Using Reviewer Information to Improve Performance of Low-Quality Review Detection

Qingliang Miao, Changjian Hu, Feiyu Xu

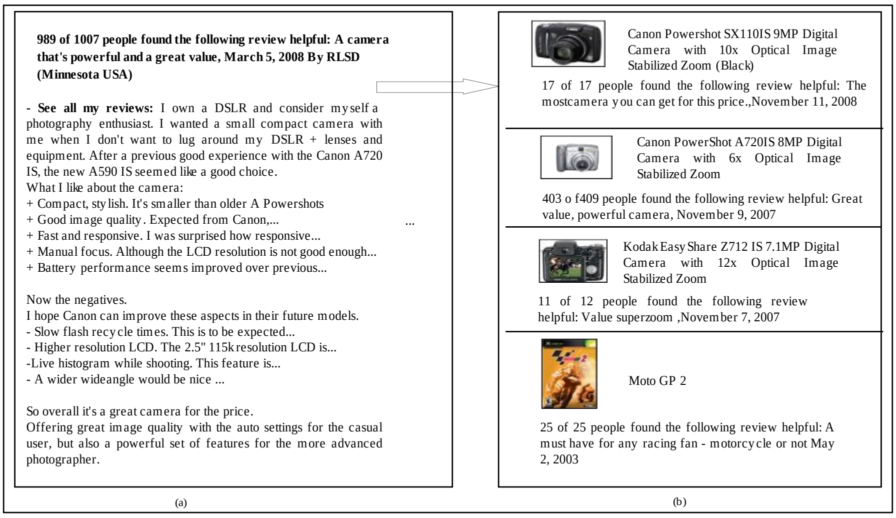
Lenovo Research, No. 6 Shangdi West Road, Haidian District, Beijing, China
{miaoql1, hucj1, fxu}@lenovo.com

Abstract. The drastic increase of user-generated contents has exhibited a rich source for mining opinions. Unfortunately, the quality of user-generated content varies significantly from excellent to meaningless, which by general estimation, causes a great deal of difficulty in mining-related applications. In the field of low-quality review detection, many previous approaches have individually detected low-quality reviews by using the intrinsic features of the review. However, no systematic study measuring the significance of reviewer information for detecting low-quality reviews has been previously done. In this paper, the importance of reviewer information when predicting review quality is studied and how to exploit it to build low-quality review detection models is determined. The experimental results on two different domains show that reviewer information does matter when modeling and predicting the quality of reviews. It is also shown that significant performance improvements can be achieved if the reviewer information is integrated with the intrinsic features of the reviews. These findings are of the essence in solving the low-quality review detection problem and in developing review-based opinion mining applications.

1 Introduction

With the dramatic development of Web 2.0, user-generated contents have become increasingly prevalent on the web. Popular user-generated contents include reviews on e-commerce websites, blogs, and web forums. However, due to the absence of editorial and quality control, user-generated contents vary greatly in quality, which in general estimation, causes problems in mining applications, such as opinion extraction [1], [2], sentiment classification [3], [4], and opinion summary [6]. In the opinion retrieval domain, the quality of a review can also be incorporated into retrieval models in the form of a prior probability [10].

Product reviews are widely used to mine customers' opinions on products. So review-based opinion mining can be more effective if low-quality reviews are preliminarily filtered. According to [6], low-quality reviews are reviews that have little or incorrect description of a product, have little or no comments on some aspects of the product, and do not provide convincing opinions with sufficient supporting evidence. Figures 1 (a) and (b), respectively, show review and reviewer information from a high-quality review. The high-quality review shown in Figure 1 (a) describes several aspects of the product, such as appearance, image quality, response, and battery, and provides convincing opinions with sufficient supporting evidence. Figure 1 (b) shows that the reviewer has published four reviews: one is about video games, and the other three are on digital cameras. Moreover, it can be seen that the helpful votes and total votes for the reviewer is very high, (17/17, 403/409, 11/12, 25/25). Based on this reviewer information, it may be concluded that the reviewer is likely to publish high-quality reviews, especially in the digital camera domain.



989 of 1007 people found the following review helpful: A camera that's powerful and a great value, March 5, 2008 By RLSD (Minnesota USA)

- See all my reviews: I own a DSLR and consider myself a photography enthusiast. I wanted a small compact camera with me when I don't want to lug around my DSLR + lenses and equipment. After a previous good experience with the Canon A720 IS, the new A590 IS seemed like a good choice.
What I like about the camera:
+ Compact, stylish. It's smaller than older A Powershots
+ Good image quality. Expected from Canon,...
+ Fast and responsive. I was surprised how responsive...
+ Manual focus. Although the LCD resolution is not good enough...
+ Battery performance seems improved over previous...

...

Now the negatives.
I hope Canon can improve these aspects in their future models.
- Slow flash recycle times. This is to be expected...
- Higher resolution LCD. The 2.5" 115k resolution LCD is...
-Live histogram while shooting. This feature is...
- A wider wideangle would be nice ...

So overall it's a great camera for the price.
Offering great image quality with the auto settings for the casual user, but also a powerful set of features for the more advanced photographer.

(a)

Canon Powershot SX110IS 9MP Digital Camera with 10x Optical Image Stabilized Zoom (Black)

17 of 17 people found the following review helpful: The mostcamera you can get for this price.,November 11, 2008

Canon PowerShot A720IS 8MP Digital Camera with 6x Optical Image Stabilized Zoom

403 o f409 people found the following review helpful: Great value, powerful camera, November 9, 2007

Kodak EasyShare Z712 IS 7.1MP Digital Camera with 12x Optical Image Stabilized Zoom

11 of 12 people found the following review helpful: Value superzoom ,November 7, 2007

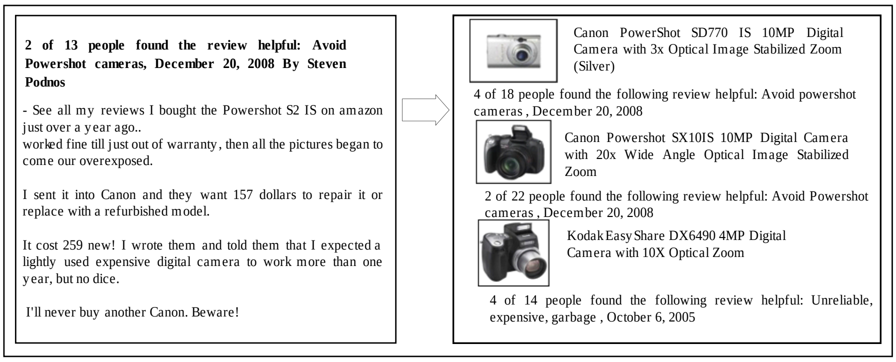
Moto GP 2

25 of 25 people found the following review helpful: A must have for any racing fan - motorcycle or not May 2, 2003

(b)

Fig. 1. A segment of review and reviewer information of a high-quality review.

