ICPE2013
The 5th International Conference on Information Capital, Property, & Ethics
Tsurumi University

Submitted on: October 30, 2013

Textual Characteristics based High Quality Online Reviews Evaluation and Detection

Hui Nie
School of Information Management, Sun Yat-sen University, Guangzhou, China.
E-mail address: issnh@mail.sysu.edu.cn

Chengying Gao
School of Software, Sun Yat-sen University, Guangzhou, China.

Zhe Rong
School of Information Management, Sun Yat-sen University, Guangzhou, China.

Abstract: *With the rapid growth of internet, a wealth of product reviews has been spread to the web. The user-generated on-line information varies greatly in quality, which making harder for review readers to identify the most useful reviews and understand the true underlying quality of a product. In this paper, we studied the problem of evaluating and detecting high-quality product reviews. We particularly examined how the textual aspect of a review affects the perceived usefulness of it. Based on a real-world data set, our results indicate that the text-specific characteristics are significantly associated with the perceived helpfulness of reviews. A review is perceived to be useful if the content of the review focusing on the given subject, with rich information and being moderately expressed in subjective ways.*

Keywords: user-generated content, text mining, product reviews, sentiment analysis

1 INTRODUCTION

With the rapid growth of internet, the way that people express themselves and interact with others has changed. They post reviews of products at commercial sites and express their viewpoints in various social media websites (Jindal 2007). The *user-generated* content contains valuable information that can be exploited for many applications. In e-commerce, particularly, there has been an increasing interest in mining opinions from reviews in recent years. A growing number of consumers wade through online product reviews to gauge their purchase decisions and many merchants are expected to focus on the on-line *word of mouth* by which they can understand the consumers' need and make further predictions for the market trends. Regardless of customers or merchants, all review readers need to seek unbiased evaluation of their target products or brands, by leveraging information from multiple reviews (Zhang 2006). However, the *user-generated* reviews are so overwhelming that make individuals harder to identify the valuable information and understand the true underlying quality of the product. Even worse, the *user-generated* reviews often vary greatly in quality due to the lack of editorial and quality control[3]. Low-quality or even spam reviews are mixed with the valuable ones, causing trouble to users who expect to obtain useful information. Obviously, if the *user-generated* contents are need to be exploited effectively, it is crucial to have a mechanism capable of assessing the quality of reviews and extracting the high-quality reviews from the high volume of original information.

In the social psychology literature, message source characteristics have been found to influence judgment and behaviour (Ghose 2011), and it has been often suggested that source characteristics might shape product attitudes and purchase propensity. Review readers, as a rule of thumb, pay more attention on review content than other information aspects when seeking the peer-reviewers' opinions, which implies, to the large extent, the vote for the *usefulness* of a review depends on the textual content. Based on the idea, we believe that deeply analyzing the textual aspects of reviews should be the most direct and effective way to detecting the quality of reviews. Therefore, we place special emphasis on how the textual aspect of a review affects the perceived usefulness of it. To address the problem, a simple but well-established framework for assessing review quality has been designed to examine the nature of *helpfulness*. A stream of NLP[1] technologies, e.g. Chinese words segmentation, POS[2] tagger and Sentiment analysis, has been fully employed to extract the corresponding text-specific characteristics. Then, we conducted the study by integrating an explanatory econometric analysis with a supervised machine learning technology, decision tree classification. The purpose of the study is maximizing the utility of online information by investigating the most influential factors for evaluating the quality of user-generated content and intending to seek an effective approach for predicting the perceived helpfulness of reviews.

2 RELATED WORK

Our research program is inspired by the works of Ghose (Ghose 2011, Ghose 2006, Ghose 2007). Ghose's work (Ghose 2011) is the first study that integrates econometric, text mining, and predictive modeling techniques to a complete analysis of the information captured by *user-generated* online reviews in order to estimate their *helpfulness* and economic impact. In addition, Jindal's research (Jindal 2007) and Zhang's work (Zhang 2006) are also relevant to our study. In these studies, review evaluation is typically viewed as a ranking or identification problem resolved with regression models or classification techniques. In the process of model training and testing, most of them used the ground-truth derived from users' votes of *helpfulness* provided by websites. And multiple information aspects, such as the numerical review data (e.g. Star-level) were investigated for building a prediction model. However, from the perspective of users' perception for reading, multiple information aspects might not specifically uncover the nature quality of a review. Hence, our study places the special emphasis on the influences exerted by the content-specific characteristics.

