

A cross-culture validation on the predictors of online review helpfulness

Yong Liu^{1}, Hongxiu Li²*

¹Department of Information and Service Economy, Aalto University School of Business, Finland

²Department of Management and Entrepreneurship, Turku School of Economics,
University of Turku, Finland

Abstract: A large amount of studies on online review has been conducted, most of which are based on studying English-speaking reviewers. However, it is unclear whether customers who speak other languages, such as Finnish, German and Japanese, would exhibit similar preference in defining the helpfulness of online review or not? There is a lack of knowledge on how different cultural backgrounds affect users' evaluation on the helpfulness of online review. The current study seeks to offer valid answer and fresh insights in this regard. Based on a collection of 57,000 online reviews, the study applied Tobit regression analysis to explore the impact of review extremity and review depth on review helpfulness across six different language groups, including English, Finnish, German, Italian, Russian and Japanese. Substantial and significant differences are found between English and non-English review groups. Specifically, review extremity affects the helpfulness counts of English review in a way significantly differing from Russian, German, Italian and Japanese reviews. In addition, the review depth has the strongest impact on the helpfulness count of English review in comparison to reviews in other languages. These findings highlight a need to discriminate English review and non-English in future research as well as in the practice of review management.

Keywords: Cultural difference, online review, user-generated content, helpfulness

1. INTRODUCTION

Online customer reviews refer to peer-generated comment on products or services that are posted online [9]. A multitude of IS studies have been conducted to understand how users evaluate and perceive an online review to be useful [9, 12]. To the best of our knowledge, most of the studies have been based on studying English reviewers while little is known on how possibly cultural difference of users would alter their preference over online reviews. The rich availability of research findings grounded on studying English-based reviews may be harmful to business managers if these English-featured findings in fact are not in line with the nature of non-English reviews. As a result, this dearth of cross-culture studies on users' preference over online review motivates the current study.

In this study, we explore and quantify the effect of different factors in shaping consumers' evaluation of review helpfulness by comparing English reviews with many other non-English reviews, including Finnish, Germany, Russian, Italian and Japanese. Based on collecting a large dataset of over 57,000 online reviews on Finnish hotels, we detect the languages used to write each review and compare the key determinants affecting the amount of helpfulness vote on a specific review.

Based on previous studies, three predictors (including overall satisfaction rating, the square of overall satisfaction rating, review depth) and one control variable (type of travel) are included in the analysis. The results confirm significant differences among the determinants of review helpfulness between different language groups. New insights are highlighted on a basis of the results.

The remainder of the paper is structured as follows. The next section presents a literature review, which is followed by research model section. Thereafter, the research model is empirically tested and the results are discussed. Section 6 highlights the implications of the research findings. Limitations of the research are

* Corresponding author. Email: yong.liu@aalto.fi(Yong Liu), hongli@utu.fi(Hongxiu Li)

presented in section seven.

2. LITERATURE REVIEW

The increasing popularity and availability of online reviews have attracted a considerable amount of research attention in academia. Within this stream of research, an important research question is what makes a review useful/helpful for online users. As a result, the amount of review helpfulness vote has been frequently adopted as an interesting dependent variable investigated by a number of previous studies. For instance, Chua and Banerjee [1] employed review reputation, review rating and review depth (the number of words in the review text) as key predictors of review helpfulness by studying the reviews on 1000 bestseller books on Amazon. The results indicated that review rating has a linear, other than a curvilinear relationship with review helpfulness. In addition, review depth not only significantly affects review helpfulness, but also moderates the relationship between review rating and review helpfulness. Yin et al. [12] detected the emotion expressed by the reviews, and found that reviews indicative of anxiety were rated more helpful than those indicative of anger. Mudambi and Schuff [9] differed reviews on experience and search products, and reported a moderating effect of product type on the effect of review extremity on review helpfulness. Based on reviews collected from Amazon UK, Korfiatis, García-Bariocanal and Sánchez-Alonso [5, p. 205] computed the reliability level of each review and found that “review readability had a greater effect on the helpfulness ratio of a review than its length”. Based on the data collected from Yelp.com, Liu and Park [8] found that real photo uploading, review length, rating and Flesch reading ease index have significant effects on review usefulness. In addition, rating has a non-linear relationship with review usefulness. This increasing availability of studies on the determinants of reviews helpfulness indicates that this is an important research topic in the IS field.

