Low-Quality Product Review Detection in Opinion Summarization

Jingjing Liu
Nankai University
Tianjin, China
v-jingil@microsoft.com

Yunbo Cao
Microsoft Research Asia
Beijing, China
yucao@microsoft.com

Chin-Yew Lin
Microsoft Research Asia
Beijing, China
cyl@microsoft.com

Yalou Huang
Nankai University
Tianjin, China
huangyl@nankai.edu.cn

Ming Zhou
Microsoft Research Asia
Beijing, China
mingzhou@microsoft.com

Abstract

Product reviews posted at online shopping sites vary greatly in quality. This paper addresses the problem of detecting low-quality product reviews. Three types of biases in the existing evaluation standard of product reviews are discovered. To assess the quality of product reviews, a set of specifications for judging the quality of reviews is first defined. A classification-based approach is proposed to detect the low-quality reviews. We apply the proposed approach to enhance opinion summarization in a two-stage framework. Experimental results show that the proposed approach effectively (1) discriminates low-quality reviews from high-quality ones and (2) enhances the task of opinion summarization by detecting and filtering low-quality reviews.

1 Introduction

In the past few years, there has been an increasing interest in mining opinions from product reviews (Pang, et al, 2002; Liu, et al, 2004; Popescu and Etzioni, 2005). However, due to the lack of editorial and quality control, reviews on products vary greatly in quality. Thus, it is crucial to have a mechanism capable of assessing the quality of reviews and detecting low-quality/noisy reviews.

Some shopping sites already provide a function of assessing the quality of reviews. For example, Amazon[1] allows users to vote for the helpfulness of each review and then ranks the reviews based on the accumulated votes. However, according to our survey in Section 3, users' votes at Amazon have three kinds of biases as follows: (1) *imbalance vote bias,* (2) *winner circle bias, and* (3) *early bird bias.* Existing studies (Kim et al, 2006; Zhang and Varadarajan, 2006) used these users' votes for training ranking models to assess the quality of reviews, which therefore are subject to these biases.

In this paper, we demonstrate the aforementioned biases and define a standard specification to measure the quality of product reviews. We then manually annotate a set of ground-truth with real world product review data conforming to the specification.

To automatically detect low-quality product reviews, we propose a classification-based approach learned from the annotated ground-truth. The proposed approach explores three aspects of product reviews, namely informativeness, readability, and subjectiveness.

We apply the proposed approach to opinion summarization, a typical opinion mining task. The proposed approach enhances the existing work in a two-stage framework, where the low-quality review detection is applied right before the summarization stage.

Experimental results show that the proposed approach can discriminate low-quality reviews from high-quality ones effectively. In addition, the task of opinion summarization can be enhanced by detecting and filtering low-quality reviews.

[1] http://www.amazon.com

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 334–342, Prague, June 2007.

The rest of the paper is organized as follows: Section 2 introduces the related work. In Section 3, we define the quality of product reviews. In Section 4, we present our approach to detecting low-quality reviews. In Section 5, we empirically verify the effectiveness of the proposed approach and its use for opinion summarization. Section 6 summarizes our work in this paper and points out the future work.

2 Related Work

2.1 Evaluating Helpfulness of Reviews

The problem of *evaluating helpfulness of reviews* (Kim et al, 2006), also known as *learning utility of reviews* (Zhang and Varadarajan, 2006), is quite similar to our problem of *assessing the quality of reviews*.

In practice, researchers in this area considered the problem as a ranking problem and solved it with regression models. In the process of model training and testing, they used the ground-truth derived from users' votes of helpfulness provided by Amazon. As we will show later in Section 3, these models all suffered from three types of voting bias.

In our work, we avoid using users' votes by developing a specification on the quality of reviews and building a ground-truth according to the specification.

2.2 Mining Opinions from Reviews

One area of research on opinion mining from product reviews is to judge whether a review expresses a positive or a negative opinion. For example, Turney (2006) presented a simple unsupervised learning algorithm in judging reviews as "thumbs up" (recommended) or "thumbs down" (not recommended). Pang et al (2002) considered the same problem and presented a set of supervised machine learning approaches to it. For other work see also Dave et al. (2003), Pang and Lee (2004, 2005).

