Maximum entropy-based sentiment analysis of online product reviews in Chinese

Conference Paper · May 2017

DOI: 10.1201/9781315210445-110

Jue (Grace) Xie
Monash University (Australia)

Maximum Entropy-based Sentiment Analysis of Online Product Reviews in Chinese

Hanqian Wu[1, a *], Jie Li[1,b] , Jue Xie[2,c]

[1] Institute of computer science and engineering, Southeast University,

Nanjing 210018，Jiangsu Province，China

[2] Southeast University - Monash University Joint Graduate School, Suzhou 215123, China

[a]email: hanqian@seu.edu.cn, [b]email:220151551@seu.edu.cn, [c]xiejue@gmail.com

Keywords: Chinese reviews, sentiment analysis, maximum entropy.

Abstract. With the explosion of Chinese online reviews, sentiment analysis emerges as an important research area. This paper proposes a supervised machine learning method based on Maximum Entropy for the classification of Chinese reviews. Maximum Entropy is a probability distribution estimation technique which has been used for a variety of Natural Language Processing tasks. Our experimental results demonstrate that the classifier based on Maximum Entropy is ideal for the classification of reviews, which can make both the accuracy and recall rate reaching considerable levels. Pre-processing of the reviews for training and testing purpose is also introduced in this paper.

Introduction

Nowadays, there are abundant Chinese reviews available in on-line documents. The majority of the online reviews are regarded as disorganized and unstructured. Merely using the artificial methods to understand their polarity is almost impossible. As a result, sentiment analysis starts to play an important role in opinion mining and product recommendation.

Sentiment classification automatically classifies a review as expressing a positive or negative opinion by mining and analyzing subjective information such as standpoint, view, attitude and moon etc. Such information is useful for manufactories to improve their products or services, and also helpful to potential customers in making purchase decisions. Thus, getting an accurate understanding of sentiment expressed in Chinese reviews is mutually beneficial.

Extensive research efforts have been made to analyzing sentiment in languages such as English and other European languages. In this paper however, we analyze the sentiment of Chinese reviews by applying a supervised classification technique of Maximum Entropy. Maximum Entropy as a supervised machine learning approach is a probability distribution estimation technique and has been extensively used for a variety of Natural Language Processing tasks. The challenging of this work is that training a Maximum Entropy model is usually a computationally intensive task.

The rest of the paper is organized as follows. Section 2 presents the related work on sentiment analysis. In Section 3, a Maximum Entropy-based model is proposed. Meanwhile Section 4 discusses the data pre-processing for sentiment analysis, which include segmentation, conjunction rules, negation handling and feature selection. Experimental results and discussions are presented in Section 5. Finally, our work is concluded in section 6 with future works.

Background of the Sentiment Analysis Methodology

In general, sentiment analysis methods can be divided into two groups, depending on which techniques they are based on, either lexicons or machine learning techniques.

Lexicon-based sentiment analysis methods, which use a dictionary to measure the polarity of reviews, have attracted wide publicity of many researchers. For these methods, a lexicon of sentiment words with their labeled polarity is required. Examples of the commonly used lexicons include: General Inquirer Lexicon, Opinion Lexicon, SentiWordNet, and etc. Turney et al.[1] calculate the sentiment orientation of sentences using Point-wise Mutual Information (PMI). It computes the sentiment orientation of the words in a subjective sentence based on pre-defined seed words. A sentence is classified into positive or negative category according to the average semantic orientation. Xiong Delan et al.[2] propose the method which is Semantic Distance-based to calculate sentence tendentiousness based on the semantic similarity calculation of HowNet. The semantic distance reflects the semantic relations between words in a sentence, which can obtain the orientation of sentences. Wen et al. [3]raise another sentiment analysis approach based on semantics where negation words and adverbs of degree are investigated. Wan [4] use a bilingual co-training approach for sentiment classification of Chinese product reviews utilizing both the English view and the Chinese view.