Actually, feature representation and selection plays a crucial role in the quality evaluation for information sources. In the related studies, Liu (Liu 2007) presented a classification-based approach for low-quality reviews detection and three aspects of reviews, namely *informativeness*, *readability* and *subjectiveness*, have been selected as metrics to evaluate the

[1] NLP: Natural Language Processing

[2] POS: Part-Of-Speech

quality of reviews. Otterbacher (Otterbacher 2009) designed as much as seventeen quality metrics on five factors, such as the relevancy, reputation and representation, to evaluate reviews. And Zhang (Zhang 2006) investigated the predicting task from text sentiment analysis point of view. A diverse set of language-specific features has been incorporated to build the prediction model and the results indicated the perceived utility of a review highly depends on its linguistic style. Therefore, the NLP-related text analysis is emphasized in the relevant researches. In Ghose's works (Ghose 2011, Ghose 2006, Ghose 2007), the researchers employed the NLP toolkit *Lingpipe*[3] to get the subjective-specific features of reviews. Zhang collected the shallow syntactic features by using existing lexical resources (Zhang 2006). And Liu adopt a sentiment analysis tool to solve the problem of subjective features extraction (Liu 2007). Since our study focus on Chinese e-commerce website, text mining techniques for Chinese language were employed to fulfil the tasks of textual analysis and opinion mining. To the best our knowledge no prior work has combined text mining with economic methods to evaluate the utility of online Chinese reviews.

3 RESEARCH DESIGN

This research is focused on investigating the influence of textual features for high-quality review identification and intents to define a valid specification to evaluate the quality of user-generated contents. In view of the research objective, the content-oriented evaluation mechanism study has been conducted and three main research questions are posed as follows:

Question1. What are the influential textual features for evaluating the quality of reviews and how could we obtain the textual features by utilizing NLP-based technologies.

Question2. How do the textual features and ways of combination exert impacts on the *helpfulness* of reviews?

Question3. How do we utilize the Machine Learning (ML) approach to discriminate between high-quality reviews and low-quality reviews in the light of text-specific features?

3.1 Metrics system for Reviews Quality Evaluation

Based on Wang's study (Wang 1996), data quality is divided into four major categories: intrinsic quality, contextual quality, representational quality and accessibility. Intrinsic quality emphasizes information have quality in their own right, such as subjectivity and objectivity. And Contextual quality focuses on the natures relevant to completeness and quantity, commonly relevant to the writhing style, such as the length of information. Generally, longer texts contain more information; however, some studies suggest tediously long texts might exert negative impact on people's reading experience. As representational quality, is commonly referred to nature of readability which can be measured by characters-to-sentences ratio or words-to-sentence ratio. Learning from Wang's analysis, a metrics system has been designed for our task of content-oriented quality evaluation, shown as table 1. As can be seen, we have incorporated 6 textual features of reviews across on three types.

Table1 The metric framework for reviews quantity evaluation

Type	Feature Variables	Explanation
Subjectivity	*Avg-Sub*	The average probability of a review being subjective
	Avg-SD	The average standard deviation of the subjectivity probability
Informativeness	*Num- wd*	The length of the review in words
	Num- sen	The length of the review in sentences

[3] LingPipe is a tool kit for processing text using computational linguistics. http://alias-i.com/lingpipe/

Topic-relevancy	*Num- wd-sen*	The words-to-sentence ratio
	Topic-relevancy	The similarity between review with product specification

Generally speaking, a high-quality product review is a reasonable mixture of subjective valuation and objective information. *Subjectivity* reflects the reviewers' viewpoints, while *objectivity* is more relevant to factual descriptions. *Topic-relevancy* tells readers what the content is about, reflecting the relevance between a product and a review description in our study case. Regarding *informativeness*, it embodies the level of understanding and the quantity of information. According to our definition, the reviews with high scores on *informativeness* are more complex and harder to be understood (Foltz 1999). Empirically, all the features should to some extend have effect on the perceptions of the on-line review readers; therefore, the features space in the statistical learning framework ought to capture these characteristics.

3.2 Text-specific features Extraction

The objective of textual analysis is to collect text-specific features; the input of analysis model is the original textual aspects of a review; and the output is a group of textual features defined in table1. As shown in Fig.1, the features extraction task is carried out in three phases, pre-processing, sentimental analysis and similarity calculation. Due to our study is for Chinese language, FundanNLP [4] is employed.

