Sentiment Classification of Chinese Reviews in Different Domain: A Comparative Study

Qingqing Zhou and Chengzhi Zhang(✉)

Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China
breeze7zhou@gmail.com, zhangcz@njust.edu.cn

Abstract. With the rapid development of micro-blog, blog and other types of social media, users' reviews on the social media increase dramatically. Users' reviews mining plays an important role in the application of product information or public opinion monitoring. Sentiment classification of users' reviews is one of key issues in the review mining. Comparative study on sentiment classification results of reviews in different domains and the adaptability of sentiment classification methods is an interesting research topic. This paper classifies users' reviews in three different domains based on Support Vector Machine with six kinds of feature weighting methods. Experiment results in three domains indicate that different domains have their own characteristics and the selection of feature weighting methods should consider the domain characteristics.

Keywords: Sentiment classification · Chinese review · Support vector machine · Feature weighting

1 Introduction

Internet is becoming more central to our lives. The coexistence of openness, virtual and sharing of Internet makes it become a new platform for people expressing their emotions or views, such as product reviews, service reviews, blog reviews and etc [1].

As the number of various users' reviews on social media websites increases dramatically, how to organize the huge amount of data from reviews effectively has become a difficult problem for us. Sentiment analysis is to determine the users' attitudes for a particular topic or something, where attitudes can be their judgment, assessment or their (speech, writing) emotional state [2]. It is different from traditional text information processing as it focuses on the emotion expressed in the text. One of key issues of text sentiment analysis is text sentiment classification. Text sentiment classification is to judge the emotional tendencies or classify types of text [3].

In this paper, we mainly study sentiment classification in three different domains based on Support Vector Machine (SVM). Experiment results show that different domains have their own characteristics which affects the selection of feature weighting methods.

The rest of this paper is organized as follows. The second section describes the related work. In section 3, key technologies of SVM-based sentiment classification are described. Section 4 presents experiment results and section 5 concludes with a discussion of future work.

X. Wang et al. (Eds.): ICMLC 2014, CCIS 481, pp. 13–23, 2014.
DOI: 10.1007/978-3-662-45652-1_2

2 Related Works

Text sentiment classification is widely used, including: social public opinion analysis [4], product quality and service evaluation from consumers, harmful information filtering, books reviews, film and television reviews, blog reviews and so on.

In 2004, AAAI successfully held the conference with the theme "explore ideas and emotions in text "which accelerates the development of sentiment classification[1]. Since then, research stage of sentiment classification is keeping on growing.

Sentiment classification consists of three steps: subjectivity classification, polarity classification and emotional intensity recognition [2]. Subjectivity Classification methods are based on supervised learning mostly. Wiebe used classifier based on Naïve Bayes with Boolean weighting in the task of subjectivity classification [5]. Complex syntactic structures can be used in the subjective classification [6][7]. Most of the current sentiment classification research focuses the polarity classification. Representative methods include: one is Turney's unsupervised method [8], another is Pang's supervised learning method [9]. Emotional intensity recognition can be implemented by the supervised learning methods. The main methods of this task can be divide into the following three categories: (1) the multi-classification method: Lin divided emotional intensity in the sentence into five levels and used LSPM to distinguish intensity [10], (2) the regression method: Pang & Li used SVM regression method to identity emotional intensity [11] and (3) the sequence annotation method: Liu put forward the sentence sentiment degree analysis model based on cascaded CRFs model [12].

Currently, performance of sentiment classification results in different domains is various. However, there is no systematic and comprehensive study about this topic. This paper studies how to use machine learning to do sentiment classification automatically and compare the performance of different domains with different feature weighting method.

3 Framework and Key Technologies of SVM-Based Sentiment Classification

3.1 Feature Selection

Reviews in Chinese must be segmented before the feature selection. This paper applies improved maximum matching word segmentation algorithm (MMSEG) [13] to segment reviews in Chinese. MMSEG is a dictionary-based word segmentation algorithm.

