

Sentiment Classification for Consumer Word-of-Mouth in Chinese: Comparison between Supervised and Unsupervised Approaches

Ziqing Zhang Qiang Ye Wenying Zheng Yijun Li

School of Management, Harbin Institute of Technology, Harbin 150001, China

Abstract

Sentiment classification aims at mining word-of-mouth, reviews of consumers, for a product or service by automatically classifying reviews as positive or negative. Few empirical studies have been conducted in comparing the different effects between machine learning and semantic orientation approaches on Chinese sentiment classification. This paper adopts three supervised learning approaches and a web-based semantic orientation approach, PMI-IR, to Chinese reviews. The results show that SVM outperforms naive bayes and N-gram model on various sizes of training examples, but does not obviously exceeds the semantic orientation approach when the number of training examples is smaller than 300.

Keywords: Sentiment classification, Supervised approach, Semantic orientation approach, Chinese

1. Introduction

Web is becoming an essential part of everyday life. More and more people now write reviews of products and services and post them online. It has provided a channel through which a consumer can learn how others like or dislike a product before buying and a manufacturer can keep track of consumer opinions on its

products. However, the number of reviews available for any given product is ever increasing. It is hard for people to decide what the majority opinion is on the Web about certain product.

Sentiment classification, also known as polarity classification or opinion mining, attempts to address this problem by automatically classifying a piece of text as expressing positive or negative sentiment. Sentiment classification has recently attracted much attention from the natural language processing community. The literature shows that two types of techniques have been utilized in its applications: machine learning and semantic orientation. The machine learning approach applied to this problem mostly belongs to supervised classification. This type of technique tends to be more accurate because the classifier is trained on a set of representative data called a corpus. In contrast, adopting semantic orientation approach to sentiment classification is unsupervised learning because it does not require prior training in order to mine the data. Instead, it uses the direction of a word's semantic orientation (positive or negative) and the strength of the semantic orientation to determine review sentiment.

Both approaches have pros and cons. Although supervised machine learning is likely to provide more accurate classification result than unsupervised semantic orientation, a machine learning model is

highly dependent upon the quality of training data, and needs retraining if it is to be applied elsewhere. Thus, the selection of sentiment classification techniques tends to be a trade-off between accuracy and generality. Chaovalit and Zhou (2005) compared supervised N-gram model with a semantic orientation approach that relies on web semantic resource (Turney 2002) using English movie reviews and found that the N-gram model performs better. To our best knowledge, it is still an open question as to which approach is better for sentiment classification of Chinese reviews.

To address the above question, we adapt both supervised and unsupervised approaches to Chinese review mining in the same domain. We focus on three machine learning approaches, i.e., support vector machines, naive bayes and statistical language model. Also, we examine the effects of seed words and search engines on the performance of a web-based semantic orientation approach. The findings will help us gain insight into the strengths and limitations of machine learning and semantic orientation techniques for Chinese opinion mining.

2. Related work

With the increasing need of information organization and knowledge discovery from text data, many studies have employed a supervised training approach in review sentiment classification. Among these methods, support vector machines, naive bayes, and statistical language model are always in the comparison list.

Support vector machines (SVM) have a good track in text classification (Yang and Liu 1999). When applied to tasks of sentiment classification, SVM were also proved to be able to achieve comparable or better performance than other counterpart techniques (Pang et al. 2002; Ye et al. 2009). Based on SVM model, Ng et al.

(2003) investigated the role of linguistic knowledge in classifying movie reviews. The experimental accuracies range from 84.5% to 90.5%, depending on the features and the datasets used in the experiments. Gamom (2004) conducted the similar research with regard to the effects of linguistic knowledge on the performance of sentiment classification, and SVM achieves an accuracy of 77.5% when using obviously positive and negative consumer feedback data (1 versus 4 on scales of 1-4). Pang et al.(2002) tried to classify movie reviews into positive and negative by using three different classifiers, namely, SVM, naive bayes and Maximum Entropy for comparison. They tested different feature combinations including unigrams, unigrams+bigrams and unigrams+POS (part-of speech) tags, etc. The results show that SVM combined with unigrams obtains the highest accuracy of 82.9%. Similarly, Dave et al. (2003) experimented more feature sets with SVM and naive bayes, and the two classifiers achieve comparable performance.

N-gram is a statistical language model widely used in natural language processing tasks. Ye et al. (2009) applied three sentiment classifiers SVM, naive bayes and N-gram model to the destination reviews. They found SVM outperforms the other two classifiers with an accuracy peak at 86.06%, when the training corpus contained 700 reviews. Cui et al. (2006) compared N-gram model with Passive Aggressive (PA) algorithm which can be viewed as an online version of a SVM. The results suggested that PA algorithm is more appropriate for sentiment classification than N-gram model.