On the other hand, some studies have been concentrating on training sentiment classifiers using machine learning techniques. Nakagawa, Inui and Kurohashi [5] introduce a dependency tree-based classification method using Conditional Random Fields with hidden variables. Zhou, Chen and Wang[6]present a novel semi-supervised learning algorithm called Active Deep Networks (ADN) and then exploit it to approach the semi-supervised sentiment classification problem.Xu Qunling [7] proposes a new sentiment orientation calculation model, which is based on the analysis of the characteristics of Chinese text. The model uses an improved point by point analysis method SO-PMI to classify category polarity using words. Li et al.[8] investigate a more common case of semi-supervised learning for imbalanced sentiment classification. In particular, it expresses the imbalanced class distribution problem via various random subspaces that are dynamically generated.

As a machine learning algorithms, Maximum Entropy is widely used for Natural Language Processing tasks, such as part-of-speech tagging, language modeling and text segmentation. Maximum Entropy combines contextual features in a principled way and allows unrestricted use of them. Moreover, Wang and Acero[9] propose that the Maximum Entropy model can obtain global optimization due to the properties of the convex objective function.

Based on the above-mentioned factors, we train a Maximum Entropy classifier to analyze the sentiment of Chinese reviews. Our experimental results show that Maximum Entropy is a technique that warrants further investigation for text classification.

Maximum Entropy Classification

Maximum Entropy is a probability distribution estimation technique that is extensively used for Natural Language Processing tasks. The intuition that motivates Maximum Entropy classification is that we should build a model that captures the frequencies of individual joint-features, without making any unwarranted assumptions. The essential principle of Maximum Entropy is that the probability distribution should be uniform when there is no pre-knowledge. In this paper， Maximum Entropy classification is used to estimate the polarity of Chinese reviews.

There is no different from other learning technique, the outputs of machine learning technique are relied on the given training data-set of input. When using Maximum Entropy classification, the first step is to get the constraints that characterize the class-specific expectations for the distribution from labeled training data-set for the model distribution.

Figure 1 is the Maximum Entropy model for Chinese reviews classification, which is composed of a training process and a testing process. In the training process, a labeled training data-set is used to derive a set of constrains for building the model that characterizes the class-specific expectations for the distribution. Finally, we use the General Iterative Scaling algorithm to find the Maximum Entropy distribution that is consistent with the given constraints. In the process of classification, the testing data-set is denoted in features, and then the review classification is obtained by using the classifier.

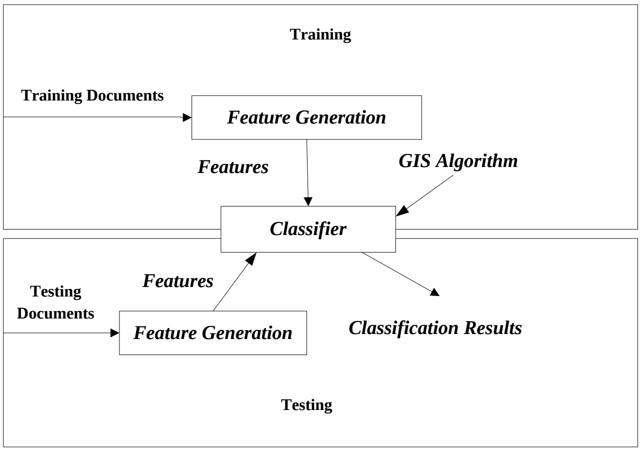

Fig 1 Model of review categorization

Data Pre-processing for Sentiment Analysis

Segmentation. There is different from European languages, Chinese sentences are made up of strings of Chinese characters without clear boundaries between words. Chinese characters are the smallest unit in Chinese sentences. Each Chinese character has a specific meaning, but we can combine two or more characters to form a word that has different meaning. For this reason, segmentation plays an important role in Chinese sentiment analysis, which belongs to the category of utilizes natural language processing techniques.

In this paper, Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS) is adopted to segmentation and Part-of-Speech tagging (POS). The purpose of segmentation and POS tagging is prepared for identifying what characters should be combined to labeling the part-of-speech. After that, we select the words, which have the specific part-of-speech using Chinese part-of-speech tagging set.

It is obvious that Chinese sentiment analysis cannot continue without a proper segmentation algorithm. The Chinese sentence must be decomposed into words before any tasks can take place. This is the first stage in pre-processing of Chinese sentiment analysis.

