Mining Millions of Reviews: A Technique to Rank Products Based on Importance of Reviews

Kunpeng Zhang, Yu Cheng, Wei-keng Liao, Alok Choudhary
{kzh980,ych133,wkliao,choudhar}@eecs.northwestern.edu
EECS department, Northwestern University
2145 Sheridan road, Evanston IL, 60208.

Abstract

As online shopping becomes increasingly more popular, many shopping web sites encourage existing customers to add reviews of products purchased. These reviews make an impact on the purchasing decisions of potential customers. At Amazon.com for instance, some products receive hundreds of reviews. It is overwhelming and time restrictive for most customers to read, comprehend and make decisions based on all of these reviews. Customers most likely end up reading only a small fraction of the reviews usually in the order which they are presented on the product page. Incorporating various product review factors, such as: content related to product quality, time of the review, content related to product durability and historically older positive customer reviews will have different impacts on the products rankings. Thus, the automated mining of product reviews and opinions to produce a re-calculated product ranking score is a valuable tool which would allow potential customers to make more informed decisions. In this paper, we present a product ranking model that applies weights to product review factors to calculate a products ranking score. Our experiments use the customer reviews from Amazon.com as input to our product ranking model which produces product ranking results that closely relate to the products sales ranking as reported by the retailer.

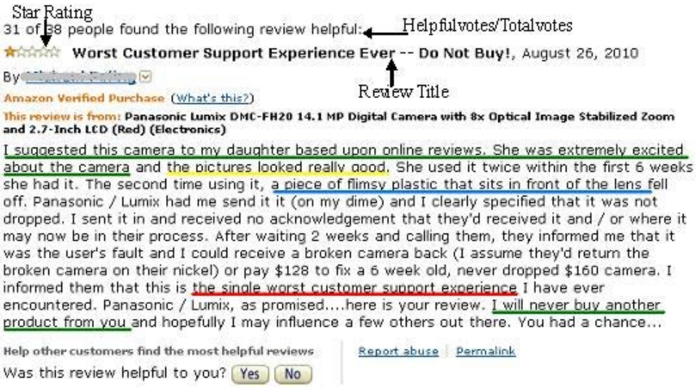
Introduction

The rapid growth in volume of product reviews for online shopping web sites drives us to analyze and mine the data in these reviews to help potential customers make informed purchase decisions. It is almost impossible for a customer to read all reviews. For instance, there are 66 SLR cameras and 85 TVs on Amazon.com. These SLR cameras and TVs each have more than 100 reviews. Some of the popular models(e.g. "Canon Digital Rebel XSi 12.2 MP Digital SLR Camera with EF-S 18-55mm/3.5-5.6 IS Lens(Black)") have more than 700 reviews. We observed that the average number of reviews for SLR cameras and TVs is 15.24 and 10.79, respectively. The average review length of products in these two categories is approximately 11 sentences. One of the challenges in analyzing these reviews is that the reviews contain complicated opinions on the quality of products, quality of customer services related to the sale and seller credibility. In this paper, we propose a computational model to mine data from these reviews that will construct a justified ranking system to help future customers make better-informed decisions.

In addition to the opinions about the product's features, reviews often include comments unrelated to the product itself. Distinguishing the content focus of these sentences is an important component in the analysis of the reviews. Figure 1 shows a typical digital camera review extracted from Amazon.com. The first sentence shows that the camera is recommended based on previous reviews. The second sentence expresses a positive opinion on the camera's quality. The sentence underscored in blue shows a negative opinion on the lens. The review also complains about the customer service of the seller. Although the review title expresses a strong negative opinion to the customer support of the seller, a potential buyer of the product cannot make a conclusion about the quality of the product from this review. Sentences/comments unrelated to product quality, such as customer service should be filtered out when measuring the product quality, otherwise it leads to a biased ranking system. In this paper, we build a filtering mechanism to preprocess review sentences/comments by using Support Vector Machine (SVM) algorithm, a very common machine learning technique (Vapnik 1995). In our experiments, this filtering method produces a better precision rate and recall rate.

In our model, we pay special attention to the following properties: the review's credibility and posting date. We consider both of these properties crucial in measuring the importance of a review. For instance, reviews receiving more helpful votes should be weighted more heavily in the product's rankings.

