Towards Assessing Online Customer Reviews from the Product Designer's Viewpoint

Mate Kovacs, Victor V. Kryssanov

Towards assessing online customer reviews from the product designer's viewpoint

Mate Kovacs[1] and Victor V. Kryssanov[2]

Ritsumeikan University, 525-8577 Kusatsu, Nojihigashi 1-1-1, Japan
gr0370hh@ed.ritsumei.ac.jp[1], kvvictor@is.ritsumei.ac.jp[2]

Abstract. Product reviews are a type of user-generated content that can be beneficial for both customers and product designers. Without quantifying the design knowledge present in product reviews, however, it is hard for the companies to integrate reviews into the design process. Several studies investigated review helpfulness in general, but few works explored the problem of review quality from the perspective of product designers. In this study, a theoretical model is presented and a system is proposed to assess the quality of online product reviews from the viewpoint of product designers. The system involves an original similarity-based metric to quantify the design information content of reviews on a continuous scale. Experiments are performed on a large number of digital camera reviews, with results indicating that the proposed system is capable of recognizing high-quality content, and would potentially assist companies in product improvement and innovation.

Keywords: eWOM · e-commerce · review quality and helpfulness · product design · information overload

1 Introduction

EWOM (Electronic Word of Mouth) can provide valuable information about customer needs, as it offers potentially useful knowledge not just for customers, but also for product designers. Presently, manufacturing evolves to become more customer-driven and knowledge-based [4]. Customer intelligence extracted from online product reviews can help manufacturers to improve their products by incorporating relevant information into the design process [8]. Typically, companies use interviews and surveys to obtain feedback from customers. Design knowledge extracted from product reviews differs from and has a complementary function to customer intelligence collected by traditional methods [29, 26]. The immense amount of reviews available at online platforms, however, makes it a challenging task for companies to obtain relevant information about product design. Popular and trending products often receive thousands of reviews from the customers, and review quality varies extensively through the large volume of reviews [13, 25]. Addressing this issue, often called *information overload*, is essential to effectively utilize customer reviews for product and service enhancement [9].

A reason for many data-mining projects being abandoned is the poor quality of the data used [12]. Review quality is seldom discussed in opinion mining studies [2], even though often most of the reviews appear practically useless from the designer's standpoint. Many e-commerce platforms introduced helpfulness-voting, where users rate other users' reviews, based on their helpfulness. These votes, however, are unavoidably influenced by the Matthew effect, as customers usually only read and vote for the top reviews, which will, thus, remain on top [22]. In fact, this kind of helpfulness score is often argued to be an unreliable measure of actual helpfulness and review quality [25, 28, 5, 3]. Another limiting factor of helpfulness-voting is the divergence between the helpfulness perceived by the customers and the helpfulness seen by the product designers [14]. Most of the studies dealt with review helpfulness only consider the customer's viewpoint, and limited work is available on quantifying design information of reviews to assist product designers and engineers.

The goal of the presented study is to reduce the information overload associated with customer reviews, and assess product review quality from the designers' standpoint in order to mine reviews that can potentially induce better design. The main contribution of this research is a theoretical model with a developed system using an original measure for quantifying review information at the design level without the need for manual feature engineering. Experiments are conducted on a large dataset of digital camera reviews, collected from Amazon US. Results obtained suggest that the proposed system can be used in practice effectively to assist companies in eliciting useful reviews.

The rest of the paper is organized as follows. Related work is presented in Section 2, while Section 3 describes the model and the system developed for assessing product review quality. Section 4 introduces the data used in this study. The experimental procedure is described in Section 5. Results obtained are interpreted, and the main findings are discussed in Section 6. Section 7 formulates conclusions and outlines future work directions.