Conjunction Rules. Ordinarily a sentence only expresses one sentiment orientation if it does not contain some conjunction words such as BUT, ONLY, ALTHOUGH, HOWEVER, WHILE and etc. These conjunction words will change the sentiment orientation of sentences. In this case, we use conjunction rules to extract the precise meaning from a given sentence.

For an example:

对于京东我一直很信赖，服务到位，但是这台电脑实在太慢了，很糟糕的一次购物，真是太失望了。(I have always trusted Jingdong because of its well-service, but the computer is too slow, it is a very bad shopping experience, really too disappointed.)

In this case, the phrase before ‘BUT’ will be cut off, and the rest of the sentence will be remained to represent the emotional polarity of the whole sentence. Although some of the sentence information is lost, the accuracy for the sentence sentiment analysis will be improved.

Conjunction words will affect the polarity of the sentence. Using conjunction rules can makes the sentence more explicit and comprehensible.

Negations Handling. Negation is a very common linguistic construction that widely appears in all languages. Therefore negation should be taken into consideration in Chinese sentiment analysis. When negation words, such as ‘not’ and ‘hardly’ occurs in the sentence, the polarity will be changed. Negation detection is used for distinguishing the factual and non-factual information that extracted from sentences.

A negation word such as ‘hardly’ will invert the polarity of the sentiment word. For example, ‘hardly know’ means ‘unknown’. Das and Chen[10]raise a technique, which uses the tag ‘NOT_’ to replace the negation word and the word following the negation word. After that, a new corpus was achieved. In Chinese, we substitute ‘不_’ for ‘NO_’ , and the list of the negative words contains ‘无’, ‘不’, ‘没有’, ‘非’, ‘没’, ‘未必’.

Feature Selection. Feature selection plays an important role in sentiment analysis. Feature selection aims to select features, which have most representative and excellent classification performance from original feature information. Therefore, the appropriate feature selection method will largely determine the quality of the final classification results. Typically, words are often used as features in text categorization. The relationship between words and categories is calculated to measure the contribution of each word to its category.

There is a huge quantity of words in data processing. The dimension of the feature space will be too large if all words are selected as the feature items. It not only increases the amount of computation, but also affects the accuracy of classification. Therefore, it is necessary to chose the appropriate feature selection method to reduce the dimension.

In feature selection, the features that are irrelevant to the emotions and the weak category correlation need to be removed to eliminate unnecessary interference. In this paper, the method of sentiment lexicon is adopted. The Bilingual Knowledge Dictionary provided by HowNet and National Taiwan University School of Dentistry are applied to distinguish the sentiment words from candidate feature words.

Experiment

The analysis data. Our corpus is derived from the Jingdong shopping site (www.jd.com), a leading online B to C trading platform in China. The corpus contains 450417 reviews: 80% positive and negative reviews are used as the training data, while 20% reviews for both polarities are used for testing purpose.

All reviews have been labeled by the starValue attribute that is rated by customers. The reviews are then manually set to positive when starValue >=3 and negative when starValue < 3.

Model Evaluation. It is very important to evaluate the experiment results. The balance of various factors is considered in the selection of evaluation indicators to make an objective and impartial evaluation. Given sentiment analysis is drawn in the category of Natural Language Processing, the evaluation indicators in information retrieval field can be used to evaluate the effectiveness of classification. In this paper, five evaluation indicators are used to evaluate the experiment results, they are Precision, Recall, F-score, Accuracy and Specificity respectively.

Experiment 1. One word may have opposite meaning in different contexts, such as "骄傲" (pride). For example:

他们可以骄傲地回首过去付出的努力。 (They can look back on their endeavors with pride.)

骄傲乃万恶之源。(Pride is the parent of all evils.)

Thus, two scenarios are taken into consideration in the phase of training. One situation is that extracted sentiment words are labeled as both positive and negative. The words here refer to the features. For instance, feature "骄傲" (pride) have both positive and negative sentiment polarities.

In this scenario, a high recall can be guaranteed. The recall means that recognition of positive reviews accounts for the proportion of the total positive reviews. However, the capacity of the classifier to recognize the negative reviews is lower, namely specificity. The result is shown in table 1 below.