Most on-line retailers use common helpful voting mechanisms. For example, at Amazon.com, each review contains two numerical values: the number of helpful votes and the number of total votes. These two numbers together are a good measure of a reviewer's credibility. On the other hand, a review posted more recently should be more



31 of 38 people found the following review helpful:

★☆☆☆☆ **Worst Customer Support Experience Ever -- Do Not Buy!**, August 26, 2010

By [redacted]

Amazon Verified Purchase (What's this?)

This review is from: **Panasonic Lumix DMC-FH20 14.1 MP Digital Camera with 8x Optical Image Stabilized Zoom and 2.7-Inch LCD (Red) (Electronics)**

I suggested this camera to my daughter based upon online reviews. She was extremely excited about the camera and the pictures looked really good. She used it twice within the first 6 weeks she had it. The second time using it, a piece of flimsy plastic that sits in front of the lens fell off. Panasonic / Lumix had me send it it (on my dime) and I clearly specified that it was not dropped. I sent it in and received no acknowledgement that they'd received it and / or where it may now be in their process. After waiting 2 weeks and calling them, they informed me that it was the user's fault and I could receive a broken camera back (I assume they'd return the broken camera on their nickel) or pay $128 to fix a 6 week old, never dropped $160 camera. I informed them that this is the single worst customer support experience I have ever encountered. Panasonic / Lumix, as promised....here is your review. I will never buy another product from you and hopefully I may influence a few others out there. You had a chance...

Help other customers find the most helpful reviews | Report abuse | Permalink

Was this review helpful to you? Yes No

Figure 1: A review example of digital camera from Amazon.com

valuable when it refers to earlier reviews. Such an instance can be implicitly seen from the example in Figure 1. Another important reason to incorporate the posting date as a factor is to compare the volume of reviews received across different product release dates. The significance of reviewing posting date will be explained in later sections. In our ranking model, we adopt a weighted scheme to take into account the review's credibility and posting date.

The star rating is another popular measure for product evaluation. As shown in Figure 1, the review is also given a score in the number of stars between 1 and 5. The average number of stars received from all reviews of a product is usually shown at the beginning of the product page next to the product name. We argue that the star rating average can be biased. Figure 2 shows the standard deviations of star rating for two product categories under TV and SLR camera. The fluctuated lines and high deviations tell us that the star rating is not reliable because each reviewer has a different grading standard. A low star rating score from a tough reviewer may still mean a good product. In addition, the average star rating score for a product with very few reviews is not statistically significant. For example, 94 out of 191 TVs in the price range of \$800 to \$1000 contain only 1 review. Furthermore, as observed on Amazon.com, a large number of products share the same star rating scores, rendering such a rating system meaningless.

The reminder of this paper is organized as followed: the related work section, the methodology section which explains our methodology and describes the impact of different factors on the ranking scores, the results section which contains the results and their analysis, the conclusion and the future work section.

Related Work

Recently, there has been a wide range of research done on customer reviews, from studying the quality of reviews to mining reviews for product evaluation. The most closely related work on product ranking is (Zhang, Narayanan, and Choudhary 2010), which proposes a feature-based product ranking system. In this system, review sentences are classified into 4 different categories: positive subjective, negative subjective, positive comparative, and negative comparative. A directed and weighted feature graph is then built by using statistics of all review sentences. The method employs a keyword strategy to identify feature sentences and the evaluation is carried out by a pRank algorithm using Amazon.com data. Their experiment results show that their rankings correspond well with rankings performed by domain experts. In (Lu, Zhai, and Sundaresan 2009), the authors study the problem of generating a "rated aspect summary" of short comments(reviews), which is a decomposed view of the overall ratings for the major aspects described in the comment. As a result, the user can gain different perspectives towards the target entity. They proposed a topic modeling method, called Structured PLSA (Hofmann 1999), modeling the dependency structure of phrases in short comments to extract major aspects. In addition, they use two unsupervised approaches (local prediction and global prediction) to predict ratings for each aspect from the overall ratings. In (Baccianella, Esuli, and Sebastiani 2009), they presented a system for automatically rating product reviews that independently rates many distinct aspects("facets") of the product based on their textual content. These systems consider all reviews with equal weights to the product rankings.