In Figure 2 (a), the reviewer gives little useful information about the product, but complains of an unsatisfactory experience with the camera. It can be seen, in Figure 2 (b), that the reviews did not receive much peer-to-peer voting, (4/18, 2/22, 4/14). Intuitively, it may be expected that the reviewer is unlikely to publish high-quality reviews.



2 of 13 people found the review helpful: Avoid Powershot cameras, December 20, 2008 By Steven Podnos

- See all my reviews I bought the Powershot S2 IS on amazon just over a year ago..
worked fine till just out of warranty, then all the pictures began to come our overexposed.

I sent it into Canon and they want 157 dollars to repair it or replace with a refurbished model.

It cost 259 new! I wrote them and told them that I expected a lightly used expensive digital camera to work more than one year, but no dice.

I'll never buy another Canon. Beware!

Canon PowerShot SD770 IS 10MP Digital Camera with 3x Optical Image Stabilized Zoom (Silver)

4 of 18 people found the following review helpful: Avoid powershot cameras , December 20, 2008

Canon Powershot SX10IS 10MP Digital Camera with 20x Wide Angle Optical Image Stabilized Zoom

2 of 22 people found the following review helpful: Avoid Powershot cameras , December 20, 2008

Kodak EasyShare DX6490 4MP Digital Camera with 10X Optical Zoom

4 of 14 people found the following review helpful: Unreliable, expensive, garbage , October 6, 2005

Fig. 2. A segment of review and reviewer information of a low-quality review

At present, most e-commerce websites allow users to evaluate the quality of the reviews by assigning helpful votes to them. For example, Amazon.com provides review readers with a mechanism for judging whether a review is helpful or not. The mechanism accumulates helpful votes from a particular review, and the number of helpful votes a review receives indicates its actual effectiveness. For convenience, in this paper, this mechanism is called peer-to-peer voting, which is a good way to assess the quality of reviews. However, the mechanism is not effective in the following cases [5]: (1) newly-written reviews cannot be evaluated immediately, because they need to accumulate peer to peer votes; and (2) low-traffic reviews and reviews with few helpful votes similarly cannot be evaluated. Therefore, it is vital to have the ability to detect low-quality reviews automatically, especially newly-written and low-traffic reviews.

The task of detecting low-quality reviews is presently treated as a binary classification problem [6] or a regression problem [5], [7]. Several methods applying only the intrinsic features to assess review helpfulness have been reported. These approaches are based either on lexical or syntactic features, along with semantic features. Information from high-quality answer findings in the question-answering community has proven to be very helpful in estimating the quality of the answers [8], [9]. In [16], Zhang presented a preliminary analysis of whether author knowledge was a powerful usefulness predictor and drew the conclusion that authorship did seem to be a powerful usefulness predictor. However, how to acquire and exploit reviewer information should be further investigated. Inspired by [8], [9], and [16], it is hypothesized that the quality of reviewer information could improve the performance of a low-quality review detection model. Therefore, one of the focuses of this study is to further demonstrate whether reviewer features have potential for predicting review quality.

The research questions described above can be summarized as follows:

1. Does the quality of the reviewer information matter for building better models to detect low-quality reviews?
2. Which reviewer features are most predictive in low-quality review detection models?
3. How do we acquire and exploit reviewer information to improve the performance of the low-quality reviews detection model?

These questions are answered by conducting an empirical study on two real world datasets, under different experimental conditions. To answer the first question, two kinds of low-quality review detection models were built, one included reviewer information and the other did not, and their detection performances were compared. To answer the second question, various reviewer features were selected as a baseline and then other reviewer features were added into the low-quality review detection models, and then their performance in detecting low-quality reviews was compared. For the third question, the derivation of the reviewer information is explained and the reviewer information is translated into reviewer features in low-quality review detection models.

This paper makes the following contributions to the study of reviewer information in low-quality review detection. First, it is systematically demonstrated that the quality of reviewer information indeed matters when detecting low-quality reviews in some domains. Second, the reviewer features that are most predictive are discovered, using

our low-quality review detection models. Third, the acquisition and exploitation of reviewer information to improve the performance of low-quality review detection models is explained. Fourth, based on empirical experiment results on electronic products and book domains, it is determined that reviewer information is more effective in the electronic product domain than in the book domain.
The rest of the paper is organized as follows: Section 2 contains the literature review. Section 3 presents our approach for the detection of low-quality reviews. In Section 4, the study's evaluation criterion is introduced. In Section 5, the proposed research questions are empirically demonstrated. Section 6 summarizes the work in this paper and calls attention to future work.

2 Literature Review

Interest in sentiment analysis has recently increased as part of a larger research effort in affective computing [17]. Many approaches on sentiment analysis and feature level-based opinion mining have been proposed.

2.1 Sentiment analysis

In the field of sentiment analysis, P. Turney [3] proposed a corpus-based approach, PMI-IR, to determine semantic orientation. Theresa Wilson et al. [18] presented the first experimental results classifying the strength of opinions. B. Pang et al. [4] adopted standard machine learning techniquesto determine whether a review is positive or negative. Moreover, S. Kim et al. [19] proposed a system to determine the Sentiment of Opinions.

