Automatic Positive Sentiment Word Extraction for Chinese Text Classification

Zhen'gang Yu
Dept. of Computer Science
Hangzhou Dianzi University
Hangzhou, China, 310085
kingc_yu@hotmail.com

Ning Zhen, Ming Xu
Dept. of Computer Science
Hangzhou Dianzi University
Hangzhou, China, 310085
{nzheng, mxu}@hdu.edu.cn

***Abstract*— Sentiment analysis aims to predict sentiment tendency automatically. Traditional methods tackling this problem are mostly based on supervised learning, but it is time-consuming and uneasy to extendable. In this paper, we provide a novel method of sentiment analysis based on un-supervised learning together with some language rules. It is no necessary to have a positive sentiment dictionary beforehand as we can build it automatically during processing the comments. By this positive sentiment dictionary, it provides an efficient way to classify the product reviews. The methodology presented is easy to extend due to its un-domain-dependency. As we can see, the experiment result obtained shows its promising application.**

Keywords*— *sentiment analysis; product reviews; unsupervised learning; Chinese

I. INTRODUCTION

Nowadays, the information on the web is much easier available than ever before. And it plays an important part in people's life. On the one hand, the potential customers usually trend to get enough information before they make their decisions. For instance, they will go to some famous product forums in which existing many related people's (people who had bought the same product or very familiar to it) comments on the product that they want to buy. On the other hand, the corporations who sell the product or services also want to reference the reviews which the customers have already made to improve their future market planning. Due to the huge number of information, it is not easy or realistic for human who manage the forum to tackle all the reviews and classify them to positive or negative manually. Under this situation, automatic sentiment analysis has significant meaning. And via this method, we can automatically extract the opinion which the text trends to.

In this paper, a new methodology is provided to classify the reviews. It uses some seed words and some language rules to determine the review's polarity, after that, new sentiment words will be selected and the positive sentiment word dictionary will be expanded dynamically at the same time. Once the positive sentiment dictionary has been formed eventually, we will use this dictionary to classify the whole reviews once again. Due to the un-domain dependency and little language restricts, the method can also be extended to other languages easily.

The remaining of the paper is arranged as follows. In section II, we introduce the previous related researches on sentiment classification. In section III, we illustrate our method in detail. The corpus used in our experiment and the experiments result will be presented in section IV. Section V draw the conclusion of our research and then propose the future work.

II. RELATED WORK

One important issue of opinion mining is to judge whether the opinion is positive or negative or objective which also say sentiment classification. And lots of work has already been done in this area.

Sentiment classification can be group into three levels: word based, sentence based and document based. Here we just focus on the document sentiment classification. The method be used for document sentiment classification can be generally categorize into lexicon-based and corpus-based.

Corpus-based methods usually are associated with machine learning. The main character of machine learning is that it considers the sentiment analysis as a classification task and always uses a labeled corpus to train a sentiment classifier. For example, [1] took use of three classical machine learning approaches to classify the reviews: Naive Bayes, Maximum Entropy and Support Vector Machines. Then they conducted their experiment on the movie reviews which had already been annotated. Since then, kinds of classification models and linguistic features have been proposed to improve the classification performance ([2], [3], [4] and [5]). Meanwhile, there are also some restricts of the supervised system. It is often the case that people have to collect annotated data in the new domain and retrain the classifier again if their research area moves to another domain. [6] showed that the accuracy in cross-domain classification would decrease because the sentiment was often expressed by different means in different domains. [7] investigated domain adaptation for sentiment classifiers for different product types.

Lexicon-based methods usually involve with sentiment lexicons and some linguistic features. The sentiment lexicons can comprise of several seed words or just a big dictionary. And the sentiment lexicons usually are utilized when researchers adopt the method of the unsupervised learning. [8] proposed a method naming PMI (point-wise mutual information). They used two human-selected seed words (the words poor and excellent), then the phrases' semantic orientation is evaluated through their association with the seed words. The final sentiment orientation of a document is

calculated as the average semantic orientation of all such phrases. [9] built three models to do the classification based on the word sentiment orientation in order to assign a sentiment category to a given sentence.