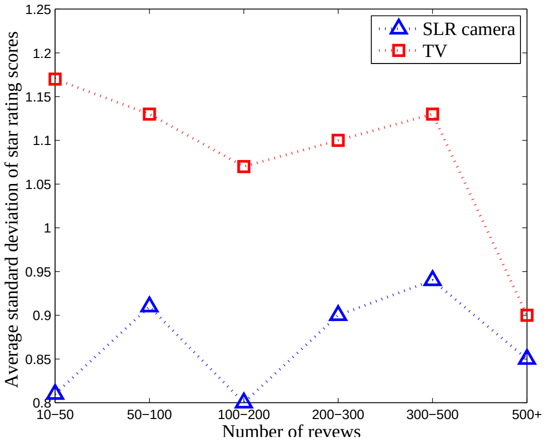


Figure 2: Standard deviation of star rating scores for products under categories of SLR camera and TV. Along X axis, the products are binned into 6 ranges based on the number of reviews they received.

The quality of reviews can have a significant impact on purchase decisions for future customers. Determining the quality of a review has been studied by many researchers (Liu et al. 2008; Danescu-Niculescu-Mizil et al. 2009; Ghose and Ipeirotis. 2007; Kim et al. 2006; Zhang and Trani 2008). In (Liu et al. 2008), a detailed analysis of the major factors affecting the helpfulness of a review is given and a nonlinear model based on radial basis functions for helpfulness prediction is proposed. In (Danescu-Niculescu-Mizil et al. 2009), authors develop a framework for analyzing and modeling opinion evaluation, using a large-scale collection of Amazon book reviews as a dataset. They found that the perceived helpfulness of a review depends not just on its content but also in subtle ways on how the expressed evaluation relates to other evaluations of the same product. Their analysis also allows them to distinguish among the predictions of competing theories from sociology and social psychology, and to discover unexpected differences in the collective opinion-evaluation behavior of user populations from different countries. In (Ghose and Ipeirotis 2007), they proposed two ranking mechanisms for ranking product reviews: a consumer-oriented ranking mechanism ranks the reviews according to their expected helpfulness, and a manufacturer-oriented ranking mechanism ranks the reviews according to their expected effect on sales. Their experiment results show that subjectivity analysis can give useful clues about the helpfulness of a review and about its impact on sales. RevRank (Tsur and Rappoport 2009) uses feature selection techniques to construct a "virtual core review" to represent the review space for finding a set of the most helpful reviews. Another related work is (McGlohon, Glance, and Reiter 2010), in which statistic- and heuristic-based models are explored for estimating the true quality of a product by aggregating reviews from multiple vendor web sites. The performance of these estimators is compared for ranking pairs of products. Our proposed model considers not only review qualities but also two weighted factors (posting date and helpfulness votes).

Other research focuses on sentiment analysis and review summarizations. In (Hu and Liu 2004), they aim to summarize all the customer reviews of a product by mining the features of the product on which the customers have expressed their opinions and whether the opinions are positive or negative. In (Pang and Lee 2004; Pang, Lee, and Vaithyanathan 2002), authors employed machine learning techniques to classify documents by overall sentiments. In addition, some researchers put their focus on feature sentence identification (Nobata, Sekine, and Isahara 2003).

Methodology

Most existing product ranking systems discussed in the related work section consider all review data equally weighted. We argue that reviews should be categoried and weighted to calculate the product ranking scores when compared with other similar reviews. We proposed a model consisting of three stages to enhance the review reliability to the product evaluation. The first stage filters out the sentences that contain comments which are unrelated to the product quality. The second stage derives weights for a review based on its helpfulness votes and age, i.e. since the posting date. The third stage calculates the product's overall ranking score. In our ranking system, the ranking score is determined by the *review contents, relevance of a review to the product quality, helpful votes and total votes from posterior customers, posting date and durability of reviews*.

Table 1: The ratios of irrelevant sentences detected.