Not all words are useful for sentiment classification, so feature selection is necessary. The representative words can be extracted from text according to feature weighting methods [14]. There are many classical feature selection methods, such as TF, DF, IG, MI, CHI etc [15]. Hwee's experiment results show that the CHI is the best feature selection method according to the performance of classification [16]. Chen compared existing feature extraction methods and proved that no feature selection method is applicable to all or most of the experimental corpora [17]. Above all, we choose the CHI method for feature selection in this paper.

[1] http://www.aaai.org/Press/Proceedings/aaai04.php

3.2 Feature Weighting Methods

The performance of feature weighting methods affects the classification accuracy directly. Classical feature weighting methods includes Boolean weights, TF, IDF, TF* IDF and so on.

Deng replaced IDF with CHI, and experiment result shows that TF-CHI performed better than TF-IDF in the task of SVM-based text classification [18]. TF-CHI can be computed via formula 1:

$$w_i = tf(t_i) * CHI(t_i, C_j) \tag{1}$$

Lan proposed a weighting method, namely TF*RF (where RF means Relevance Frequency) and experiment result proves that it has better performance than TF*IG and other methods [19]. RF can be calculated through formula 2:

$$\mathrm{rf} = \log\left(2 + \frac{\mathrm{a}}{\mathrm{c}}\right) \tag{2}$$

Where, *a* denote co-occurrence number of the feature and class, *c* means the number of feature appears but class does not appear.

The feature weighting methods include Boolean weight, TF, log (TF), TF * IDF, TF * CHI and TF * RF is used in this paper. These methods are compared according to the performance of sentiment classification.

3.3 Parameters Selection and Optimization of SVM Model

LIBSVM[2] is used to classify the reviews in this paper. Two most important parameters in LIBSVM are C and γ. The C is the sample misclassification penalty factor. The larger the C is, the smaller the error tolerates [14].Whether the C is to too large or too small will affect the generalization ability of the model. The γ is the parameter come from the RBF kernel function. The larger the γ is, the more support vectors. The number of support vectors affects the speed of the model training and prediction directly.

3.4 Classification Results Determination

Classification results in this paper consist of two parts: predicted category and membership degree. The larger the membership degree is, the greater confidence that the sample belongs to the class [14]. Membership degree's is computed by the following formula:

$$\mathrm{M} = \frac{\sum S_i}{2*K} + \frac{K_s}{2*K} \tag{3}$$

Where, Si denotes the score of support discrimination classes, Ks means the number of support discrimination classes, K means the number of all categories. Membership degree is used to improve accuracy rate as the credibility of using category labels alone as the classification results is low. The membership degree algorithm in the paper is the one-against-one algorithm [20].

[2] http://baike.baidu.com/view/598089.htm

4 Experiments and Result Analysis

4.1 Experimental Data

We use blog reviews of ScienceNet[3], hotel reviews of Ctrip[4] and book reviews of Dangdang[5] as experiment data, each type of data contains training set and test set as shown in table 1. Table 2 shows experiment samples.

Table 1. Experiment data set

Domain	Training set	Positive	Negative	Test set
Blog Reviews	950	600	350	2,800
Hotel Reviews	1,000	500	500	3,633
Book Reviews	1,000	500	500	2,870

Table 2. Experiment data sample

Category		Samples
Blog Reviews	**Positive reviews**	写得好！顶一个！
	negative reviews	你是帮倒忙的吧，以后写文章不要这样。。。
Hotel Reviews	**Positive reviews**	环境很好，地点很方便，服务也很好，下回还会住的！
	negative reviews	缺点：太多了;1,设施太陈旧了，地毯到处都是黑污渍，房间有股怪味，虽然3星级不能要求太多，但也不能比经济型酒店还差吧…
Book Reviews	**Positive reviews**	这本书不错，读过以后，才知道该如何面对与解决自己或周边关系存在的问题。这是学校里没有的。当然，这是要靠自己去领悟书中思想…
	negative reviews	这是一本捡别人漏沟水的书，非常后悔买了这么一本毫无看头的书，请大家别再上当了！！！！

[3] http://blog.sciencenet.cn/
[4] http://hotels.ctrip.com/
[5] http://www.dangdang.com/

4.2 SVM Model Training

We use MMSEG algorithm to segment the reviews in Chinese, CHI method for feature selection, six different kinds of feature weighting methods, namely: Boolean weight, TF, log (TF) TF-IDF, TF- RF, TF-CHI, LIBSVM for model training. We train models in turn and get 18 different models.