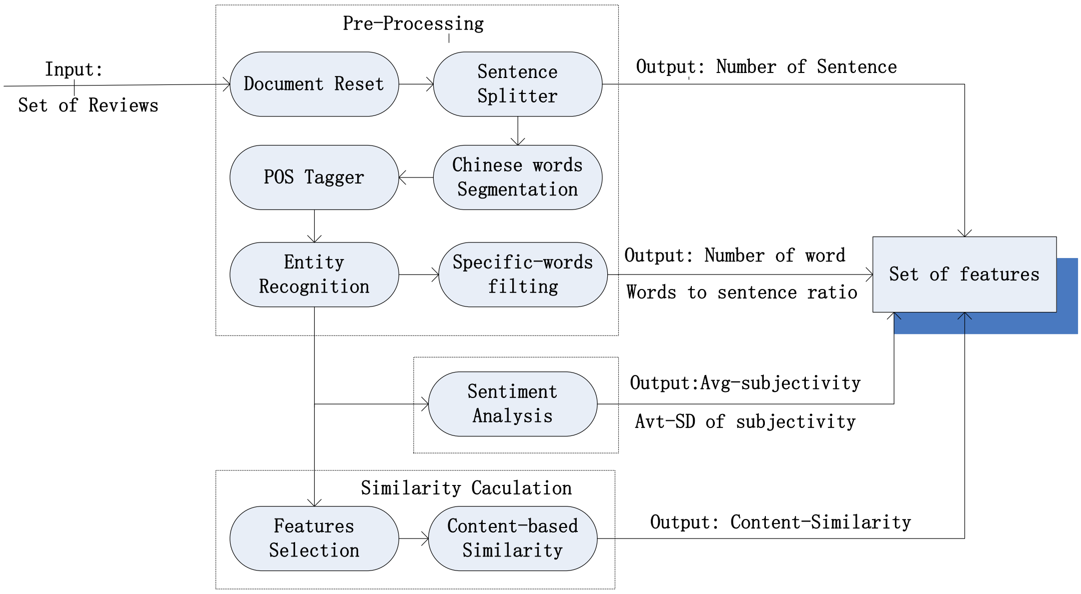


Figure 1. Framework of Textual Analysis processing

In the stage pre-processing, the fundamental text processing and textual features extraction are performed. *Sentence Splitter* is used to segment the text into sentences according to the segmentation mark; then, by using *Chinese word segmentation*, the sentences are further spitted into the Chinese words. Afterwards, each word is annotated with a part-of-speech tag and an entity type respectively by *POS Tagger* and *Entity Recognition*. Finally, *Words filter* extracts the feature words for the further analysis. Adjectives and Adverbs in the text, for

[4] FundanNLP is a tool kit for Chinese natural language processing. http://code.google.com/p/fudannlp/

instance, are chosen for sentiment analysis, meanwhile, text-specific features relevant to *informativeness* (e.g. *Num- sen* and *Num- wd*) are also obtained in the stage.

Regarding *Subjectivity* features, the average probability of a review being subjective and the average standard deviation of the subjectivity probability designed in the work of Ghose (Ghose 2011) are adopted to describe sentiment factors implied in a review's textual content. The formula of the average probability of a review is showed as follow:

$$Avg\_Sub(R) = \frac{1}{n}\sum_{i=1}^{n} Pro_{Subjectivity}\,(Sentence_{\mathrm{i}})$$

Where n is the number of sentences in review R and $Sentence_i$ (i=1,2,…,n) is the sentences that appear in review R and $Pro_{Subjectivity}(Sentence_{\mathrm{i}})$ represents the probability of $Sentence_i$ being subjective. Two types of information, the objective information, listing the characteristics of the product and the subjective information, in which the reviewers give a very personal description of a product, were presented. Empirically, a helpful review should present several aspects of a product and provides convincing opinions with enough evidence as well, which means both types of information should be included in a review. Since a review may be a mixture of objective and subjective sentences, the standard deviation $Avg\_SD(R)$ of the subjectivity probability for the review has been defined to describe the statue of mixture:

$$Avg\_SD(R) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Pro_{Subjectivity}(Sentence_i) - Avg\_Sub(R))^2}$$

We employed a machine learning based approach to predicting the probability of a sentence being subjective. By manually annotating the training set, a subjectivity classifier based on linear regression has been constructed. Performance of the classifier was evaluated by 10-fold cross-validation on the test set, which is promising as the accuracy of the classifier is above 0.85.

Topical relevancy embodies the semantic content of a review. A helpful review can be commonly taken as the main reference that users need to read before making their purchase decision on a product, which indicates that the content of a review should be written closely around a given product. The similar situation is conceivable between a customer review and an editorial review; the latter approximates a relatively objective and authoritative view of the product. Based on the understanding, similarity between customer reviews and the official evaluation for products has been adopted to quantify the feature of *topical relevancy*. The standard cosine similarity in Vector Space Model (*VSM*) with *TF*IDF* term weighting are used. The processing framework is showed as figure 2.