Product Categories	Recall Rate	Precision rate
SLR Camera	89.53	78.46
TV	91.22	82.86

Filtering Mechanism

A relevant sentence is either an overall or feature-based comment on a product. It evaluates at least one aspect of a product and provides convincing opinions. The most commonly seen irrelevant review sentence is about the customer service of the seller. In our study, we consider the task of differentiating the irrelevant sentences a binary classification problem. We use a Supporting Vector Machine (SVM) (Vapnik 1995), a well-known supervised machine learning algorithm, to train a hypothesis function, h. Each review sentence becomes trained data in the form of $\{sentence, h(sentence)\}$ pair. Given a review sentence, we first construct its feature vector $\mathbf{X}$, which is fed to the SVM to generate a relevance score based on the linear regression model:

$$h(\mathbf{X}) = \beta^T \mathbf{X} + b \tag{1}$$

where β is the coefficient vector of weights and b is the intercept. The training set contains 1000 sentences collected manually. The value of $h(\mathbf{X})$ indicates the probability that the sentence is relevant or not. The details of extracting feature vectors are given in the next section.

Three types of features are used to the classification task:

- *Brand-level (PL) feature* includes the product brand names, e.g. Nikon or Canon, and the model names (e.g. 550d or D90). This feature counts the number of lexical matches in each sentence.
- *Semantic-level (SL) feature* are the subjective and objective words (positive or negative) describing products.
- *Product-level (FL) feature* is the number of product specification attributes, such as camera pixels and lens, mentioned in the sentences and the number of words related to the customer services, such as shipping and customer support.

We use 10-fold cross validation on the training set and the precision and recall rates are shown in Table 1. Recall rate is the ratio of the number of irrelevant sentences detected by our model to the total number of irrelevant ones filtered manually. Precision rate is the ratio of the number of irrelevant sentences detected by our model to the total number of sentences.

Helpfulness Vote

The perceived importance of a review depends on not only its quality but also the helpfulness votes cast by the posterior customers. The helpfulness of a review is determined by

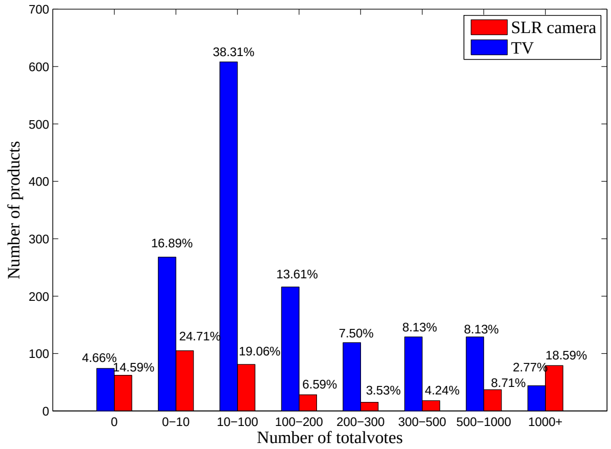


Figure 3: Product distribution based on the number of total votes per review. X-axis is split into 8 bins, representing the number of total votes received by a review. The percentages indicate the distribution within the product categories, i.e. SRL camera or TV.

the total number of helpful votes and the number of total votes. From the Amazon data, we observe that almost every review has helpful votes (X out of Y people found the following review helpful).

However, as indicated in (Zhang and Varadarajan 2006) and (Liu et al. 2007), the votes can be biased and using the votes solely to assess the usefulness of a review is subject to these biases. Figure 3 shows the distributions of the products based on the number of total votes received. The chart also tells us the following: a surprisingly large number of reviews receive less than 10 votes, 21.5% and 39.3% in SLR camera and TV, respectively. Most reviews receive 10-200 votes, 25.65% and 52.22% for SLR camera and TV, respectively. Some reviews have more than 200 votes. To ensure the robustness of our model using such voting information, we assign higher weights to the reviews with more votes. Reviews with less than 10 votes will be ignored. The function calculating the helpfulness weights is as follows:

$$H(r,p) = \begin{cases} 0 & \text{if } Y(r,p) < 10 \\ \frac{H_v(r,p)}{T_v(r,p)} & \text{if } 10 \le Y(r,p) \le 200 \\ \frac{H_v(r,p)}{T_v(r,p)} \cdot \delta & \text{if } Y(r,p) > 200 \end{cases} \tag{2}$$

where r is a review of product p. $H_v(r,p)$ is the number of helpful votes cast on the review (r,p), $T_v(r,p)$ is the total number of total votes cast on review (r,p), and δ ($\delta > 1$) is the gaining factor for the reviews that receive more than 200 votes.