Nonetheless, to the best of our knowledge, previous studies are dominantly grounded on studying English-based reviews. The rich availability of English-review-based findings offer limited or skeptical business insights when applying to the context of reviews for non-English-speaking users, such as those from non-English speaking Europe or Asia countries, such as Germany, Italy, Finland and Japan. In this light, a handful of across-culture studies on online reviewing behavior provide a warning. For instance, Koh, Hu, & Clemons [4] reported that, because collectivist societies tend to emphasize more on harmony, they incline to write less extremely negative reviews than individualist societies. Note that previous studies have highlighted that customers from distinctive cultural background tend to behave differently. Fang, Zhang, Bao, and Zhu [2] suggested that US consumers tend to offer a large number of online reviews than Chinese consumers while online review in US online platform is more likely to be voted by readers in comparison to online reviews in Chinese platform. In marketing research, Hofstede's cultural dimensions theory has been widely applied and cited to explore the behavioral difference among consumers originating from different cultural backgrounds [c.f. 3]. In this regard, the work of Leidner and Kayworth [7] provided an excellent summary on the linkages between IT use and culture.

In the current study, the degree of impacts for the key determinants of review helpfulness are computed and further contrasted across different language groups, including English, Finnish, German, Russian, Italian and Japanese reviews. The identified languages serve as a proxy of the cultures of online users who vote for the reviews. Given that the study investigated six intricately different and profound cultures, in this study, we do not seek to explicitly introduce each culture and theorize the difference between every two cultures. Instead, we will treat language of reviews as a proxy of cultural background and to quantify its moderating effect on the hypothesized relationship. In other words, the assumed moderating effect is exploratory in nature.

3. RESEARCH MODEL

3.1 Review extremity

In previous studies, review extremity has been frequently employed as a key predictor of review helpfulness [9]. Specifically, review extremity is measured by the existence of either very high or very low rating on services or products. For instance, in a rating scale of 1-5, 1 (terrible) or 5 (excellent) represent extremely negative or positive view, while a rating of 3 (average) reflects a moderate view [c.f. 9].

Many online review forums, such as TripAdvisor, provide a numeric rating with regard to reviewers' overall feeling towards a service or a product. In TripAdvisor, rating is posted on the top of the text of each review, which is capable of catching the immediate attention of readers [1]. Previous studies suggested that readers may favor extreme reviews than those exhibiting a moderate rating or perspective [see. 8], even though some studies suggest that a moderating rating could be more helpful in certain contexts [see. 9]. Albeit a curvilinear relationship is in general assumed, it is difficult to determine whether the value peaks at extreme ratings or moderate ratings [1]. Therefore, in line with the work of Chua and Banerjee [1], the current study assumes a general curvilinear relationship between review extremity and review helpfulness:

H1. Review extremity has a curvilinear relationship with the helpfulness of a review.

3.2 Review depth

Review depth refers to the amount of various information provided by a review text. It is typically measured by the length of the review text [8, 9]. The longer the review text is, the more extensive and elaborative the review will be [8]. Mudambi and Schuff [9] noted that in-depth or long reviews tend to be diagnostic for readers, which assist readers in making purchase decision. A lengthier review tends to be more elaborative, which helps alleviate users' uncertainty about the product or service quality [8, 9]. The reduced perceived uncertainty in turn raises consumers' confidence in purchase process [8, 9]. The study of Chua and Banerjee [1] also found that length of review affects review helpfulness vote. In line with above-mentioned studies, a positive relationship between review depth and the helpfulness of review is assumed.