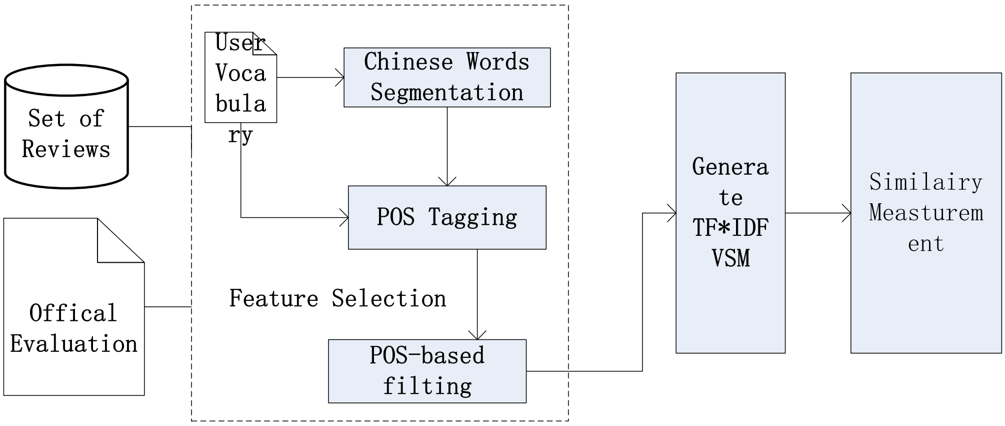


Figure2. Processing framework for topical relevancy measurement

3.3 Explanatory Econometric analysis model

Once we have derived the textual features of each review, we aim to look into how the textual features and ways of combination exert impacts on the quality of reviews. Here, the quality of a review refers to the perceived *helpfulness* of the review since the high-quality reviews are commonly with high score of *helpfulness*. So, in our explanatory econometric model, *helpfulness* is the dependent variable and the textual features variables are the predictors. Before presenting the model, we need test the following hypotheses:

Hypothesis 1. All else equal, a change in the subjectivity levels in a review will be associated with a change in the helpfulness of that reviews.

Hypothesis 2. All else equal, a change in the Informativeness of a review will be associated with a change in the helpfulness of that review.

Hypothesis 3. All else equal, a change in the *topical relevancy* of a review will be associated with a change in the *helpfulness* of that review.

According to the hypothesis above, a linear specification for our *helpfulness* estimation has been defined as:

$$R_{helpfulness} = \alpha + \beta_1 Ln\left(R_{Num\_wd}\right) + \beta_2 R_{Num\_sen} + \beta_3 R_{Num\_wd\_sen} + \beta_4 R_{Avg\_sub} + \beta_5 R_{Avg\_SD} + \beta_6 R_{Topic\_relevancy} + \varepsilon$$

The unit of observation in our analysis is a review *R*. The dependant variable $R_{helpfulness}$ is the radio of helpful votes to total votes received for a review. Independent variables, R_{Avg_sub} and R_{Avg_SD} are used to capture the level of subjectivity in a review; R_{Num_wd}, R_{Num_sen} and $R_{Num\_wd\_sen}$ quantify the degree of Informativeness; and $R_{Topic\_relevancy}$ embodies the semantic content feature of a review. In our model, the influence of features derived from the textual aspect of a review is emphasized; we do not consider all possible information aspects (e.g. features about reviewers). Actually, from the view of review readers, the content, containing the most specific information, might be their major concern when they determining whether a review is useful or not.

3.4 The Rule-based prediction model

Although the explanatory study can uncover what kinds of factor influence the perceived *helpfulness* of a review, another objective is to examine, given an existing review, how well we can predict it's *helpfulness* on the basis of the content. The *helpfulness* of each review in our data set is defined by the votes of the peer customers. In our predictive framework, we attempt to build a binary prediction model that can classify a review as helpful or not.

Therefore, the first step, the continuous variable *helpfulness* (*helpfulness* $\in$ (0,1)) is converted into a binary one. The threshold ? is set to mark all reviews that have *helpfulness* $\geq \ \tau$ as helpful and others are not helpful. Then, as a rule-based approach, the decision tree model has been selected to perform the task of prediction; since classification rules can make people better understand the result. In the process of constructing the tree, the splitting feature is specifically selected based on the *information gain* which being interpreted as the informational value of creating a branch on the feature. Basically, higher the information gain of an attribute is, more influential the attribute is for classification; so, the predictive ability of attribute can be examined according to the structure of tree.

For better understanding how decisions being made, given the decision tree, we also need to convert the numeric attributes into nominal variables. A clustering algorithm is specifically used. For instance, according to R_{Avg_sub} and R_{Avg_SD}, the *subjectivity* is converted into a nominal variable with three-styles: *highly subjective*, *moderately subjective* and *weakly subjective*. *Topical relevancy* is ranked into *strong*, *less strong* and *weak*, and the *informativeness* is also categorized into three types according to the length of the review in words: *large, medium and small*.

4 EXPERIMENTS & ANALYSES

4.1 Data Collection

As the most commercially marketable IT-specific website in China, www.zol.com.cn has been chosen to be our source of data. By using a web spider toolkit --- Locoy[5], we collected totally 1,569 original reviews about an *Moto mobile phone* (*ME525*) from the website. A parsing program has been developed to automate the data extraction. Textual aspects of reviews are the major target data. Certainly, the total number of voters and the number of useful voters has been extracted, used as the ground-truth to approximate the target value of the regression model and the decision tree model as well.