4.3 Experiment Results Evaluation Method

The evaluation indicators include: Precision, Recall and F1 value [21]. The contingency table for results evaluation of classification is shown in table 3. contingency table for results evaluation of classification

Table 3. Contingency table for results evaluation of classification

		Prediction Classification	
		P(Positive)	N(Negative)
Real classification	P	TP(Ture ositive)	FN(False Negative)
	N	FP(False Positive)	TN(Ture Negative)

These evaluation indicators are calculated via the following formula respectively:

(a) Recall $$\mathrm{R} = \frac{TP}{TP+FP} \quad (4)$$

(b) Precision $$\mathrm{P} = \frac{TP}{TP+FN} \quad (5)$$

(c) F_1 value $$F_1 = \frac{2*P*R}{P+R} \quad (6)$$

For all classes:

(a) Macro Recall $$\mathrm{MacroR} = \frac{1}{n}\sum_{j=1}^{n} R_j \quad (7)$$

(b) Macro Precision $$\mathrm{MacroP} = \frac{1}{n}\sum_{j=1}^{n} P_j \quad (8)$$

(c) MacroF_1 value

$$\mathrm{Macro}F_1 = \frac{2*MacroP*MacroR}{MacroP+MacroR} \quad (9)$$

4.4 Experiment Results Analysis

In order to analyze the experiment results, we annotated the polarity of testing corpus manually. The testing set of blog reviews contains 2,100 positive reviews and 700 negative reviews. The test set of hotel reviews contains 1,702 positive reviews and 1,921 negative reviews. The test set of book reviews contains 1,420 positive reviews and 1,450 negative reviews.

4.5 Results Analysis of the Same Corpus with Different Feature Weighting

(a) Blog reviews of ScienceNet

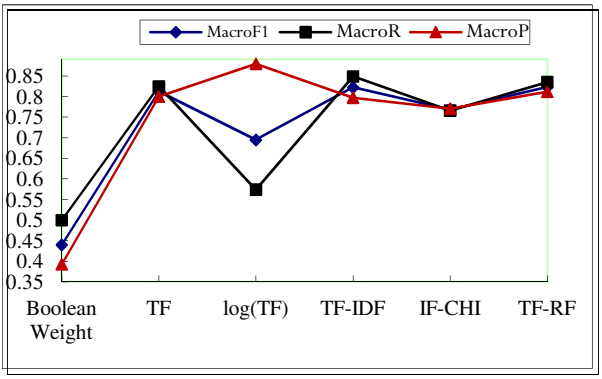

Fig. 1. Blog reviews of ScienceNet

From figure 1 we can find that for blog reviews of ScienceNet, TF-RF has the best classification performance, followed by TF-IDF, Boolean weighting method does worst.

(b) Hotel reviews of Ctrip

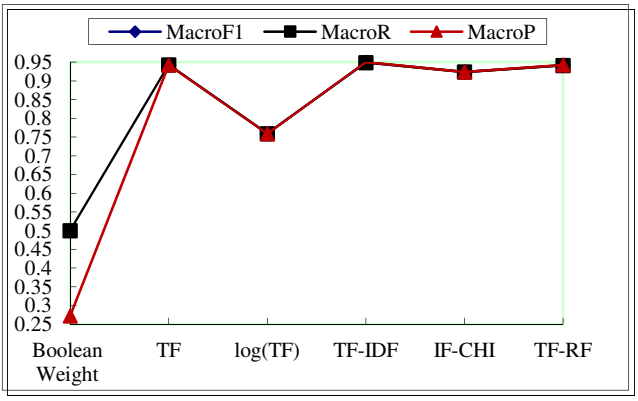

Fig. 2. Hotel reviews of ctrip

Figure 2 shows that for hotel reviews of Ctrip, the best classification is done by TF-IDF, followed by TF, Boolean weighting method performs worst.

