

# ON THE HELPFULNESS OF PRODUCT REVIEWS

## *An Analysis of Customer-to-Customer Trust on eShop-Platforms*

Georg Peters and Vasily Andrianov  
*University of Applied Sciences - Muenchen*  
*Department of Computer Science and Mathematics*  
*Lothstrasse 34, 80335 Munich, Germany*

Keywords: Consumer-to-Consumer Trust, eShops, Helpfulness of product reviews.

Abstract: In the last decade the market share of online stores in the retail sector has risen constantly and partly replaced traditional face-to-face shops in cities and shopping malls. One reason is that the cost structure of online shops is lower than of classic shops since the latter have to finance physical stores and sales personnel. On the one hand, this often leads to a strategic cost advantage and results lower selling prices. On the other hand, online stores normally do not provide personal consulting services as in traditionally face-to-face shops. However, online shops have established different forms of product consulting to compensate the missing personal advice of the sales persons in a physical shop - examples are product related hotlines or online chatrooms. An even cheaper possibility is to establish a recommendation system where previous buyers are invited to write reviews on a product. Some eShops even provide some kind of cascading system: a product review written by a customer can be classified as helpful or not by other customers. In our research we focus on this second cascade. The objective of our paper is to analyze if there are structures or rules that make product reviews written by customers helpful for other customers.

## 1 INTRODUCTION

In the last decade the market share of online stores in the retail sector has risen constantly and partly replaced traditional face-to-face shops in cities and shopping malls. One reason is that the cost structure of online shops is lower than of classic shops since the latter have to finance physical stores and sales personnel.

On the one hand, this often leads to a strategic cost advantage and results lower selling prices of products. On the other hand, online stores normally do not provide personal consulting services as in traditionally face-to-face shops.

However, online shops have established different forms of product consulting to compensate the missing personal advice of the sales persons in a physical shop - examples are product related hotlines or online chatrooms.

An even cheaper possibility is to establish a recommendation system where previous buyers are invited to write reviews on products. Therefore - besides the reputation of a product or a company and reviews in neutral consumer magazines, like Stiftung Warentest in Germany or Which in the UK - recommendations by former buyers influence the decision

which product will be bought by a consumer. The non-commercial communication between consumers about goods and services is known as word of mouth or electronic word of mouth if it is primarily electronically based (Arndt, 1967; Westbrook, 1987).

Some eShops even provide some kind of cascading system: a product review written by a customer can be classified as helpful or not by other customers. In our research we focus on this second cascade. In a previous study Peters et al. (Peters et al., 2007) analyzed the German Amazon shopping platform to find out if there are hidden rules of thumb for writing helpful product reviews. They applied basic statistical methods and found some weak indicators how to write helpful reviews.

The objective of our present research is to apply more advanced statistical methods and analyze if there are structures or rules that make product reviews helpful for other customers. In contrast to the previous study of Peters et al. (Peters et al., 2007) we analyze data of Amazon's U.S. based shop system.

The paper is organized as follows. In Section 2 we present the results of our study on the helpfulness of product reviews on eShop platforms. The paper concludes with a summary in Section 3.

## 2 ANALYSIS AND RESULTS

### 2.1 Preliminaries

Amazon is as a pioneer of the Internet. It was founded as an Internet based bookshop more than 10 years ago. Today, its product range is similar to the product range of a classic department store covering kitchen ware, watches, CD & DVD, computer hard- as well as software, cloth besides many others products. Besides its own department store business Amazon runs an Internet-based shop platform for third-party retailers and therefore functions as some kind of Internet mall.

In the year 2007 Amazon generated net sales of USD 14,835 millions up from USD 10,711 millions in the previous year which is an increase of more than 38% (Amazon, 2008). This makes Amazon to the world's leading Internet retailer and shopping platform.

On its shopping sites Amazon provides its customers a C2C communication platform where reviews and pictures of products can be exchanged. Amazon constantly adds new C2C functionality like the recently introduced video messages. The C2C information of mostly former buyers shall help potential buyers to make their decisions which product to choose. So these information influence the level of trust a potential buyer experiences for a product. However, since the information sources, the authors of the reviews, are normally not personally known or even anonymous to the potential buyer she/he faces another challenge: Can I trust the information sources and their comments on a certain product.