2 Related work

Some of the related literature formulate a classification problem of review helpfulness (as an aspect of review quality) [7, 11, 15], and other works treat it as a regression or ranking problem [3, 24, 16]. Most of the studies dealt with several types of features, such as product features (key attributes of products), sentiment values (e.g. positive or negative), linguistic cues (e.g. the number of nouns, grammatical rules, review length, readability features, etc.), and user information (e.g. reviewer reputation, gender). Qazi et al. [18] considered also the review type to develop a model for helpfulness prediction. The authors conducted experiments on 1500 hotel reviews with results suggesting that the number of concepts in a review, and the type of the review (regular, comparative, or suggestive) influences review helpfulness. Saumya et al. [21] found that besides features extracted from review texts, customer question-answer data improves the prediction of review helpfulness, as perceived by the customers. Krishnamoorthy [7] proposed a

helpfulness prediction model based on review metadata, linguistic features, and also review subjectivity.

While the research on assessing review quality for product designers has been limited, there are still a number of notable studies dealing with the subject. Liu et al. [14] estimated the helpfulness of product reviews from the product designer's perspective, utilizing only the review text itself. The authors conducted an experiment to better understand what are the determinants of review helpfulness for product designers. Based on the results, four categories of features were identified to be important. These are linguistic features (e.g. the number of words), product features (attributes of a specific product), information quality-based features (e.g. the number of referred products), and information-theoretic features (e.g. review sentiment). The authors used regression to predict the helpfulness of reviews, and found that extracting these features using only the review content can help with identifying helpful reviews. Yagci and Das [26] argue that design intelligence helpful to both designers and customers can be extracted from product reviews. In their work, sentence-level opinion polarity determination (with categories of negative, neutral, and positive) was used together with noun-adjective and noun-verb association rules to extract the probable cause of a certain opinion. In a later work, the authors introduced the design-level information quality (DLIQ) measure for assessing the volume and quality of design knowledge of product reviews [27]. Reviews were evaluated based on content (the total number of words), complexity (the total number of sentences and nouns), and relevancy (the total number of nouns matching predefined design features), and promising results were obtained for assisting businesses in product development.

Most of the previous work do not differentiate between the helpfulness seen by customers and by product designers. Furthermore, nearly all of the related studies required manual feature engineering to obtain product features, and used noun and noun phrase matching to extract them from the reviews. One of the biggest issues of such methods is that the same feature can be expressed in various ways (explicitly or implicitly, with different words and phrases, etc.), and pattern matching approaches cannot account for such cases. Moreover, the presence of a specific word does not necessarily guarantee high-quality content, as word context plays a critical role in its interpretation. In the next section, an approach capable of dealing with these issues is introduced.

3 Proposed approach

3.1 Theoretical model

The approach proposed in this study builds on the assumption that any word of a language can appear in any kind of document. More formally, depending on for whom the document is targeted (target population) and what is the subject of the document (domain), words of a language appear with certain probabilities, and have weights indicating their importance. Let us denote the set of all documents d as $D = \{d_1, d_2, ..., d_n\}$, and the vocabulary of all words

w for language L as $L = \{w_1, w_2, ..., w_m\}$. Let us define a set of scalar weights $\Phi_V^T = \{\varphi_{V_1}^T, \varphi_{V_2}^T, ..., \varphi_{V_z}^T\}$, indicating the importance of words w for the domain V for a target population T. As follows,

$$(\forall V)(\forall T)[(\forall w)(w \in L) \wedge (\forall d)(d \in D)(\Diamond(w \in d)) \wedge (\exists \varphi_V^T)(\varphi_V^T \in \Phi_V^T)(w \rightarrow \varphi_V^T)], \quad (1)$$

where $\Diamond$ is the modal operator of possibility. For example, there is a chance that one encounters the word *sensor* in any kind of text, but if the domain is *digital cameras*, and the target population is the *product designer community*, this word has a high importance. If the domain-target combination is *programming-high school students*, the word *sensor* could still be important, but probably does not carry the same weight. On a similar note, the word *traditional* is probably not important for *digital cameras-product designer community*, but that does not mean it carries no information whatsoever about V and T, especially in the appropriate context. In fact, the underlying pragmatics will always have an impact on the interpretation of words. Word meaning in a natural language is defined by the context, and a huge part of the context depends on the domain and the target population. Thus, V and T also function as indicators of word meaning.