4.2 Experiments and Analyses

Based on the data set described above, we firstly approximate the predictive variable $R_{helpfulness}$, as follow:

$$R_{helpfulness} = \frac{R_{Num\_helpful\_voters}}{R_{Num\_total\_voters}}$$

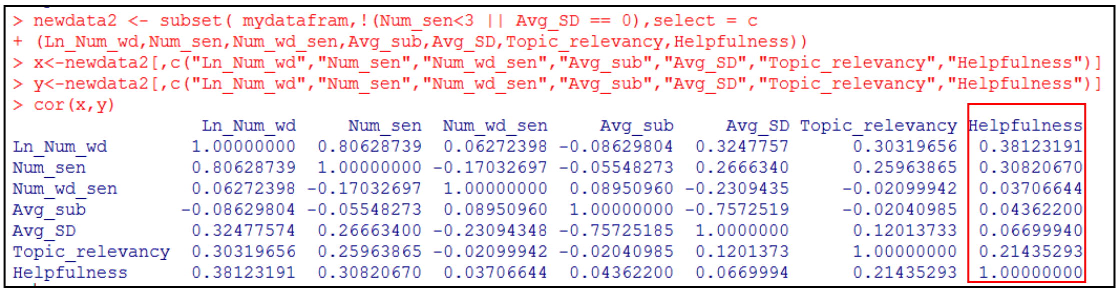
Then, we obtained the relevant predictors by running a stream of textual analysis (Section 3.2). Two major studies have been conducted to investigate the quality of reviews. The first one (Section 3.3) investigates the influence exerted by textual features, attending to determine the most predictors for assessing the *helpfulness* of a review by using regression analysis. The second one (Section 3.4) is designed for validating the reliability of regression model and attempt to obtain an effective rules for detecting high quality reviews.

4.2.1 Estimating the quality of reviews

The experiment is conducted on the platform of ***R*** Language[6]. The correlations between the predictive variables $R_{helpfulness}$ and a group of independent variables have been examined, as shown in table 2. More distinct results are obtained by use of the regression analysis, as shown in table3.

[5] http://www.locoy.com/

[6] R is a language and environment for statistical computing and graphics. http://www.r-project.org/



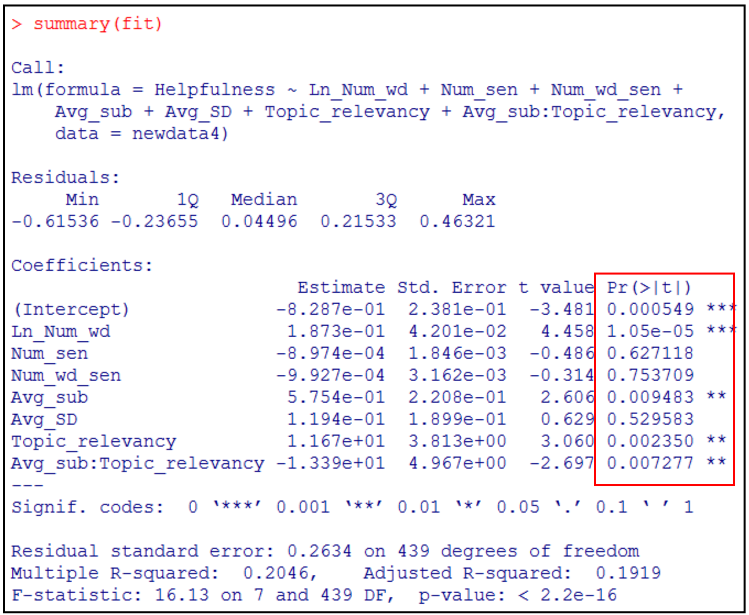
```
> newdata2 <- subset( mydatafram,!(Num_sen<3 || Avg_SD == 0),select = c
+ (Ln_Num_wd,Num_sen,Num_wd_sen,Avg_sub,Avg_SD,Topic_relevancy,Helpfulness))
> x<-newdata2[,c("Ln_Num_wd","Num_sen","Num_wd_sen","Avg_sub","Avg_SD","Topic_relevancy","Helpfulness")]
> y<-newdata2[,c("Ln_Num_wd","Num_sen","Num_wd_sen","Avg_sub","Avg_SD","Topic_relevancy","Helpfulness")]
> cor(x,y)
                  Ln_Num_wd     Num_sen  Num_wd_sen     Avg_sub     Avg_SD Topic_relevancy Helpfulness
Ln_Num_wd        1.00000000  0.80628739  0.06272398 -0.08629804  0.3247757      0.30319656  0.38123191
Num_sen          0.80628739  1.00000000 -0.17032697 -0.05548273  0.2666340      0.25963865  0.30820670
Num_wd_sen       0.06272398 -0.17032697  1.00000000  0.08950960 -0.2309435     -0.02099942  0.03706644
Avg_sub         -0.08629804 -0.05548273  0.08950960  1.00000000 -0.7572519     -0.02040985  0.04362200
Avg_SD           0.32477574  0.26663400 -0.23094348 -0.75725185  1.0000000      0.12013733  0.06699940
Topic_relevancy  0.30319656  0.25963865 -0.02099942 -0.02040985  0.1201373      1.00000000  0.21435293
Helpfulness      0.38123191  0.30820670  0.03706644  0.04362200  0.0669994      0.21435293  1.00000000
```

