Summer 5-27-2016

A Model-Driven Method for Quality Reviews Detection: An Ensemble Model of Feature Selection

Hongwei Wang
School of Economics and Management, Tongji University, Shanghai, 200092, China, hwwang@tongji.edu.cn

Yuan Meng
School of Economics and Management, Tongji University, Shanghai, 200092, China

Pei Yin
School of Business, University of Shanghai for Science and Technology, Shanghai, 200093, China

Jin Hua
Saic General Motors Sales Co., LTD, Shanghai, 201206, China

A Model-Driven Method for Quality Reviews Detection: An Ensemble Model of Feature Selection

Hongwei Wang[1], *Yuan Meng*[1], *Pei Yin*[2], *Jin Hua*[3]
[1]School of Economics and Management, Tongji University, Shanghai, 200092, China
[2] School of Business, University of Shanghai for Science and Technology, Shanghai, 200093, China
[3]Saic General Motors Sales Co., LTD, Shanghai, 201206, China

Abstract: With the rapid growth of e-commerce and user-generated content online, the increasing product online reviews have significant influence on both buyers and sellers. However, among the thousands of online reviews, only the reviews of high-quality matters to the market, thus quality reviews detection rises in response to the requirement of retrieving authentic feedbacks from consumers. In this paper, a state-of-the-art ensemble model, gradient boosting decision trees (GBDT), is applied to select useful features for quality evaluation of online reviews. Firstly, four types of features are extracted based on information adoption theory. Then, the GBDT model is adopted to select useful features for quality reviews detection. At last, comparative experiments are conducted through online reviews of searching goods, based on two baseline models such as Decision Tree and Logistic Regression, and the results show that GBDT model achieves a better performance in detecting reviews of high-quality. This research indicates that product attributes, reviewer characteristics and objectiveness of reviews are key ingredients in high quality reviews.

Keywords: review quality, ensemble model, feature selection, information adoption theory, gradient boosting decision trees (GBDT)

1. INTRODUCTION

User-generated content such as online reviews have become a prevalent means of communication between customers and business in recent years, generating big influence on product market. However, among the vast amount of online reviews, only quality reviews conveys authentic feedbacks from consumers. Therefore, quality reviews detection becomes a topic of interest for both academia and industry.

A commonly-used measurement of review quality by most e-commerce websites is the ‘helpful votes’ of each review based on the evaluations of whether one review is helpful or not to its readers. But this method is somewhat subjective and often comes with manipulation and bias. Therefore, researchers are determined to develop more objective methods to assess the quality of reviews.

There are mainly two methods to evaluate review quality, the econometrical model based method and the machine learning method. Econometrical model based method provides interpretable results but typically tends to be insufficient in feature selection resulting in poor performance, while machine learning method normally achieves better performance but hardly provide evidence to support feature selection. As a result, this paper is motivated to tackle the problem using ensemble model, which closes the gap of accuracy and interpretability.

The remainder of this paper is organized as follows. The existing literatures are reviewed in ‘Literature review’ section and the method is introduced in ‘Proposed Approach’ section. Experiments on quality reviews detection are conducted and the results are estimated in ‘Experiments and Evaluations’ section. The paper is concluded with possible further research in ‘Conclusion’ section.

* Corresponding author. Email: hwwang@tongji.edu.cn(Hongwei Wang)

2. LITERATURE REVIEW

2.1 The impact of different types of features on review quality

The existing researches analyze the influence of several types of features on review quality including linguistic features, semantic features, product attributes, meta-data features and information features. For example, Reference[1] verifies that the linguistic features such as the number of words adjectives and sentences has impact on the quality of reviews. High quality reviews should evaluate a certain amount of product attributes so that conveys substantial information for readers[2]. The metadata features, such as each reviewer's rating and its variance to average rating, are also important for evaluating to the helpfulness of reviews[3]. Reference[4] argues that interactive Electronic Word-of-Mouth (EWOM) systems positively influence the quality of reviews on website. Besides that, information gain of each review is of help to assess review quality[1][5].