3.2 Review quality assessment

Based on the theoretical model proposed in Section 3.1, the inferential problem dealt with in this study is to approximate Φ_V^T for $T =$ *product designer community*, and for a certain domain V. Then, review quality would be estimated by measuring the distance between a review and Φ_V^T. In the presented study, technical documents of domain V are analyzed for the definition of Φ_V^T, as texts, which are in the interest of product designers are assumed to contain a high volume of technical content. Fig. 1 gives an overview of the proposed system for assessing review quality. There are two types of inputs involved, a database of technical documents, and a database of reviews, both from the same domain V.

Term dictionary formulation A collection of technical documents representing the product domain is cleaned from "unwanted" content (e.g. author information, bibliography, etc.), and tokenized to build a corpus of preprocessed sentences. To obtain an approximation of the set Φ_V^T, the sentences are used to select a large number of words with weights attached to them, based on their importance. All words are first lemmatized to obtain their dictionary form (e.g. studying, study, studies all becomes study), and stopwords are eliminated. The word weights constituting the set Φ_V^T are calculated, based on the mean sentence-wise term frequency-inverse document frequency (tf-idf) scores of the words. The statistical measure tf-idf is usually used to evaluate document-wise word importance. In this study, sentences are treated as individual documents, and the sentence-wise scores are averaged. Hence, the weight φ of word w is given by

$$\varphi(w) = \frac{1}{N} \sum_{i=1}^{N} f_{s_i}(w) \ln \left(\frac{N}{f_w(s)} \right), \quad (2)$$

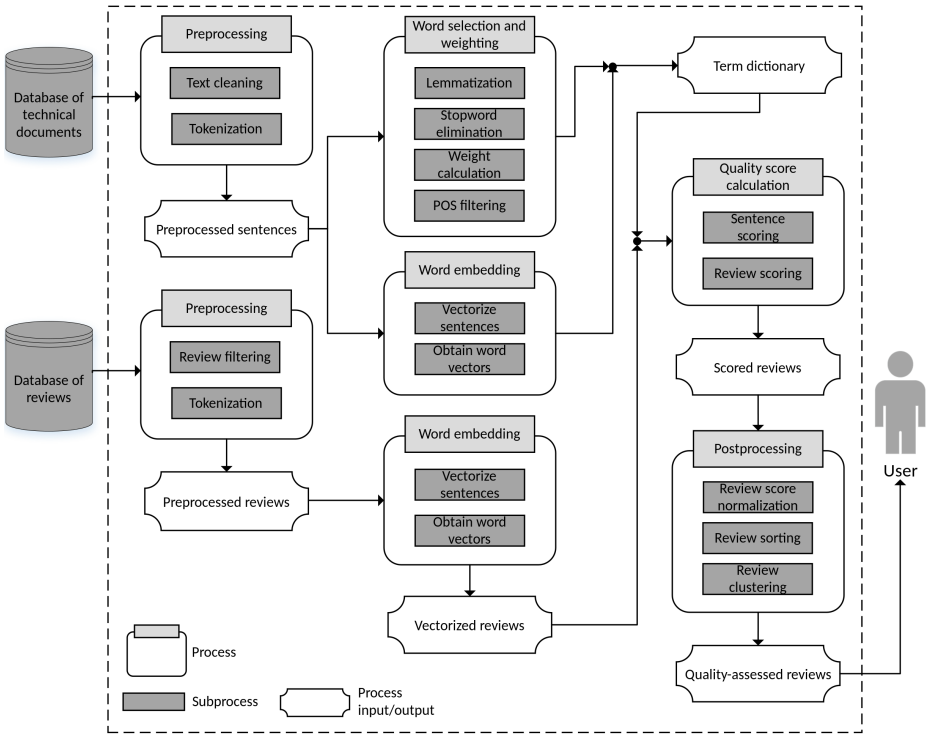

Fig. 1: The structure of the system proposed for assessing review quality

where $f_{s_i}(w)$ is the frequency of word w in sentence s_i, $f_w(s)$ is the number of sentences containing w, and N is the total number of sentences in the corpus. As all word tokens are weighted by the above equation, irrelevant Part of Speech (POS) are to be filtered out from the sentences. In previous work, few nouns and noun phrases were typically considered for candidate features. In the proposed system, all adjectives, adverbs, verbs, nouns and noun phrases are considered with their corresponding importance weights.