Another area of research on opinion mining is to extract and summarize users' opinions from product reviews (Hu and Liu, 2004; Liu et al., 2005; Popescu and Etzioni, 2005). Typically, a sentence or a text segment in the reviews is treated as the basic unit. The polarity of users' sentiments on a product feature in each unit is extracted. Then the aggregation of the polarities of individual sentiments is presented to users so that they can have an at-a-glance view on how other experienced users rated on a certain product. The major weakness in the existing studies is that all the reviews, including low-quality ones, are taken into consideration and treated equally for generating the summary. In this paper, we enhance the application by detecting and filtering low-quality reviews. In order to achieve that, we first define what the quality of reviews is.

3 Quality of Product Reviews

In this section, we will first show three biases of users' votes observed on Amazon, and then present our specification on the *quality of product reviews*.

3.1 Amazon Ground-truth

In our study, we use the product reviews on digital cameras crawled from Amazon as our data set. The data set consists of 23,141 reviews on 946 digital cameras. At the Amazon site, users could vote for a review with a "helpful" or "unhelpful" label. Thus, for each review there are two numbers indicating the statistics of these two labels, namely the number of "helpful" votes and that of "unhelpful" ones. Kim et al (2006) used the percentage of "helpful" votes as the measure of evaluating the "quality of reviews" in their experiments. We call the ground-truth based on this measure as "Amazon ground-truth".

Certainly, the ground-truth has the advantage of convenience. However, we identify three types of biases that make the Amazon ground-truth not always suitable for determining the quality of reviews. We describe these biases in details in the rest of this section.

3.1.1 Imbalance Vote Bias

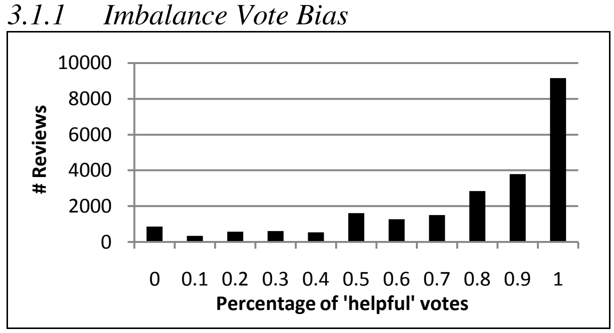

Figure 1. Reviews' percentage scores

At the Amazon site, users tend to value others' opinions positively rather than negatively. From Figure 1, we can see that a half of the 23,141 reviews (corresponding to the two bars on the right of the figure) have more than 90% "helpful" votes, including 9,100 reviews with 100% "helpful" votes. From an in-depth investigation on these highly-voted reviews, we observed that some did not really have as good quality as the votes hint. For example, in Figure 2, the review about *Canon PowerShot S500* receives 40 "helpful" votes out of 40 votes although it only gives very brief description on the product features in its second paragraph. We call this type of bias "imbalance vote" bias.

This is my second Canon digital elph camera. Both were great cameras. Recently upgraded to the S500. About 6 months later I get the dreaded E18 error. I searched the Internet and found numerous people having problems. When I determined the problem to be the lens not fully extending I decided to give it a tug. It clicked and the camera came on, ready to take pictures. Turning it off and on produced the E18 again. While turning it on I gave it a nice little bump on the side (where the USB connector is) and the lens popped out on its own. No problems since.

It's a nice compact and light camera and takes great photos and videos. Only complaint (other than E18) is the limit of 30-second videos on 640x480 mode. I've got a 512MB compact flash card, I should be able to take as much footage as I have memory in one take.

Figure 2. An example review

3.1.2 Winner Circle Bias

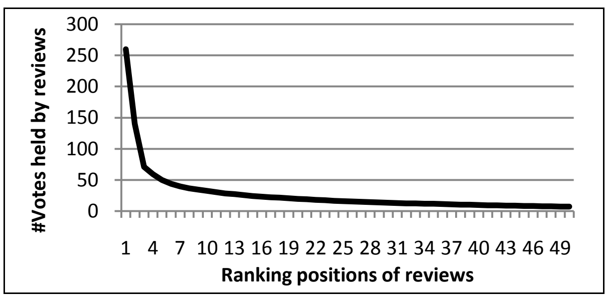

Figure 3. Votes of the top-50 ranked reviews

There also exists a bootstrapping effect of "hot" reviews at the Amazon site. Figure 3 shows the "helpful" votes for the top 50 ranked reviews. The numbers are averaged over 127 digital cameras which have no less than 50 reviews. As shown in this figure, the top two reviews hold more than 250 and 140 votes respectively on average; while the numbers of votes held by lower-ranked reviews decrease exponentially. This is so-called the "winner circle" bias: the more votes a review gains, the more default authority it would appear to the readers, which in turn will influence the objectivity of the readers' votes. Also, the higher ranked reviews would attract more eyeballs and therefore gain more people's votes. This mutual influence among labelers should be avoided when the votes are used as the evaluation standard.

