Research on Quality Analysis Model of Additional Comments Based on Text Analysis

Yali Dong, Suyang Wang and Shixuan Ye

Department of Management, Wuhan University of Technology, Wuhan, Hubei, 430000, China

*Corresponding author’s e-mail: 2501639769@qq.com

Abstract. Based on the idea of text analysis, this study analyzes the relevance of text topics, the emotional consistency of fist comments and additional comments, and the information contained in the text, then get out of the online quality assessment model. The calculation results can reflect the quality of additional comments, helping consumers and businesses to identify important information and make correct judgments.

1. Introduction

In recent years, e-commerce has been developing rapidly, and more and more consumers would like to choose online shopping. The 34th "Statistical Report on the Development of China’s Internet Network" showed that users’ reviews are the main reference factor for online shopping users to make purchasing decisions. As an important form of online word-of-mouth, the comment system has become a major reference source for consumers' online shopping and it affects consumers' purchasing decisions. However, because of the cashback reward to favorable comments provided by the merchants, the trust of online comments is greatly reduced. Therefore, Taobao stipulates that: Within 180 days (inclusive) from the date of successful transaction, the buyer can add comments after making the Taobao credit evaluation, and the content of additional comments can not be modified. So when consumers browse comments and look for commodity information, they will pay special attention to the contents of additional comments. But in order to obtain more profits, the merchants come up with ways to give money to praise and even hire online water army to give praise. This has undoubtedly caused great trouble to consumers. Therefore, how to “get the fake and take the truth, get the rough and take the fine” and identify high-quality additional comments has become the subject of this research.

2. Related researches

For the question of the quality of comments, the researches of domestic and foreign scholars mainly focus on the initial comments and their main topics are the following two aspects.

2.1. Identify factors that affect the quality of online reviews

Kim S et al. [1] use SVM regression method to predict the quality of comments from five text features: structure, morphology, syntax, semantics and metadata. The experimental results show that comment length, unary grammar and product rating are the key factors to judge the usefulness of comment. Hyunmi Baek [2] believes that the inconsistency of comments, the ranking of reviewers, the number of comments, etc. have a certain impact on the credibility of comments. The research by Huang Tingting...
et al. [3] shows that whether the content of the review is closely related to the function of the product is an important indicator for judging the quality of the comment.

2.2. Study and analyze the model of online comment quality

Tang Xiaobo et al. [4] propose a topic-oriented high-quality comment mining model, which combines existing useful votes with other weighting indicators to automatically extract high-quality comments under each topic. Li Zhiyu [5] starts with the study of influencing factors of online comment effectiveness and establishes a commentary utility index system. Liu Jingjing et al. [6] note the problem of evaluation bias caused by the cumulative vote, so they use the data labeled manually and evaluate the quality of comments from three aspects: information content, subjectivity and readability.

Based on the carding of above theories and models, it provides an idea for constructing an additional comments quality evaluation index system, which is helpful to the construction of this model.

3. Theory and method

3.1. LDA model theory

LDA (Latent Dirichlet Allocation) is a document theme generation model, which can be used to identify hidden subject information in large document sets or corpora. LDA gains the relationship between vocabularies, documents and topics in the corpus by generating a probability model. The LDA generation process of a document is as follows [7]:

- Get the total number of all words in the document \( N \sim \text{Poisson}(\beta) \);
- Get the distribution of the document on the topic \( \theta_{m} \sim \text{Dir}(\alpha) \);
- Traverse each word from 1 to \( n \). First, get the probability distribution of each topic \( \sim \text{Multinomial}(\theta_{m}) \). Then, get the probability distribution of topics and words \( \sim \text{Multinomial}(\Phi_{j}) \).

Among them, \( \alpha \) and \( \beta \) are priori parameters, \( \theta \) and \( \Phi \) are estimated parameters. The experiment uses Gibbs sampling to estimate \( \theta \) and \( \Phi \), and predicts the subject probability of the current word while keeping the subject distribution of other words unchanged, and indirectly obtains their values by marginalizing \( \theta \) and \( \Phi \) [7]:

\[
P(Z_{n} = j | Z_{-n}, W_{m,n}, \alpha, \beta) \propto \frac{C_{vK}^{W-n} + \beta_{vn} \Sigma_{v=1}^{V} (C_{vK}^{W-n} + \beta_{vj}) \times C_{MK}^{d-n} + \alpha_{dj}}{\Sigma_{k=1}^{K} (C_{MK}^{d-nk} + \alpha_{dk})}
\]

(1)

In the formula, \( C_{vK}^{W-n} \) and \( C_{MK}^{d-n} \) represent matrix quantities of dimension \( V \times K \) and \( M \times K \) respectively, \( V \) is the number of words, \( K \) is the number of topics and \( M \) is the number of documents. \( C_{vK}^{W-n} \) is the frequency that the word \( w \) assigned to the subject \( j \); \( C_{MK}^{d-n} \) is the number of words assigned to the subject \( j \) in the document \( d \).