The preprocessed sentences are also used to create word embeddings for representing the words in a continuous vector space. Most word embedding methods are not contextualized, meaning that the vector representations of words are static (one vector per word). However, as word meaning in a natural language is context-dependent, word vectors should change, depending on the context. Embeddings from Language Models (*ELMo*) [17] uses an unsupervised, pre-trained deep bidirectional long short-term memory (BiLSTM) neural network model to compute contextualized word vectors in the following way:

$$ELMo_w = \gamma \sum_{j=0}^{L} b_j h_{w,j}, \tag{3}$$

where $h_{w,j}$ is the output of the jth layer L of the network for word w. The weight b_j is learned for each individual task (input sentence), and normalized by the softmax function. The parameter γ scales the word vectors for optimization. Since the model is initially pre-trained on a large amount of data, the network requires a single sentence to output context dependent vector representations of words. The ELMo embeddings obtained for all words are averaged for every unique word lemma remaining after POS filtering to acquire vectors describing V and T in the most accurate way possible. The selected words and their weights with their corresponding vector representations are stored together in a dictionary. These words will further be referred to as "terms".

Calculating review scores The other input of the system is a database of product reviews. First, one-word and non-English reviews are filtered out from the database, and the remaining texts are tokenized. Next, the preprocessed reviews are vectorized with the same method as the technical documents (ELMo). As the embedding vectors are context dependent, even when product features or components are described in different ways, the corresponding phrases and sentences will still have similar vectors. The vector representations of reviews are used together with the term dictionary to compute the quality scores of the reviews. The procedure of score calculation is specified by Algorithm 1. For each

Algorithm 1 Calculate review quality scores

```
procedure CALCULATE SCORES(reviews,terms)
    initialize array corpus_scores
    for all review ∈ reviews do
        initialize array review_scores
        for all sentence ∈ review do
            initialize sentence_score ← 0
            for all word ∈ sentence do
                if word ∉ stopwords then
                    initialize word_pertinence ← 0
                    for all term ∈ terms do
                        cos(θ) = (v_term · v_word) / (||v_term|| ||v_word||)
                        word_pertinence ← word_pertinence + φ_term^(1−cos(θ))
                    end for
                    sentence_score ← sentence_score + word_pertinence
                end if
            end for
            insert sentence_score into review_scores
        end for
        insert review_scores into corpus_scores
    end for
end procedure
```

review, the scores are calculated first on the word level, then on the sentence

level and, lastly, on the review level. Cosine similarities are calculated between vectors $\mathbf{v}$ of the observed word w and terms t in the term dictionary to assess their closeness. The similarity scores computed are subtracted from 1 to get the cosine distance between the observed word and the terms, that is used as the exponent of term weights for the word scores. This means, each word w will receive as many scores as the number of terms t in the term dictionary, which are then summed up to compute the *word pertinence*:

$$pertinence(w) = \sum_{k=1}^{t} \varphi_k^{1-cos(\mathbf{v}_k, \mathbf{v}_w)}. \tag{4}$$

With Equation 4, contributions of relevant words to the total score are much higher than irrelevant ones, and the dependence of word similarity the on context gets addressed. Sentence and review scores are defined as cumulated word pertinences and sentence scores, respectively. Thus, the final score of review r is obtained as

$$score(r) = \sum_{i=1}^{s} \sum_{j=1}^{w} \sum_{k=1}^{t} \varphi_k^{1-cos(\mathbf{v}_k, \mathbf{v}_j^i)}. \tag{5}$$

The reason for keeping the sentence-level scores is that knowing what sentences s contributed most to the final review score is often useful in practice.