3.1.3 Early Bird Bias

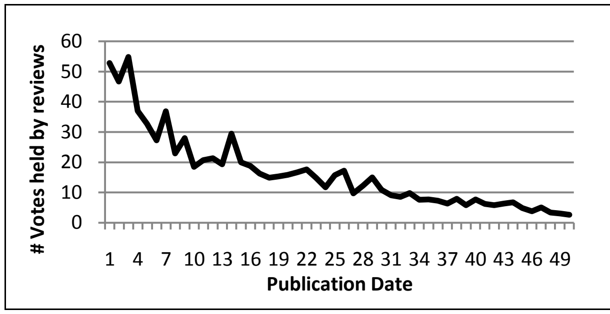

Figure 4. Dependency on publication date

Publication date can influence the accumulation of users' votes. In Figure 4, the *n*'*th* publication date represents the *n*'*th* month after the product is released. The number in the figure is averaged over all the digital cameras in the data set. We can observe a clear trend that the earlier a review is posted, the more votes it will get. This is simply because reviews posted earlier are exposed to users for a longer time. Therefore, some high quality reviews may get fewer users' vote because of later publication. We call this "early bird" bias.

3.2 Specification of Quality

Besides these aforementioned biases, using the raw rating from readers directly also fails to provide a clear guideline for what a good review consists of. In this section, we provide such a guideline, which we name as the specification (SPEC).

In the SPEC, we define four categories of review quality which represent different values of the reviews to users' purchase decision: "best review", "good review", "fair review", and "bad review". A generic description of the SPEC is as follows:

A *best* review must be a rather complete and detailed comment on a product. It presents several aspects of a product and provides convincing opinions with enough evidence. Usually a best review could be taken as the main reference that users only need to read before making their purchase decision on a certain product. The first review in Figure 5 is a *best* review. It presents several product features and provides convincing opinions with sufficient evidence. It is also in a good format for readers to easily understand. Note that we omit some words in the example to save the space.

A *good* review is a relatively complete comment on a product, but not with as much supporting evidence as necessary. It could be used as a strong and influential reference, but not as the only recommendation. The second review in Figure 5 is such an example.

A *fair* review contains a very brief description on a product. It does not supply detailed evaluation on the product, but only comments on some aspects of the product. For example, the third review in Figure 5 mainly talks about "*the delay between pictures*", but less about other aspects of the camera.

A *bad* review is usually an incorrect description of a product with misleading information. It talks little about a specific product but much about some general topics (e.g. photography). For example, the last review in Figure 5 talks about the topic of "*generic battery*", but does not specify any digital camera. A bad review is an "unhelpful" review that can be ignored.

Best Review:

I purchased this camera about six months ago after my Kodak Easyshare camera completely died on me. I did a little research and read only good things about this Canon camera so I decided to go with it because it was very reasonably priced (about $200). Not only did the camera live up to my expectations, it surpassed them by leaps and bounds! Here are the things I have loved about this camera:

BATTERY - this camera has the best battery of any digital camera I have ever owned or used. ...

EASY TO USE - I was able to ...

PICTURE QUALITY - all of the pictures I've taken and printed out have been great. ...

FEATURES - I love the ability to quickly and easily ...

LCD SCREEN - I was hoping ...

SD MEMORY CARD - I was also looking for a camera that used SD memory cards. Mostly because...

I cannot stress how highly I recommend this camera. I will never buy another digital camera besides Canon again. And the A610 (as well as the A620 - the 7.0MP version) is the best digital camera I've ever used.