Table 1 Result from Experiment 1

indicator	precision	recall	F-score	accuracy	specificity
result	0.9806535	0.9777033	0.97917616	0.95956135	0.31958762

Experiment 2. Another scenario is that the labeled sentiment words are either positive or negative. For instance, feature "骄傲" (pride) is either of the positive or negative category.

In this situation, the capacity of the classifier to recognize the negative reviews is improved, which has increased by approximately 40%. But the indicators of recall and accuracy are reduced by 15 percentage points. The result is shown in table 2 below.

Table 2 Result from Experiment 2

indicator	precision	recall	F-score	accuracy	specificity
result	0.9890121	0.81236035	0.8920245	0.8087564	0.68162525

From the experimental results, we can see that the performance of the classifier is better in the second scenario, where the capacity of the classifier to recognize positive reviews and negative reviews is well matched. In summary, the indicators of precision, accuracy, and F-score are approximately 90% in both scenarios. We can conclude that the classifier based on the Maximum Entropy can get satisfying classification performance on the Chinese reviews.

Conclusion

Maximum Entropy is a technique that is widely used for Natural Language Processing. Its overriding principle is that probability distributions should be estimated from the training data-set.

In this paper, we present the approach for sentiment analysis of Chinese reviews based on Maximum Entropy. Specifically, the paper analyzes the process of sentiment analysis for short-text based reviews. Our approach is mainly developed in three aspects of pre-processing, feature extraction and classification arithmetic based on the Maximum Entropy model.

The experimental results demonstrate that the classifier is ideal for the classification of Chinese reviews, which can yield a high accuracy and recall.

Future Work

At present, few researchers take the strength of user sentiments into consideration in sentiment classification. In the future work we will estimate the sentiment strength according to the strength of adverbs and adjectives in the reviews.

Another area for future research is to make a further comparison of Maximum Entropy to the other classification algorithms. Ongoing work includes direct comparisons of Maximum Entropy to Ripper, Support Vector Machines and K-Nearest Neighbor.

Given the increasingly rampant and sophisticated opinion spamming in social media and the major challenge it presents for opinion spam detection, another interesting area for future research is to detect the spam reviews from Chinese online reviews.

References

[1]Turney, Peter 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 (ACL-2002).2002.

[2]Xiong Delan, Juming Cheng, and Shengli Tian. The research of sentence sentiment tendency based on HowNet. Computer engineering and Applications 44(22): 143-145, 2008

[3]Wen B. et al. (2010) Text Sentiment Classification Research based on Semantic Comprehension. Computer Science 37(6):261-264.

[4]Wan X (2011) Bilingual co-training for sentiment classification of Chinese product reviews. Computational Linguistics 37 (3):587-616.

[5]Nakagawa, Tetsuji, Kentaro Inui, and Sadao Kurohashi. Dependency treebased sentiment classifi cation using CRFs with hidden variables. in Proceedings of Human Language Technologies: The 20 10 Annual Conference of the North American Chapter of the ACL (HAACL-2010).2010.

[6]Zhou, Shusen, Qingcai Chen, and Xiaolong Wang. Active deep networks for semi-supervised sen timent classification. in Proceedings of Coling 2010: Poster Volume. 2010.

[7]Xu Qunling. A new model of Chinese text sentiment computing [J]. Computer application technology and software, 2011,6.

[8]Li, Shoushan, Zhongqing Wang, Guodong Zhou, and Sophia Yat Mei Lee. Semi-Supervised Learning for Imbalanced Sentiment Classification. In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-2011). 2011.

[9]Wang, Y. Y., and Acero, A. 2007. Maximum Entropy Model Parameterization with TF*IDF Weighted Vector Space Model. IEEE Automatic Speech Recognition and Understanding Workshop, 213-218. Kyoto, Japan: Institute of Electrical and Electronics Engineers, Inc.

[10]Das, S., and Chen, M. 2001. Yahoo! for Amazon:Extracting market sentiment from stock message boards. 8th Asia Pacific Finance Association Annual Conference (APFA).