Postprocessing The computed raw quality scores are between 0 and theoretically, infinity. This means that normalization is necessary to obtain easily interpretable results, and to establish lower and upper bounds for the quality assessment of future reviews. Additionally, reviews are to be sorted in a descending order to help the end-user choosing high-quality reviews. Rather than defining a threshold value for when a review would become helpful, 1-dimensional K-means is used to cluster the review scores into potentially meaningful groups, and help the user with the elicitation of high-quality reviews. The number of clusters k is determined by the elbow method [20]. As k increases, there will be a point where the improvement of the model starts declining. At that point, the sum of squared distances of the datapoints to the nearest cluster creates an "elbow of an arm" when plotted. The location of this elbow point indicates the optimal number of k.

4 Data

In the presented study, the shared domain V of technical documents and product reviews is *digital cameras*. The reviews used are part of Amazon review data [6]. The data includes 7,824,482 reviews from the "Electronics" category, written by customers of Amazon.com in the period from 1996 to 2014. Reviews of digital cameras and closely related products (e.g. lenses, battery chargers, etc.) were selected, using product subcategory tags and product IDs. Reviews with sentences longer than 50 tokens are presumably reviews without punctuation, that would

bias the review score and the overall results. Such reviews, for that reason, were eliminated from the dataset. The final review database used in the study consists of 300,170 product reviews, with 1,315,310 sentences in total. The technical text database was created from Wikipedia articles. Wikipedia contains a large quantity of technical information [23], and the articles are publicly available and downloadable from Wikimedia dumps[1]. All Wikipedia articles published until the 1st of Oct., 2018 were downloaded, and 1039 articles related to digital photography terminology, techniques, equipment, and product descriptions were extracted, using wikipedia tags and other metadata. Unrelated parts of the articles (e.g. "References", "See also", "History", etc.) were later removed from the technical document database. The ELMo language model[2] used in this study has been pre-trained on the 1 Billion Word Benchmark dataset [1] for word vectors of 1024 dimensions.

5 Results

From the total of 24,134 sentences in the technical document database, 13,166 unique word lemmas were derived into the term dictionary. To give a few examples, the top 20 terms obtained are as follows: *mount, sensor, model, focus, shutter, series, aperture, flash, photography, system, frame, light, design, zoom, mode, speed, exposure, format, specification, iso.* To unbiasedly compute the

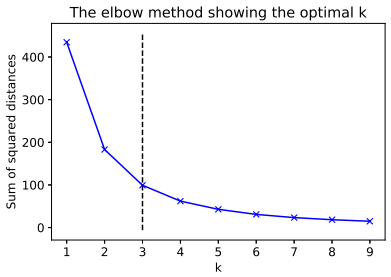

Fig. 2: Sum of squared distances for different number of k

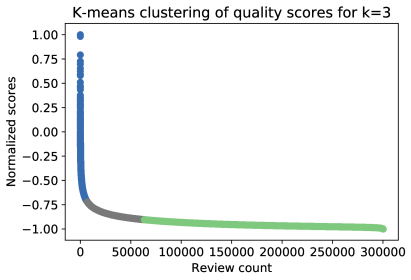

Fig. 3: Distribution of normalized and clustered review quality scores, with different colors indicating the three clusters

elbow point of the sum of squared distances for different number of K-means clusters, the algorithm called "Kneedle" [20] is used. Fig. 2 illustrates that the elbow point was detected at $k = 3$. Three clusters were, therefore, used to group the review quality scores. Fig. 3 shows the distribution of the sorted (descending order) review quality scores normalized between [-1,1], clustered by K-means.

[1] https://dumps.wikimedia.org

[2] The pretrained ELMo model was obtained from AllenNLP (https://allennlp.org).

The first cluster contains 6942 reviews with scores [1.0,-0.72], the second includes 56,550 reviews with scores (-0.72,-0.9], and the third has the rest of 236,678 reviews with scores (-0.9,-1.0]. Finally, Fig. 4 gives the distribution of sentence-wise quality scores normalized between -1 and 1.