Good Review:

The Sony DSC "P10" Digital Camera is the top pick for CSC. Running against cameras like Olympus stylus, Canon Powereshot, Sony V1, Nikon, Fuji, and More. The new release of 5.0 mega pixels has shot prices for digital cameras up to $1000+. This camera I purchased through a Private Dealer cost me $400.86. The Retail Price is Running $499.00 to $599.00. Purchase this camera from a wholesale dealer for the best price $377.00. Great Photo Even in dim light w/o a flash. The p10 is very compact. Can easily fit into any pocket. The camera can record 90 minutes of mpeg like a home movie. There are a lot of great digital cameras on the market that shoot good pictures and video. What makes the p10 the top pick is it comes with a rechargeable lithium battery. Many use AA batteries, the digital camera consumes theses AA batteries in about two hours time while the unit is on. That can add continuous expense to the camera. It's also the best resolution on the market. 6.0 megapix is out, though only a few. And the smallest that we found. Also the best price for a major brand.

Fair Review:

There is nothing wrong with the 2100 except for the very noticeable delay between pics. The camera's digital processor takes about 5 seconds after a photo is snapped to ready itself for the next one. Otherwise, the optics, the 3X optical zoom and the 2 megapixel resolution are fine for anything from Internet apps to 8" x 10" print enlarging. It is competent, not spectacular, but it gets the job done at an agreeable price point.

Bad Review:

I want to point out that you should never buy a generic battery, like the person from San Diego who reviewed the S410 on May 15, 2004, was recommending. Yes you'd save money, but there have been many reports of generic batteries exploding when charged for too long. And don't think if your generic battery explodes you can sue somebody and win millions. These batteries are made in sweatshops in China, India and Korea, and I doubt you can find anybody to sue. So play it safe, both for your own sake and the camera's sake. If you want a spare, get a real Canon one.

Figure 5. Example reviews

3.3 Annotation of Quality

According to the SPEC defined above, we built a ground-truth from the Amazon data set. We randomly selected 100 digital cameras and 50 reviews for each camera. Totally we have 4,909 reviews since some digital cameras have fewer than 50 unique reviews. Then we hired two annotators to label the reviews with the SPEC as their guideline. As the result, we have two independent copies of annotations on 4,909 reviews, with the labels of "best", "good", "fair", and "bad". Table 1 shows the confusion matrix between the two copies of annotation. The value of the *kappa* statistic (Cohen, 1960) calculated from the matrix is 0.8142. This shows that the two annotators achieved highly consistent results by following the SPEC, although they worked independently.

Annotation 1	Annotation 2				
	best	*good*	*fair*	*bad*	total
best	294	44	2	0	340
good	66	639	113	0	818
fair	0	200	1,472	113	1,785
bad	1	2	78	1,885	1,966
total	361	885	1,665	1,998	4,909

Table 1. Confusion matrix bet. the annotations

In order to examine the difference between our annotations and Amazon ground-truth, we evaluate the Amazon ground-truth against the annotations, with the measure of "error rate of preference pairs" (Herbrich et al, 1999).

$$ErrorRate = \frac{|incorrect\ preference\ pairs\ |}{|all\ preference\ pairs|} \quad (1)$$

where the "*preference pair*" is defined as a pair of reviews with a order. For example, a *best* review and a *good* review correspond to a preference pair with the order of "*best* review preferring to *good* review". The "*all preference pairs*" are collected from one of the annotations (the annotation 1 or the annotation 2) by ignoring the pairs from the same category. The "*incorrect preference pairs*" are the *preference pair*s collected from the Amazon ground-truth but not with the same order as that in the *all preference pairs*. The order of the *preference pair* collected from the Amazon ground-truth is evaluated on the basis of the *percentage* score as described in Section 3.1.

The error rate of preference pairs based on the annotation 1 and that based on the annotation 2 are 0.448 and 0.446, respectively, averaged over 100 digital cameras. The high error rate of preference pairs demonstrates that the Amazon ground-truth diverges from the annotations (our ground-truth) significantly.

To discover which kind of ground-truth is more reasonable, we ask an additional annotator (the third annotator) to compare these two kinds of ground-truth. More specifically, we randomly selected 100 preference pairs whose orders the two kinds of ground-truth don't agree on (called incorrect preference pairs in the evaluation above). As for our ground-truth, we choose the Annotation 1 in the new test. Then, the third annotator is asked to assign a preference order for each selected pair. Note that the third annotator is blind to both our specification and the existing preference order. Last, we evaluate the two kinds of ground-truth with the new annotation. Among 100 pairs, our ground-truth agrees to the new annotation on 85 pairs while the Amazon ground-truth agrees to the new annotation on 15 pairs. To confirm the result, yet another annotator (the fourth annotator) is called to repeat the same annotation independently as the third one. And we obtain the same statistical result (85 vs. 15) although the fourth annotator does not agree with the third annotator on some pairs.