Assign random themes to the words in the document to form the initial Markov chain, and then assign them the subject again according to formula (1) in order to obtain the state of the next Markov chain, and iterate continuously until the Markov chain becomes stable. Finally, use the Gibbs sampling algorithm to estimate the \( \theta \) value and \( \Phi \) value for each word [7]:

\[
\phi_{w_{m}}^{(z=j)} = \frac{C_{vK}^{W-n} + \beta_{vn}}{\Sigma_{v=1}^{V} (C_{vK}^{W-n} + \beta_{vj})}
\]

(2)

\[
\theta_{m}^{(z=j)} = \frac{C_{MK}^{d-n} + \alpha_{dj}}{\Sigma_{k=1}^{K} (C_{MK}^{d-nk} + \alpha_{dk})}
\]

(3)

In this way, the probability distribution of each topic and the probability distribution of the topic and the word can be obtained, thereby we can obtain a "topic-word" probability distribution matrix and a "document-topic" probability distribution matrix.

3.2. Grey correlation analysis

In 1981, Professor Deng Julong, a Chinese cybernetic expert, first proposed the concept of the gray system. Later, he published many papers and monographs on the gray system and established the gray
system theory. The grey system theory puts forward the concept of gray correlation analysis for each subsystem, and intends to seek the numerical relationship between subsystems (or factors) in the system with certain methods. Therefore, the grey correlation analysis provides a quantitative measure for the development and change of a system, which is very suitable for dynamic process analysis. The grey correlation method is also an effective way to determine weights. It aims to determine the weight value by making a reasonable processing of the existing actual data in order to achieve a scientific description of the system development.

4. Design of the indicator system
For the online comments given by online consumers, this study attempts to establish a reasonable text quality evaluation system oriented to comment content. We can use it to evaluate the quality of users’ reviews, identify high-quality additional comments quickly and improve the utility value of online comments.

4.1. Data collection
Taobao is currently the largest C2C e-commerce website in China, which has rich online comment resources. This study uses web crawler technology to get a large number of initial comments and additional comments from the website and store them in the corresponding documents.

4.2. Data preprocessing
The data obtained from the website is mixed. To ensure data quality and reduce the time wasted on invalid data, the original data needs to be preprocessed. Specific operations include the followings:
- Delete the duplicate data. Delete the data whose initial comment is “This user did not fill out the comment” and whose comment content only contains special characters, numbers or pictures.
- Use the NLPIR-ICTCLAS Chinese word segmentation system to the text word segmentation and part of speech tagging.
- Delete the stop words in the comments.
- Save the processed data

4.3. Relevance of comment topic
A study by Titov [8] shows that the higher the relevance of a comment to a topic, the greater its contribution to the topic and the higher the quality of the comment. Therefore, the first indicator in the model constructed in this study is the relevance of the text content to the topic. Referring to the research of Tang Xiaobo [9], we can confirm the idea of judging the relevance of commentary topic:

Each comment in the comment text set corresponds to a document in the corpus. The topic distribution of each document can be represented as a vector, and the similarity between documents can be obtained by calculating the distance between documents [10]. Then in the "Comment-Subject" matrix of the corresponding LDA model, the topic distribution of each comment can be expressed as a vector

$$\theta_i = (p_{t_1}^{(i)}, p_{t_2}^{(i)}, ..., p_{tk}^{(i)})$$

where $\theta_i$ represents the subject distribution of the $i$th comment, $p_{tk}^{(i)}$ represents the probability of the $i$th comment on topic $k$, and $k$ is the number of topics. If the vector space of the two comments is closer, the similarity between the two comments is higher. Then, take each topic’s comment set as a document, and combine the content in the next comment of the theme as a sentence. So we can calculate the similarity between each sentence in the document and other sentences by using the cosine angle, therefore to obtain the similarity of the comment set under a certain theme, which is recorded as Relativity.