(c) Book reviews of Dangdang

From Figure 3 we can find that that for book reviews of Dangdang, TF-IDF has best classification performance, followed by TF, Boolean weighting method classifies worst.

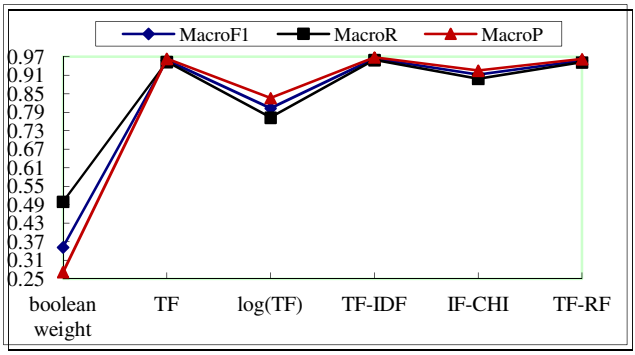

Fig. 3. Book reviews of Dangdang

4.6 Results Analysis of Different Corpus with the Same Feature Weighting

Figure 4 shows that IF-IDF has optimal classification performance in the three different reviews corpus, followed by TF-RF and TF, TF-CHI and Log (TF) perform in general, the classification performance of Boolean weighting method is the worst.

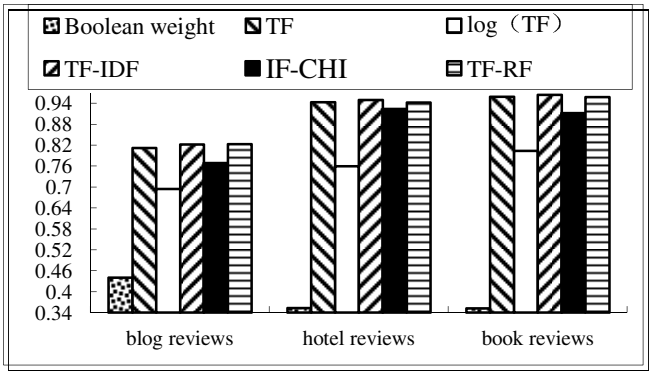

Fig. 4. Three reviews corpus evaluation

We speculated the results have the following reasons:

(a) The average length of reviews

We calculated the average length of three kinds of reviews, and got that the average length of blog reviews is 14.61 words, hotel reviews' average length is 85.87 words and book reviews' average length is 104.03 words. We speculate that TF, TF-IDF, TF-RF, Log(TF) and Boolean weight are linearly related to the average length of reviews, the first four is positive correlation, the last one is negative correlation. TF-CHI is curve related to the average length of reviews, and it forms a convex function.

(b) The ratio of positive and negative reviews

We calculated the ratio of positive and negative reviews of the three domains, we got that the ratio of blog reviews is 1:0.33, hotel reviews' ratio is 1:1.13, book reviews' ratio is 1:1.02. We speculate that TF, TF-IDF, TF-RF, Log(TF) and Boolean weight are curve

related to the ratio of reviews, the first four form convex functions, the last one forms a Concave function. TF-CHI is linearly related to the ratio of reviews and it is positive correlation.

(c) Other reasons

In addition to reasons above, sentiment classification is associated with some emotional factors such as its own characteristics of each domain, the review language characteristics and viewpoint holders. It is difficult to do quantitative comparison of these reasons' effect.

4.7 Results Analysis of the Same Corpus with Different Threshold

As IF-IDF, TF-RF and TF perform better than several other feature weighting methods, we analyze the performance of the three methods in different thresholds only.