In practice, we treat the reviews in the first three categories ("*best*", "*good*" and "*fair*") as high-quality reviews and those in the "*bad*" category as low-quality reviews, since our goal is to identify low quality reviews that should not be considered when creating product review summaries.

4 Classification of Product Reviews

We employ a statistical machine learning approach to address the problem of detecting low-quality products reviews.

Given a training data set $D = \{x_i, y_i\}_1^n$, we construct a model that can minimize the error in prediction of *y* given *x* (generalization error). Here $x_i \in X$ and $y_i = \{high\ quality\ , low\ quality\}$ represents a product review and a label, respectively. When applied to a new instance *x*, the model predicts the corresponding *y* and outputs the score of the prediction.

4.1 The Learning Model

In our study, we focus on differentiating low-quality product reviews from high-quality ones. Thus, we treat the task as a binary classification problem.

We employ SVM (Support Vector Machines) (Vapnik, 1995) as the model of classification. Given an instance *x* (product review), SVM assigns a score to it based on

$$f(x) = w^T x + b \quad (2)$$

where *w* denotes a vector of weights and *b* denotes an intercept. The higher the value of *f(x)* is, the higher the quality of the instance *x* is. In classification, the sign of *f(x)* is used. If it is positive, then *x* is classified into the positive category (high-quality reviews), otherwise into the negative category (low-quality reviews).

The construction of SVM needs labeled training data (in our case, the categories are "high-quality reviews" and "low-quality reviews"). Briefly, the learning algorithm creates the "hyper plane" in (2), such that the hyper plane separates the positive and negative instances in the training data with the largest "margin".

4.2 Product Feature Resolution

Product features (e.g., "image quality" for digital camera) in a review are good indicators of review quality. However, different product features may refer to the same meaning (e.g., "*battery life*" and "*power*"), which will bring redundancy in the study. In this paper, we formulize the problem as the "resolution of product features". Thus, the problem is reduced to how to determine the equivalence of a product feature in different forms.

In (Hu and Liu, 2004), the matching of different product features is mentioned briefly and addressed by fuzzy matching. However, there exist many cases where the method fails to match the multiple mentions, e.g., "*battery life*" and "*power*", because it only considers string similarity. In this paper we propose to resolve the problem by leveraging two kinds of evidence: one is "surface string" evidence, the other is "contextual evidence".

We use *edit distance* (Ukkonen, 1985) to compare the similarity between the surface strings of two mentions, and use *contextual similarity* to reflect the semantic similarity between two mentions.

When using contextual similarity, we split all the reviews into sentences. For each mention of a product feature, we take it as a query and search for all the relevant sentences. Then we construct a vector for the mention, by taking each unique term in the relevant sentences as a dimension of the vector. The cosine similarity between two vectors of mentions is then present to measure the contextual similarity between two mentions.

4.3 Feature Development for Learning

To detect low-quality reviews, our proposed approach explores three aspects of product reviews, namely informativeness, subjectiveness, and readability. We denote the features employed for learning as "learning features", discriminative from the "product features" we discussed above.

4.3.1 Features on Informativeness

As for informativeness, the resolution of product features is employed when we generate the learning features as listed below. Pairs mapping to the same product feature will be treated as the same product feature, when we calculate the frequency and the number of product features. We apply the approach proposed in (Hu and Liu, 2004) to extract product features.

We also use a list of product names and a list of brand names to generate the learning features. Both lists can be collected from the Amazon site because they are relatively stable within a time interval.

The learning features on the informativeness of a review are as follows.