Research works on Chinese sentiment analysis have also been focused ([10], [11] and [12]) and most such work adopted similar lexicon-based or corpus-based methods. [13] proposed an approach that can extract seed words automatically. Our approach is similar to theirs as we also use some seed words to select new seed words automatically. Different to theirs, we have taken advantage of some negative language features and we use the slide window to extract the candidate word while theirs has to statistic the frequent of the word occurring in both negative corpora and positive corpora.

III. OUR APPROACH

To determine a review's polarity fast and automatically, in this section, we present our approach which belongs to unsupervised learning. Unsupervised approaches seem better than supervised approaches in that they can extend to other domains easily and they don't need annotated corpus as supervised approaches do.

Generally speaking, most sentiment words are adjectives, and the pre-processing of word segmentation will be facilitated to find the adjective. But in our approach, there is no need to segment the Chinese text into words as most NLP researchers did. The main reasons why we don't take word processing are:

- Most words have multi-POS. For instance, "好" (*good*) sometimes is an adjective but sometimes it also may be a verb or an adverb. Therefore, the words' POS may be not the right one in some cases.
- The Chinese "word" after segmentation may be not a meaningful word in fact.

The approach addressed here can be described as follows: Firstly, we determine the sentiment orientation of the review by some Chinese language traits and the seed word, such as "好", "不错" (*not bad*), etc. Then we use a slide window with fixed length to extract the word behind the adverb as positive sentiment word if the current comment's polarity is positive. The new sentiment word together with the old ones in the positive sentiment dictionary will be utilized to judge the next comments' polarity iteratively. We also use a negative sentiment dictionary to judge the comment if the comment does not contain any positive word. If there is a negative word which defined in our negative sentiment dictionary appears in the comment and no

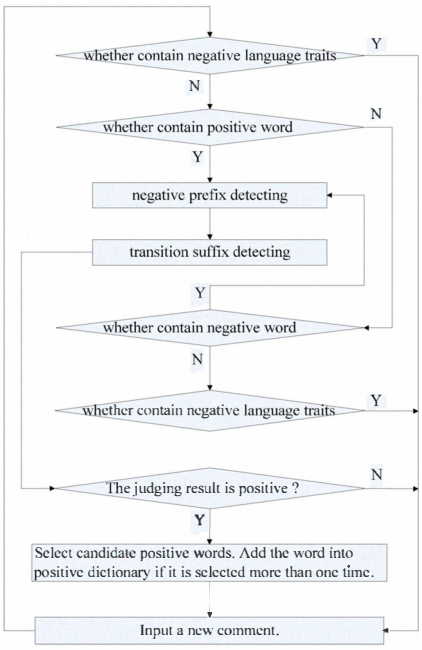

Figure 1 The process of our approach

contrary word such as "但" (*but*) follows the negative word, thus the review is determined as negative. Then we will extract positive words to form the positive word dictionary. After the whole comments have been handled, we use the new formed positive word dictionary as the new initial seed words to do the whole process again. That means we use 2 iterations.

Fig. 1 above shows the process in detail.

3.1 Basic Presumptions and Observation

We assume that in the products comments, people who are reading them have already known what the topic is and in most cases there are only one topic they talk about. In this paper, we do not care about the case that the original of the review's truth. That is, we assume all the product comments are reliable, no matter it belongs to positive or negative.

The observation which is also a common phenomenon that people are accustomed to use adverb together with positive word such as "非常好!" (*Very good!*) , "相当出色!" (*Quite excellent!*), etc. Therefore, if the polarity of the comment is positive, then we can identify the positive word by use of the adverb.

3.2 Negative Language Traits in the Comments

As we know, people are free to express their opinions in the online forum due to its virtual character. So the words they used in the comment are usually un-official or similar to oral language. For example, they often use exclamation "唉" or "哎" to express their disappointments about something. And when the quality of the product or service is too bad, people will use three or even more exclamation mark "!" to show their crazy furious. Thus, once these negative language

traits[1] appear, we can draw the conclusion that this comment is negative under most situations.