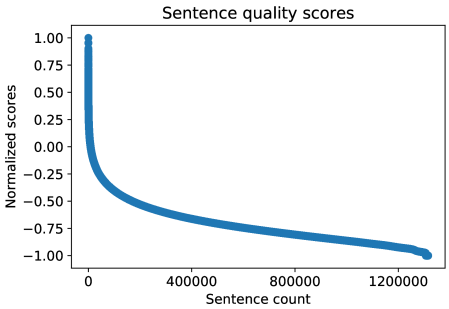

Fig. 4: Distribution of the normalized sentence quality scores

5.1 Comparision with human assessment

In order to predict the quality of a future review accurately, and to apply the system used in this study effectively in industrial settings, the system should utilize initially as many reviews as possible. However, this makes validation of the system a challenging task. As the scores define a ranking among the reviews, the three K-means clusters were labeled according to their ranks (1, 2, or 3, in the order of decreasing quality). 400 reviews were chosen randomly from each cluster to create the validation set of 1200 reviews.

The validation procedure used in this study is as follows. One review is chosen randomly from each cluster of the validation set, resulting in three reviews with a certain ranking among them established by the system. These reviews are then independently ranked by a human annotator (one of the authors), and Kendall's rank correlation coefficient is computed between the two rankings. The tau coefficient of Kendall measures the ordinal association between two rankings in the range of [-1,1] by

$$\tau = \frac{n_c - n_d}{n(n-1)/2}, \tag{6}$$

where n is the number of elements in each ranking, n_c is the number of concordant pairs, and n_d is the number of discordant pairs. $\tau = -1$ indicates a perfect inverse association, 1 implies 100% agreement, and 0 assumes no correlation between the rankings. Finally, the two-sided p-value for $H_0 : \tau = 0$ is computed to estimate the significance of the statistic. This process is repeated until every review is included in exactly one correlation calculation (400 iterations). The

final correlation score is computed by taking the mean of all individual τ, and the p-values are combined by Fisher's method to obtain the significance of the averaged coefficients. The results have been obtained are as follows: the averaged correlation between the two rankings $\tau = 0.827$, significant at $p < 0.001$.

6 Discussion

Unsurprisingly, words in the term dictionary with higher weights are important features of domain V. However, there are words in the top few hundred terms, which are not necessarily features in the strict sense, nevertheless are quite significant, owing to their meaning and expected context in V. Examples of such terms include *capture, interchangeable, back, short, depth, integrate, dark, compensate,* etc.

Camera brands, series names, and product description-like facts about measurements (e.g. Canon, uhd, ias, dx, d500, ev, mm) are often encountered in product reviews. For example, the sentence "Has a superb AF-S DX NIKKOR 16-80mm f/2.8-4E ED VR lens" is not very useful by itself, but the mere presence of these specification-like technical words would increase review score significantly. For this reason, the stopword list used in this study had to be extended with camera brand names, and all non-English words were also considered as stopwords.

The distribution of review scores (Fig. 3) reflects the fact that the number of reviews useful for the product designer community is very limited. The range of scores in the cluster with the highest review qualities is rather extensive. A reason for this is that reviews of topmost quality are highly detailed and particularly in-depth, yet extremely rare to encounter. The score range of the K-means clusters indicates that reviews with scores, for instance, 0 or -0.6 are still useful. Reviews in the second cluster can still be helpful for the product designers, but generally, these have a shorter review length, compared to the first cluster. The number of words and sentences in a review have been found to strongly correlate with review quality and helpfulness in the literature [10, 28, 19]. The same is observed in the results obtained in this study. As wordy reviews usually discuss more aspects of the product, these have high scores. While one could still find some useful information in the reviews from the third cluster, such reviews are very short, thus have low overall quality scores. A similar tendency can be observed for the sentence quality (Fig. 4), but the distribution of sentence scores is significantly more balanced. Accordingly, the transition between "high quality" and "low quality" is much more smooth and gradual, compared to the case of review qualities. This supports the validity of the idea of assessing quality scores not just on the review, but also on the sentence level. A few examples of review sentences with their corresponding scores are given in the next paragraph.