Another type of approaches uses semantic orientation of a word in order to determining review sentiment. Research on predicting the semantic orientation (SO) of adjectives was initiated by Hatzi-vassiloglou and McKeown (1997). Se-

mantic orientation has both direction (positive or negative) and intensity (mild or strong). In the present work, the PMI-IR approach taken by Turney (2002) is used to derive semantic orientation for selected phrases in the text. Two word phrases conforming to particular part-of-speech templates that represent possible descriptive combinations are used (see Table 1). Once the desired phrases have been extracted from the text, each one is assigned an SO value. The SO of a phrase is determined based upon the phrase's pointwise mutual information (PMI) with the positive and negative seed words. PMI between two terms, $term_1$ and $term_2$ is defined as follows:

$$PMI(term_1, term_2) = \log_2 \left(\frac{p(term_1 \& term_2)}{p(term_1)p(term_2)} \right) \quad (1)$$

where $p(term_1 \& term_2)$ is the probability that $term_1$ and $term_2$ co-occur.

The SO for a phrase is the difference between its PMI with the word "excellent" and its PMI with the word "poor." The method used to derive these values takes advantage of the possibility of using the World Wide Web as a corpus. The probabilities are estimated by querying the AltaVista Advanced Search engine for counts. The search engine's NEAR operator, representing occurrences of the two queried words within ten words of each other in a text, is used to define co-occurrence. The final SO equation is:

$$SO(phrase) = \log_2 \left(\frac{\text{hits}(phrase \text{ NEAR } excellent) \text{ hits}(poor)}{\text{hits}(phrase \text{ NEAR } poor) \text{ hits}(excellent)} \right) \quad (2)$$

A review is then classified as recommended/not recommended if the average sentiment orientation of its phrases is positive/negative.

	First word	Second word	Third word (Not extracted)
1	adjective	noun	anything
2	adverb	adjective	not noun
3	adjective	adjective	not noun
4	noun	adjective	not noun
5	adverb	verb	anything

Table 1: Two-word phrase patterns.

3. Data

We constructed a corpus by retrieving travel destination reviews from the Ctrip website (URL: <http://www.ctrip.com>). A crawler was developed by Java to download all traveler reviews of ten Chinese popular tourist cities, all of which were written prior to January 23, 2010.

Three annotators were trained to label these reviews. This resulted in a total of 763 positive and 457 negative reviews when the neutral ones were discarded. The 854 reviews (538 positive and 316 negative examples) posted before January 15, 2010 were used as training data, and the 366 reviews (225 positive and 141 negative examples) posted between January 15 and January 23 in 2010 were used as held-out test data.

4. Experiment setup

In the SVM and naive bayes experiments, we used CAS's SVMCLS 2.0 implementation of a support vector machine classifier and our own implementation of a naive bayes classifier (Ye et al. 2009). ICTCLAS was employed for Chinese word segmentation and part-of-speech. Then, we used information gain method for feature selection. Word presence rather than word frequency was chosen to represent a document for probability estimation. An N-gram classifier was built using LingPipe DynamicLMClassifier implementation (Carpenter, 2005). This

classifier depends on a character-based N-gram language model with a generalized form of Witten-Bell smoothing. DynamicLMClassifier accepts training events of categorized character sequences. Training is based on a multivariate estimator for the category distribution and dynamic language models for the per-category character sequence estimators. We set $N = 4$ in the character-based N-gram model. In the light that the performance of supervised learning approaches relies on the amount of training data, this paper examined the effects of different amounts of training data on the classification accuracy, precision and recall of the classifier.

For PMI-IR semantic orientation approach, since none of existing search engines has a Near operator, as in (Ye et al. 2006), we used an AND operator instead which returns documents that contain query words without restriction to their distance. Our preliminary experiments showed that the performance of PMI-IR is affected by the seed words and search engine, and zero is not appropriate value for classifying reviews. Here, we chose “Hao” (good) and “Cha” (bad) as seed words and Baidu search engine, and set 0.086 as sentiment classification threshold for travel destination reviews.

In addition, to investigate the effects of seed words and search engine on PMI-IR, we experimented with a variety of seed words and search engines to examine the effectiveness of different combinations. A Chinese movie review corpus of 108 positive examples and 105 negative examples was created using the method that is described in section 3. To determine representative seed words for movie data, ten college students were asked to list a few positive and negative words to describe their attitude to a movie. Then we ranked two types of words according to their appearance frequency, respectively. The most frequent positive word and negative

word were “JingCai” (excellent) and “ShiWang” (disappointed). The fifth most frequent ones were “BuCuo” (not bad) and “ShiBai” (failure). Those words were combined into four seed word pairs. The search engines for comparison were three popular ones in China, namely Baidu, Google and yahoo.