3.3 Initial Positive Seed Word

In positive product reviews, "好" is always more often used without negation ([7], [13]). So we can use "好" as an initial seed word. Besides, we find that the word "不错" is also an important seed word to show people's satisfaction feeling which [13] did not proposed. In fact, their approach can't select "不错" as the candidate positive seed word because it has a prefix negation word "不" (*no*). In the product comments, we also observed that when the product support some functions, the whole polarity of the sentence is always be treated as positive although it should have no sentiment. So we also accept the word "支持" (*support*) into the initial positive seed set.

Once the final positive word dictionary is created (See section 3.8), then the whole dictionary together with the three initial seed words will be treated as the new initial seed words.

3.4 Prefix Negation Word Detecting

The purpose of negation detecting is to determine whether the positive word is preceded by any negation2 within a slide window. Here we define the width of slide window is five characters (space is excluded) before the positive word.

Since the checking process is over, the comments' polarity will be identified. Here we define P_i is the polarity of the comment before checking and P_e after. The number of the negation is defined as N_{neg}. Then P_e can be calculated according to the following equation:

$$P_e = (-1)^{N_{neg}} \times P_i \quad (1)$$

In (1), the value of P_i is 1 if the comment before checking is positive and -1 otherwise. The value of P_e has the same meaning. That is to say, value 1 indicates the polarity of the comment after negation checking is positive and -1 negative.

3.5 Transition Word Detecting

If a transition word such as "但" appears in a sentence, it is usual the case that the sentiment orientation between the part behind of transition word and the part ahead of it is opposite. Therefore, the polarity will be converted when a transition word is detected. And the final sentiment is same to the behind part under most situation.

In our methodology, just one transition word "但" is adopted due to its frequently occurring. And similar to prefix negation word detecting, we also use a slide window with a fixed width to determine whether the transition word is in. According to the experiment performance, we use the width of 15 characters after the sentiment word.

3.6 Negation Word Detecting

This step will be executed if there are no positive words detected. We create a customized negation word dictionary (Appendix A) which contains the most frequently used negation words. Once there is a negation word found, the prefix negation word detecting and the transition word detecting are needed.

3.7 Candidate Positive Word Selecting

Candidate positive word selecting is used to enlarge the initial positive seed vocabulary into a comprehensive one.

The prerequisites to select candidate positive word are:

- The orientation of the comment after judging is positive. It assures that the candidate words we extract are also positive.
- The comment at least has one adverb which defined in our adverb dictionary[3].

Once the above two conditions are met, then the two characters after the specific adverbial will be extracted as candidate positive word if it is not contained in the negation word dictionary defined in Appendix A. In the next iteration, the new selected candidate positive word together with the old ones will be utilized to classify the next comment until no more comments available.

3.8 Positive Word Dictionary Building

When all comments have been classified, some positive sentiment words need to be cut. Each candidate word that occurs at least twice in the corpus will be accepted to the positive word dictionary. This restricts some meaningless words to participate so that the precision gets improved. The positive word dictionary formed eventually can be seen in Appendix B.

IV. EXPERIMENTS

4.1 Data Set

To compare our approach's performance with the previous unsupervised method, especially concerning to Chinese corpus, we adopt the corpus[4] [13] adopted which is derived from the website IT168. In this corpus, all the reviews were tagged by their authors as either positive or negative overall. It contains a total of 7982 reviews distributed between 10 product types.

---

[1] We use three groups of negative language traits that can be observed obviously:
① "唉", "哎"(means 'sigh'), "经常"(the potential meaning: go bad very frequently), "才.....就"(the potential meaning: go bad very soon), "只好"(means 'have to').
② Two question marks("??") or more, four exclamation marks("!!!!") or more, four dots ("….") or more.
③ The sentence contains "太"(means 'very') but not contains "不"(means 'no') in a fixed window.

[2] We use only three frequently negations in daily life: "不"(bu),"没"(mei) and "避免"(bimian).

[3] We use five frequently occurring adverbials: "十分"(shifen), "非常"(feichang), "相当"(xiangdang), "特别"(tebie), "较"(jiao).

