

Assessing Quality of Product Reviews

Yunbo Cao¹

Microsoft Research Asia,
5F Sigma Center, Zhichun Road, Haidian District, Beijing, China. 100080
yunbo.cao@microsoft.com

1 Introduction

In the past few years, there has been an increasing interest in mining opinions from product reviews [3][4][5]. 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 and noisy reviews.

Some shopping sites already provide a function of assessing the quality of reviews. For example, Amazon² allows users to vote for the helpfulness of each review and then ranks the reviews based on accumulated votes. However, according to our survey, users' votes at Amazon have three kinds of biases as follow:

- 1) Imbalance vote bias -- users tend to value others' opinions positively rather than negatively. As the result, most reviews are considered as high-quality ones;
- 2) Winner circle bias -- the more votes a review gains, the more default authority it would appear to readers, which in turn will influence the objectivity of the readers' votes. As the result, a few reviews receive most of users' votes;
- 3) Early bird bias -- the earlier a review is posted, the more votes it will get. Therefore, some high-quality reviews may get fewer users' vote because of later publication.

Existing studies [2][6] used these users' votes for training ranking models to assess the quality of reviews, which therefore are subject to these biases.

In our research, we identify 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.

2 Specification on Quality of Product Reviews

Besides these aforementioned biases, using review readers' rating directly also fail 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 the different values of reviews to users' purchase decision: 'best review', 'good review', 'fair review', and 'bad review'. A generic description of the SPEC is as follows:

¹ Joint work with Jingjing Liu and Chin-Yew Lin.

² <http://www.amazon.com>

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 purchasing decision on certain product.

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.

A fair review contains a very brief description on a product. It doesn't supply detailed evaluation on the product, but only comments some aspects of the product.

A bad review is usually an incorrect description of a product with misleading information. It talks little about a specific product but much on some general topics (e.g. photography).

3 Our Annotation

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.

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. In total, we have 4,910 reviews because some digital cameras have fewer than 50 reviews. Then we hired two annotators to label the reviews with SPEC as their guideline. As the result, we have two independent copies of annotations on 4,910 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 [1] calculated from the matrix is 0.8082. This shows that the two annotators achieved highly consistent results by following the SPEC, although they worked independently.

Table 1. Confusion matrix between the annotations

Annotation 1	Annotation 2				total
	<i>best</i>	<i>good</i>	<i>fair</i>	<i>bad</i>	
<i>best</i>	294	44	4	0	342
<i>good</i>	67	639	116	0	822
<i>fair</i>	1	200	1,450	115	1,766
<i>bad</i>	1	2	89	1,888	1,980
total	363	885	1,659	2,003	4,910

4 Conclusion

In this research, we studied the standard for assessing quality of 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; (2) we developed a new ground-truth by proposing a specification on quality of product reviews.

References

1. Cohen, J.: A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement* 20: 37–46, 1960.
2. Kim, S., Pantel, P., Chklovski, T. and Pennacchiotti, M.: Automatically Assessing Review Helpfulness. *EMNLP'06*.
3. Liu, B., Hu, M., and Cheng, J.: Opinion Observer: Analyzing and Comparing Opinions on the Web. *WWW '05*.
4. Pang, B., Lee, L., and Vaithyanathan S.: Thumbs Up? Sentiment Classification Using Machine Learning Techniques. *EMNLP'02*.
5. Popescu, A. and Etzioni, O.: Extracting Product Features and Opinions from Reviews. *HLT-EMNLP'05*.
6. Zhang, Z. and Varadarajan, B.: Utility Scoring of Product Reviews. *CIKM'06*