2.2 Review quality evaluation models

There are mainly two kinds of models to evaluate review quality, the econometrical model and the machine learning model. The econometrical model based method often takes the metadata features (e.g. review ratings and reviewers' identifications) or the linguistic features (e.g. the number of words and sentences) as independent variables, and takes the ratio of 'helpful votes' to total votes as a proxy dependent variable of review quality, in order to analyze the influence of these features on the quality of reviews. For example, Reference[6] uses a multiple linear regression model to analyze the helpfulness of DVD product reviews, and the results shows that the reviewers' characteristics and linguistic features of a review have significantly positive impacts on the review's helpfulness. Similar researches are conducted in [7]-[9]. On the other hand, the machine learning method considers the quality evaluation of reviews as a two-class classification problem, and classifies reviews into the ones of high-quality and the ones of low quality by manually annotating training set, training classifier to determine the category of testing set. Commonly used machine learning methods are support vector machine, support vector regression, decision tree, logistic regression, etc, as in [10]-[13]. In summary, although the machine learning methods performs better on review quality evaluation than the econometric models, they rarely analyzes the extent of influence that features have on the evaluation.

2.3 Comments on related works

Based on reviewing literatures above, the performance of review quality evaluation is heavily relied on selecting useful features, but there is no unanimous conclusion on the features determining the quality of reviews. Besides that, just few researches explain the contribution that each feature makes to determine review quality and in what way it contributes. Furthermore, using the ratio of helpful votes to total votes as the proxy for review quality is somewhat heuristic.

Therefore, this paper is motivated to tackle these problems above and proposes a method based on an ensemble model instead of a single model to select features and detect high quality reviews.

3. PROPOSED APPROACH

3.1 Basic procedure of the proposed approach

A state-of-the-art ensemble method, gradient boosting(GB), is utilized and regression trees are employed as base learners. First of all, based on information adoption theory, four types of features are put forward to measure the quality of review, including linguistic features, semantic features, information features and reviewer relevant features. Then, GB models are used to select key features and analyze their contributions in the process. At last, comparative experiments are conducted with baseline methods. The basic procedure of the proposed approach is shown in Fig. 1.

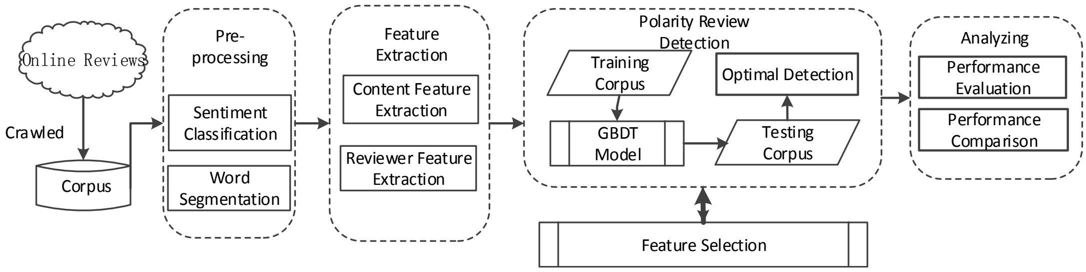

Figure 1. The basic procedure of the proposed approach

3.2 Feature extraction

According to the Information Adoption Theory proposed by Sussanman[14] in 2003, information content and information source reliability are the two direct factors affecting the perceived helpfulness of information for recipients in the context of online information communication. And the level of the perceived helpfulness of information directly affects the recipients' adoption of information. As a result, information content including linguistic features, semantic features and information features and information source reliability such as reviewer identification features are employed to measure the quality of reviews.

(1) Linguistic features(F1) extraction

Linguistic features are text features regarding words and sentences in reviews.. After preprocessing, the total number of words, sentences, adjectives, adverb words and verb words in a review and the average sentence length of the review are extracted as linguistic features.

(2) Semantic features(F2) extraction

Reviews are always mixed with objective descriptions and subjective opinions. And it is obvious that the more similar the review is to product description, the more objective the review will be. Subjective opinion is the positive or the negative sentiment towards the product expressed in a review[15], hence is quite different to the description of a product. In addition, the sentiment of a review is determined by the polarity words in the review in this paper.