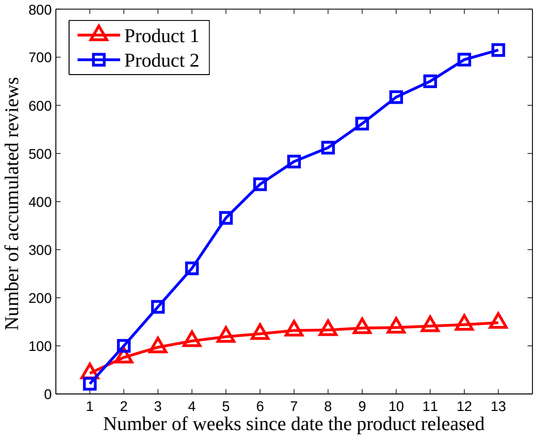


Figure 4: Number of reviews keeps increasing along with time moving forward. Product 1 and product 2 have different release dates. We use number of weeks for X-axis scale.

Age of Review and Durability

In our model, reviews posted more recently receive higher weights in assessing their importance. We assume newer reviews are likely written by customers who read some of the reviews posted earlier. These reviews shall receive higher credits because they are more influential to potential customers. Without adding weights to the newer reviews, they would contribute less to the ranking score, as they are "young" and likely receive less votes. Another important consideration is newer releases/versions of the same products. Correspondingly, the number of reviews for a product version released earlier is likely higher than the product version released recently. In order to balance the contributions to the ranking scores among the similar products and minimize the effects from large volumes gaps, we reduce the importance of older reviews and increase the weight for newer reviews. Figure 4 shows the growing trends of the review numbers during a 13-week period since the product release. We use the following exponential equation 3 to model the age importance of a review.

$$T(r,p) = e^{\beta(t_r - t_0)+d} \tag{3}$$

where $T(r,p)$ is the estimated weight, t_0 is the product p release date, t_r is the published date of review r, β controls the decay rate of $T(r,p)$, and d is an initializing factor. Note that β and d have different values when calculating the age weights for products from different categories.

Sentence Splitter and Part-Of-Speech Tagging

A customer review typically consists of several sentences. It is not uncommon to see multiple positive and negative opinions of a product in a single review. For example, a review of a digital camera may use a few sentences to praise the picture quality and others to criticize the weight and color of the camera. It is not easy to determine the sentiment orientation of such a review as a whole. To simplify this problem, we split reviews into sentences. The sentences are then assigned positive or negative sentiments. In this study, we do not consider sentences expressing both positive and negative sentiments. We use MXTERMINATOR (Jeffrey and Ratnaparkhi 1997) to split reviews into sentences.

Most sentiment bearing words are adjectives. These adjectives determine if a sentence is subjective, and whether it expresses positive or negative sentiment or both. In order to help us identify the sentiment orientation of a sentence, we use the part-of-speech information. At first, CRFTagger (Ratnaparkhi 1996), a java-based conditional random field part-of-speech(POS) tagger for English is employed to label each word. Then, each sentence is saved along with the POS tag information. An example of a review sentence for the digital camera domain after part-of-speech tagging is given below.

> It/PRP 's/VBZ very/RB easy/JJ to/TO learn/VB
> and/CC very/RB light/JJ weight/NN too/RB

[1]

Score Function

A product ranking score is calculated based on the review contents by incorporating two factors: helpfulness vote and review age. The core of our model is a sentiment analysis engine for each relevant sentence. We only consider positive and negative sentiments in this work. Although some researchers suggest a supervised machine learning algorithm to infer sentiment orientation, we use a very simple yet powerful method to accomplish this task.