Evidently, there is an intersection between reviews important for the customers and those useful for the product designers. Product designers and more experienced customers would for example, both appreciate such a review sentence: "There are a few design and function annoyances (the silly sliding lens

door, proprietary rechargable battery rather than AA batteries, and difficulty in achieving intended effects with large apertures in aperture priority mode) but overall this is a great little camera that produces great images" (computed score: 0.8). Likewise, there are reviews and sentences which are generally irrelevant, like "This camera is awesome" (computed score: -1.0). Reviews and sentences dealing with delivery, retailers, Amazon, etc. can be helpful for the customers, but not so for the designer community. Therefore, these sentences have lower scores, such as "I then received a bill in the mail for more than the camera was worth and when I contacted them about this they said it was b/c I did not send in the proper paperwork" (computed score: -0.75). Unfortunately, this kind of reviews can be excessively long, and so their overall review scores can be higher than a short review with at least one piece of useful information. Individual sentence scores can help to reveal such cases, and assist the user to properly evaluate reviews. Reviews dealing with existing problems would help designers to improve the product, e.g. "The focus ring is a little on the narrow side but usable and it took a little time to get used to the zoom and focus rings being reversed (zoom on far end of lens - focus closer to camera body), opposite of the Canon lenses" (score 0.5). On the other hand, reviews praising some attributes of a product could be used for product innovation and customer need assessment, for instance, "Super fast lens, great telephoto reach, numerous creative modes, intuitive and easy-to-use features are attributes of this camera and makes this an attractive alternative to carrying & switching different lenses for different photo shoots, or different subject compositions" (computed score: 0.56).

Even if a review involves only a small amount of relevant content, the information present could still be extremely significant. Thus, it can happen that a shorter review discussing only one attribute of a product is more useful than an in-depth review. Usually, this was the reason for the discrepancies between the human and system rankings. 52.88% of the ranking differences occured between ranks 1 and 2, 45.19% between ranks 2 and 3, and 1.93% between ranks 1 and 3. Nevertheless, the obtained value of the correlation coefficient τ suggests that the system proposed in this study can efficiently differentiate between high- and low quality reviews. This suggests that besides eliciting potentially helpful reviews for product designers, the system can be used to obtain high-quality datasets for other data mining purposes, such as sentiment analysis, text summarization, etc.

7 Conclusions

In this work, the problem of online review quality was examined from the product designer's viewpoint. The presented study offers contributions both conceptually and methodologically to the field of review quality estimation. In order to deal with the information overload of online product reviews, a theoretical model was proposed, and a system was developed to quantify design knowledge in reviews without human involvement. Experiments were conducted on a large number of digital camera reviews from Amazon US, with results indicating that the system

would potentially help companies improving their products, and to focus on customer-driven product innovation.

Future work should extend this study by using more technical documents for term dictionary development (such as product manuals), to obtain a better approximation of Φ_V^T. As sentence-wise quality assessment is more practical than focusing on entire reviews, a sentence-level review analysis tool could be developed to assist product designers in a user-friendly manner. Furthermore, a more refined evaluation of the proposed system is necessary to examine the validity of in-cluster rankings.