In order to measure the effect of the mixture of opinions to the review quality, PosDegree and DevPos are defined as semantic features to represent the sentiment mixing level of a review. And ObjDegree and DevObj are defined as semantic features to represent objectivity of a review.

- Sentiment mixing level of a review

Consumers usually praise some aspects while criticize other aspects of a product, thus making reviews a mixture of positive and negative comments. In this paper, the percentage of positive sentences in a review is denoted as the review's PosDegree. And the larger the percentage is, the more positive the review will be. On the contrary, the smaller the percentage is, the more negative the review will be. In other words, PosDegree represents the valance of the mixed sentiments of a review. r^+ is used to denote positive sentences in a review, *total(r)* is used to denote the sum of positive and negative sentences, and PosDegree of review *r* is calculated as follows.

$$PosDegree(r) = \frac{count(r^+)}{total(r\ )} \tag{1}$$

Since the average PosDegree of product *p* reflects a steady value of the sentiment mixing level of all *p*'s reviews, the deviation of PosDegree to the average PosDegree of *p* is used to represent the sentiment mixing level of review *r*, and DevPos of review *r* is calculated as follows.

$$DevPos(r) = | PosDegree(r) - Avg(\sum_{r \in R} PosDegree(r) | \tag{2}$$

A supervised learning method, Bernoulli NB model(accuracy reach to 92.5% in this paper), is employed to determine the polarity of a sentence.

- objectivity of a review

As mentioned above, the objectivity of a review is determined by the similarity degree of the review to merchant's product description. Since a review is composed of several sentences either are subjective or objective, the objectivity of a review is predicted at the sentence level.

The tf-idf method is adopted to obtain the text vector space of sentence *s* and description text *d*, and the cosine similarity algorithm is employed to calculate the similarity between the two .Set a threshold value, and the *s* with greater value than the threshold is taken as an objective sentence, denoted as s^+.

Thus, we get each review's objectivity, ObjDegree, as follows.

$$ObjDegree(r) = \frac{count(s^+)}{total(r)} \tag{3}$$

Similarly, the average *ObjDegree* of product *p* reflects a steady value ofthe objectivity of all *p*'s reviews. Thus, the deviation of *ObjDegree* to the average *ObjDegree* of *p* is used to reflect the objectivity of review r, and DevObj of review *r* is calculated as follows.

$$DevObj(r) = | ObjDegree(r) - Avg(\sum_{r_i \in R} ObjDegree(r) | \tag{4}$$

(3) Information features(F3) extraction

The information features include the amount of information of a review, review's timeliness, review's depth, and etc. Most related researches utilize information entropy to measure the amount of information contained in reviews. In addition, Jelinek[16] gives definition to information perplexity based on information entropy. Therefore, both information entropy and information perplexity are introduced as the information features to measure the amount of information of a review. The higher the value of a review's entropy and perplexity are, the more distinct the review will be from the other reviews. Suppose a review *r* is composed of word $w_1, w_2, \ldots w_n$, and $p(w_i)$ is the occurrence probability of w_i in the corpus, the entropy and the perplexity of review *r* are determined as follows.

$$Entropy(r) = -\sum_{w_i \in r} p(w_i) \log_2 p(w_i) \tag{5}$$

$$Perplexity(r) = 2^{Entropy(r)} \tag{6}$$

After preprocessing of corpus, a unigram language model is trained using the Natural Language Tool Kit package in Python to calculate the entropy and the perplexity of each review.

As consumers always give comments to different attributes of a product in the reviews, the more product attributes mentioned in a review, the more detailed information the review will have. As a result, the frequency of product attributes is also adopted to measure the amount information of a review. At first, based on the method of extracting product attributes proposed by Hu[17], a product attributes set is built based on association rules. Then, point mutual information(PMI) method is introduced to filter incorrect product attributes in the preliminary set , such as " problem", "aspect", "condition", "reason". In the experiment, 20 representative words are manually selected from the preliminary set to constitute a seed set of product attributes. We filter incorrect product attribute based on the sum of PMI value of every word in the preliminary set.

Furthermore, the number of days from a review's publishing date to the experiment date is also collected as a feature to measure timeliness of a review.