[4] This corpus can be available at: http://www.informatics.sussex.ac.uk/users/tz21/coling08.zip

We evaluate the overall results using the accuracy formulated as follows:

$$Accu = \frac{Num_{correct}}{Num_{total}} \quad (2)$$

where $Num_{correct}$ is the number of reviews that are correctly classified and Num_{total} is the total number of reviews in the corpora.

We also measure the performance of our approach by F1-measure in positive reviews.

4.2 Experiment 1

At the beginning, we use the three seed words "好", "不错" and "支持" to run our system. After all the comments have been processed, a positive word dictionary will be created (Section 3.8). Then we use the positive word dictionary whole as the seed words to run our system again. The result shows in the Fig. 2. The macro-averaged accuracy with our positive word dictionary is 86.23% while the result without it is 83.61%.

4.3 Experiments 2

We use the positive sentiment word vocabulary which we created in Section 3.8 (Appendix B) as the initial seed word to run our full system. Fig. 3 shows the results measured by accuracy.

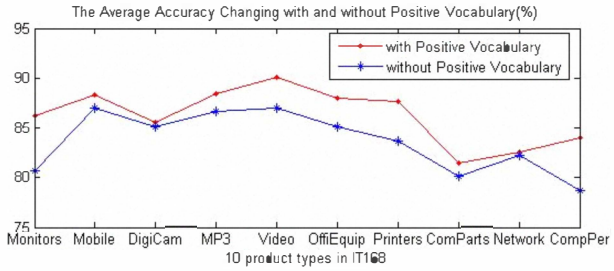

Figure 2 Results with the whole positive vocabulary as the seeds

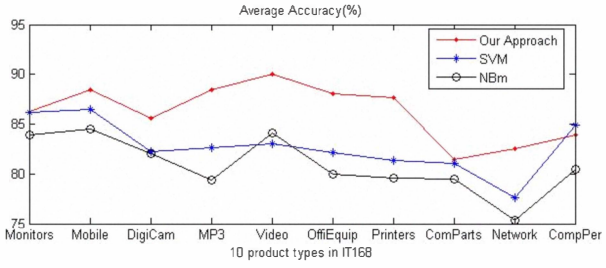

Figure 3. Average Accuracy of Our Approach compared with SVM and NBm

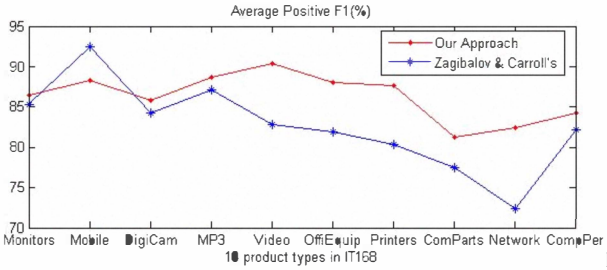

Figure 4. Average Positive F1 Measure of Our Approach compared with Zagibalov & Carroll's

In the experiment, we compare the average accuracy of our approach with the supervised method using SVM (Support Vector Machines) and NBm (Naive Bayes multinomial)[5]. The result shows that our approach performs better than SVM and NBm in most cases. And the macro-averaged accuracy across all 10 corpora is 86.23% comparing to 82.78% (NBm) and 80.89% (SVM).

4.4 Experiments 3

We also used the positive sentiment word vocabulary created in Section 3.8 to run our full system to compare with the system in [13] which is also an unsupervised system. Fig. 4 shows the result measured by positive comments' F1 value. Different to their iteration process, our approach only need 2 iterations. Although the recall will be increased as the number of iteration grows, the precision and the whole accuracy will be decreased. To balance the performance, we just use 2 iterations.

The outcome indicates that in most cases our system performs better than the system in [13]. The positive macro-average F1 value of our system is 86.32% which is better than theirs' 82.72%. The curve tendency overall is much more gentle. This fact indicates that our system can run stably. One important reason lies in that we adopt three seed words while they only use one word "好". The word "不错" we adopted is one of the most frequently word occurring in positive sentence. But in [13], "不错" was not selected as positive word in that there was a negation "不" in it. Thus, the overall performance was greatly affected. The comparison between ours system and theirs is obvious especially in the "Networking" sub-corpora which existing many "不错".