2.2 Feature level-based opinion mining

There are some approaches for mining product opinion at product feature levels, based on product reviews, and they are usually classified as unsupervised- and supervised-based methods. Representative works of the unsupervised-based method include [20], [21], and [2]. In [20] and [21], M. Hu and B. Liu's work is performed in three steps: (1) mining the product features and opinions, (2) identifying the opinion orientation, and (3) summarizing the mining results. Popescu et al. [2] proposed a web-based feature extraction method. In their method, each noun phrase is given a pointwise mutual information score between the phrase and part discriminators associated with the product class. The score is computed by the "KnowItAll" system. Qi Su et al. [23] mainly studied the problem of extracting implicit features from customer reviews; they proposed a feature-based pointwise mutual information algorithm. Carenini et al. [22] proposed a more sophisticated method based on several similarity measures. Their system merges each discovered feature to a feature node in the user-defined taxonomy. The similarity measures are defined based on string similarity, while synonyms and other distances are measured using WordNet. Bin Shi and Kuiyu Chang [24] proposed an "opinion first, feature second" approach. They manually built a hierarchical product feature concept model using product domain knowledge, and extracted product features based on the concept model. Ronen Feldman, Moshe Fresko, et al. [25] presented a study in extracting comparison information. Representative works describing supervised-based methods include [26] and [27]. Rayid Ghani, Katharina Probst, et al. [26] viewed the product features extraction problem as a classification problem, using single-view and multi-view semi-supervised learning algorithms. Bo Wang and Houfeng Wang [27] considered the fact that product properties and opinion words usually co-occur with high frequency in product review articles and proposed to bootstrap both of them using cross-training.

2.3 Quality assessment of user-generated contents

When mining opinions from reviews or other user-generated contents, it is important to consider whether or not individual reviews are helpful or useful [17]. In the past few years, there has been an increasing interest in automatically assessing the quality of user-generated contents, including product reviews on e-commerce websites, weblogs, question-answer communities, and web forums.

In the field of assessing review helpfulness and detecting low-quality reviews, a representative work is offered by Kim et al. [5], which considered the task as a ranking problem and solved it with regression models. In their experiments, they adopted an SVM regression model and used structural features, lexical features, syntactic features, semantic features, and meta-data features in the process of regression model training. The peer-to-peer voting information, which was derived from Amazon.com, was used as ground-truth. Based on their experimental results, they found that the most useful features included the length of the review (structural feature), the unigrams of the review (lexical feature), and the product rating of the review (meta-data feature). Zhang and Varadarajan [7] proposed a framework that integrated polarity and the utility of the reviews. Within this framework, they also used regression models to predict the utility of the reviews. More specifically, they adopted simple linear regression and ϵ-support vector regression to rank reviews according to utility. In their experiments they found that shallow syntactic features, such as proper nouns and numbers of modal verbs, account for most predicting power of the regression model. Zhang [16] defined a new task in text sentiment analysis, which adds usefulness scoring to opinion extraction to improve product review ranking services and helps shoppers and vendors leverage information from multiple sources. Ghose and Ipeirotis [11] proposed a review ranking mechanism that combines econometric analysis with text mining techniques, and they found that reviews which include a mixture of subjective and objective elements are considered more helpful by users. In addition, they observed that for feature-based goods, such as electronics, users prefer reviews to contain mainly objective information with a small amount of subjective information. However, for experience goods, such as movies, users prefer personalized, highly-sentimental opinions.

There are some other studies that treated the low-quality review detection problem as a binary classification problem. Liu's work [6] may be the most representative research in this area. In their work, they defined a standard specification to measure the quality of product reviews and proposed several intrinsic review features to train a model. They found that sentence level features, word level features, and product characteristic level features were most effective in their experiments. More importantly, they argued that three types of biases, including imbalance vote bias, winner circle bias, and early bird bias, exist in peer-to-peer voting evaluation standard [5], [7]. Therefore, they hired four annotators to label the reviews manually. In addition, they applied the low-quality review detection approach to enhance opinion summarization and yielded better performance. In other words, the importance of low-quality review detection was validated in their work. The approach used in the present study is different from [5], [6], [7], and [11]. First, our focus is to further demonstrate whether reviewer information has potential for detecting low-quality reviews. In addition, the reviewer features that are most effective in a low-quality review detection model are discovered.

Some approaches for finding high-quality answers in question-answer communities were also proposed. Agichtein et al. [8] introduced a general classification framework for combining the evidence from different sources of information and investigated methods for exploiting intrinsic content quality and community feedback to automatically identify high quality content. Jeon et al. [9] presented a framework to use non-textual features to predict the quality of documents. To the best of our knowledge, no systematic study measuring whether reviewer information matters for building better models predicting review quality has been previously conducted, which is the focus of this paper. Weimer and Gurevych [13] studied the problem of predicting the quality of web forum posts, and they built a system which learns from human ratings by applying SVM classification. Surface, lexical, syntactic, forum specific, and similarity features were used in the learning process. They tested the model on three datasets and found that surface and forum-specific features are more useful.

3 The Low-Quality Review Detection Approach

Review quality evaluation is an interesting problem, which has many potential applications. For example, it can be used as a pre-processing procedure for review ranking algorithms. In essence, the approach of this study is to exploit features that are intuitively correlated with the quality of user- generated contents, and then train a model to mine the relationship between them. Based on the mined knowledge, the quality of user-generated contents can be evaluated.

3.1 Problem definition

As previously discussed, low-quality reviews are reviews that have little or incorrect description of a product, have little or no comments on some aspects of the product, and do not provide convincing opinions with sufficient supporting evidence [6]. In other words, low-quality reviews do not provide enough useful information to users. The core of this research includes two main issues: features learning and model selection. The first issue concerns the features that should be selected to model the quality of the reviews, and the second one concerns the learning algorithms that are effective to model the quality of the reviews. In this paper, it is assumed that there are two kinds of reviews in the review space: high-quality reviews and low-quality reviews. Under this assumption, low-quality review detection is treated as a binary classification problem. Formally, given a training data set of high-quality reviews and low-quality reviews, $T = \{ f_i, Y_j \}, i = 1 \ldots n; j = 1, 2$, statistical machine learning approaches are adopted to learn classification models that can maximize the accuracy in the classification of Y_i given $f_i, i = 1 \ldots n$ where $f_i, i = 1 \ldots n$ represents learning features and $Y_j, j = 1, 2$ stands for high-quality and low-quality, respectively. When a new review comes, the classification model automatically assesses high-quality or low-quality to the review.