4.4. Emotional consistence of comments
Zhou Hong et al. [11] point out that the emotional consistency of the initial comments and the additional comments has a great impact on consumers’ information adoption. Therefore, the second indicator in the model constructed in this study is the emotional consistency of the initial and additional comments.
This study uses polarity scores to measure the emotions of comments. It is stipulated that the score is positive for positive emotions and negative for negative emotions, and the emotional polarity scores are calculated with reference to Wang Qianqian [12]:

- Extract the product attribute words and calculate their weights. In this study, the ICTCLAS word segmentation system is used for word segmentation and part-of-speech tagging, and then the tagged phrases are classified into parts of speech to extract product attribute words. Then the weight of an attribute word is determined according to the frequency of its occurrence in the whole comment corpus.
- Calculate the emotional score of the product attribute words. This study calculates the emotional scores of product attribute words with the method proposed by Zhang Jing and Jin Hao [13].
- Calculate the score of comment text. Suppose there are m comments, in which the i (0 < i < m) comments have y (0 < y <= n) emotional words. And the words’ weights are: \( w_1, w_2, w_3, ..., w_y \), emotional scores are \( x_1, x_2, x_3, ..., x_y \), then the polarity score \( E \) of the ith comment is \( E = w_1 x_1 + w_2 x_2 + w_3 x_3 + ... + w_y x_y \). It also stipulates that E is a good rating between 1 and 0.667, a middle rating between 0.666 and 0.334, and a negative rating between 0.333 and 0
- Calculate the consistency score. Wang Changzheng [14] believes that in the comments with additional comments, inconsistent comments can give readers a higher sense of usefulness than consistent comments. Therefore, this study stipulates that if the initial comment contradicts the additional comment sentiment tendency, the consistency score \( \text{Consistence} = 1 \), otherwise \( \text{Consistence} = 0.5 \).

4.5. Information amount of comments

It is generally believed that the more information a comment contains, the higher the quality of the comment. Therefore, the third indicator in the model constructed in this study is the amount of information contained in the additional comments. It is generally agreed that the amount of information contained in a comment is proportional to the number of comments. So, this study stipulates that the comments which have less than 10 Chinese characters, information amount \( \text{Information} = 0.1 \), 20–30 Chinese characters, \( \text{Information} = 0.2 \), and so on, 90–100 Chinese characters, \( \text{Information} = 0.9 \), 100 Chinese characters or more, \( \text{Information} = 1 \). 

4.6. The construction of model

Based on the above indicators, this study constructed a multiple linear regression model:

\[
Q = w_1 \times \text{Relativity} + w_2 \times \text{Consistence} + w_3 \times \text{Information} \tag{4}
\]

The weight of each indicator \( w_i = (w_1, w_2, w_3) \) (0 < \( w_i < 1 \) \( \sum w_i = 1 \)) is determined by the questionnaire method and the gray correlation analysis method. Referring to the research of Zhou Linlin et al. [15], the specific steps are as follows:

- Get the weight of each indicator by the questionnaire survey method
  \[
  X = (x_{ij})_{n \times m}, i=1, 2, ..., n; j=1, 2, ..., m
  \tag{5}
  \]
- Determine the reference sequence \( X_0 \), then select the largest weight value from it and assign the value to each indicator as the reference weight value.
- Calculate the distance between each indicator sequence \( X_1, X_2, ..., X_m \) and the reference sequence \( X_0 \)
  \[
  b_j = \frac{1}{\Sigma_{i=1}^{n} (x_0(i) - x_j(i))^2}, j=1, 2, ..., m
  \tag{6}
  \]
- Calculate the weight of each indicator
  \[
  w_j = \frac{1}{1 + b_j}
  \tag{7}
  \]
- Calculate the normalized weight of each indicator
  \[
  w'_j = \frac{w_j}{\Sigma_{j=1}^{m} w_j}
  \tag{8}
  \]
5. Experimental results and analysis

5.1. Data collection and processing
This study selects online comments of AL MI speaker sold by Xiaomi flagship store as the research object, and uses the reptile software to obtain 1880 data from Taobao. After preliminary screening, 900 valid data are left and divided into 600 training set and 300 test set. Use NLPIR-ICTCLAS Chinese Word Segmentation System to segment and part-of-speech tagging all commentary texts, the specific effect is shown in Table 1. Filter the stop words of texts, and the specific effect is shown in Table 2.

| Types       | Content                                                                 |
|-------------|-------------------------------------------------------------------------|
| Initial     | Sound/NN quality/NN is/VBZ good/JJ, there/EX are/VBP many/DT favorite/JJ |
| comment     | songs/NNS, /PUN operas/NNS can/MD be/VB repeated/VPN after/IN downloading/VBG, too/RB good/JJ to/T0 use/VB. You/PRP can/MD also/RB use/VB this/DT speaker/NN to/TO tell/VV stories/NNS and/CC sing/VB children/NNS ‘s/POS songs/NNS. I/PRP bought/VBD it/PRP during/PRP the/DT activity/NN period/NN. It/PRP ‘s/VB a/DT lot/NN of/IN discount/NN. The/DT quality/NN of/IN MI/NNP is/VBZ very/RB good/JJ. Many/DT products/NNS used/VBD at/IN home/NN are/VB under/IN MI/NNP. |
| Additional  | Customer/NN service/NN attitude/NN is/VBZ very/RB good/JJ. They/PRP are/VB |
| comment     | patient/JJ and/CC professional/JJ.                                    |

