# The Impact of Readability on the Usefulness of Online Product Reviews: A Case Study on an Online Bookstore

N. Korfiatis[1], D. Rodríguez[1], and M.A. Sicilia[2]

[1] Department of Informatics
Copenhagen Business School (CBS)
DK-2000, Frederiksberg, Copenhagen, Denmark
nk.inf@cbs.dk
[2] Department of Computer Science
University of Alcalá
28805 Alcalá de Henares Madrid, Spain
{daniel.rodriguezg,msicilia}@uah.es

**Abstract.** Online product reviews is an important advantage for consumers of experience goods in online marketplaces and act as a useful source of information during the purchase of a good. Furthermore in some online marketplaces consumers have the opportunity to evaluate how helpful a review was by using a binary evaluation interface provided by the online marketplace. This results to the usefulness score of a review which is calculated as a fraction of helpful votes over the total votes that this review has received. Our early results indicate that the usefulness score of a particular review is affected in a significant way by the qualitative characteristics of the review as measured by readability tests applied to a large dataset of reviews collected from the UK section of the popular online marketplace *Amazon*.

**Keywords:** Readbility tests, Online services.

## 1 Introduction

One of the most profound advantages of online marketplaces and electronic commerce to the product choice process and purchase is the ability to encapsulate and promote the opinions of their customers for the products that they have purchased. This has led to a massive amount of product reviews accessible online, which a consumer may use for an informed decision about the product or the service that is considering to purchase.

The importance of online reviews on the choice process of a certain product by an individual on the internet has been a subject of several recent studies in the literature [1,2,3]. Due to the digital nature of delivery, most of the goods available for purchase online are experience goods [4]. An experience good is a product or a service which quality and utility for a consumer can only be determined upon consumption. This provides that in order for a consumer to

M.D. Lytras et al. (Eds.): WSKS 2008, LNAI 5288, pp. 423–432, 2008.

make a decision for the purchase of this good or service, she/he has to rely on previous experiences which will provide an indicator whether this good or service is worthy for purchase or not. One example of an experience good can be the case of a book where the utility that the consumer perceives by reading such a book can be extracted only after reading it. Therefore, in experience goods such as books, the producers (e.g. publishers) often use the reviews by authoritative sources such as literature experts to provide an opinion and endorse the book, so that consumers trusting these sources will continue to the purchase of this good.

However the inclusion of a prior experience to the promotion of an experience good can actually pose a problem for the consumer mainly due to following factors:

- The cost for the producers of publishing experiences by previous consumers and especially in cases where negative views might reach new customers. Such a cost makes the producers not willing to do so [5].
- The obvious search costs that arise for a consumer in order to search acquire and evaluate the prior experiences.
- The variance between the different versions of the same good which may confuse a consumer (e.g. an mp3 player with a large set of characteristics vs. a simpler mp3 player).

The development of Internet marketplaces where consumers can establish interaction has undoubtedly has affected the way a review –as an expression of prior experience– influences the way consumers make a choice about a product or a service based on prior experiences [6,7]. First and foremost, the use of online mechanisms for the reporting and categorization of reviews by a product or a service in conjunction with the development of modern search engines has eliminated the search costs for the consumers. Online marketplaces such as the popular bookstores *Amazon.com*[1] and *Barnes and Noble*[2] provide the ability for a consumer to read a series of reviews about the product that a consumer is interested in purchasing. Furthermore, apart from the description of the experience deriving by the purchase of the reviewed product a consumer is able also to rate the usefulness of the product usually by rating on a standard Likert scale.

The later comes into connection with an important field in marketing literature, which has to do with the referral value of a specific product. In particular in "word of mouth" scenarios consumers refer to a product or a service to fellow consumers usually by enthusiasm (if they are satisfied, or regret if they are unsatisfied). The extent to which the referral value of a specific customer might affect another one still remains an issue to identify.

In this study we assume that a review submitted by an individual reflects his/her experience from the product usage. Furthermore the review text acts as a "justification" of the rating so the potential buyer can evaluate if the review was fair or not. In addition, most online marketplaces use a way of meta-rating on

[1] `http://www.amazon.com/`
[2] `http://www.barnesandnoble.com/`


**Fig. 1.** The interface of the review evaluation mechanism that we use in this study

unfair reviews where interested buyers can evaluate how helpful was this review during the decision-making process of purchasing a good. Again on that case, the review acts as the main source of evaluation of the usefulness of this specific review by other consumers. Reviews by individual consumers often express a personal view of their experience with the product and might differ in such from the expectations of the interested buyer. For example it might be that someone expected a book to contain more action elements; however, an interested buyer might is not be interested in that specific characteristic. Nevertheless in order to evaluate the usefulness of the review someone has to read it first. Therefore the style and the readiness of a review might actually play a role on how its usefulness is evaluated.