(4) Reviewer reliability features extraction

As the opinion holder of a review, reviewer's reliability represents the source reliability of the review. Thus reviewer's ranking, the number of reviews with the 'helpful votes' above five, and reviewer's average 'helpful vote' rate of all one's reviews are extracted as reviewers' relevant features to measure the reliability degree of them..

To sum up, 17 features belonging to those four categories above are extracted. The specific meaning and abbreviation of each feature is illustrated in table 1.

Table 1. Extracted features and abbreviation

Feature set	Feature implication	Abbreviation
Liguistic features F1	number of words in a review	Nwords
	number of sentences in a review	Nsents
	average sentence length: （Nwords/Nsents）	Averlen
	number of adjectives in a review	Nadj
	number of adverbs in a review	Nadv
	number of verbs in a review	Nverb
Semantic features F2	positive opinion level	PosSenti
	sentiment mixing level	DevPos
	objective level	ObjSenti
	objectivity level	DevObj
Information features F3	number of product attributes	AttriFreq
	entropy of each review	Entropy
	perplexity of each review	Perplexity
	number of days from publish date to experiment date (the logarithm value)	Timeliness
Reviewer reliability features F4	reviewer ranking	ReviewerRank
	reviewer's average helpful vote rate	AverHelpRate
	number of the same reviewers' review (number of helpful votes more than five)	VoteAbove5

3.3 Feature selection based on Gradient Boosting Model

Gradient Boosting (GB) algorithm is an ensemble of models, originally proposed by Friedman[18], which is one of the most effective machine learning models for predictive analytics. To train GB model, this paper consider *m* reviews with extracted features as input measurements and their corresponding reviews qualities as outcomes in the form of $(x_1,y_1),\ldots,(x_m,y_m)$ pairs, where each $\{x_i\}_{i=1}^{m}$ is a vector containing extracted features for review *i*. Based on that, the unknown function $y = F(x^*)$ is utilized to predict the quality of a new review. The unknown function *F* is estimated by minimizing a loss function *L* defined over training set as shown in equation (7).

$$F = \arg\min_F \sum_{i=1}^{m} L(y_i, F(x_i)) \tag{7}$$

The gradient of the loss function *L* for each training point x_i at the iteration step *n* is given by equation (8).

$$g_{i,n}(x_i) = \nabla_{F_{n-1}(x_i)} L(y_i, F_{n-1}(x_i)), \qquad 1 \le i \le m. \tag{8}$$

In order to generalize the gradient to other *x*, a regression tree $h(x, a_n)$ is chosen to produce $h_n = \{h(x_i, a_n)\}_1^m$ most parallel to $-g_n \in R^m$. The regression tree is obtained from equation (9)

$$a_n = \arg\min_{a,\beta} \sum_{i=1}^{m} \left[-g_n(x_i) - \beta h(x_i, a) \right]^2 \qquad (9)$$

Where a_n denotes the parameter of the regression tree $h(x, a_n)$ and β denotes the learning rate, which determines the contribution of each tree to the approximation. The regression trees that are most highly correlated with $-g_n(x)$ over the data distribution are estimated and used to update the approximation of F_n as shown in equation (10).

$$F_n(x) = F_{n-1}(x) + \gamma_n h(x, a_n) \qquad (10)$$

Where γ_n is the optimal length, which is obtained by equation (11).

$$\gamma_n = \arg\min_{\gamma} \sum_{i=1}^{m} L(y_i, F_{n-1}(x_i) + \gamma h(x_i, a_n)) \qquad (11)$$

In order to optimize the performance of GBDT model, an internal validation is carried out on testing data to find the optimal GB model parameters. Besides that, using the internal mechanism of GB, the relative importance of extracted features are compared and ranked based on the prediction result.

4. EXPERIMENTS AND EVALUATIONS

4.1 Experiment setup

Mobile reviews in Chinese are collected from the Amazon website (Http://www.amazon.cn) as corpus using a crawler. Besides that, the metadata of each review (such as reviewer's relevant information and helpful vote) and product description are also crawled to build the experimental data set. After removing duplicate and feature missing reviews, the experimental data set contains 1421 pieces of mobile phone reviews across 10 different brands.