H2. Review depth positively relates to the helpfulness of a review.

As specified in the section 2, we assumed a moderating effect of culture on H1 and H2. Hofstede's cultural dimensions theory, as one of the most important and widely cited social science theories, suggest that differences in human behavior frequently attribute to their cultural difference [c.f. 3]. Based on the theory, we assume culture to be a significant factor moderating consumers' decision-making process when deciding the helpfulness of a review. We admit that the assumed relationship is exploratory in nature, other than being explanatory, due to complexity to compare six intricate cultures with a restricted paper length. Thus, based on the above discussion, the theorized model is presented as follow:

$$\text{Model: Review helpfulness} = \beta_1 * \text{Rating} + \beta_2 * \text{Rating}^2 + \beta_3 * (\text{Review depth}) + \epsilon$$

4. DATA ANALYSIS

Research data is collected from TripAdvisor with regard to reviews on hotels in Finland. Specifically, over 57,000 reviews on Finnish hotels that were posted before February 5, 2016 were collected. The reviews reflects the accommodation experience of 14,329 business trips, 16,636 couple trips, 11,744 family trips, 6,011 friends trips and 4,293 solo trips. About 4,100 reviews did not specify their travel types. Language-detection of each review was performed through the use of the R textcat package [6] and MySQL database. Our analysis focused on six most popular languages that were used in writing reviews on Finnish hotels, including English (N = 34,869), Russian (N = 6,459), Finnish (N = 3,945), Italian (N = 2,439), German (N = 2,355) and Japanese (N = 1,806).

A large amount of reviews did not attract any helpfulness vote, indicating that the dependent variable,

amount of helpfulness vote, is left censored. As a result, Tobit regression, a regression method designed to handle censored dependent variable, is adopted to testify the validity of H1 and H2. Note that Tobit regression has been utilized in previous studies on review helpfulness as well [see. 5, 9] [5, 9]. Rating on hotel was z-score standardized in order to avoid multicollinearity issue between rating and rating2. Review depth is measured by the length of a review.

Coefficient values for each predictor were firstly computed via Tobit regression, as shown in Table 1. The travel type is utilized as a controlled variable in the regression model. Thereafter, the equality of regression coefficients between different language groups are tested via the formula proposed by Paternoster, Brame, Mazerolle and Piquero [10]. Particularly, the coefficients of each language group are contrasted with the coefficient value of English review group. As shown in Table 2, a contrast of coefficients between English and other language groups is performed while the analysis reported significant differences of the coefficient values between language groups. Note that apparent difference is observed as well between non-English reviews, such as between Russian and Japanese reviews.

The results of Tobit regression analysis are visualized in Figure 1. In Figure 2, the curvilinear relationship between rating and review helpfulness for each language groups are visualized as well, except for Japanese group due to the insignificant relationship. Both Figures 1 and 2 are produced via the use of R package ggplot2 [11].

Table 1. Results of Tobit regression (*: p-value < 0.001)**

	All samples	Finnish	English	Russian	German	Italian	Japanese
Rating	0.344***	0.557***	0.429***	n.s.	0.181*	0.222***	n.s.
Rating2	0.339***	0.323***	0.390***	0.314***	0.247***	0.220***	n.s.
In(Review depth)	1.736***	1.296***	1.892***	0.956***	0.724***	1.164***	0.755***
Travel type	Controlled	Controlled	Controlled	Controlled	Controlled	Controlled	Controlled
Pseudo-R ²	16.1%	14.3%	18.3%	6.8%	8.8%	12.8%	9.9%

Table 2. Comparing the equality of regression coefficient between different language groups

	Finnish to English	Russian to English	German to English	Italian to English	Japanese to English
Rating	n.s.	-0.429***	-0.248**	-0.207*	-0.429***
Rating2	n.s.	n.s.	-0.143**	-0.170***	-0.390***
In (Review depth)	-0.596***	-0.936***	-1.168***	-0.728***	-1.137***

Note: i) *: p-value < 0.05; **: p-value < 0.01; ***: p-value < 0.001; n.s.: not significant;
ii) insignificant coefficient is replaced with a value of zero in regression equality test.