3.2 The low-quality review detection model

As previously discussed, the core of the low-quality review detection model is how to identify the features and how to learn the detection models. For the first issue, previous studies have proved that reviewing intrinsic information is important to model review quality; however whether reviewer information is helpful to model review quality is unknown. Product reviews often involve personal experience, knowledge, and interests; therefore, both the intrinsic information in the review and the reviewer information are taken into consideration. For the second issue, three classification algorithms (Adaboost, C4.5, and SVM) are adopted to learn the low-quality review detection models. In particular, reviews and reviewers' information was collected from Amazon.com, and then both review and reviewer features were extracted as learning features. After features extraction, reviews are labelled as high-quality class or low-quality class. Classification algorithms are then adopted to learn the detection models. Finally, the learned detection models are evaluated using a test dataset.

3.3 The learning features

Many previous studies have detected low-quality reviews by using intrinsic review features. One of the focuses in this paper is to further demonstrate whether reviewer information matters for building better models to detect low-quality reviews. If the reviewer information is effective in modeling review quality, which reviewer information is more effective? In the approach here proposed, both review and reviewer features are taken into consideration. User-generated contents are created by millions of end-users; therefore, the quality of the contents is closely correlated with the end-users. "Good" reviewers write "good" reviews. Reviewers' personal experience, knowledge, interests, and reputation are closely related to review quality; therefore, reviewer information can be useful in constructing and optimizing review quality models. In [5], [6], [7], researchers have reported which are the effective review features for detecting low-quality reviews, therefore, these effective review features are adopted as a baseline.

Review features

Three categories of review features are chosen, including surface features, structure features, and shallow syntactic features.

Surface and Structure Features:

F1: The total number of tokens in a syntactic analysis of a review [5].

F2: The number of sentences in a review [6].

F3: The average length of sentences [6].

F4: The number of sentences with product features [6].

F5: The number of products in a review [6].

F6: The number of brand names in a review [6].

F7: The number of product features in a review [6].

F8: The total frequency of product features in a review [6].

F9: The average frequency of product features in a review [6].

F10: The number of paragraphs in a review [6].

F11: The average length of paragraphs in a review [6].

Shallow Syntactic Features:

F12: Proper nouns: reference to existing, maybe technical concepts [7].

F13: Modal verbs: reflection of certainty, confidence, mood, etc., which are all instances of modality [7].

F14: Interjections: signals of emotion [7].

F15: Comparative and superlative adjectives: indicators of comparison [7].

F16: Comparative and superlative adverbs: also indicators of comparison [7].

F17: wh-determiners, wh-pronouns, possessive wh-pronouns, wh-adverbs: wh-words signify either questions or other interesting linguistic constructs, such as relative clauses [7].

Features F1, F2, F3, F5, F6, F10, and F11 can be easily derived from the review itself. Features F4, F7, F8, and F9, cannot be directly obtained. An integration strategy [12] is adopted to mine the product features, and then they are computed. For features F12 to F17, Stanford POS tagger is used; the POS taggers of F12 to F17 are {NNP}, {MD}, {UH}, {JJR}, {JJS}, and {WDT, WP, WP$, WRB}.

Reviewer features

In order to exploit reviewer information, it has to be translated into reviewer features. Eight reviewer features are introduced in this paper.

F18: The total number of reviews the reviewer has written.

Our hypothesis is that if a reviewer has published many reviews, his reviews are likely high-quality.

F19: The sum of helpful votes the reviewer has received.

Our hypothesis is that if a reviewer has received many helpful votes from other people, then his reviews are likely high-quality.

F20: The total votes the reviewer has received.

F21: The average total votes the reviewer has received.

F22: The average helpful votes the reviewer has received.

Note that since the helpful votes and total votes information will be used in Section 4 to label the reviews as high-quality or low-quality, features F19 to F22 do not contain the helpful votes and total votes information of the review that is to be classified.

F23: The reviewer's domain authority.

Our hypothesis is that if a reviewer has published many reviews in a domain, he or she is authoritative in that domain. Under certain conditions, a reviewer may write many reviews that belong to different domains. In this case, different weights are assigned to different domains. For example, a reviewer writes electronic products reviews and he\she also writes book and movie reviews. When the review in the electronic product domain is classified, more weight will be assigned to it.

F24: The rating score the reviewer assesses in a review.

The rating score is from one star to five stars.

F25: The Kullback-Leibler distance between the rating score and the average rating score given by all reviewers.

In order to derive the features F18 to F25, reviewers' information is collected, including total reviews, ranting, helpful votes, and total votes from Amazon.com.

4 Evaluation Criterion

As previously discussed, there are two kinds of evaluation methods: peer-to-peer voting evaluation [5], [7], [16], and manual annotation evaluation [6]. In [5], Kim et al. made use of peer-to-peer voting information to evaluate the quality of reviews and defined a review helpfulness function as:

$$h(r \in R) = \frac{rating_{+}(r)}{rating_{+}(r) + rating_{-}(r)}$$

where $rating_{+}(r)$ is the number of people that will find a review helpful and $rating_{-}(r)$ is the number of people that will not find the review helpful. In [6], Liu et al. argued that there are three kinds of biases in peer-to-peer voting evaluation: imbalance vote bias, winner circle bias, and early bird bias. Therefore, manual annotation evaluation was adopted in this study. The definitions of high-quality reviews and low-quality reviews are as follows [6]:

High-quality review: A review should contain a complete or relatively complete comment on a product and features of the product. Moreover, it should provide convincing opinions with sufficient supporting evidence. It provides practical information to the user.

Low-quality review: A review provides little useful information or gives misleading information. It does not help the user in making a decision.

5 Experiment Setup

In this section, datasets are first introduced, and then the experimental results in answer to the research questions are provided. In our experiment, three popular learning models were adopted: Adaboost, C4.5, and SVM.

5.1 Dataset

Two product types were chosen, namely electronic products and books. We selected 1,756 electronic products and 2,035 books as seeds. Then, the ASIN of each product was transferred to Amazon Web Services API and 72,072 electronic product reviews and 92,212 book reviews were obtained. Using Amazon Web Services API, 41,722 and 44,588 pieces of reviewers' information were obtained in the electronic products and book domains, respectively. Figures 3 and 4 show the distribution characteristics of the reviews and reviewers. Figure 3 shows that a large number of products get very few reviews and a small number of products get a large number of reviews. Therefore, the products with less than 50 reviews were dropped, resulting in 414 electronic products with 41,360 reviews and 477 books with 44,653 reviews. Figure 4 indicates that a large number of reviewers write only a few reviews, and a few reviewers write a large number of reviews. In order to test the effectiveness of reviewer features, the reviewers who only wrote two reviews were removed. Finally, a dataset or 29,645 electronic products reviews and 29,776 book reviews was obtained. Six people were employed to annotate the data, according to the criterion proposed in Section 4. The statistics of the dataset are shown in Table 1.