Table 2 Correlation analysis for $R_{helpfulness}$ with the relevant independent variables (Screenshot)

From the table 2, we can see that $ln(R_{Num\_wd})$ and R_{Num_sen} have the relatively strong correlation with $R_{helpfulness}$ compared to other variables; the corresponding correlation coefficients come to 0.3812 and 0.3082 respectively. And correlation between $R_{Topic\_relevancy}$ and $R_{helpfulness}$ is 0.2143, being on the second place. Moreover, we notice that there is no significantly relevant between the subjective features of a review with its *helpfulness* as we intuitively expected, e.g., correlation between R_{Avg_sub} and $R_{helpfulness}$ is only about 0.0436.

From the table 3, it is clearly shown that the change in $Ln(R_{Num\_wd})$ or $R_{Topic\_relevancy}$ of a review leads to the change of $R_{helpfulness}$ and significantly. Regarding sentimental factors, R_{Avg_sub} has positive influence on the *helpfulness* of the review in statistical significance; but, R_{Avg_SD} does not exert significantly impact on the *usefulness*. No significant influence also comes to the other two independent variables, $R_{Num\_wd\_sen}$ and R_{Num_sen}. Beyond that, in the process of optimizing the regression specification, we further detected the impacts led by the interaction of two variables. The product term of R_{Avg_sub} and $R_{Topic\_relevancy}$ is significantly put negative influence on *usefulness* of a review, suggesting the change rate of the *usefulness* might decrease with the increase of topical relevancy and subjective level. Namely, the relationship between sentimental factors and *usefulness* of a review is also associated with its content. A review with rich content and high topic correlation can be perceived to be useful respectively; but a review might not be recognized as being much *useful* if it's over emotional even if the content is closely related with product-specific subject.

Since not all features are statistically significant with the *helpfulness* of a review, we obtain the best two models according to the value of *adjusted* R^2 by exploring models on all possible feature subsets. As shown in Figure 4, the best two models are on the top row, being described by the variables $Ln(R_{Num\_wd})$, R_{Avg_sub}, R_{Avg_SD}, $R_{Topic\_relevancy}$ and R_{Avg_sub} .



```
> summary(fit)

Call:
lm(formula = Helpfulness ~ Ln_Num_wd + Num_sen + Num_wd_sen +
    Avg_sub + Avg_SD + Topic_relevancy + Avg_sub:Topic_relevancy,
    data = newdata4)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61536 -0.23655  0.04496  0.21533  0.46321

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -8.287e-01  2.381e-01  -3.481 0.000549 ***
Ln_Num_wd                1.873e-01  4.201e-02   4.458 1.05e-05 ***
Num_sen                 -8.974e-04  1.846e-03  -0.486 0.627118
Num_wd_sen              -9.927e-04  3.162e-03  -0.314 0.753709
Avg_sub                  5.754e-01  2.208e-01   2.606 0.009483 **
Avg_SD                   1.194e-01  1.899e-01   0.629 0.529583
Topic_relevancy          1.167e+01  3.813e+00   3.060 0.002350 **
Avg_sub:Topic_relevancy -1.339e+01  4.967e+00  -2.697 0.007277 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2634 on 439 degrees of freedom
Multiple R-squared:  0.2046,    Adjusted R-squared:  0.1919
F-statistic: 16.13 on 7 and 439 DF,  p-value: < 2.2e-16
```

Table 3. Result of regression analysis (Screenshot)

$R_{Topic\_relevancy}$ and with the highest value of *adjusted* R^2(about 0.2**)**, which indicates the two models can account for up to 20 percent of *adjusted* R^2 and the major influential factors for evaluating *usefulness* of reviews are *Num_wd*, *Avg_sub*, *Avg_SD* , and *Topic_relevancy* respectively.