5. DISCUSSION

Based on the Table 1, we found that both rating and review depth have significant impacts on review helpfulness across most language groups, except for Japanese subsamples. The findings are in line with a number of previous studies, and provide a good support for H1 and H2. In other words, reviews with extreme ratings (either very positive or negative) are more likely to be voted by most customers, but not by Japanese customers. Across all subsamples, review depth is a significant predictor of review helpfulness. This indicates that, lengthier a review is, more likely the review will be voted to be helpful.

Nonetheless, by comparing the equality of regression coefficients, significant differences can be found between English and other language subsamples. Specifically, across six different language subsamples, the perceived helpfulness of English reviews are more heavily affected by review depth than that of Finnish, German, Russian, Italian and Japanese reviews. German customers seem to be much less affected by review

depth than customers speaking other languages that were investigated.

The effects of rating on review helpfulness vary a lot among different language groups. Especially, there seems to be an insignificant relationship between rating and review helpfulness for Japanese reviews, not matter the testing relationship is linear or curvilinear. In other words, rating does not affect Japanese customers' vote on review helpfulness, albeit review depth does.

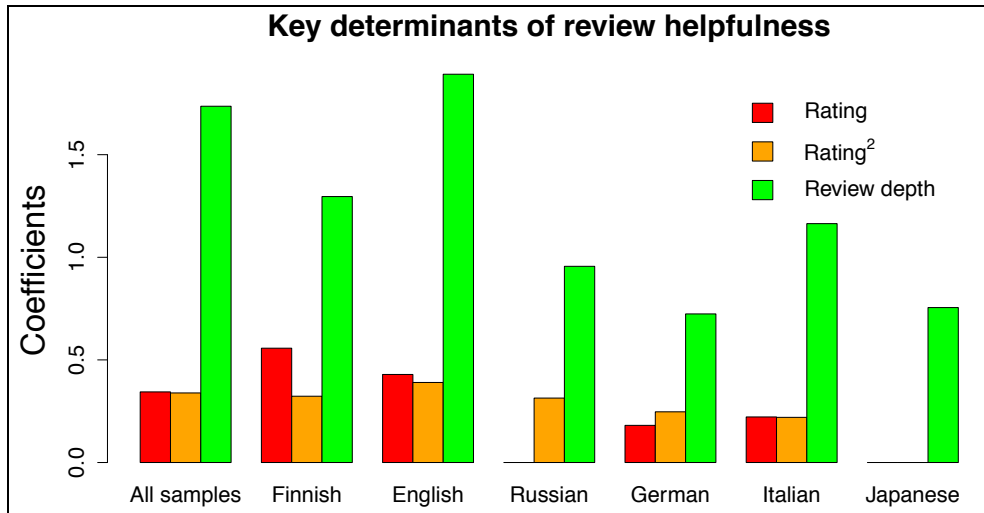


Figure 1. Summary of coefficients value for predictors

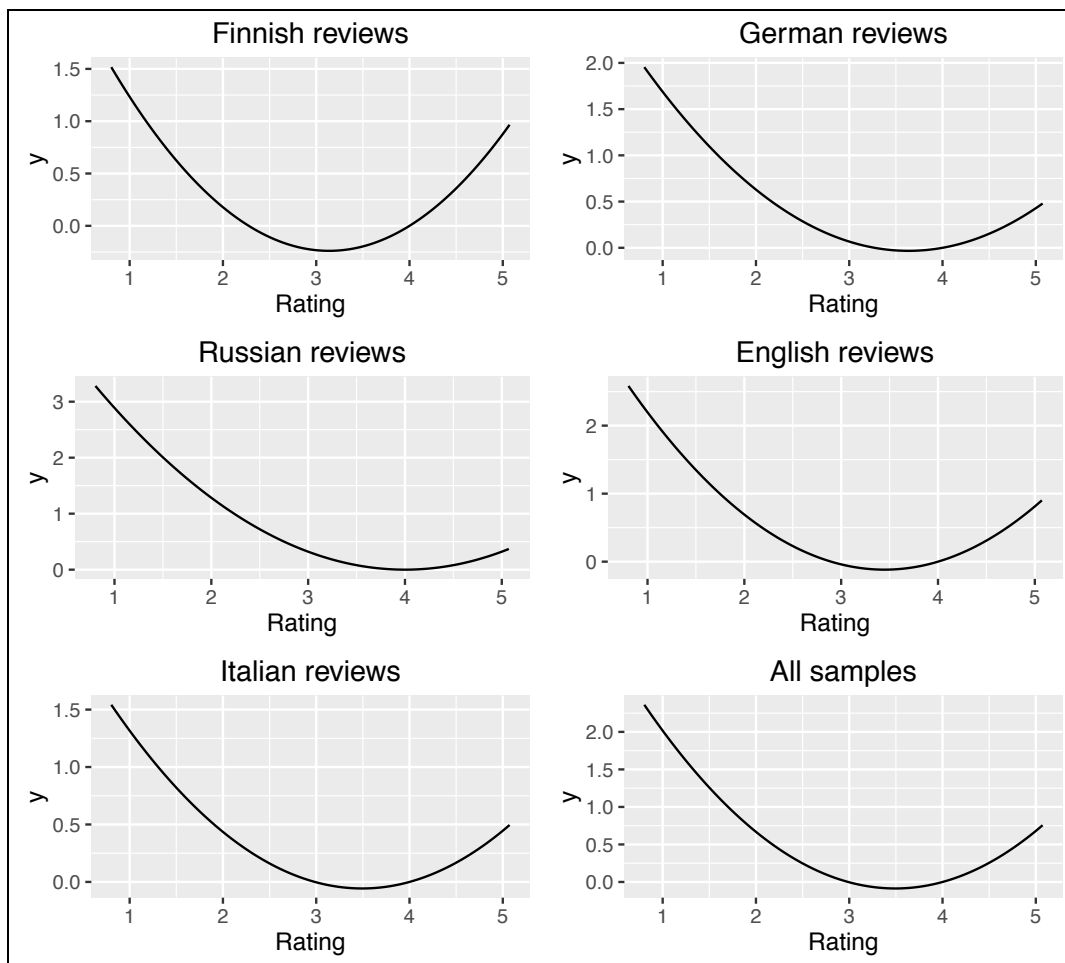


Figure 2. Visualizing the nonlinear impact of ratings on review helpfulness.

Further, even though Finnish, English, Italian, German and Russian subsamples reported significant curvilinear effects of rating on review helpfulness, the effects in fact differ a lot in terms of tendency to vote highly positive reviews. As shown in Figure 2, after controlled the difference in review depth, very positive Finnish reviews have a good possibility to be voted to be helpful, which is followed by English and Italian reviews. However, it seems to be not the case for Russian customers. Very negative reviews exhibit a dominant chance for Russian customers to vote helpful in comparison to very positive reviews, after the effect of review length is controlled.

6. IMPLICATIONS

The study has a number of implications for both research and practice. From a research perspective, the findings of the present study highlight a need to extend online review research to non-English review contexts. It is a fact that current research on online reviews is dominantly drawn from studying English reviews that are mostly obtained from the U.S. customer market. The significant differences in predictors between English and non-English review indicated a research gap in the field. On one hand, applying results obtained from studying English review to understand customers from other language backgrounds may be inappropriate. On the other hand, a dearth of studies on the review adoption behavior of non-English speaking customers exhibits an underestimated research issue, which deserves more research attention in the future. For instance, the behavioral difference in review utilization between Russian and other European customers, and between European and Asian users, exhibit a possibly fertile ground for more research endeavor.