Table 1. The statistics of the datasets.

Domain	#Product	#Reviews	#High Quality	#Low Quality
Electronic Products	414	29645	11877	17768
Books	477	29776	11604	18172

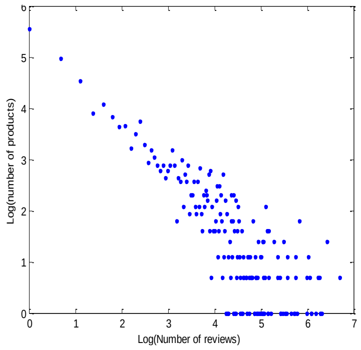


Fig. 3. The log-log plot of the number of reviews to the number of products

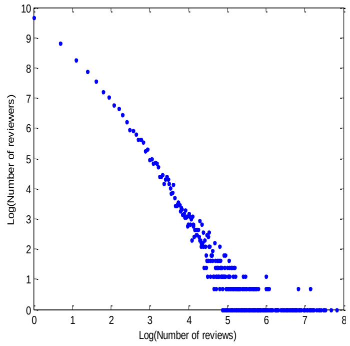


Fig. 4. The log-log plot of the number of reviews to the number of reviewers

5.2 Performance measures

Six performance measures were used in these experiments, including Accuracy, Precision, Recall, F-Measure, RUC, and ROC curves, to evaluate the effects of the reviewer information.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad precision = \frac{TP}{TP+FP} \quad recall = \frac{TP}{TP+FN} \quad F - measure = \frac{2\, precision * recall}{precision + recall}$$

TN is the number of negative examples correctly classified (True Negatives), FP is the number of negative examples incorrectly classified as positive (False Positives), FN is the number of positive examples incorrectly classified as negative (False Negatives) and TP is the number of positive examples correctly classified (True Positives). ROC curves can be thought of as representing the family of best decision boundaries for relative costs of TP and FP [14]. On an ROC curve the X-axis represents False Positive Rate = FP/(TN+FP) and the Y-axis represents True Positive Rate = TP/(TP+FN). The AUC (area under the ROC curve) is a useful metric for classifier performance, as it is independent of the decision criterion selected and prior probabilities [14].

5.3 Does reviewer information matter?

The aim of this study is to experimentally demonstrate whether reviewer information matters when predicting review quality. Therefore, two detection models were constructed and their predictive performances were compared. One model included reviewer information and the other model did not.

Table 2. The experiment results for the electronic product domain using Adaboost

Features	Accuracy	Precision	Recall	F-Measure	AUC
Review-features	0.734	0.755	0.822	0.787	0.80
Reviewer-features	0.782	0.801	0.846	0.823	0.864
All-features	0.838	0.861	0.870	0.866	0.913

Table 3. The experiment results for the electronic product domain using C4.5

Features	Accuracy	Precision	Recall	F-Measure	AUC
Review-features	0.732	0.743	0.845	0.791	0.768
Reviewer-features	0.765	0.791	0.825	0.808	0.831
All-features	0.812	0.842	0.847	0.844	0.858

Table 4. The experiment results for the electronic product domain using SVM

Features	Accuracy	Precision	Recall	F-Measure	AUC
Review-features	0.766	0.774	0.862	0.815	0.742
Reviewer-features	0.790	0.809	0.850	0.829	0.775
All-features	0.838	0.916	0.803	0.856	0.846

Table 2, 3 and 4 show the experiment results for the electronic products domain using Adaboost, C4.5, and SVM.

Table 2, illustrates that reviewer features are more effective than review features on all measures. Comparing review features and reviewer features, the improvements for Accuracy, F-Measure, and AUC are 4.8%, 3.6%, and 6.4%, respectively. When training the learning model using both review features and reviewer features, the performance is more improved than when using review features only. Accuracy, F-Measure, and AUC are improved by 10.4%, 7.9%, and 11.3%, respectively, when using review and reviewer features together. From Table 3, it can also be seen that reviewer features perform better than review features. Comparing review features and reviewer features, the improvements of Accuracy, F-Measure, and AUC are 3.3%, 1.7%, and 6.3%, respectively; when using reviewer features and review features together, Accuracy, F-Measure, and AUC improved by 8.0%, 5.3%, and 9.0%, respectively. In Table 4, comparing review features and reviewer features, the improvements of Accuracy, F-Measure, and AUC are 2.4%, 1.4%, and 3.3%, respectively; when using reviewer features and review features together, Accuracy, F-Measure, and AUC improved by 7.2%, 4.1%, and 10.4%, respectively.

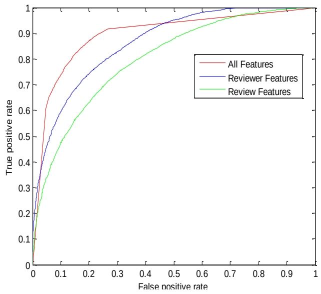


Fig. 5. ROC curves on different features of the electronic product domain using Adaboost

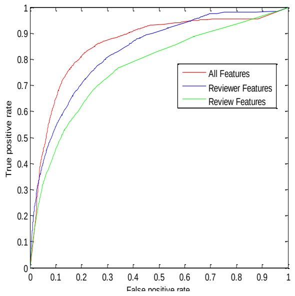


Fig. 6. ROC curves on different features of the electronic product domain using C4.5

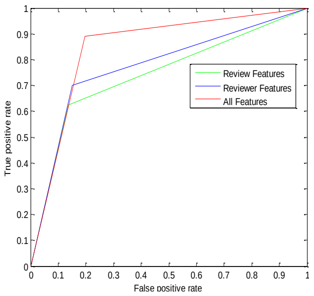


Fig. 7. ROC curves on different features of the electronic product domain using SVM

Figure 5 shows the ROC curves on different features of the electronic product domain using Adaboost. The ROC curve of reviewer features is above the ROC curve of review features, therefore, reviewer features perform better than review features. Used together, reviewer features and review features provide the best detection performance. Figure 6 shows the ROC curves on different features of the electronic product domain using C4.5. From Figure 6, the same conclusion can be drawn. Figure 7 shows the ROC curves on different features of the electronic product domain using SVM. Again, the same conclusion can be drawn. From the above analysis, it can clearly be seen that reviewer information indeed matters when predicting review quality in the electronic product domain.