In order to get the actual quality of reviews as much as possible, both automatic labeling based on 'helpful votes' and manual annotation are adopted to determine whether a review is of quality or not. At first, the reviews with the number of 'helpful votes' above or equal to 5 and the ratio of 'helpful votes' to total votes above or equal to 0.75is automatically labeled as high-quality. And then, five researchers are invited to manually annotate the rest of reviews with 'helpful votes' less than 5, the reviews with more than 4 researchers annotated as high-quality is determined to have high quality, and the rest are annotated as low quality reviews. Finally, 801 high quality reviews and 620 low quality reviews are annotated.

80% of the experiment data are used as training set and the rest are used as test set. 10 fold cross verification is employed to test the three models. And the area under the ROC curve (AUC) and the accuracy of predicting quality reviews are utilized to evaluate their performances.

4.2 Comparison on relative importance of features

As shown in Fig.2, of all the features, the frequency of product attributes (AtttriFreq) is of most importance to quality reviews detection, followed by reviewer's ranking (ReviewerRank), the number of verb words review (Nverb), average helpfulness of the reviewer's all reviews(AverHelpRate). In contrast, adverb words number (Nadverb), sentence number(Nsents), sentiment mixing level of each review (PosSenti) are of minimal importance to the task.

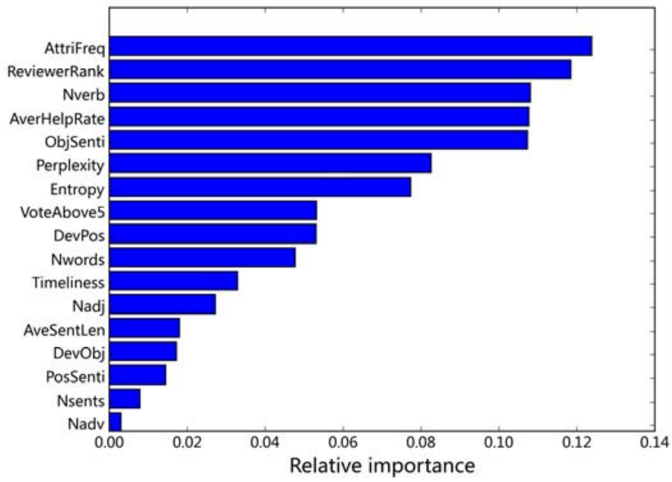

Figure 2. Comparison on relative importance of features

4.3 Influences of key features on review quality

To analyze the influence of each key feature on review quality, the partial dependence of a single feature to model is introduced and the influence of the other features is marginalized out. The top feature in each feature set as shown in Fig.2 is selected to be analyzed, and the results are shown in Fig.3.

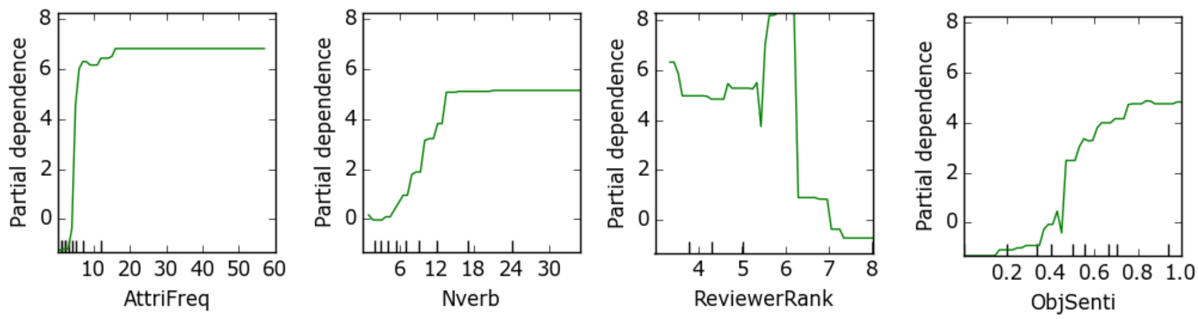

Figure 3. Partial dependence of key features

The result shows that the positive influence of the frequency of product attributes on review quality increases until the number reaches 20, and the impact stays the same after that. This indicates that with a limited number of product attributes mentioned in a review, the more attributes commented, the more information provided for readers and the more perceived usefulness. But after the number reaches a threshold, overly-crowded product attributes in one review might means too much for readers to digest.

