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