In order to further demonstrate whether reviewer features can improve low-quality review detection performance in other domains, a comparative trial was conducted in the book domain.

Table 5. The experiment results for the book domain using Adaboost

Features	Accuracy	Precision	Recall	F-Measure	AUC
Review-features	0.657	0.692	0.792	0.738	0.682
Reviewer-features	0.720	0.754	0.804	0.778	0.797
All-features	0.721	0.763	0.789	0.775	0.798

Table 6. The experiment results for the book domain using C4.5

Features	Accuracy	Precision	Recall	F-Measure	AUC
Review-features	0.658	0.694	0.786	0.737	0.632
Reviewer-features	0.733	0.751	0.840	0.793	0.787
All-features	0.747	0.770	0.835	0.801	0.810

Table 7. The experiment results for the book domain using SVM

Features	Accuracy	Precision	Recall	F-Measure	AUC
Review-features	0.639	0.673	0.795	0.729	0.595
Reviewer-features	0.698	0.705	0.868	0.778	0.65
All-features	0.636	0.632	0.965	0.764	0.543

Table 5, 6 and 7 show the experiment results for the book domain using Adaboost, C4.5, SVM. Table 5 shows that reviewer features perform better than review features. Comparing review features and reviewer features, the improvements of Accuracy, F-Measure, and AUC are 6.3%, 4.0%, and 11.5%, respectively. When training the learning model using both review features and reviewer features together, Accuracy, F-Measure, and AUC improved by 6.4%, 3.7%, and 11.6%, respectively. Accuracy and F-Measure improved less than AUC. From Table 6, it can be seen that reviewer features perform better than review features on all measures, and the improvements of Accuracy, F-Measure, and AUC are 7.5%, 5.6%, and 15.5%, respectively. When using reviewer features and review features together, Accuracy, F-Measure, and AUC improved by 8.9%, 6.4%, and 17.8%, respectively. Table 7 shows that reviewer features perform better than review features on all measures, and the improvements of Accuracy, F-Measure, and AUC are 5.9%, 4.9%, and 5.5%, respectively. Surprisingly, when using reviewer and review features together, only recall improved by 17%.

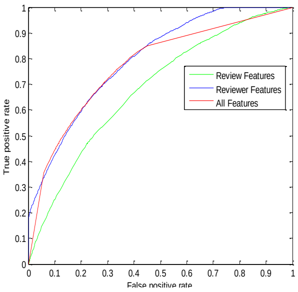


Fig. 8. ROC curves on different features of the book domain using Adaboost

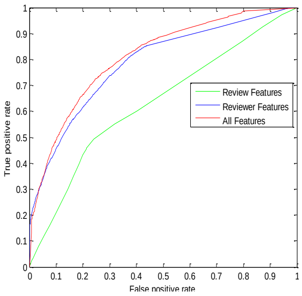


Fig. 9. ROC curves on different features of the book domain using C4.5

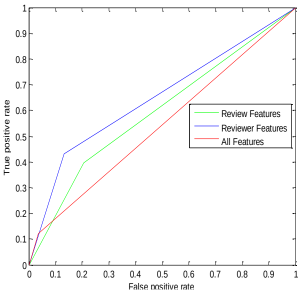


Fig. 10. ROC curves on different features of the book domain using SVM

Figure 8 shows the ROC curves on the different features of the book domain using Adaboost. Figure 8 shows that the ROC curve of reviewer features is above the ROC curve of review features, therefore, reviewer features perform better than review features. Reviewer and review features used together result in the best detection performance. Figure 9 shows the ROC curves on different features of the book domain using C4.5. From Figure 9, the same conclusion can be drawn: reviewer features improve the performance of the detection model. Figure 10 shows the ROC curves on different features of the book domain using SVM, and indicates that reviewer features perform best.

5.4 Which reviewer features are most predictive?

In this experiment, reviewer features F19 and F20 were used as a baseline, and other reviewer features were incrementally added to the training process. The experiment results of the electronic product domain and the book domain, using Adaboost, C4.5, and SVM are shown in Table 8 to Table 13.

Table 8. The results for reviewer features of the electronic product domain using Adaboost

Reviewer Features	Accuracy	Precision	Recall	F-Measure	AUC
F19,20	0.718	0.766	0.763	0.765	0.784
F19,20,21,22	0.748	0.787	0.795	0.791	0.809
F19,20,21,22,18	0.760	0.787	0.821	0.804	0.845
F19,20,21,22,18,23	0.767	0.789	0.835	0.811	0.851
F19,20,21,22,18,23,24	0.780	0.802	0.841	0.821	0.861
F19,20,21,22,18,23,24,25	0.782	0.801	0.846	0.823	0.864

Table 9. The results for reviewer features of the electronic product domain using C4.5

Reviewer Features	Accuracy	Precision	Recall	F-Measure	AUC
F19,20	0.719	0.782	0.736	0.759	0.772
F19,20,21,22	0.749	0.791	0.79	0.79	0.795
F19,20,21,22,18	0.755	0.782	0.82	0.801	0.819
F19,20,21,22,18,23	0.759	0.788	0.818	0.802	0.825
F19,20,21,22,18,23,24	0.764	0.789	0.826	0.807	0.832
F19,20,21,22,18,23,24,25	0.765	0.791	0.825	0.808	0.831

Table 10. The results for reviewer features of the electronic product domain using SVM

Reviewer Features	Accuracy	Precision	Recall	F-Measure	AUC
F19,20	0.734	0.783	0.769	0.776	0.725
F19,20,21,22	0.762	0.787	0.827	0.807	0.746
F19,20,21,22,18	0.778	0.793	0.852	0.822	0.760
F19,20,21,22,18,23	0.781	0.795	0.856	0.824	0.763
F19,20,21,22,18,23,24	0.782	0.797	0.854	0.824	0.764
F19,20,21,22,18,23,24,25	0.790	0.809	0.850	0.829	0.775