➢ **Sentence level (SL)**
- The number of sentences in the review
- The average length of sentences
- The number of sentences with product features

➢ **Word level (WL)**
- The number of words in the review
- The number of products (e.g., DMC-FZ50, EX-Z1000) in the review
- The number of products in the title of a review
- The number of brand names (e.g., Canon, Sony) in the review
- The number of brand names in the title of a review

➢ **Product feature level (PFL)**
- The number of product features in the review
- The total frequency of product features in the review
- The average frequency of product features in the review
- The number of product features in the title of a review
- The total frequency of product features in the title of a review

4.3.2 Features on Readability

We make use of several features at paragraph level which indicate the underlying structure of the reviews. These features include,
- The number of paragraphs in the review
- The average length of paragraphs in the review
- The number of paragraph separators in the review

Here, we refer to the keywords, such as "Pros" vs. "Cons" as "paragraph separators". The keywords usually appear at the beginning of paragraphs for categorizing two contrasting aspects of a product. We extract the nouns and noun phrases at the beginning of each paragraph from the 4,909 reviews and use the most frequent 30 pairs of keywords as paragraph separators. Table 2 provides some examples of the extracted separators.

Separators		***Separators***	
Positive	***Negative***	***Positive***	***Negative***
Pros	Cons	The Good	The Bad
Strength	Weakness	Thumb up	Bummer
PLUSES	MINUSES	Positive	Negative
Advantages	Drawbacks	Likes	Dislikes
The upsides	Downsides	GOOD THINGS	BAD THINGS

Table 2. Examples of paragraph separators

4.3.3 Features on Subjectiveness

We also take the subjectiveness of reviews into consideration. Unlike previous work (Kim et al, 2006; Zhang and Varadarajan, 2006) using shallow syntactic information directly, we use a sentiment analysis tool (Hu and Liu, 2004) which aggregates a set of shallow syntactic information. The tool is a classifier capable of determining the sentiment polarity of each sentence. We create three learning features regarding the subjectiveness of reviews.

- The percentage of positive sentences in the review
- The percentage of negative sentences in the review
- The percentage of subjective sentences (regardless of positive or negative) in the review

5 Experiments

In this section, we describe our experiments with the proposed classification-based approach to low-quality review detection, and its effectiveness on the task of opinion summarization.

5.1 Detecting Low-quality Reviews

In our proposed approach, the problem of assessing quality of reviews is formalized as a binary classification problem. We conduct experiments by taking reviews in the categories of "best", "good", and "fair" as high-quality reviews and those in the "bad" category as low-quality reviews.

As for classification model, we utilize the SVMLight toolkit (Joachims, 2004). We randomly divide the 100 queries of digital cameras into two sets, namely a training set of 50 queries and a test set of 50 queries. For the two copies of annotations, we use the same division. We use the training set from "annotation 1" to train the model and apply the model to the test sets from both "annotation 1" and "annotation 2", respectively. Table 3 reports the accuracies of our approach to review classification. The accuracy is defined as the percentage of correctly classified reviews.

We take the approach that utilizes only the category of features on sentence level (SL) as the baseline, and incrementally add other categories of features on informativeness, readability and subjectiveness. We can see that both the features on word level (WL) and those on product feature level (PFL) can improve the performance of classification much. The features on readability can still increase the accuracy although the contribution is much less. The features on subjectiveness, however, make no contribution.

Feature Category		Annotation1	Annotation2
Informative-ness	SL	73.59%	72.81%
	WL	80.41%	79.15%
	PFL	83.30%	82.37%
Readability		83.93%	82.91%
Subjectiveness		83.84%	82.96%

Table 3. Low-quality reviews detection

We also conduct a more detailed analysis on each individual feature. Two categories of features on "title" and "brand name" have poor performance, which is due to the lack of information in the title and the low coverage of brand names in a review, respectively.

5.2 Summarizing Sentiments of Reviews

One potential application of low-quality review detection is the opinion summarization of reviews.

The process of opinion summarization of reviews with regards to a query of a product consists of the following steps (Liu et al, 2005):

1. From each of the reviews, identify every text segment with opinion in the review, and determine the polarities of the opinion segments.
2. For each product feature, generate a positive opinion set and a negative opinion set of opinion segments, denoted as $\mathrm{POS}(f)$ and $\mathrm{NOS}(f)$.
3. For each product feature, aggregate the numbers of segments in $\mathrm{POS}(f)$ and $\mathrm{NOS}(f)$, as opinion summarization on the product feature.

In this process, all the reviews contribute the same. However, different reviews do hold different authorities. A positive/negative opinion from a high-quality review should not have the same weight as that from a low-quality review.

We use a two-stage approach to enhance the reliability of summarization. That is, we add a process of low-quality review detection before the summarization process, so that the summarization result is obtained based on the high-quality reviews only. We are to demonstrate how much difference the proposed two-stage approach can bring into the opinion summarization.