In this paper, we evaluate how the style and the comprehension of a review as depicted by a readability test, might affect the usefulness of a review –the number of people that found this review useful out of the total number of people that read and evaluated this review. In order to investigate this issue, we employ the use of readability metrics applied on a dataset of reviews with their meta-evaluations collected by the bookstore section of the website of Amazon in UK[3].

The major objective of this study is the evaluation of the impact that the qualitative characteristics of a review might have to consumers that are interested in buying a product or a service from an online Web store. Early results indicate that apart from the review score of a particular review, a consumer also evaluates its importance by how this review is close to his/her communication code which is denoted by the way the review has been written.

## 2 A Background on Readability Tests

The concept of readability describes in general terms the cognitive effort that is needed by an individual to understand and comprehend a piece of text [8]. In a more formalized way, a readability test consists of a formula which is the result of a linear regression applied to subjects regarding the reading ease of different pieces of text that were asked to comprehend using specific instruments. The objective of a readability test is to provide an indication on a scale of how difficult is the comprehension of a piece of the text by readers in conjunction with the linguistic characteristics of a text. In that case a readability test can

[3] http://www.amazon.co.uk/

only provide us with an indication on how understandable is this piece of text based on its syntactical elements and style.

As such we assume that the attention that a review might receive by the interested buyers of the particular product, can in a large extend be associated with its readability. On our case the assessment of a review by a readability test provides us with an indication whether someone who evaluated how useful a particular review was, actually comprehended this piece of text. On the other hand we might expect that the fact that some reviews were not considered helpful might have been affected by the readability of the content as well.

However the use of a readability test has some weaknesses which we should take into consideration during the analysis of the results in this study. In particular the result of a readability formula cannot tell us whether the content of the review expresses personal views on the product and/or contains some gender, social class or even cultural bias. In order to avoid the case of a selection effect due to the cultural background we collected the reviews only from the U.K. store of the online marketplace in order to maximize the number of native English speakers and as much cultural homogeneity of the population as possible.

**Table 1.** Readability Tests Used

| ***Readability Measure*** | ***Score Range*** | ***Measurement Implications*** |
|---|---|---|
| Gunning-Fog Index | 1-12 | Indicates the grade level of the education scale. The lower the grade the more readable the text |
| Flesch Reading Ease Index | 0-100 | Scores above 80% make the text understandable by literally everyone. As the value of the index decreases the comprehensiveness of the text becomes more difficult. |

Table 1 lists the two readability texts that we selected for our analysis. The tests - Gunning's Fog index and the Flesch Reading Ease Index - evaluate the readability of a text by consistently decomposing the text into its basic structural elements which then combine using the empirical regression formula. An important issue of a readability test is that it can be used to evaluate texts of certain length since the comprehension of a text by a reader has also other cognitive properties which are beyond the scope of this study. The logic behind the calculation and the norms of these instruments is described in the sections bellow.

### 2.1 The Gunning Fog Index

Gunning's Fog index [9] literally produces a measure of how comprehensible is a piece of text by an individual with a high school education. In order to calculate the Fog score for our data we followed the following steps:

For each review we calculated the average number of words per review sentence on a 100+ word review passage. This gives as the average sentence length ($L$).

We then obtained the number of the difficult words ($D$) that is words that have more than three letters by excluding proper nouns, compound words and common suffixes. We finally added the average sentence length to the number of the difficult words. The following equation describes the empirical relation in the Fog Index:

$$Fog = 0.4 \times \left( \frac{Words}{Sentence} + 100 \times \left( \frac{N(complex\_words)}{N(words)} \right) \right)$$

An obvious difficulty on measuring the Fog index for a given text is the evaluation of the number of complex words. In our analysis we considered a word as complex when it has more than two syllables.