Table 11. The results for reviewer features of the book domain using Adaboost

Reviewer Features	Accuracy	Precision	Recall	F-Measure	AUC
F19,20	0.672	0.805	0.611	0.695	0.732
F19,20,21,22	0.725	0.765	0.793	0.779	0.787
F19,20,21,22,18	0.727	0.755	0.816	0.785	0.803
F19,20,21,22,18,23	0.727	0.755	0.816	0.785	0.803

F19,20,21,22,18,23,24	0.721	0.751	0.814	0.781	0.798
F19,20,21,22,18,23,24,25	0.720	0.754	0.804	0.778	0.797

Table 12. The results for reviewer features of the book domain using C4.5

Reviewer Features	Accuracy	Precision	Recall	F-Measure	AUC
F19,20	0.671	0.792	0.625	0.699	0.729
F19,20,21,22	0.725	0.765	0.794	0.779	0.773
F19,20,21,22,18	0.730	0.758	0.819	0.787	0.787
F19,20,21,22,18,23	0.730	0.758	0.819	0.787	0.787
F19,20,21,22,18,23,24	0.733	0.750	0.845	0.794	0.788
F19,20,21,22,18,23,24,25	0.733	0.751	0.840	0.793	0.787

Table 13. The results for reviewer features of the book domain using SVM

Reviewer Features	Accuracy	Precision	Recall	F-Measure	AUC
F19,20	0.659	0.741	0.677	0.707	0.654
F19,20,21,22	0.709	0.748	0.789	0.768	0.686
F19,20,21,22,18	0.701	0.713	0.853	0.777	0.658
F19,20,21,22,18,23	0.701	0.713	0.853	0.777	0.658
F19,20,21,22,18,23,24	0.705	0.721	0.844	0.778	0.666
F19,20,21,22,18,23,24,25	0.698	0.705	0.868	0.778	0.650

Table 8 to Table 13 show that F18, F19, F20, F21, and F22 greatly improve performance, while other reviewer features are less effective. At first, F19 and F20 achieved 78.4% and 73.2%, respectively, on the AUC measure in the electronic product domain and the book domain. When F21 and F22 were added, Accuracy, F-Measure, and AUC improved. F18 also improves performance. Note that when computing feature F19 to F22, the helpful vote and total vote information of the review that is to be classified is not taken into account. Based on the experiment results, the hypothesis that features F19 to F22 reflect the probability that a reviewer writes high-quality reviews is validated; in other words, a reviewer has received many helpful votes, and his reviews are likely in the high-quality classification.

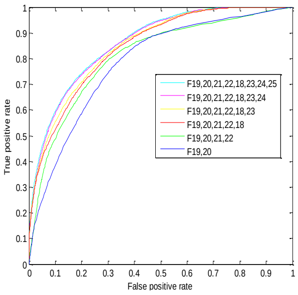


Fig. 11. ROC curves on different reviewer features in the electronic products domain using Adaboost

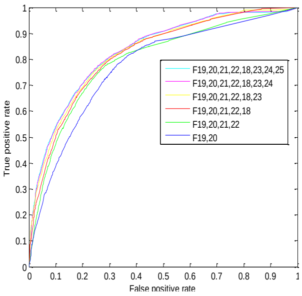


Fig. 12. ROC curves on different reviewer features in the electronic products domain using C4.5

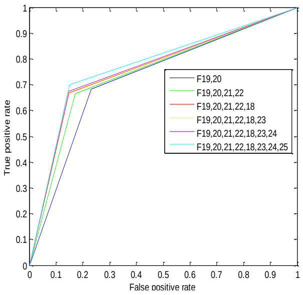


Fig. 13. ROC curves on different reviewer features in the electronic products domain using SVM

Figures 11 to 16 show the ROC curves on different reviewer features using Adaboost, C4.5, and SVM in the electronic product domain and the book domain. It can be seen that features F19 and F20 achieved a relatively high AUC, and the AUC measure increased when features F21 and F22 were added. Reviewer feature F18 also improved the AUC measures, which validates the hypothesis that when a reviewer has published many reviews, his reviews are likely to be in the high-quality classification. Surprisingly, AUC is not improved by feature F23 (reviewer's domain authority) as much as by F19 to F22.

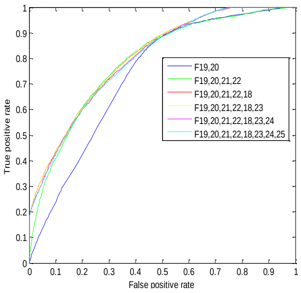


Fig. 14. ROC curves on different reviewer features in the book domain using Adaboost

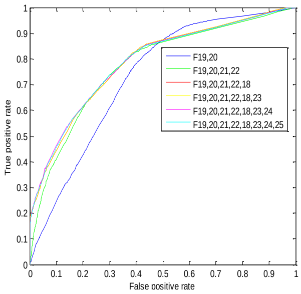


Fig. 15. ROC curves on different reviewer features in the book domain using C4.5

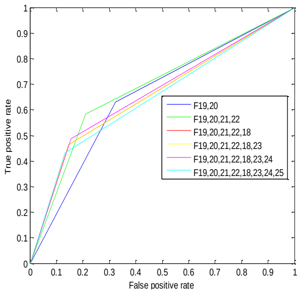


Fig. 16. ROC curves on different reviewer features in the book domain using SVM

5.5 Effectiveness analysis in different domains

From the experiment results in Section 5.3, it can be seen that reviewer information indeed matters when predicting review quality in low-quality review detection models. However, reviewer features are more predictive in the electronic product domain than in the book domain. In the electronic product domain, Accuracy, F-Measure, and AUC achieved 83.8%, 86.6%, and 91.3%, respectively, while in the book domain, the best results for Accuracy, F-Measure, and AUC were 74.7%, 80.1%, and 81.0%, respectively. Through analyzing the reviewers' information in the electronic product domain and the book domain, two phenomena were discovered. First, reviewers are likely to write only a few reviews on the electronic product domain, while reviewers in the book domain usually submit many book reviews. The reason for this phenomenon is that consumers usually do not have to purchase many electronic products in the course of a lifetime, while books are usually purchased frequently. Second, reviews in the electronic product domain usually receive more peer-to-peer scoring than reviews on books. The reason for this phenomenon might be that when people are going to purchase electronic products, they often read the reviews for reference and assess the reviews as helpful or not while they are at the review site. However, when people are in the market for books, they seldom read reviews, therefore, book reviews gain less peer-to-peer scoring than electronic product reviews. Based on the conclusion that F21 (the average total votes the reviewer has received) and F22 (the average helpful votes the reviewer has received) are the most effective features; we propose the hypothesis that reviewer information is more effective when the reviews provide a great deal of peer-to-peer information.

