. 4.

In Tab. 4, the average precision reaches 90% which is much beyond the precision of the baseline. It means that our method performance well in simplified Chinese.

5 Conclusion

In this paper, we propose to build a classifier which is able to predict the reader emotion with emotion dictionary on news articles in simplified Chinese. We tune the Naive Bayesian conditional probability formula and add a new coefficient which stands for the weight of the emotion dictionary. From the experiment results, it is obvious that the system performances better with emotion dictionary. That high accuracy means that the classifier can seize the emotion tendency correctly for each news article and classify news article according to the emotion tendency precisely.

After investigating the experiment results, we find that most wrongly-classified samples are more likely to be judged as angry, which means the emotions of news articles have a large component of anger. This may lead to some mental illnesses or social instabilities. This situation results from the news itself. The purpose of reporters of writing news articles is to attract the reader's attention. People may pay less attention to boring news, instead they may prefer to news with strong emotions, angry or surprised.

Our work still has some limitations. First, the sample set contains 800 news articles in all. If we can download much more documents, the precision will get a rise. Second, since the articles are all long text, the classifier cannot work well on short text dataset. We will continue to work on the emotion classification on micro-blog. Third, our classifier can only predict reader emotion into four categories. For the other emotions, we cannot make the prediction because of the lack of the training dataset.

In the future, we may improve our research in two possible ways. For the classifier itself, we will refine training method to increase the accuracy and make some update of emotion dictionary. Web users create many new web words each day, it is necessary to keep the developing step. On the other hand, we would like to use our classifier into other systems such as web personality analysis system.

Acknowledgments. The authors gratefully acknowledges the generous support from NSFC (61070115), Institute of Psychology(113000C037), Strategic Priority Research Program (XDA06030800) and 100-Talent Project(Y2CX093006) from Chinese Academy of Sciences.

References

1. Bhowmick, P.K., Basu, A., Mitra, P.: Classifying emotion in news sentences: When machine classification meets human classification. International Journal on Computer Science and Engineering, 98–108 (2010)
2. Bracewell, D.B., Minato, J., Ren, F., Kuroiwa, S.: Determining the Emotion of News Articles. In: Huang, D.-S., Li, K., Irwin, G.W. (eds.) ICIC 2006. LNCS (LNAI), vol. 4114, pp. 918–923. Springer, Heidelberg (2006)
3. Christopher, H.S., Manning, D., Raghavan, P.: Introduction to Information Retrieval. Cambridge University Press (2008)
4. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2nd edn. China Machine Press (2008)
5. Zhang, H.: Ictclas chinese segmentation tool (2010)
6. Zhou, L., He, Y., Wang, J.: Survey on research of sentiment analysis. Computer Applications 4, 2725–2728 (2008)
7. Hsin, K., Lin, Y., Chen, H.H.: Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, vol. 9, pp. 136–144 (2008)
8. Lin, K.H.-Y., Yang, C., Chen, H.-H.: What emotions do news articles trigger in their readers? In: SIGIR 2007 Proceedings, vol. 2, pp. 733–734 (2007)
9. Quan, C., Ren, F.: Construction of a blog emotion corpus for chinese emotional expression analysis. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 8, pp. 1446–1454 (2009)
10. Sina.com. Sina society moodrank (2010), `http://news.sina.com.cn/society/`
11. Tokuhisa, R., Inui, K., Matsumoto, Y.: Emotion classification using massive examples extracted from the web. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 881–888 (2008)
12. Weare, C., Lin, W.Y.: Content analyis of the world wide web: opportunities and challenges
13. Wikipedia. Naive bayes classifier, `http://en.wikipedia.org/wiki/Naive_Bayes_classifier`
14. Wikipedia. Support vector machine, `http://en.wikipedia.org/wiki/Support_vector_machine`
15. Hu, Y., Zhou, X., Ling, L., Wang, X.: A bayes text classification method based on vector space model. Computer and Digital Engineering 32, 28–30 (2004)
16. Zhang, Y., Li, Z., Ren, F., Kuroiwa, S.: A preliminary research of chinese emotion classification model. IJCSNS International Journal of Computer Science and Network Security, 127–132 (2008)
17. HIT-IRLab. Extended tongyici cilin, `http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm`
18. HowNet, `http://www.keenage.com/`
19. Mei, J., Zhu, Y., Gao, Y., Yin, H.: TongYiCi CiLin. Shanghai CiShu Press (1996)
20. Ning, Y., Zhu, T., Wang, Y.: Affective word based chinese text sentiment classification. In: Proceedings of 5th International Conference on Pervasive Computing and Applications, ICPCA 2010 (2010)

21. Lin, K.H.-Y., Yang, C., Chen, H.-H.: What emotions do news articles trigger in their readers? In: SIGIR 2007 Proceedings, vol. 2, pp. 733–734 (2007)
22. Bhowmick, P.K., Basu, A., Mitra, P.: Classifying Emotion in News Sentences: When Machine Classification Meets Human. International Journal on Computer Science and Engineering 2(1), 98–108 (2010)
23. Ning, Y., Li, A., Zhu, T.: Are Online Mood Labels A True Reflection of Our Experiences? In: Proceedings of 2011 3rd Symposium on Web Society (SWS 2011), pp. 21–26 (2011)