### 2.2 The Flesch Reading Ease

The Flesch Reading Ease index [10] is a readability test which uses as a core linguistic measure that is based on syllables per word and words per sentence in a given text. The Flesch test is used to evaluate the complexity of the text to determine the number of years of education which are needed for someone to understand the examined text. The following equation describes the calculation of the Flesch score for a given text:

$$FK = 0.39 \times \left( \frac{total\_words}{total\_sentences} \right) + 11.8 \times \left( \frac{total\_syllables}{total\_words} \right) - 15.59$$

The variables *total_words*, *total_sentences* and *total_syllables* denote the total number of words, sentences and syllables found in a text respectively. For calculating the Flesch score of a particular review we decomposed the text into sentences, then words and finally into syllables which were combined using the constants presented in the formula above. It can be easily be implied from the mathematical expression that the sorter is the number of words per sentence is, the better the readability score that the Flesch test will give.

## 3 Analysis and Results

Having provided a background on the readability tests used, we analyzed the reviews stored in our dataset to test whether the readability tests can actually provide us with an indication on how the qualitative characteristics of a review influence its usefulness for a consumer.

### 3.1 Data Collection and Definition of Variables

In order to apply the readability tests presented in Section 2, we developed a Web crawler to capture the content of the book section of Amazon UK. The crawler consisted of two parts: (i) a Web client to randomly pick items from the frontpage of the bookstore; and (ii) a client to the Web service interface provided by Amazon (AWS) where data from the particular item was collected.

**Table 2.** Main Components of the Initial Dataset Collected by Using the Web Crawler

| Variable Code | Variable description |
|---|---|
| productid | The id of the product that this review is written for. It is used to control for the publication date and other product characteristics |
| summary | The summary / title of the review |
| content | The actual content of the review. To be used for content analysis |
| revieworder | The order that the review appears on the product review page. |
| reviewpage | The page that the review appears (default setting is five reviews per page) |
| rating | The rating that this review justifies, measured in a Likert scale. |
| totalvotes | The number of total votes that have been given to this review |
| helpfulvotes | The number of votes that consider this review helpful. |
| reviewerid | The id of the customer used to control if the customer is a professional reviewer or not |

The list of books was stored in a relational database which we used for further processing of the reviews expressed in each individual book page. We omitted from the database those books that the publication date was older than 6 months or had no rating. We also excluded books at special offers or discounts to control for price effects.

As aforementioned, the reason for selecting Amazon U.K. to obtain the dataset used in this study is the case of language homogeneity among reviewers and consumers which might play a role in the comprehension of a text. In addition this is important because readability tests are useless in case a reader is not a native speaker of the language that the text is written. This is due to the fact that many languages differ in syntactical form and the style of the language is written in the review might be totally different from the reader's native language.

Another issue with the dataset that we had collected was the case of by-passing promotion backed items such as bestsellers. Since these items are more accessible to the visitors of the online bookstore/future customers there is always the case of a selection bias towards the more visible items. This may result to a high exposure of the product reviews that are more recent in contrast with those that are older. In order to avoid that the web crawler was keeping a list of the frequency of the items that were displayed in the frontpage and randomly chose items listed by categories.

Our dataset contains in total seven variables and two identifiers. The *reviewerid* provides the identifier of the customer in the online bookstore's central database in order to group reviews performed by customer (a customer may have submitted reviews for more than one book). The *productid* is the unique product identifier provided for this product. With this identifier we can group the reviews by product and check for variances between products of different categories.

We define the usefulness ratio of a review ($UR$) as the fraction of the votes that considered this review helpful *(helpfulvotes)* divided with the total number of readers that evaluated the usefulness of the review *(totalvotes)*.