From a perspective of practice, managers of online review sites should pay a sufficient attention to the demand of non-English speaking users. As shown in our study, non-English customers may have different preferences in terms of defining the helpfulness of online reviews. When tailoring interface for users who set up different language preference, special design efforts may be needed for website managers. Some users such as Russian exhibit a preference to vote for negative reviews; this may suggest a tendency for Russian users to utilize online review sites to eliminate bad alternatives. On the other hand, it is possible that, comparing to other customers, Russian users are loss-averse in hotel-choosing – they tend to deselect a hotel that has a very negative comment, and prefer a hotel that did not receive extremely negative review at all. Thus, a lack of extremely negative reviews may be a key criterion for many Russian users to select a hotel, albeit more research is needed to further valid this possible explanation. For Finnish users, it is possible that they use online review sites to appreciate both advantages and disadvantages of a hotel in order to form their decision. Users in general exhibit a clear preference over lengthier and extreme reviews. Thus, given the same rating on a product, a review with a longer text should be ‘topped’ in the list since it tends to be perceived more useful for users.

7. LIMITATIONS

The paper has a number of limitations. First, the study is based on studying in Finnish accommodation sector. Only a couple of language groups have been explored in the current study while languages like Chinese, Swedish and French have not been studied. Second, it is unclear whether there is any difference in review preference by comparing domestic versus international guests. Third, some users may use both English reviews and reviews of their native language in online review sites. Compared to the users who tend to read review written in their native language, these users may have different demographic features and are more international, thus exhibiting different preference over review helpfulness. Our data cannot discriminate the review helpfulness vote for bilingual users. Therefore, English subsamples may contain a mix of different language users albeit non-English subsamples tend to be more grounded on ‘local’ language users.

REFERENCES

- [1] Chua, A.Y.K. and Banerjee, S. Understanding Review Helpfulness as a Function of Reviewer Reputation, Review Rating, and Review Depth. *Journal of the Association for Information Science and Technology* 66, 2 (2014), 354–362.
- [2] Fang, H., Zhang, J., Bao, Y., and Zhu, Q. Towards effective online review systems in the Chinese context: A cross-cultural empirical study. *Electronic Commerce Research and Applications* 12, 3 (2013), 208–220.
- [3] Hofstede, G. and Bond, M.H. Hofstede's Culture Dimensions An Independent Validation Using Rokeach's Value Survey. *Journal of Cross-Cultural Psychology* 15, 4 (1984), 417–433.
- [4] Koh, N.S., Hu, N., and Clemons, E.K. Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications* 9, 5 (2010), 374–385.
- [5] Korfiatis, N., García-Bariocanal, E., and Sánchez-Alonso, S. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications* 11, 3 (2012), 205–217.
- [6] Kurt Hornik, Mair, P., Rauch, J., Geiger, W., Buchta, C., and Feinerer, I. The textcat Package for n-Gram Based Text Categorization in R. *Journal of Statistical Software* 52, 6 (2013), 1–17.
- [7] Leidner, D.E. and Kayworth, T. Review: A Review of Culture in Information Systems Research: Toward a Theory of Information Technology Culture Conflict. *MIS Quarterly* 30, 2 (2006), 357–399.
- [8] Liu, Z. and Park, S. What makes a useful online review? Implication for travel product websites. *Tourism Management* 47, (2015), 140–151.
- [9] Mudambi, S.M. and Schuff, D. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly* 34, 1 (2010), 185–200.
- [10] Paternoster, R., Brame, R., Mazerolle, P., and Piquero, A. Using the correct statistical test for the equality of regression coefficients. *Criminology* 36, 4 (1998), 859–866.
- [11] Wickham., H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2009.
- [12] Yin, D., Bond, S., and Zhang, H. Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews. *MIS Quarterly* 38, 2 (2014), 539–560.