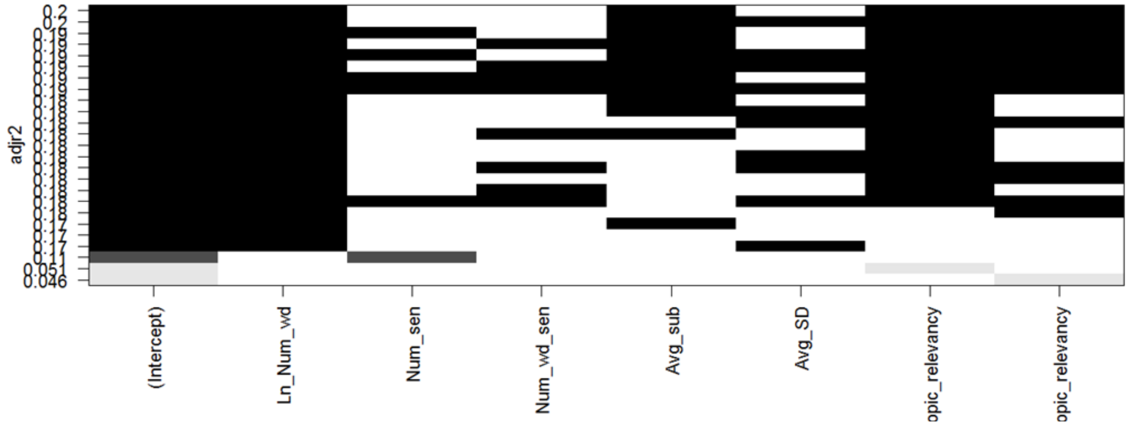


Figure 4. Rank the explanatory models based on the *adjusted R*^{2}

Furthermore, relative importance for the predictor variables has also been measured. Relative importance can be thought of as the contribution each predictor makes to R^2, both alone and in combination with other predictors. An approach for measuring the metric, which closely approximates the average increase in R^2 obtained by adding a predictor variable across all possible sub_models, has been employed in our experiment. As shown in figure 5, it is clear *Ln_Num_wd* has the greatest relative importance, followed by *Topic_relevancy* and subjective factors, in that order.

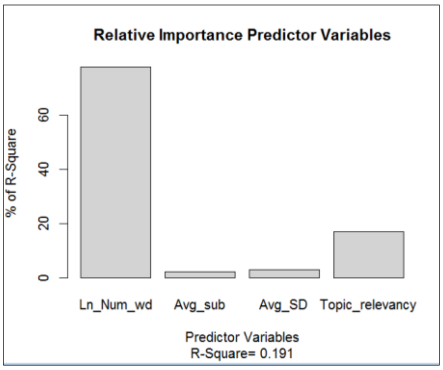


Figure 5. Relative Importance for the predictor variables

4.2.2 Detect the quality of reviews by Decision Tree

According to the analysis in Section 3.4, we use the decision tree to examine whether, given an existing review, how well can we predict the *helpfulness*, i.e., of a review that was not included in the data used to train the predictive model. The influential factors obtained from the optimal model in section 4.2.1 have been used as the features to build the classifier and Rapid Miner[7] was employed to conduct the classification experiment. The evaluation results are based on stratified 10-fold cross validation on our experimental data set. The resulting performance of the classifier is satisfactory. The classification accuracy comes to 82.6%.

Another interesting result is the tree model obtained. As shown in Fig. 6, we can see that *Num_wd* is the first attribute being tested, followed by the *topic relevancy* and the *sentimental types* in that order. Based on the decision tree approach, the result indicates *Num_wd* is the biggest attributor to the classification, namely, it exert the most significant influence on detecting useful reviews. The second most powerful classification features is *topic relevancy*, then followed by *subjectivity* associated with predictors R_{Avg_sub} and

[7] Rapid Miner: The world-leading open-source system for data mining. http://rapid-i.com/content/view/181/190/

R_{Avg_SD}. The result is consistent with the results of regression analysis in section 4.2.3. Beyond that, routing down the tree according to the values of the attributes tested in successive nodes, we can find a set of easy-understanding rules for determining whether a review is helpful or not. For example,

Rule1: if (*Num*_wd = *large* **and** *Topic_relevancy* = *Strong* **and** *Sentimental style* = **(***moderate* **or** *weak*) then (**The review is *helpful***)

Rule2: if (Num_wd = *large* **and** Topic_relevancy = *Strong* **and** Sentimental style = *high*) then (***The review is not helpful***)