As shown in Fig.3, most reviewers are inactive and have the same ranking (specifically, most reviews are given ranking 8 by the Amazon website). When a reviewer is ranked between 5 and 6, the reviewer's influence on review quality reaches the highest, whereas when a reviewer is ranked over 6, its influence is decreasing rapidly to a non-obvious degree.

Since ObjSenti measures the objectivity of each review, the result indicates that the more objective the review is, the more helpful the review will be. And Fig.3 also shows that the number of verbs has a linear increasing effect on review quality when it is less than 15, and after that, the influence remains unchanged.

4.4 The performance of GBDT model on quality reviews detection

As shown in Fig.4, the deviance of training set is decreasing as the progress of gradient boosting is proceeding. The AUC also indicates that the performance of predicting quality reviews is improving as the number of trees increases and gains the best results as the trees number reach 1997 and tree depth is set at 3. Fig.5 illustrates that the ROC curve for quality reviews detection based on GBDT model. The ROC curve is close to the upper left of axis, indicating a good prediction result

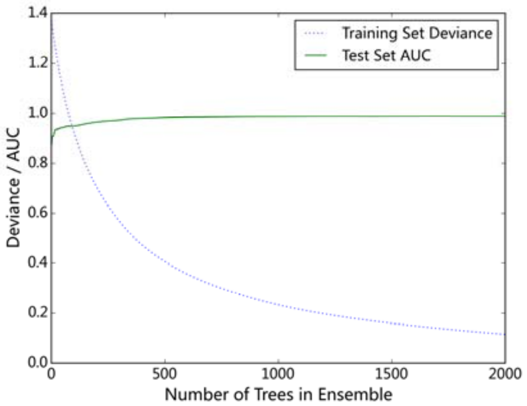

Figure 4. AUC versus ensemble size of GBDT model

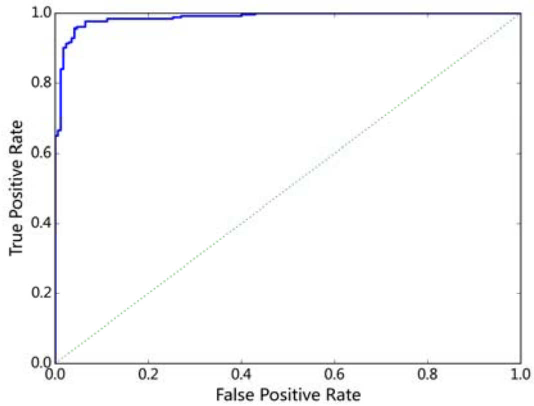

Figure 5. ROC curve of GBDT model

4.5 Comparative experiments

Comparative experiments have been conducted based on two baselines and the approach proposed in this paper, in order to justify the effectiveness of the proposed method in obtaining a better performance in quality review detection. And the two baselines are conventional logistic regression and Decision Tree.

The performances of two baselines and the proposed approach on different feature sets are compared based on the combination of four feature sets such as F1, F1+F2, F1+F2+F3, F1+F2+F3+F4. 10-fold prediction accuracy and AUC value are listed in table 2, in which DT、LR、GB are abbreviations of Decision Tree、Logistic Regression and Gradient Boosting Decision Tree.

Table 2. Comparison of two baselines and the proposed approach

Feature Sets	Accuracy			AUC		
	DT(%)	LR(%)	GB(%)	DT(%)	LR(%)	GB(%)
F1	67.00	75.67	75.67	69.18	75.76	75.60
F1+F2	70.50	76.00	76.23	70.30	75.78	76.45
F1+F2+F3	77.67	78.67	83.33	77.60	78.67	83.19
F1+F2+F3+F4	85.67	87.67	**94.67**	84.61	88.08	**94.53**

As shown in table 2, the proposed approach outperforms DT model based method and LR model based method in both prediction accuracy and AUC value. In addition, the best performance of quality review detection is achieved using all four feature sets as F1+F2+F3+F4.