References

1. Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P.: One billion word benchmark for measuring progress in statistical language modeling. Computing Research Repository (CoRR) pp. 1–6 (2013)
2. Chen, C.C., Tseng, Y.D.: Quality evaluation of product reviews using an information quality framework. Decision Support Systems **50**(4), 755–768 (2011)
3. Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., Lee, L.: How opinions are received by online communities: A case study on Amazon.Com helpfulness votes. In: Proceedings of the 18th International Conference on World Wide Web. pp. 141–150 (2009)
4. Ferreira, F., Faria, J., Azevedo, A., Marques, A.L.: Product lifecycle management in knowledge intensive collaborative environments. International Journal of Information Management **37**(1), 1474–1487 (2017)
5. Ghose, A., Ipeirotis, P.G.: Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering **23**(10), 1498–1512 (2011)
6. He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of the 25th International Conference on World Wide Web. pp. 507–517 (2016)
7. Krishnamoorthy, S.: Linguistic features for review helpfulness prediction. Expert Systems with Applications **42**(7), 3751–3759 (2015)
8. Ku, Y.C., Wei, C.P., Hsiao, H.W.: To whom should i listen? finding reputable reviewers in opinion-sharing communities. Decision Support Systems **53**(3), 534–542 (2012)
9. Lee, H., Choi, K., Yoo, D., Suh, Y., Lee, S., He, G.: Recommending valuable ideas in an open innovation community: A text mining approach to information overload problem. Industrial Management & Data Systems **118**(4), 683–699 (2018)
10. Lee, S., Choeh, J.Y.: Predicting the helpfulness of online reviews using multilayer perceptron neural networks. Expert Systems with Applications **41**(6), 3041–3046 (2014)
11. Liu, H., Hu, Z., Mian, A.U., Tian, H., Zhu, X.: A new user similarity model to improve the accuracy of collaborative filtering. Knowledge-Based Systems **56**, 156–166 (2014)
12. Liu, Q., Feng, G., Wang, N., Tayi, G.K.: A multi-objective model for discovering high-quality knowledge based on data quality and prior knowledge. Information Systems Frontiers **20**(2), 401–416 (2018)

13. Liu, Y., Huang, X., An, A., Yu, X.: Modeling and predicting the helpfulness of online reviews. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. pp. 443–452. ICDM '08 (2008)
14. Liu, Y., Jin, J., Ji, P., Harding, J.A., Fung, R.Y.K.: Identifying helpful online reviews: A product designer's perspective. Computer-Aided Design **45**(2), 180–194 (2013)
15. Malik, M., Hussain, A.: Helpfulness of product reviews as a function of discrete positive and negative emotions. Computers in Human Behavior **73**, 290–302 (2017)
16. Mukherjee, S., Popat, K., Weikum, G.: Exploring latent semantic factors to find useful product reviews. In: Proceedings of the 2017 SIAM International Conference on Data Mining. pp. 480–488 (2017)
17. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2227–2237 (2018)
18. Qazi, A., Shah Syed, K.B., Raj, R.G., Cambria, E., Tahir, M., Alghazzawi, D.: A concept-level approach to the analysis of online review helpfulness. Computers in Human Behavior **58**(C), 75–81 (2016)
19. Salehan, M., Kim, D.J.: Predicting the performance of online consumer reviews. Decision Support Systems **81**(C), 30–40 (2016)
20. Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops. pp. 166–171 (2011)
21. Saumya, S., Singh, J.P., Baabdullah, A.M., Rana, N.P., Dwivedi, Y.K.: Ranking online consumer reviews. Electronic Commerce Research and Applications **29**, 78–89 (2018)
22. Singh, J., Irani, S., Rana, N., Dwivedi, Y., Saumya, S., Roy, P.: Predicting the "helpfulness" of online consumer reviews. Journal of Business Research **70**, 755–768 (2017)
23. Talukdar, P.P., Cohen, W.W.: Crowdsourced comprehension: Predicting prerequisite structure in wikipedia. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. pp. 307–315 (2012)
24. Tang, J., Gao, H., Hu, X., Liu, H.: Context-aware review helpfulness rating prediction. In: Proceedings of the 7th ACM Conference on Recommender Systems. pp. 1–8 (2013)
25. Tsur, O., Rappoport, A.: Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In: Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009 (2009)
26. Yagci, I.A., Das, S.: Design feature opinion cause analysis: a method for extracting design intelligence from web reviews. International Journal of Knowledge and Web Intelligence **5**(2), 127–145 (2015)
27. Yagci, I.A., Das, S.: Measuring design-level information quality in online reviews. Electronic Commerce Research and Applications **30**, 102–110 (2018)
28. Yang, Y., Yan, Y., Qiu, M., Bao, F.: Semantic analysis and helpfulness prediction of text for online product reviews. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics ACL. pp. 38–44 (2015)
29. Yu, X., Liu, Y., Huang, X., An, A.: Mining online reviews for predicting sales performance: A case study in the movie domain. IEEE Transactions on Knowledge and Data Engineering **24**(4), 720–734 (2012)