To form the positive and negative word sets for our methodology, we manually pick a set of very common adjectives/adverbs as a seed list. These word sets are added to the collection of sentiment oriented words from Mpqa citecorpus. Finally, the word sets are augmented with synonyms and antonyms extracted by using a WordNet (WordNet 2010) search. The resulting positive and negative word sets contain 1974 positive words and 4605 negative words. When the positive and negative word sets are compare to Mpqa citecorpus, they are very similar. A simple rule to identify a positive sentiment sentence is to see if a sentence contains a adjective/adverb word (identified by using POS) from the positive word set. Negative sentiment sentences are handled similarly. However, a sentence may have a negative qualifier (eg. this is not a good camera). In this case, the orientation is reversed. We have manually constructed a set of 42 commonly used negation words, such as not, don't, hasn't, never, etc. With this approach, we achieved a precision rate of 82% and a recall rate of 80%.

For reviews containing more than one relevant sentence, we label each with a positive, negative, or neutral tag. The difference between the number of positive and negative sentences determines the polarity of a review. The weights derived from the helpfulness vote and review age are then applied to the ranking score calculation. The ranking score of a product is the sum of all weighted scores of individual reviews.

The following equation computes the ranking score S of product p.

$$S(p) = \frac{\sum_{all\ r} Polarity(r,p)\cdot T(r,p)\cdot H(r,p)}{\sum_{all\ r} H(r,p)\cdot\sum_{all\ r} T(r,p)} \qquad (4)$$

where $Polarity(r,p) = Pos(r,p) - Neg(r,p)$. $Pos(r,p)$ and $Neg(r,p)$ are the numbers of positive and negative sentences in review r of product p, respectively.

Experiment Results

The data used in our experiments is collected from Amazon.com. This data includes the information about the products and their reviews. In particular, we use data from two very popular electronic categories, SLR camera and TV. The rankings are calculated only on products in the same categories and within the same price ranges. We select products with price range from \$500 to \$700 for both categories. Table 2 describes the statistical data about the reviews used in our experiments.

Evaluation and Analysis

The use of customer reviews for product ranking is still a subjective problem, there has not been a commonly recognized method for validating a ranking system. For each product, Amazon.com provides "sales rank", a number indicating how well the product is sold in its category. We consider this a fair indicator to justify a ranking system. In our evaluation, we use two measures to quantify the effectiveness of our ranking model: 1) correlation between our ranking results and the Amazon's sales rank, and 2) Mean Average Precision (MAP), a popular measure used in information retrieval for evaluating ranking accuracy (Turpin and Scholer 2006). The Spearman correlation function (Correlation) ρ is

$$\rho(\vec{s_a},\vec{s_b}) = 1 - \frac{6\cdot\sum_{all\ i}(s_{ai}-s_{bi})^2}{n(n^2-1)} \qquad (5)$$

where $\vec{s_a}$ and $\vec{s_b}$ are the vectors of size n for ranking from our approach and sales rank, respectively. The value of ρ is between 0 and 1. A ρ value closer to 1 means the better correlation of our model and the Amazon's sales rank. However, the Spearman correlation puts equal emphasis on all elements in the vector and does not reflect the quality of the top ranked products. An alternative is the MAP measurement. We pick the top products ranked by our

[1] PRP: Personal pronoun,VBZ: Verb, past participle, RB: Adverb, JJ: Adjective, VB: Verb, CC: Coordinating conjunction, NN: Noun, singular or mass

Table 2: Statistics of review data and their corresponding sentences for SLR camera and TV within the price range from $500 to $700. Symbol '#' represents the number. Symbol '%' represents the percentage.

Category	SLR camera	TV
# of products	252	245
# of reviews	9932	3256
# of sentences	96006	28748
# of irrelevant sentences(filtered out)	3080	1656
# of total votes	108995	21447
# of helpful votes	83829	16983
% of positive sentences	41.39%	39.46%
% of negative sentences	20.37%	21.24%

Table 3: Correlations with the Sales Rank generated by different ranking methods under SLR camera category.

Method	Correlation	MAP@10(annotator 1)	MAP@10(annotator 2)
Baseline	0.5140	0.5525	0.5238
S_1	0.5484	0.6214	0.6573
S_2	0.6365	0.8756	0.8926
$S_2(w/Filter)$	**0.6380**	**0.9012**	**0.9137**

model and ask two experienced annotators in the SLR camera and TV to give their ranking orders. The MAP is calculated from these product rankings.