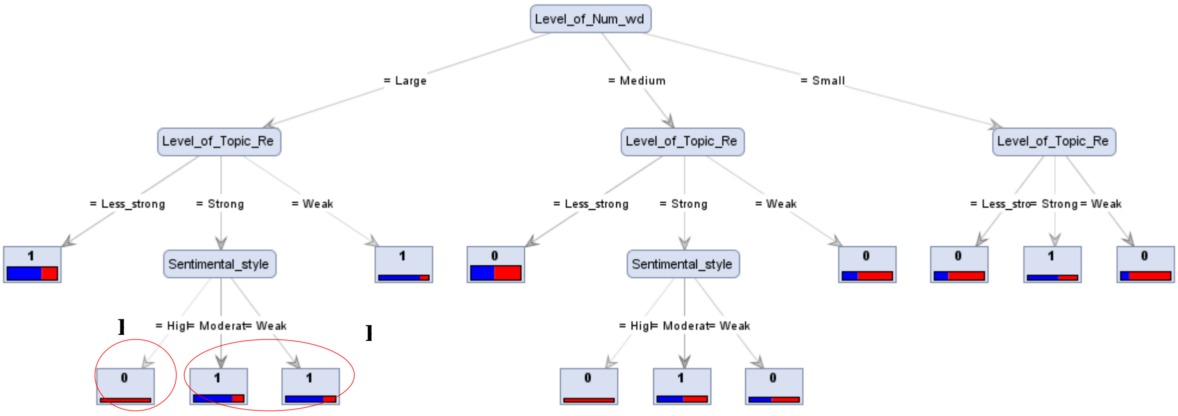


Figure 6. The Decision Tree Model for detecting the helpfulness of reviews (screen shot)

Rule1 implies when the content of a review is informative and detailed on a specific subject interested by review readers and being moderately expressed in subjective ways, it is commonly perceived to be helpful; whereas, ***Rule2*** reveals reviews with high topical relevancy but over-subjective content regardless of being positive or negative, tend to be perceived as excessive assessments and have a significant possibility of being identified as uselessness, which can be explained by the negative influence exerted by the interaction of R_{Avg_SD} and R_{Topic_sim} in the regression model in 4.2.1. Clearly, classification rules derived from the tree model make review readers get more intuition about the detection procedures.

5 CONCLUSIONS

In this paper, we studied the problem of evaluating and detecting high-quality product reviews. We particularly examined how the textual aspect of a review affects the perceived usefulness of it. To address the problem, a simple but well-established framework for assessing review quality has been designed to examine the nature of *helpfulness*. A stream of NLP technologies, e.g. Chinese words segmentation, POS tagger and sentiment analysis, has been fully employed to extract the corresponding text-specific features. We conduct the study by combining an explanatory econometric analysis and a supervised machine learning technology, decision tree classification.

Based on a real-world data set, our econometric analysis reveals that the text-specific characteristics, including quantity of content, topical relevancy and extent of subjectivity, are significantly associated with the perceived helpfulness of the review. The optimal model demonstrates a review is perceived being useful if the content of the review focusing on the given subject, with sufficient information and being expressed in moderately subjective ways. Additionally, by using the decision tree classifier, we also examine the relative importance of the three broad feature categories: *subjectivity*, *informativeness* and *topic relevance*, and find *informativeness* plays the most important role in detecting the *helpfulness* of reviews. Beyond that, the good performance of the tree-based classifier further indicates the high-quality reviews can be discriminated from the low-quality ones only by examining their textual features.

In summary, our study suggests we can quickly estimate the quality of a review by performing an automatic stylistic analysis according to its textual content and sentimental characteristics and we can straightforwardly indentify the product reviews expected to be helpful to the online-customers and show them at first time on the commercial websites without any biases resulted from the lacking of voters.

REFERENCES

Foltz,P.W., Laham, D. and Landauer, T. K. (1999). Automated essay scoring: applications to educational technology. World Conference on Educational Multimedia, Hypermedia and Telecommunications, 1999(1): 939-944.

Jindal,N. and Liu,B. (2007). Analyzing and detecting review spam, 7th IEEE International Conference on Data Mining (ICDM '07), 547-552.

Ghose,A. and Ipeirotis,P.G. (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.

Ghose, A. and Ipeirotis,P.G.(2006). Designing ranking systems for consumer reviews: The impact of review subjectivity on product sales and review quality. Proceedings of the 16th Annual Workshop on Information Technology and Systems, 303-310.

Ghose, A. and Ipeirotis,P. G.(2007). Designing novel review ranking systems: predicting the usefulness and impact of reviews, Proceedings of the 9th international conference on Electronic commerce, 3-310.

Liu,J. and Cao,Y. (2007).Low-quality product review detection in opinion summarization, Proc. of 2007 Joint Conference on empirical methods in NLP and CNLL, 334-342.

Otterbacher, J. (2009). “Helpfulness” in online communities: a measure of message quality, Proceedings of the 27th international conference on Human factors in computing systems (CHI '09), 955-964.

Wang, R.Y. and Strong, D.M. (1996). Beyond accuracy: what data quality means to data consumers, Journal of Management Information System, 12(4), 5-33.

Zhang, Z. and Varadarajan, B. (2006). Utility scoring of product reviews, International Conference on Information and Knowledge Management (CIKM '06), 51-57.