We note that in the "Mobile phones" sub-corpora, our system's positive F1 measure is 88.29% while the other system's is 92.41%. The phenomenon can be attributed to the following reasons: Firstly, the algorithm presented in this paper does not considering the comparative sentences. That is, if the comment contains much more than one objects or products, and if the other product(s) is better than the current product such as "A 比 B 好"(A is better than B), then our system's result will be positive although the truth is negative. Because the sentence contains "好" and exclude pre-negative word and does not contain transition word, thus the outcome

[5] We reference the result in *Zagibalov & Carroll's* paper which the result is got by use of WEKA 3.4.1.

of comment's polarity is positive. Secondly, in our approach if the comment does not contain any positive word or negative word within the dictionary we defined or no obvious language traits we defined in Section 3.2, then the algorithm will classify this comment to be positive no matter it belongs to negative or positive.

V. CONCLUSIONS AND FUTURE WORK

The unsupervised approach is an efficient way to tackling the problem of domain dependency as it does not depend on the annotated data.

The approach we proposed in this paper takes advantage of what the unsupervised approach has. Besides this, we also observed that there are many negative language traits in the online product reviews. Once these traits appear, the polarity of the comment is negative under most situations.

What's more, we have also built up a positive sentiment word dictionary while processing the comments. With this dictionary, we can locate more positive reviews and the accuracy also will be improved which indicated in Section 4.2. The remaining experiments also show that our system performs well: nine out of ten results showed that our approach outperforms the supervised approach. And the macro-average F1 measure of our system beats the most nearly unsupervised system near 5 percent.

In the near future, we plan to improve our system in the following aspects:

- To find an efficient way of classifying the comment that contains more than one product. Usually people will compare each other among these products, and the whole orientation relies on the comparing results. In our current system, we did not consider this problem thoroughly.
- To find a feasible methodology of judging the comment that does not contain any positive word, negative word or obvious negative language trait.

It is also likely that we will extend our system to other languages other than Chinese and other genres such as newspaper reviews that including some quantity of evaluative language.

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of Zhejiang Province of China under Grant No. Y1090114. Great appreciation is addressed.

REFERENCES

[1] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", in Proceedings of EMNLP. 2002.

[2] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts", in Proceedings of ACL. 2004.

[3] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources", in Proceedings of EMNLP. 2004.

[4] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", in Proceedings of HLT/EMNLP. 2005.

[5] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification", in Proceedings of ACL. 2005.

[6] Engström, Charlotte, "Topic dependence in sentiment classification", Unpublished MPhil dissertation, University of Cambridge.

[7] J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, Boom-boxes and Blenders, "Domain Adaptation for Sentiment Classification", in Proceedings of ACL. 2007.

[8] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", in Proceedings of ACL. 2002.

[9] SM. Kim and E. Hovy, "Determining the sentiment of opinions", in Proceedings of COLING. 2004.

[10] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", in Proceedings of HLT/EMNLP. 2005.

[11] Q. Ye, W. Shi and Y. Li, "Sentiment classification for movie reviews in Chinese by improved semantic oriented approach", in Proceedings of 39th Hawaii International Conference on System Sciences. 2006.

[12] J. Li and M. Sun, "Experimental study on sentiment classification of Chinese review using machine learning techniques", in Proceeding of IEEE NLPKE. 2007.

[13] T. Zagibalov and J. Carroll, "Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text", in Proceedings of COLING. 2008.