In this context Amazon provides the possibility to rate a product review as helpful. So, product reviews that are mostly considered as helpful can be regarded as good reviews. The objective of this study is to analyze correlations between product review features and the acceptance of a review defined by the *helpfulness* parameter. The analysis is conducted within product categories as well as between different categories.

For our study we choose product categories that have been used in previous studies on eMarketing and customer behavior (e.g. (Wang and Head, 2007)) already. They are listed below:

- Books
- Digital cameras
- Hardware
- Mobile phones
- Software.

For our analysis the following parameters of the product reviews were selected and group into three categories:

- **Review based Parameters.** Text length; absolute and relative numbers of syllables, words, sentences, paragraphs, punctuation marks, signal words (superlative and customer-support-groups), figures, personal pronouns; product rating given by the reviewer, helpful votes (*HelpfulV*), total votes (*TotalV*), review acceptance ( $Acceptance = \frac{HelpfulV}{TotalV}$ ), readability indexes (Flesch and SMOG), presence of listed passages in the review.
- **Reviewer based Parameters.** Number of reviews, average rating of all author's reviews, presence of the 'real name' badge, presence of a reviewer's photo.
- **Product based Parameters.** Average product rating (*avProdRat*), product rating given by the reviewer (*ProdRat*), relative product rating ( $relProdRat = \frac{ProdRat}{avProdRat}$ ), total reviews on product, presence of the product name in the review.

As shop systems Amazon was chosen. Besides its leading position as online shop a main reason for selecting Amazon was that it provides easy access to its data.

### 2.2 Hypotheses

Based on the parameters as given above we defined hypotheses which can roughly be grouped into the following categories:

- Letter and Word Analysis
- Semantic Analysis
- Product-based Analysis
- Reviewer-based Analysis

Our hypotheses are defined in the following paragraphs.

#### 2.2.1 Letter and Word Analysis

In this category we subsume hypotheses that count letters and words without caring about their meaning.

**Text Length.** The impact of the text length of a review on its helpfulness is difficult to predict. On the one hand, a long review may be regarded as more sophisticated and comprehensive than a short review; on the other hand, a short review may be clearer and more precise.

This motivates to the following hypothesis.

**H1b.** The impact of the text length on the helpfulness gets weaker for longer texts.

**Punctuation Marks.** Punctuation marks help to structure a text and give it an intonation. Therefore punctuation marks should have a positive impact on the helpfulness of the review.

This motivates to the following hypotheses.

**H1c.** The frequency of punctuation marks correlates positively with the helpfulness of the review.

**H1e.** The frequency of the question marks correlates positively with review acceptance.

**H1d.** The frequency of the exclamation marks correlates positively with review acceptance.

**Paragraph Density.** We define 'paragraph density' as average paragraph size (in number of words) in a review. Long paragraphs might be more difficult to read and therefore have a negative impact on the helpfulness of the review.

This motivates to the following hypothesis.

**H1g.** Reviews with a high paragraph density are less helpful than reviews with a low paragraph density.

## 2.2.2 Semantic Analysis

Automated analysis of semantics is still an ongoing research topic. Therefore, we restrict our analysis to some very basics.

**Readability.** According to DuBay (DuBay, 2004) 'readability is what makes some texts easier to read than others'. McLaughlin who suggested the SMOG readability index defines readability as 'the degree to which a given class of people find certain reading matter comprehensible and conclusive' (McLaughlin, 1969).

By 1981, over 200 readability indexes were proposed (Klare, 1981) which can be grouped into two categories:

- Semantic aspects, e.g. related to vocabulary
- Syntactic aspects, e.g. average sentence length.

In our analysis we select two popular readability indexes, a variation of Flesch's (Flesch, 1974) famous readability index, the Flesch-Kincaid Grade Level, and SMOG readability index.