To the best of our knowledge, no previous work has attempted to solve the product ranking problem discussed in this paper. Our earlier work and (McGlohon, Glance, and Reiter 2010) consider all sentimental sentences equally weighted when calculating the ranking scores. In addition, unlike our model that preprocesses the data to filter out irrelevant sentences, these other works use all sentences. We use the method used in these other works as the baseline for our evaluation. Our experiments show that filtering out irrelevant sentences produces more accurate ranking scores. The ranking results based on the correlation and MAP measurements are given in Table 3 and 4, respectively. The method with the best performance in each category is highlighted. S_1 is the method that adds the weighted helpfulness votes. In method S_2, the weighted review helpfulness and its age are included. $S_2(w/Filter)$ method is based on S_2 with the filtering. As shown in both tables, $S_2(w/Filter)$ method generates the best correlation. However, the improvement of $S_2(w/Filter)$ over S_2 for SLR cameras is not as significant as for the TVs. This is because the number of filtered sentences for SLR cameras is much smaller than for the TVs. Through further investigation of the improvement from S_1 to S_2, we could also see that the posting date factor has a big impact on product rankings. In addition, we also compute the MAP for ranking orders of top 10 products from our ranking system and annotators. The higher MAP@10 values show that $S_2(w/Filter)$ can more accurately retrieve the top 10 products based on sentiment analysis of weighted review sentences than other models. Since the customer's interest is more attracted by top ranked products, the$S_2(w/Filter)$ generates more useful ranking results from a user's perspective.

Effects of Individual Features

Because a product often has multiple features, different customers may be interested in only a subset of specific features of a product. In this paper, we would like to examine and find the features that most affect the overall ranking. In order to obtain the contribution of each feature to the ranking score, we compute the correlation $\rho(\vec{r_f}, \vec{r_o})$ between the ranking based on feature f, $\vec{r_f}$, and overall ranking, $\vec{r_o}$. The feature-specific ranking is calculated by running the same algorithm on the sentences that contain only the words about the given feature. The feature sentences are extracted by using keywords which contain all synonyms of the features.

For example, if we extract all sentences related to lens, we use "lens", "wide angle", "normal range", "zoom", "optical" as search keywords. All keywords are collected from the product descriptions at Amazon.com and the Consumer Report (ConsumerReports). This approach allows us to collect the desired sentences with 70% accuracy. Table 5 shows the correlations for top features of SLR cameras and TVs. This correlation also compares overall ranking and feature specific ranking. From the table, we could see that lens (zoom) and picture quality are leading factors that contribute to the overall quality for SLR cameras and TVs, respectively.

Conclusion and Future Work

In this paper, we present a novel approach (model) to rank products by analyzing the sentiments of reviews and considering how a review's helpfulness votes and its posting date impact the product's ranking. We develop a filter mechanism to remove sentences that do not relate to the product itself. The weight of a review's helpfulness is calculated based on the number of helpful votes and the total votes received from posterior reviewers. We use an

Table 4: Correlations with the Sales Rank generated by different ranking methods under TV category.

Method	Correlation	MAP@10(annotator 1)	MAP@10(annotator 2)
Baseline	0.3725	0.4827	0.4659
S_1	0.3752	0.5739	0.5846
S_2	0.5610	0.7525	0.7833
$S_2(w/Filter)$	**0.6010**	**0.8018**	**0.8406**

Table 5: Individual feature contributions to the overall product ranking.

SLR camera	$\rho(\vec{r_f}, \vec{r_o})$	TV	$\rho(\vec{r_f}, \vec{r_o})$
Lens	**0.8241**	Picture Quality	**0.7813**
Size	0.7411	Size	0.5246
Flash	0.6735	Setup	0.4236
Exposure	0.5919	Input	0.3097
Instruction	0.4309	Control	0.1986
Timer	0.3714	Connect	0.0292
Video	0.3601	Ease of use	0.0021
Battery	0.1696		

exponential function to model the weight of a review's age. Our experiment results demonstrate a good correlation between our proposed model to the Sales Rank (see Table 5: Individual feature contributions to the overall product ranking) reported by Amazon.com. We believe that our model can be used for other E-commerce sites for ranking sellers. Additionally, we believe that our model can be used to analyze campaign topics with comments on social network communities, such as Facebook, Twitter, Blogs, etc.