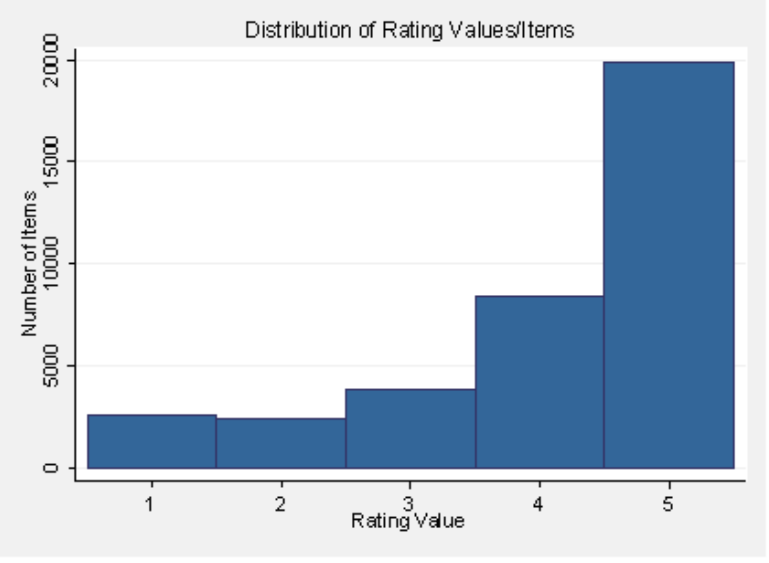


**Fig. 2.** Distribution of rating values among items in our dataset

Thus we have the main unit of analysis for our study defined as $UR = helpfulvotes/totalvotes$. However since the number of total votes that a review had (that is the minimum amount of readers) may affect the consistency of the metric we need to keep control of the exposure of this review since some reviews at a certain period of time receive more exposure than others. Typically the system displays first the most recent reviews that were submitted for the book under review).

In our study the particular exposure of a Review was measured by keeping a set of two variables for the pagination results. In particular the variable *reviewpage* is informative on whether this review was at the first, second or third page at the time the review was retrieved. Subsequently the same case was for the *revieworder* which controls the display order for a particular review in a particular page. Combining the two variables (*reviewpage, revieworder*) into a new compositional variable we are able to control for the review exposure on the website during time. For example if a review appears in page 2 and was ordered as third in the page then the order number is 23 and so forth. It is generally assumed that reviews which appear on the top of a page obtain a much higher exposure than a review that appears at the bottom since visitors' attention get to be captured by elements that are displayed in the beginning of the space under the product description.

On the other hand, regarding the actual exposure of a review and in particular the amount of people that read the review we don't have a variable that justifies that. However in order to hold our analysis in an acceptable level we make the assumption that the total number of people that evaluated the usefulness of this

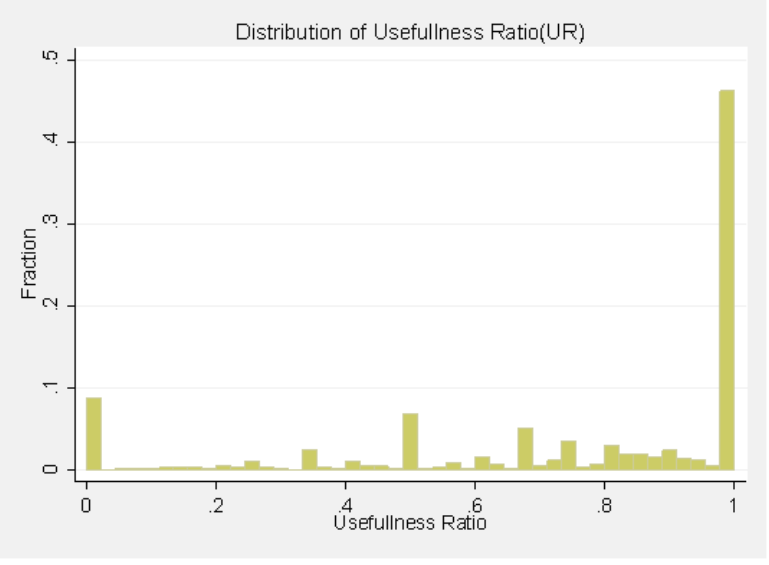


**Fig. 3.** Distribution of Usefulness Scores Plotted by Density ($N = 37,221$)

particular review is the minimum number of the readers that this review had. With that way we get an indication whether this review has been read by a high amount of visitors since it is assumed that those two numbers are positively correlated.

Our dataset consists of 38,366 reviews where the total votes (*totalvotes*) were greater than zero which means that the reviews on our dataset have been evaluated for their usefulness at least once.

Fig. 3 provides an overview of the distribution of usefulness scores on our dataset by rating. It is interesting to note that around 47% (total of: 17,695) of the reviews have received a perfect score by the readers which provides that around half of the reviews were very highly acclaimed by their readers resulting for this group of particular reviews, the number of helpful votes to be the same with the potential buyers that have read it. On the other hand we find that approximately 9% of the votes (total of: 3,292) of the reviews were found totally non-useful for their readers receiving an absolute zero (0) of helpful votes. As it can be observed in Fig. 3 much of the variance on the usefulness score happens between the 0.8 and perfect (1).