The Flesch-Kincaid Grade Level is defined as follows:

$$FKGL = 0.39 \cdot \frac{words}{sent} + 11,8 \cdot \frac{syll}{words} - 15.59$$

while the SMOG index (McLaughlin, 1969) is defined as:

$$SMOG = 1.043 \cdot \sqrt{\frac{complexwords \cdot 30}{sent}} + 3.1291$$

To analyze the correlation between readability and the helpfulness of a review we apply the relative deviation from the means of the indexes:

$$relRD1 = \frac{FKGL - MeanFKGL}{MeanFKGL}$$

and

$$relRD2 = \frac{SMOG - MeanSMOG}{MeanSMOG}.$$

We propose that reviews are more helpful when they are easier to read in comparison reviews with low readability indexes.

This motivates to the following hypothesis.

**H2a.** The review readability positively correlates with the helpfulness of a review.

**Personal Pronouns.** A high frequency of personal pronouns of the first and second person may make the reader more involved in the review. So, in this case the review helpfulness is expected to be above average. The same applies to the usage of personal and relative pronouns of the second person.

This motivates to the following hypotheses.

**H2b.** The more personal pronouns (I, me, we and us) and relative pronouns (my, our, mine and ours) are used, the more helpful the review is.

**H2c.** The more a reviewer uses personal pronouns (you) as well as relative pronouns (your, yours) the more helpful the review is.

**Numbers.** Numbers are normally used to describe details of a product and maybe considered as objective. This objectiveness may be perceived positively by the readers and result in trust in the author of the review.

This motivates to the following hypothesis.

**H2f.** The more numbers are used in a review the more helpful the review is.

## 2.2.3 Product-based Analysis

In this Section we define hypotheses that are related to the reviewed product itself.

**Relative Product Rating.** We assume that extreme product ratings, in numbers of stars, are less helpful in comparison to average product ratings. Extreme product ratings are defined to be far away from average. For example, the average number of stars for a product is four. Then, a product reviewer is considered as extreme when she/he awards the product just one star. This deviation from the average may

be regarded as less helpful review than a mainstream judgement.

This motivates to the following hypotheses.

**H3a1.** The relative product rating correlates with review helpfulness.

**H3a2.** The product acceptance correlates with review helpfulness.

**Product Rating and Review Attention.** Here, we focus on the relationship of the product rating and the helpfulness of the review. We assume that critical reviews are more helpful than encomiums.

When a critical review is an exception to the common opinion about this product it will get more attention (Homer and Yoon, 1992; Park and Lee, 2008) since the reader is already aware of the positive sides of the product and is looking for its downside to obtain a comprehensive evaluation (Bone, 1995).

The number of feedbacks on a review indicates the attention it gets. Extreme reviews probably get more attention than mainstream reviews. However, that might have a negative impact on the acceptance of the review.

This motivates to the following hypotheses.

**H3b.** The helpfulness of negative reviews (one to three stars) is higher than the helpfulness of positive reviews (four and five stars).

**H3d.** The more often a review has been evaluated the less helpful it is.

## 2.2.4 Reviewer-based Analysis

In this Section we develop hypotheses that are centered around the reviewer of a product.

**Previous Review Helpfulness of the Reviewer.** Normally, the authors of the product reviews are not personally known by the reader. Authors even have the possibility to hide personal information like their real names, e-mail addresses etc. and publish their opinions using a pseudonym.

So, the reader of a review may regard an author who was trusted by previous readers already also as more trustful as an author without a comparable good reputation. For example, McKnight et al. (McKnight et al., 2002) claim that the reputation plays an essential role when a person is identified as a trustworthy one.

A high degree of trust can be obtained by signals that indicate a high expertise of the reviewer. So, Bone (Bone, 1995, p. 220) found that 'the influence

of WOM was stronger when provided by an expert than when provided by a non-expert.'<sup>1</sup>

Therefore, an experienced reviewer with many well accepted reviews might probably has a head-start over a novice with respect to trust.

This motivates to the following hypotheses.