5. Results and discussion

5.1. Comparison between supervised and unsupervised classification approaches

Figure 1 illustrates the performance of the SVM classifier on various sizes of training sets. The classification accuracy, precision and recall for positive and negative reviews increase with more training examples. When training sets have 420 or more examples, all ratios reach 90% except the recall of negative reviews. A possible reason for low recall of negative reviews is that the number and average length of negative training reviews are less than those of positive reviews in our training corpus and thus the extracted negative features are relatively not sufficient. The recall of negative reviews is improved to 88.65% when 854 training examples are used.

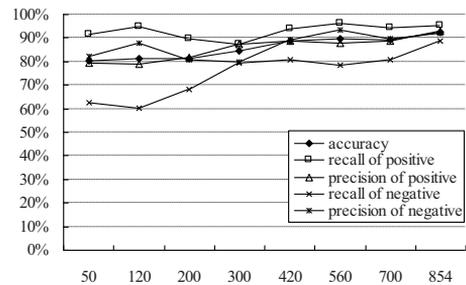


Fig. 1: The classification accuracy, precision and recall of the SVM classifier.

Figure 2 shows classification accuracy, precision and recall of the naive bayes

classifier on various sizes of training sets. The results indicate that the stability and effectiveness of naive bayes are inferior to SVM. The naive bayes classifier achieves best performance with 700 training examples, whereas due to over-training its performance begins to decline with more training data.

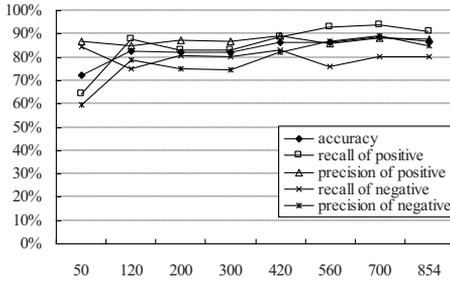


Fig. 2: The classification accuracy, precision and recall of the naive bayes classifier.

Figure 3 shows classification accuracy, precision and recall of the N-gram classifier on various sizes of training sets. The results indicate that the effectiveness of N-gram is inferior to SVM and naive bayes. Its performance is improved with more examples in the training data sets. As shown in Figure 1 to Figure 3, the N-gram classifier needs more training data than SVM and naive bayes to gain good performance.

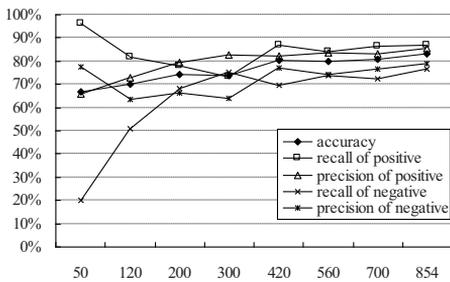


Fig. 3: The classification accuracy, precision and recall of the N-gram classifier.

Using “Hao” and “Cha” as seed words and Baidu search engine, PMI-IR achieves an accuracy of 81.42%. The precision and recall of positive reviews are 86.17% and 83.11%, and the precision and recall for negative reviews are 74.50% and 78.72% with a test corpus of 366 reviews.

Table 2 compares the accuracies of PMI-IR and those of SVM on various sizes of training examples. The results indicate that SVM outperforms PMI-IR with a large size of training corpus, whereas when the number of training examples is smaller than 300, there is no significant difference between the two classifiers. The results confirm our expectation that given infinite resources we can train a classifier using a supervised algorithm that outperforms unsupervised or weakly-supervised approaches. Nevertheless, acquisition of data and model training can be costly and time-consuming.

Training examples	SVM Accuracy	PMI	<i>p</i>
50	80.33%		0.7070
120	81.42%		1.0000
200	81.42%		1.0000
300	84.43%	0.8142	0.2799
420	88.80%		0.0051**
560	89.62%		0.0016**
700	89.07%		0.0035**
854	92.62%		0.0000**

Table 2: Comparison between accuracies of SVM and PMI-IR, ** denotes $p < .01$.

5.2. Effects of seed words and search engine on PMI-IR approach

Table 3 illustrates the average sentiment orientation of the 108 positive and 105 negative movie reviews with each combination of seed word pair and search engine. There seem no patterns about the average SO values achieved by the combinations of seed word pair and search engine. The reason is that the numbers of pages returned by three search engines

are quite different even with the same query, according to Equation (2), resulting in considerable variation of SO values of a phrase and a review. Thus, it is unreasonable to adopt zero to classify positive reviews and negative reviews. To set appropriate threshold for each combination, we manually selected three positive reviews and three negative reviews, and then use PMI-IR to calculate the average SO of the six documents. The average SO gained by each combination was used as the corresponding classification threshold.