Appendix A. The Negation Words Adopted

差 (bad, poor)	耗 (to consume, to cost, to waste)
不 (no, not)	破 (broken)
无 (no, without)	只 (only, merely)
坏 (broken)	麻烦 (inconvenient, boring)
烂 (means the quality or service is very bad)	容易 (likely to be…)
贵 (expensive)	明显 (obvious)
次 (not good)	问题 (problem)
没 (no, without)	难看 (ugly)
缺 (to lack, to be deficient)	严重 (serious)
欠 (to be lacking)	垃圾 (rubbish)
慢 (slow, to postpone)	毛病 (a trouble, a shortcoming, a fault)
晕 (faint)	

Appendix B. Positive Seed Words Automatically Selected for each Corpus

Corpus	*Seed*	*Corpus*	*Seed*
Monitors	好(good) , 不错 (not bad), 支持 (support) 出色 (excellent), 时尚 (stylish, popular) 适合 (adapt to), 清晰 (clarity, clear) 舒服 (comfortable), 专业 (professional) 方便 (convenient), 犀利 (incisive), 简洁 (concise)	Video cameras and lenses	好(good) , 不错 (not bad), 支持 (support) 出色 (excellent), 清晰 (clarity, clear) 清楚 (clear), 人性 (humanity), 精细(exquisite) 方便 (convenient), 简单 (simple), 满意(satisfied) 自然 (naturally), 轻巧(light and handy) 灵活 (flexible), 令人 (cause someone to)
Mobile Phones	好(good) , 不错 (not bad), 支持 (support) 出色 (excellent), 合适 (suitable), 清晰(clarity, clear) 满意 (satisfied), 实用 (practical), 漂亮 (good-looking) 方便 (convenient), 时尚 (stylish, popular) 人性 (humanity), 好用 (useful), 流畅 (fluent) 流行 (popular), 齐全 (complete), 细腻 (delicate) 小巧 (compact), 顺畅 (smooth), 响亮 (sonority) 丰富 (plentiful), 符合(conform to), 舒服 (comfortable) 喜欢 (be fond of), 鲜艳 (bright), 适合 (adapt to) 合理 (reasonable), 强大 (powerful), 舒适 (comfortable) 准确 (accurate), 精彩 (splendid), 理想 (perfect) 精致 (fineness), 耐磨 (wear proof)	Digital cameras	好(good) , 不错 (not bad), 支持 (support) 出色 (excellent), 满意 (satisfied) 专业(professional), 方便(convenient) 小巧 (compact), 丰富 (plentiful) 喜欢 (be fond of), 合理 (reasonable) 适合 (adapt to), 突出(outstanding), 简洁 (concise), 实用 (practical), 人性 (humanity), 省电 (saving electricity) 耐用 (durable), 全面 (comprehensive) 优秀 (excellent), 时尚 (stylish, popular) 迅速 (rapid), 高性 (high-performance)
Network-ing	好(good) , 不错 (not bad), 支持 (support) 简单 (simple), 稳定 (stable)	Printers	好(good) , 不错 (not bad), 支持 (support) 出色 (excellent), 清晰 (clarity, clear),丰富 (plentiful) 满意 (satisfied), 简单 (simple), 方便 (convenient) 人性(humanity), 便宜 (cheap), 适合 (adapt to)
MP3 Players	好(good) , 不错 (not bad), 支持 (support) 时尚 (stylish, popular), 流畅 (fluent), 出色 (excellent) 突出(outstanding), 小巧 (compact), 齐全(complete) 紧密 (compact), 轻巧 (light and handy) 清晰 (clarity, clear), 丰富 (plentiful), 舒服(comfortable) 满意 (satisfied), 适合 (adapt to), 精细 (exquisite) 舒适 (comfortable), 漂亮 (good-looking) 精致 (fineness), 省电 (saving electricity) 吸引 (attractive), 简单 (simple)	Computer peripherals	好(good) , 不错 (not bad), 支持 (support) 出色 (excellent), 舒服(comfortable) 喜欢 (be fond of), 舒适 (comfortable) 紧密 (compact), 清晰 (clarity, clear) 适合 (adapt to), 顺手 (conveniently)
Computer parts	好(good) , 不错 (not bad), 支持 (support) 稳定 (stable)	Office equipment	好(good) , 不错 (not bad), 支持 (support) 出色 (excellent), 满意 (satisfied), 齐全(complete) 方便(convenient), 便宜(cheap) , 适合 (adapt to) 清晰(clarity, clear), 丰富 (plentiful)