**H4a.** The higher an average acceptance of previous reviews of a reviewer is, the higher the acceptance of his current or next review is.

**H4b.** A review by an author who has written many reviews before is more helpful than a review written by an author who has written a small number of reviews.

## 2.3 Analysis and Results

### 2.3.1 Some Fundamental Descriptive Statistics of the Data

On average the longest reviews (*TotalWords*) can be found in the product categories music, software and digital cameras. In the category digital cameras we have the highest number of (*Numbers*). *Punctuation* is higher than the average in music reviews. Personal pronouns of first-person (*PP1*) are most frequently used in digital cameras reviews.

The best readability (average of *relRD1* and *relRD2*) can be found in reviews on music and cell phones. In contrast to that the worst readability indexes occurred in reviews on software and digital cameras.

The most positive average product ratings (*avProdRat*) are in the categories books and music. The worst in software and cell phones.

On average, the helpfulness of reviews is between 77% and 79% for all product categories except for music (61%) and software (70%).

### 2.3.2 Correlations

All variables are significantly correlated.

In general the correlations appear to be weak. There is no single features that has a dominate influence on the helpfulness of a review.

### 2.3.3 Factor Analysis

In total we analyzed 24 features (review-based, reviewer-based and product-based). To reduced these features we apply factor analysis, in particular Principal Components Analysis (PCA) with a varimax rotation.

<sup>1</sup>WOM: words-of-mouth.

We use the graphical method 'scree test' (Cattell, 1966) to define the number of factors to be extracted. The eigenvalues are depicted in a line plot in Fig. 1.

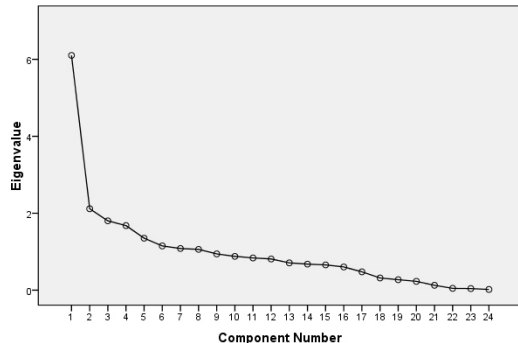


Figure 1: Eigenvalues over Component Numbers.

Cattell suggests to find the point where the smooth decrease of eigenvalues appears to level off to the right of the plot. To the right of this point, one finds only 'factorial scree'.

According to this criterion we retain five factors. The first factor is characterized by high loadings on the review variables that reflect text size and complexity. The second factor is characterized by high loadings reflecting review attention. The third factor is related to product ratings in the review. The fourth factor contains variables referring to the review helpfulness. The fifth factor sums up variables representing the emotional context. Five variables had loadings less than 0.2 and were dropped from subsequent analysis.

The correlation analysis of the retained factors and review helpfulness show that the factors are independent - as it is a central objective of the data reduction.

The factors were tested on correlations with the helpfulness of the reviews. Along the lines with our analysis of all features we obtain correlations with the helpfulness criterion on comparable levels of significance and strengths.

## 2.4 Limitations of the Study

The main limitations of the study are as follows:

*Source of Test Data.* For our study we only analyzed data of the Amazon shop. Other eShop platforms may provide different data. We also do not have any influence on the role of Amazon as moderator that may do not publish all or change product reviews.

*Spelling Errors in Reviews.* The spelling errors in reviews as well as other linguistic irregularities are not considered.

*Dictionaries for Word Groups 'Superlatives' and 'Customer Support'.* Dictionaries for superlatives

may not include all significant signal words.

*Social Structure of Reviewers and Readers.* Another limitation is that product reviews can only be submitted by registered Amazon customer. These customer group may not be representative for average Internet users. We also restrained from any analysis of the demographics of the users.

*Review Manipulations.* There is also a possibility, that some reviews are 'fakes', e.g. written by companies trying to promote their products.

*No Cause-Effect Relations Findings.* Correlation analysis do not provide insides into cause-effect relations.