(a) Blog reviews of ScienceNet

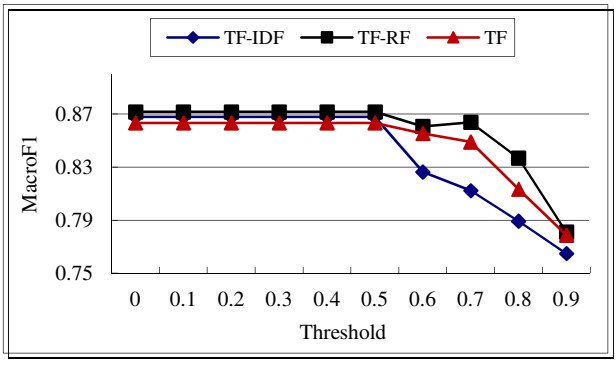

Fig. 5. Blog reviews of ScienceNet

From figure 5 we can find that TF-IDF, TF-RF and TF perform best when threshold less than or equal to 0.5, where TF-RF has optimal performance, followed by TF-IDF.

(b) Hotel reviews of Ctrip

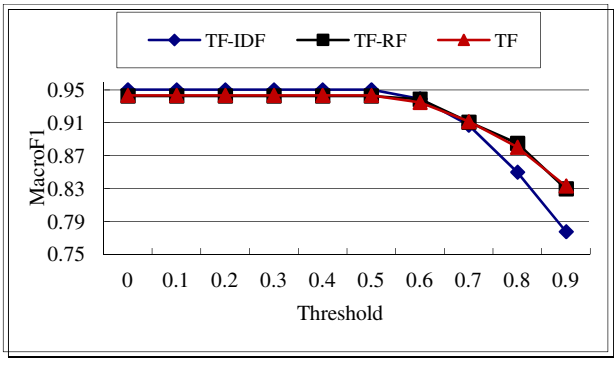

Fig. 6. Hotel reviews of Ctrip

Figure 6 shows that TF-IDF, TF-RF and TF perform best when threshold less than or equal to 0.5, in which TF-IDF has best performance, followed by TF.

(c) Book reviews of Dangdang

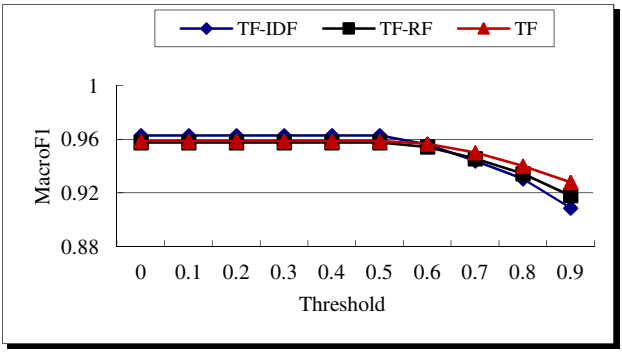

Fig. 7. Book reviews of Dangdang

From figure 7 we can see that TF-IDF, TF-RF and TF classify best when threshold less than or equal to 0.5, where TF-IDF has optimal performance, followed by TF.

Based on the above threshold performance, we can draw the conclusion that TF-IDF, TF-RF and TF perform best when threshold less than or equal to 0.5.

5 Conclusion and Future Works

5.1 Conclusion

Sentiment classification based on SVM is performed and classification results are compared in different domains. By analyzing the experiment results, adopting feature weighting method TF-RF attained the optimal performance for blog reviews of ScienceNet while TF-IDF works best for hotel reviews of Ctrip and book reviews of Dangdang. Overall, we found that feature weighting method TF-IDF performed the best, followed by TF, TF-RF. For different thresholds, we concluded that TF-IDF, TF-RF and TF classified best when thresholds are less than or equal to 0.5.

We concluded that different domains have their own characteristics, the domain characteristics are needed to be taken into account for the selection of feature weighting methods.

5.2 Future Works

In the future, our aim is to improve the performance of sentiment classification by the following approaches.

(a) We will expand the amount of experiment data.