5. CONCLUSION AND DISCUSSION

This paper is focused on feature selection for quality review detection. The proposed approach in this research firstly extracts four types of features are extracted based on information adoption theory, then a gradient boosting decision tree model is introduced to feature selection, and finally comparative experiments are conducted through DVD online reviews in Chinese from the Amazon website, with two baseline models such as Decision Tree and Logistic Regression. The experiment results demonstrate that the proposed approach achieves a better performance in detecting quality reviews. Furthermore, it is also indicated that the frequency of product attributes mentioned in a review, the reviewer's rank, the objectivity of a review, and the number of verbs in a review are the most essential features in determining the quality of the review.

Further research will be conducted in the following aspects. First of all, more features should be extracted to for quality review detection, for example, comparative reviews on different products may provide additional information for readers. Besides that, the experiments in this research are conducted merely on corpus of searching goods, the reviews of experience goods (e.g. restaurant and etc.) will be collected and experimented on in the future.

ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China under Grant 70971099, 71371144 and Shanghai philosophy and social science planning projects 2013BGL004.

REFERENCES

[1] Liu Y., Jin J., et al. (2013). Identifying helpful online reviews: A product designer’s perspective. Computer- Aided Design,, 45:180-194

[2] Chen C.C., Tseng Y.-D.(2011).Quality evaluation of product reviews using an information quality framework.Decision Support Systems , 50:755-768

[3] Kim S.M., Pantel P., et al. (2006). Automatically Assessing Review Helpfulness. In: Proceeding of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney: ACL Press, 423-430

[4] Yoo C.W., Kim Y.J., Sandersc G.L.(2015). The impact of interactivity of electronic word of mouth systems and E-Quality on decision support in the context of the e-marketplace. Information & Management, 52(4):496-505

[5] Zhang R, Tran T. (2011).An information gain-based approach for recommending useful product reviews. Knowledge and Information Systems, 26(3): 419–434

[6] Ghose A, Ipeirotis G P, et al. (2011).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

[7] Hao Y.Y., Ye Q., Li Y.J. (2010). Research online impact factors of customer reviews usefulness based on movie reviews data.Journal of Management Sciences in China, 13(8):78-96

[8] Gao Y.，Li H., Shi H.B.(2012).The study of factors influencing online review votes.China Management Informationization,2012, 15(17):88-91

[9] Yan J. Y., Zhang L., Zhang L.(2012). An Empirical Study of the Impact of Review Content on Online Reviews Helpfulness in E-commerce.Information Science, 30(5): 713-719

[10] Yang S.(2013). The Imp act Mechanism of Information Quality and Community Status on Perceived Usefulness for User-Generated Product Reviews——Tobit Regression Analysis.Management Review, 25(5):136-143

[11] Yin G..P.(2012). How do consumers think what kind of reviews are more useful——Based on the effects of social factors. Management World, 2012(12): 115-124

[12] Liu J J, Cao Y B, Lin C Y, et al. (2007) .Low-quality product review detection in opinion summarization. Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague: ACL Press, 334-342.

[13] Zuccala1 A., Someren V.M., Bellen V.M. (2014). A machine-learning approach to coding book reviews as quality indicators: Toward a theory of mega-citation. Journal of the Association for Information Science and Technology, 65(11): 2248-2260.

[14] Sussman S W, Siegal W S.(2003). Infomational Influence in Organizations: An Integrated Approach to Knowledge Adoption. Information System Research, 2003, 14:4-65

[15] Pang B, Lee L.(2004). A Sentiment Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceeding of the 42nd Annual Meeting of the Association for Computational Linguistic. Morristown, NJ, USA: ACL Press, 271-278

[16] Wu J.(2012). The Beauty of Mathematics..Beijing: The People's Posts and Telecommunications Press, 60-64

[17] Hu M.Q. , Liu B.(2004). Mining and Summarizing Customer Reviews. Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, USA.New York, NY, USA:ACM Press, 168-177

[18] Friedman, J. (2000). Greedy boosting approximation: a gradient boosting machine. Annals of Statistics, 29(5):1189–1