		JingCai-ShiWang	JingCai-ShiBai	BuCuo-ShiWang	BuCuo-ShiBai
Baidu	Pos	-0.315	-0.576	0.100	-0.131
	Neg	-0.585	-0.855	0.060	-0.221
Google	Pos	-2.440	-1.479	-1.559	0.094
	Neg	-2.636	-1.766	-1.493	-0.006
Yahoo	Pos	0.405	-2.063	-1.095	-3.136
	Neg	0.314	-2.147	-0.755	-3.214

Table 3: Average sentiment orientation of positive and negative reviews with each combination of seed word pair and search engine.

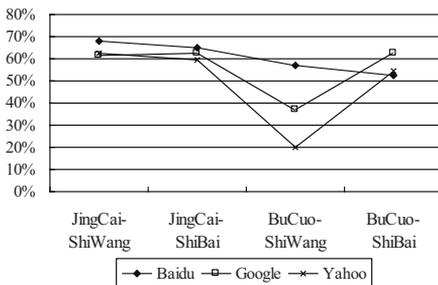


Fig. 4: Classification accuracy with each combination of the seed word pair and search engine.

With the rest of 105 positive and 102 negative reviews as test data, Figure 4 illustrates the classification accuracy using different seed words and search engines. The findings suggest that the effectiveness of PMI-IR approach is influ-

enced by the seed words and search engine. Seed word pairs “JingCai-ShiWang” and “JingCai-ShiBai” do a better job than “BuCuo-ShiWang” and “BuCuo-ShiBai” for calculating semantic orientation of movie reviews. Baidu search engine obtains relatively consistent performance.

6. Conclusion

This paper adapts three supervised learning approaches, namely, SVM, naive bayes and N-gram model, and an unsupervised semantic orientation approach to Chinese reviews. We look at the effects of training set size on the performance of those supervised learning approaches. The results show that as the amount of training data increases, the accuracies of SVM and N-gram classifiers are improved, whereas the naive bayes classifier is subject to over-training when the training examples reach 854. The SVM classifier outperforms naive bayes and N-gram model on various sizes of training examples, but SVM are not obviously better than PMI-IR when the number of training examples is smaller than 300. The results confirm our expectation that the machine learning approach is more accurate but requires a significant amount of data to train the model. In comparison, the semantic orientation approach is slightly less accurate but is more efficient to use in real-time applications. We also investigate the effects of seed words and search engines on PMI-IR and found that the effectiveness of PMI-IR relies on the seed words and search engine. Consequently, representative seed word pairs “JingCai-ShiWang” and “JingCai-ShiBai” are more appropriate for Chinese movie review mining. Baidu search engine does a better job than Google and Yahoo for calculating semantic orientation of Chinese phrases.

Acknowledgments

This study was funded by National Science Foundation of China (70890080-70890082, 70771032).

Reference

- [1] Carpenter, B., "Scaling High-order Character Language Models to Gigabytes," *Proc. of the 2005 association for computational linguistics software workshop*, pp. 1-14, 2005.
- [2] Chaovalit, P., Zhou, L., "Movie Review Mining: a Comparison between Supervised and Un-supervised Classification Approaches," *Proc. of the 38th Hawaii International Conference on System Sciences*, pp. 1-9, 2005.
- [3] Cui, H., Mittal, V., Datar, M., "Comparative Experiments on Sentiment Classification for Online Product Reviews," *Proc. of AAAI-06, the 21st National Conference on Artificial Intelligence*, 2006.
- [4] Dave, K., Lawrence, S., Pennock, D. M., "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *Proc. of 12th international conference on World Wide Web*, pp. 519-528, 2003.
- [5] Gamon, M., "Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis," *Proc. of the 20th Intl. Conf. on Computational Linguistics*, pp. 841-847, 2004.
- [6] Hatzivassiloglou, V., McKeown, K., "Predicting the Semantic Orientation of Adjectives," *Proc. of the 35th Annual meeting of the Association for Computational Linguistics*, pp. 174-181, 1997.
- [7] Ng, V., Dasgupta, S., Arifin, S. M. N., "Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews," *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 611-618, 2006.
- [8] Pang, B., Lee, L., Vaithyanathan, S., "Thumbs up? Sentiment Classification Using Machine Learning Techniques," *Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing*, pp. 79-86, 2002.
- [9] Turney, P., "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proc. of Association for Computational Linguistics 40th Anniversary Meeting*, pp. 417-424, 2002.
- [10] Yang, Y. M., Liu, X., "A Re-examination of Text Categorization Methods," *Proc. of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, pp. 42-49, 1999.
- [11] Ye, Q., Shi, W., Li, Y., "Sentiment classification for movie reviews in Chinese by improved semantic oriented approach," *Proc. of the 39th Hawaii international conference on system sciences*, 2006.
- [12] Ye, Q., Zhang, Z., Law, R., "Sentiment Classification of Online Reviews to Travel Destinations by Supervised Machine Learning Approaches," *Expert Systems with Applications*, 36(3), pp.6527-6535, 2009.