However, for the purpose of building a reliable ranking system, there may be additional future work needed, as follows:

- Leveraging additional factors which impact product rankings. Currently, our method is based on two properties of reviews to rank products. Other factors such as reviewer credibility and the order in which the review is presented on a product page may also impact the importance of reviews. A review written by authors who have higher credentials may have a higher weight. Also, reviews shown on the product first page may be read by more customers and consequently will receive more helpfulness votes.
- Developing a better strategy to calculate product ranking scores. In our current system, we give all products features the same weight when calculating the overall rankings. Since some product features may be more desirable to customers, weighting specific product features based on importance to the customer to obtain a customized ranking system is also necessary.
- Identifying sarcastic sentences to make sentiment recognition more accurate. Sarcasm is a sophisticated form of speech widely used in online communities (Tsur, Davidov, and Rappoport 2010). Since the sentiments of sarcastic sentences are usually opposite to the literal meanings, using the keyword strategy will not properly filter them out.
- Filtering out spam reviews to make the data cleaner. Spam reviews are intentionally generated to affect customer's purchase decisions. By using heuristics found in the literature, potential spam reviews can be removed.
- Expanding data from other sources. There are many other retailers providing a similar mechanism which allows customers to express their opinions on products or services. Including reviews from multiple sources may increase our dataset and will generate a more statistically significant ranking system.

Acknowledgement

This work is supported in part by NSF award numbers: OCI-0724599, CNS-0830927, CCF-0621443, CCF-0833131, CCF-0938000, CCF-1029166, and CCF-1043085 and in part by DOE grants DE-FC02-07ER25808, DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, and DE-SC0005340.

References

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2009. Multi-facet rating of product reviews. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Series*, 461–472.

ConsumerReports. `http://www.consumerreports.org/cro/index.htm`.

Correlation, S. `http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient`.

Danescu-Niculescu-Mizil, C.; Kossinets, G.; Kleinberg, J.; and Lee, L. 2009. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *18th International World Wide Web Conference*, 141–150.

Ghose, A., and P., I. G. 2007. Designing novel review ranking systems: Predicting the usefulness and impact of reviews. In *Proceedings of the Ninth International Conference on Electronic Commerce*, 303–310.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *UAI, Stockholem.*

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discivery and Data Mining*, 168–177.

Jeffrey, R., and Ratnaparkhi, A. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*.

Kim, S.; Pantel, P.; Chklovski, T.; and Pennacchiotti, M. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 423–430.

Liu, J.; Cao, Y.; Lin, C.-Y.; Huang, Y.; and Zhou, M. 2007. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL(Poster)*, 334–342.

Liu, Y.; Huang, X.; An, A.; and Yu, X. 2008. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 443–452.

Lu, Y.; Zhai, C.; and Sundaresan, N. 2009. Rated aspect summarization of short comments. In *18th International World Wide Web Conference*, 131–140.

McGlohon, M.; Glance, N.; and Reiter, Z. 2010. Star quality: Aggregating reviews to rank products and merchants. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 114–121.

Nobata, C.; Sekine, S.; and Isahara, H. 2003. Evaluation of features for sentence extraction on different types of corpora. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, 29–36.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL*, 271.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*.

Ratnaparkhi, A. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*.

Tsur, O., and Rappoport, A. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Third International AAAI Conference on Weblogs and Social Media*.

Tsur, O.; Davidov, D.; and Rappoport, A. 2010. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *PProceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Turpin, A., and Scholer, F. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, 11–18.

Vapnik, V. 1995. The nature of statistical learning theory. In *Springer*.

WordNet. 2010. `http://wordnet.princeton.edu`.

Zhang, R., and Trani, T. 2008. An entropy-based model for discovering the usefulness of online product reviews. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 759–762.

Zhang, Z., and Varadarajan, B. 2006. Utility scoring of product reviews. In *CIKM*, 51–57.

Zhang, K.; Narayanan, R.; and Choudhary, A. 2010. Voice of the customers: Mining online customer reviews for product feature-based ranking. In *3rd Workshop on Online Social Networks*.