6. Conclusion

In this paper, three research questions were considered: 1) Does reviewer information matter for building better models to detect low-quality reviews?, 2) Which reviewer features are most predictive in the detection models?, and 3) How do we acquire and exploit reviewer information to improve low-quality review detection performance? Besides the intrinsic features of the review, some reviewer features are proposed in the low-quality review detection model. Experiment results on two real world datasets show promising results, from which the following conclusions are derived: First, reviewer information indeed matters when detecting low-quality reviews. Moreover, greater improvement can be achieved when simultaneously using both review features and reviewer features. Secondly, the average helpful votes and the average total votes a reviewer gains from peer-to-peer voting are the most effective features in the reviewer features set. Through effective analysis in different domains, we propose the hypothesis that reviewer information is more effective when the reviewer's reviews provide a great deal of peer-to-peer information. In future work, we plan to apply the proposed model to other user-generated contents.

References

1. Bing Liu, Minqing Hu, Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web". In Proceedings of the 14th international World Wide Web conference (WWW-2005), May 10-14, 2005, in Chiba, Japan, pp: 342-351.
2. A.-M. Popescu, O. Etzioni, "Extracting Product Features and Opinions from Reviews". In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP'05), 2005, pp: 339-346.
3. Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the Meeting of the Association for Computational Linguistics (ACL'02), pp: 417-424.
4. Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP'02), 2002, pp: 79-86.
5. Soo-Min Kim, Patrick Pantel, Tim Chklovski, Marco Pennacchiotti, "Automatically Assessing Review Helpfulness". In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP'06), 2006, pp: 423-430.
6. Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, Ming Zhou, "Low-Quality Product Review Detection in Opinion Summarization". In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP'07), 2007, pp: 334-342.
7. Zhu Zhang, Balaji Varadarajan, "Utility Scoring of Product Reviews". In Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management (CIKM'06), 2006, pp: 51-57.

8. Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, Gilad Mishne, "Finding High-Quality Content in Social Media". In Proceedings of the First ACM International Conference on Web Search and Data Mining (WSDM'08), 2008, pp: 183-194.
9. Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, Soyeon Park, "A Framework to Predict the Quality of Answers with NonTextual Features". In Proceedings of 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR'06), 2006, pp: 228-235.
10. Yun Zhou, W. Bruce Croft, "Document Quality Models for Web Ad Hoc Retrieval". In Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management (CIKM'05), 2005, pp: 331-332.
11. Anindya Ghose, Panagiotis G. Ipeirotis, "Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews". In Proceedings of the ninth international conference on Electronic Commerce (ICEC'07), 2007, pp: 303-310.
12. Qingliang Miao, Qiudan Li, Ruwei Dai. "An Integration Strategy for Mining Product Features and Opinions". In Proceedings of the 2008 ACM CIKM International Conference on Information and Knowledge Management (CIKM'08), 2008, pp: 1369-1370.
13. Markus Weimer, Iryna Gurevych, "Predicting the Perceived Quality of Web Forum Posts". In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07), 2007, pp: 643-648.
14. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research 16 (2002), pp. 321-357.
15. Philip S. Yu, Xin Li, Bing Liu. "Adding the Temporal Dimension to Search - A Case Study in Publication Search". In Proceedings of the 2005 IEEE/WIC/ACM Conferences on Web Intelligence (WI'05), 2005, pp. 543-549.
16. Zhu Zhang. "Weighing Stars: Aggregating Online Product Reviews for Intelligent E-commerce Applications". Intelligent Systems, IEEE, Sept.-Oct. 2008, Volume: 23, Issue: 5, pp. 42-49.
17. Bo Pang and Lillian Lee. "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1–2 (2008) 1–135, DOI: 10.1561/1500000001.
18. Theresa Wilson, Janyce Wiebe, and Rebecca Hwa (2004). Just how mad are you? Finding strong and weak opinion clauses. In Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-2004), pp.761-769.
19. S. Kim and E. Hovy. Determining the Sentiment of Opinions. In Proceedings of the Intl. Conf. on Computational Linguistics (COLING'04), 2004, pp. 1367-1373.
20. M. Hu and B. Liu. Mining Opinion Features in Customer Reviews. In Proc. of the 19th National Conf. on Artificial Intelligence (AAAI'04), 2004, pp. 755-760.
21. M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In Proc. of ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD04), 2004, pp. 168-177.
22. Giuseppe Carenini, Raymond T.Ng and Ed Zwart. Extracting Knowledge from Evaluative Text. In Proceedings of third International Conference on Knowledge Capture (K-CAP 05), 2005, pp. 11-18.
23. Qi Su, Kun Xiang et al. Using Pointwise Mutual Information to Identify Implicit Features in Customer Reviews. In Processings of the 21st International Conference on the Computer Processing of Oriental Languages (ICCPOL 2006), Volume 4285, 2006.
24. Bin Shi and Kuiyu Chang. Mining Chinese Reviews. In Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM 2006), 2006, pp. 585-589.
25. Ronen Feldman, Moshe Fresko et al. Extracting Product Comparisons from Discussion Boards. In Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM 2007), 2007, pp. 469-474.

26. Rayid Ghani and Katharina Probst et al. Text Mining for Product Attribute Extraction. In ACM SIGKDD Explorations Newsletter Volume 8, Issue 1, 2006, pp. 41-48.
27. Bo Wang and Houfeng Wang. Bootstrapping both Product Properties and Opinion Words from Chinese Reviews with Cross-Training. In Processings of 2007 IEEE/WIC/ACM International Conference on Web Intelligence, 2007, pp. 259-262.