## 2.5 Discussion of the Results

The study shows that there are weak, however significant correlations between our tested features and the helpfulness of the reviews. Therefore, it is not possible to give a simple rule of thumb how to write helpful reviews.

However, some indicators that may help to get a positive feedback on a review can be derived out of the Tables referred to in the previous Sections.

The main reason for this result of our analysis is that a document is more than 'bag-of-words' (Whitelaw and Patrick, 2004) and our analysis was limited to formal, mainly non-semantic aspects. More sophisticated methods, like concepts based on computer linguistics, may provide further evidence. Such methods - utilized for semantic analysis (Scott and Matwin, 1998) and syntactic analysis (Carr and Estival, 2002) - and other text classification models may provide further insides.

Our study shows and proofs again the present big gap between formals methods and natural language. From a pure IT perspective that might be considered as a pity; however, from a more human centered perspective this result is positive since natural language still remains a hideaway that is difficult to addressed by information technology. For IT it will remain one of the big challenges in the foreseeable future.

## 3 CONCLUSIONS

The hierarchical structure in the retail market, experts in shops recommend products to their customers, has changed significantly towards a flat structure where customers exchange their experience on products. In this context trust plays an important role since customers have to trust the judgement of often anonymous authors of product reviews.

Therefore, in this study we investigate the relationship between text features and the helpfulness of reviews on an online shopping platform. We found statistically significant but weak correlations. The results are positive in the sense that, presently, it seems to be impossible to automatically write helpful reviews and that human language is still too complex to be fully mapped into basic computer linguistic concepts.

Based on the results of our present analysis we are planning to apply more advanced computer linguistic concepts to disclose further structures of helpful reviews.

## REFERENCES

- Amazon (2008). Amazon.com - Annual Report 2007.
- Arndt, J. (1967). Role of product-related conversations in the diffusion of a new product. *Journal of Marketing Research*, 4:291–295.
- Bone, P. (1995). Word-of-mouth effects on short-term and long-term product judgements. *Journal of Business Research*, 32:213–223.
- Carr, O. and Estival, D. (2002). Text classification of formatted text documents. In *Proceedings of the 2002 Australian Natural Language Processing Workshop*.
- Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1:245–276.
- DuBay, W. (2004). *The Principles Of Readability*. Impact Information.
- Flesch, R. (1949, 1974). *The art of readable writing*. New York.: Harper.
- Homer, P. and Yoon, S. (1992). Message framing and the interrelationships among ad-based feelings, affect, and cognition. *Journal of Advertising*, 21(1):19–33.
- Klare, G. (1981). Readability indices: do they inform or misinform? *Information Design Journal*, 2:251–255.
- McKnight, D., Choudhury, V., and Kacmar, C. (2002). Developing and validation trust measures for e-commerce: an integrative typology. *Information Systems Research*, 13:334–359.
- McLaughlin, G. (1969). Smog grading a new readability formula. *Journal of reading*, 22:639–646.
- Park, C. and Lee, T. (2008). Information direction, website reputation and ewom effect: A moderating role of product type. *Journal of Business Research*.
- Peters, G., Damm, M., and Weber, R. (2007). Consumer-to-consumer trust in e-Commerce - Are there rules for writing helpful product reviews. In *ICEIS 2008 - Proceedings of the Tenth International Conference on Enterprise Information Systems, Volume SAIC*, pages 61–66.
- Scott, S. and Matwin, S. (1998). Text classification using wordnet hypernyms. In Harabagui, S., editor, *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference.*, pages 38–44. Association for Computational Linguistics, Somerset, New Jersey.
- Wang, F. and Head, M. (2007). How can the web help build customer relationships? *Information & Management*, 44:115–129.
- Westbrook, R. (1987). Product/consumption-based affective responses and postpurchase processes. *Journal of Marketing Research*, 24:258–270.
- Whitelaw, C. and Patrick, J. (2004). Selecting systematic features for text classification. In *Proceedings of the Australian Language Technology Workshop 2004*. Australian Speech Science & Technology Association Inc., Macquarie University, Sydney.