(b) We will increase the equality of experiment data in three domains.

(c) We will add more feature selection algorithms, such as IG and MI.

(d) We will replace polarity classification with multi-classification for Emotion intensity recognition.

Acknowledgements. This work was supported in part by the Fundamental Research Funds for the Central Universities (No. 30920130132013), National Natural Science Foundation of China (No.70903032) and Project of the Education Ministry's Humanities and Social Science (No. 13YJA870020).

References

1. Wang, S.: Text sentiment classification research on web-based reviews. Shanghai University, Shanghai, pp. 1–5 (2008) (in Chinese)
2. Wang, H., Liu, X., Yin, P., Liao, Y.: Web text sentiment classification research. Scientific and Technical Information **29(5)**, 931–938 (2010) (in Chinese)
3. Zhang, Y.: Text sentiment classification research. Beijing Jiaotong University, Beijing, pp. 1–10 (2010) (in Chinese)
4. Agrawal, R., Rajagopalan, S., Srikant, R,. Xu, Y.: Mining newsgroups using networks arising from social behavior. In: Proceeding of the 12th WWW Conference, Budapest, Hungary, pp. 529–535 (2003)
5. Wiebe, J.M., Bruce, R.F., O'Hara, T.P.: Development and use of a gold-standard data set for subjectivity classifications. In: Proceedings of the Association for Computational Linguistics, pp. 246–253. College Park, Maryland (1999)
6. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 105–112. ACL, USA (2003)
7. Wiebe, J., Wilson, T., Bruce, R.F., Bell, M., Martin, M.: Learning subjective language. Computational Linguistics **30**(3), 277–308 (2004)
8. Turney, P.D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 417–424 (2002)
9. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 79–86. ACL, USA (2002)
10. Lin, W.H., Wilson, T.,Wiebe, J.,et al.: Which side are you on? Identifying perspectives at the document and sentence levels. In: Proceedings of the 10th Conference on Computational Natural Language Learning, NY, USA, pp. 109–116 (2006)
11. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of Conference on Association for Computational Linguistics,. Michigan, pp. 115–124 (2005)
12. Ni, X., Xue, G.R., Ling, X., et al.: Exploring in the weblog space by detecting informative and affective articles. In: Proceedings of the 16th International Conference on World Wide Web, pp. 281–290. ACM (2007)
13. Tsai, C.H.: MMSEG: a word identification system for mandarin Chinese text based on two variants of the maximum matching algorithm (2000). http://www.geocities.com/hao150/mmseg/
14. Zhang, Z.: Tmsvm Reference Documents. tmsvm.googlecode.com/svn/trunk/Tmsvm Reference Documents (v1.1.0).docx Accessed: (May 1, 2013) (in Chinese)
15. Yang, Y., Pederson, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the Fourteenth International Conference, pp. 412–420 (1997)
16. Ng, H.T., Goh, W.B., Leong, K.: Feature selection, perceptron learning, and a usability case study for text categorization. ACM SIGIR Forum **31**(SI), 67–73 (1997)

17. Chen, T., Xie, Y.: Feature dimension reduction methods for text classification. Scientific and Technical Information **24**(6), 690–695 (2005)
18. Deng, Z.-H., Tang, S.-W., Yang, D.-Q., Zhang, M., Li, L.-Y., Xie, K.-Q.: A comparative study on feature weight in text categorization. In: Proceedings of the 6th Asia-Pacific Web Conference, Hangzhou, China, pp. 588–597 (2004)
19. Lan, M., Tan, C.-L., Low, H.-B.: Proposing a New TermWeighting Scheme for Text Categorization. In: Proceedings of AAAI Conference on Artificial Intelligence, Boston, Massachusetts, pp. 763–768 (2006)
20. Liu, B., Hao, Z., Xiao, Y.: Interactive iteration on one against one classification algorithm. Pattern Recognition and Artificial Intelligence, **21(4)**:425–431 (2008) (in Chinese)
21. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw Hill Book Co. (1983)