We use the best classification model trained as described in Section 5.1 to filter low-quality reviews, and do summarization on the high-quality reviews associated to the 50 test queries. We denote the proposed approach and the old approach as "two-stage" and "one-stage", respectively. Due to the limited space, we only give a visual comparison of the two approaches on "image quality" in Figure 6. The upper figure shows the summarization of positive opinions and the lower figure shows that of negative opinions. From the figures we can see that the two-stage approach preserves fewer text segments as the result of filtering out many low-quality product reviews.

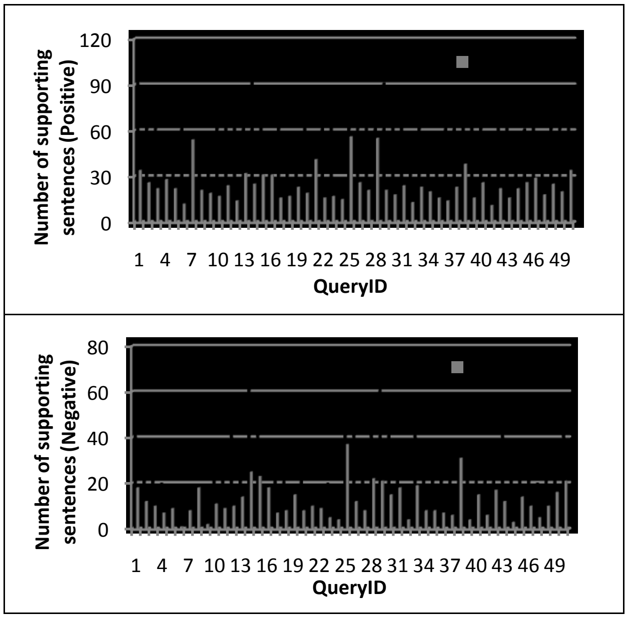

Figure 6. Summarization on "image quality"

To show the comparison on more features in a compressed space, we give the statistic ratio of change between two approaches instead. As for the evaluation measure, we define "*RatioOfChange*" (*ROC)* on a feature *f* as,

$$\mathrm{ROC}(f) = \frac{\mathrm{Rate}_{\mathrm{one-stage}}(f) - \mathrm{Rate}_{\mathrm{two-stage}}(f)}{\mathrm{Rate}_{\mathrm{one-stage}}(f)} \quad (3)$$

where Rate $_*(f)$ is defined as,

$$\mathrm{Rate}_*(f) = \frac{|\mathrm{POS}(f)|}{|\mathrm{POS}(f)| + |\mathrm{NOS}(f)|} \quad (4)$$

Table 4 shows some statistic results on *ROC* on five product features, namely "image quality"(IQ), "battery", "LCD screen" (LCD), "flash" and "movie mode" (MM). The values in the cells are the percentage of queries whose *ROC* is larger/smaller than the respective thresholds. We can see that a large portion of queries have big changes on the values of *ROC*. This means that the result achieved by the two-stage approach is substantially different from that achieved by the one-stage approach.

%Query	***RatioOfChange*** **(+)**					
	>0.30	**>0.25**	**>0.20**	**>0.15**	**>0.10**	**>0.05**
IQ	2%	4%	4%	10%	14%	22%
Battery	10%	14%	18%	30%	38%	50%
LCD	12%	18%	20%	22%	24%	28%
Flash	6%	10%	16%	20%	26%	42%
MM	6%	8%	8%	12%	18%	26%
%Query	***RatioOfChange*** **(-)**					
	<-0.30	**<-0.25**	**<-0.20**	**<-0.15**	**<-0.10**	**<-0.05**
IQ	4%	6%	10%	14%	18%	44%
Battery	2%	4%	4%	10%	14%	22%
LCD	4%	4%	8%	12%	22%	28%
Flash	4%	6%	8%	16%	18%	28%
MM	8%	10%	16%	18%	34%	42%

Table 4. *RatioOfChange* on five features

There is no standard way to evaluate the quality of opinion summarization as it is rather a subjective problem. In order to demonstrate the impact of the two-stage approach, we turn to external authoritative sources other than Amazon.com as the objective evaluation reference. We observe that CNET[2] provides a professional "*editor's review*" for many products, which gives a rating in the range of 1~10 on product features. 9 digital cameras out of the 50 test queries are found to have the editor's rating on "image quality" at CNET. We use this rating to compare with the results of our opinion summarization. We rescale the *Rate* scores obtained by both the one-stage approach and the two-stage approach into the range of 1-10 in order to perform the comparison.