### 3.2 Results

Table 3 presents the inter-correlation matrix obtained from running inter-item correlations between the items in our dataset. The scores were obtained by doing a pair- wise correlation between the variables and asking for a confidence interval of 1% ($p < 0.01$). By looking the sign of the coefficients in the first column we

obtained some interesting information. In particular the higher is the exposure of the review (*reviewexposure*); the lower is the usefulness of the review. In fact by looking more carefully at the relation between the usefulness ratio and the exposure of a review we observe that the more exposed is the review, the less helpful votes the review will take where at the same time, the coefficient of total votes is positive.

**Table 3.** The Inter-correlation Matrix between Elements in our Dataset ($*p < 0.01$)

| | ***ur*** | ***reviewexposure*** | ***rating*** | ***wordcount*** | ***fogscore*** | ***fleschscore*** |
|---|---|---|---|---|---|---|
| ***ur*** | 1.0000 | | | | | |
| ***reviewexposure*** | -0.1210* | 1.0000 | | | | |
| ***rating*** | 0.2717* | -0.0449* | 1.0000 | | | |
| ***wordcount*** | 0.1585* | 0.0248* | 0.0040 | | | |
| ***fogscore*** | 0.1158* | -0.0348* | -0.0148* | 1.0000 | | |
| ***fleschscore*** | 0.0982* | -0.0368* | -0.0317* | 0.9621* | 1.0000 | |

All three variables that are connected with the qualitative characteristics of an online product review have been found positively correlated and significant ($*p < 0.01$). The coefficients received for the simplest qualitative characteristic which indicates the review length *(wordcount)* actually provides that the perceived usefulleness of a review is affected by almost 16% while the performance of the review text affects a review by 11% and 9% respectively. As can be seen from table 2 the three factors account for almost 35% of the usefulness score of a review providing that the qualititative characteristics indeed might play a role to the perceived usefulness of a review by a consumer.

## 4 Conclusions and Further Research

The early results presented in this paper indicate that qualitative characteristics of online product reviews indeed play a role on the evaluation of the usefulness of a particular review by a consumer. The use of tools such as readability tests provided a way to evaluate the importance of these characteristics by employing simple statistical analysis methods which however need to be evaluated more for robustness. One of the limitations of this study is to assess whether the review was written in a way that was expressing a personal opinion about a product or a service. Consumers tend to associate themselves with other consumers that express a more personal experience about the product which might influence the consumers' choice process of the good. In addition a further limitation of this study is to check the actual reliability of the readability tests by cross validating whether the tests actually measures the readability of a review written on a website since the readability tests do not take into account usability factors (e.g. the position of the text on the screen etc).

## Acknowledgements

We would like to thank the Copenhagen Business School, University of Alcal and the autonomic community of Madrid (CCG07-UAH-TIC-1588) for their financial support.

## References

1. Pavlou, P., Dimoka, A.: The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. Information Systems Research 17(4), 392–414 (2006)
2. Chevalier, J., Mayzlin, D.: The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research 43(3), 345–354 (2006)
3. Hu, N., Pavlou, P., Zhang, J.: Can online reviews reveal a product's true quality? empirical findings and analytical modeling of online word-of-mouth communication. In: Proceedings of the 7th ACM conference on Electronic Commerce, pp. 324–330 (2006)
4. Nelson, P.: Information and consumer behavior. Journal of Political Economy 78(2), 311 (1970)
5. Richins, M.: Negative word-of-mouth by dissatisfied consumers: A pilot study. Journal of Marketing 47(1), 68–78 (1983)
6. Clemons, E., Gao, G., Hitt, L.: When online reviews meet hyperdifferentiation: A study of the craft beer industry. Journal of Management Information Systems 23(2), 149–171 (2006)
7. Dellarocas, C.: The digitization of word-of-mouth: Promise and challenges of online reputation mechanisms. Management Science 49(10), 1407–1424 (2003)
8. Zakaluk, B., Samuels, S.: Readability: Its past, present, and future. International Reading Association, New Yark (1988)
9. Gunning, R.: The fog index after twenty years. Journal of Business Communication 6(2), 3 (1969)
10. Flesch, R.: How to Test Readability. Harper (1951)