Figure 7 provides the visual comparison. We can see that the result achieved by the two-stage approach has a much better (closer) resemblance to CNET rating than one-stage approach does. This indicates that our two-stage approach can achieve a more consistent summarization result to the professional evaluations by the editors. Although the CNET rating is not the absolute standard for product evaluation, it provides a professional yet objective evaluation of the products. Therefore, the experimental results demonstrate that our proposed approach could achieve more reliable opinion summarization which is closer to the generic evaluation from authoritative sources.

[2] http://www.cnet.com

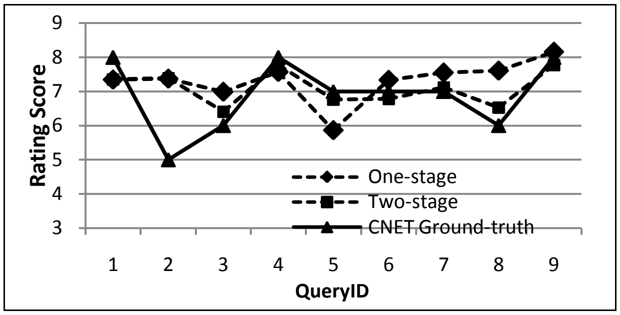

Figure 7. Comparison with CNET rating

6 Conclusion

In this paper, we studied the problem of detecting low-quality product reviews. Our contribution can be summarized in two-fold: (1) we discovered **three types of biases** in the ground-truth used extensively in the existing work, and proposed a specification on the quality of product reviews. The three biases that we discovered are *imbalance vote bias*, *winner circle bias*, and *early bird bias*. (2) Rooting on the new ground-truth (conforming to the proposed specification), we proposed a classification-based approach to low-quality product review detection, which yields better performance of **opinion summarization**.

We hope to explore our future work in several areas, such as further consolidating the new ground-truth from different points of view and verifying the effectiveness of low-quality review detection with other applications.

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* 20: 37–46.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *WWW'03*.

Harris Drucker, Chris J.C., Burges Linda Kaufman, Alex Smola and Vladimir Vapnik. 1997. Support vector regression machines. *Advances in Neural Information Processing Systems.*

Christiane Fellbaum. 1998. WordNet: an Electronic Lexical Database, MIT Press.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 1999. Support Vector Learning for Ordinal Regression. *In Proc. of the 9th International Conference on Artificial Neural Networks.*

Minqing Hu and Bing Liu. 2004a. Mining and Summarizing Customer Reviews. *KDD'04*.

Minqing Hu and Bing Liu. 2004b. Mining Opinion Features in Customer Reviews. *AAAI'04*.

Kalervo Jarvelin & Jaana Kekalainen. 2000. IR: evaluation methods for retrieving highly relevant documents. *SIGIR'00*.

Nitin Jindal and Bing Liu. 2006. Identifying Comparative Sentences in Text Documents. *SIGIR'06*.

Nitin Jindal and Bing Liu. 2006. Mining comparative sentences and relations. *AAAI'06*.

Thorsten Joachims. SVMlight -- Support Vector Machine. http://svmlight.joachims.org/, 2004.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, Marco Pennacchiotti. 2006. Automatically Assessing Review Helpfulness. *EMNLP'06*.

Dekang Lin. 1998, Automatic retrieval and clustering of similar words. *COLING-ACL'98.*

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. *WWW '0*5.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *ACL'04*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *ACL'0*5.

Bo Pang and Lillian Lee, and S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *EMNLP'02*.

Ana-Maria Popescu and O Etzioni. 2005. Extracting product features and opinions from reviews. *HLT-EMNLP'05.*

Peter D. Turney. 2001. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *ACL'02*

Esko Ukkonen. 1985. Algorithms for approximate string matching. *Information and Control*, pp. 100 – 118.

Vladimir N. Vapnik. 1995. The Nature of Statistical Learning Theory. Springer.

Zhu Zhang and Balaji Varadarajan. 2006. Utility Scoring of Product Reviews. CIKM'06