5.2. The calculation of review topic relevance
Use Java language to generate LDA probability model for comment text after word segmentation. Set the number of topics to 6, and the number of loop iterations to 100. The "topic-word" matrix and the "comment-theme" matrix are obtained through LDA clustering, as shown in Table 3 and Table 4 respectively.

| Types: Topic 1: Acoustic Fidelity | Types: Topic 2: Appearance and Material | Types: Topic 3: Brand Effect |
|----------------------------------|----------------------------------------|----------------------------|
| distinguish 0.002                | light 0.004                             | MI 0.015                   |
| loud 0.005                       | convenient 0.005                        | brand 0.006                |
| clear 0.005                      | excellent 0.006                         | trust 0.009                |
| good 0.009                       | quality 0.010                           | mi fans 0.006              |
| nice 0.011                       | awesome 0.008                           | public praise 0.004        |
| speaker 0.003                    | beautiful 0.003                         | certified product 0.005   |
| ……                               | ……                                     | ……                        |

| Types: Topic 4: Function         | Types: Topic 5: Logistics and Customer service | Types: Topic 6: Price |
|----------------------------------|-----------------------------------------------|-----------------------|
| Xiao Ai 0.014                    | fast 0.008                                    | economical 0.006      |
| Bluetooth 0.009                  | attitude 0.009                                | cost performance 0.003|
| talk 0.006                       | patient 0.007                                | expensive 0.005       |
| intellect 0.006                  | speed 0.007                                  | cut price 0.007       |
| many 0.006                       | service 0.008                                | price differences 0.007|
| powerful 0.004                   | low 0.004                                    | cost-efficient 0.004  |
| ……                               | ……                                           | ……                    |
Table 4. Review topic probability distribution matrix (partial)

| Number | Topic 1     | Topic 2     | Topic 3     | Topic 4     | Topic 5     | Topic 6     |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|
| 1      | 0.16666425  | 0.16666628  | 0.16662933  | 0.16667862  | 0.16621959  | 0.16714121  |
| 2      | 0.12321197  | 0.15479721  | 0.41699658  | 0.22294736  | 0.00193321  | 0.08001342  |
| 3      | 0.16621331  | 0.16660569  | 0.16666297  | 0.16459897  | 0.16666699  |             |
| 4      | 0.00248055  | 0.00028331  | 0.35312477  | 0.00298001  | 0.0000012   | 0.63495133  |
| 5      | 0.00798814  | 0.00788733  | 0.00799455  | 0.56468921  | 0.00799862  |             |

It is stipulated that in the "comment-topic" matrix, the maximum number of similarities between each comment and the topic should be taken as the theme relevance of the comment.

5.3. Calculation of indicator weights

In this study, 500 electronic questionnaires are distributed to people of all ages, and they score the importance of additional comments' three indicators: relevance, consistency and information amount. Finally, 385 valid questionnaires are collected. The questionnaires are divided into four categories: under 20 years old, 20-30 years old, 30-40 years old, and over 40 years old. The data of the collected questionnaires are compiled and calculated, and the weights of the indicators are shown in Table 5. It is stipulated that in the "comment-topic" matrix, the maximum number of similarities between each comment and the topic should be taken as the theme relevance of the comment.

Table 5. Indicator weights determined by the questionnaire survey method

| Index     | Under-20 | 20–30 | 30–40 | over 40 |
|-----------|----------|-------|-------|---------|
| Relativity| 0.33     | 0.54  | 0.45  | 0.39    |
| Consistence| 0.18     | 0.38  | 0.33  | 0.18    |
| Information| 0.49     | 0.08  | 0.21  | 0.43    |

According to the steps described in Section 4.6, introduce the gray correlation method, and then the weights of the three indicators are finally determined to be (0.38, 0.31, 0.31). So the model of this study is:

\[ Q = 0.38 \times \text{Relativity} + 0.31 \times \text{Consistence} + 0.31 \times \text{Information} \]  

(9)

5.4. Analysis of Model experiment results

After the experiment, the results calculated by the model are obtained, and the Q value is defined as a low-quality comment in the interval \([0, 0.5]\), and a high-quality comment in the interval \([0.5, 1]\). Applying the model and conclusion to the test set, we can get the accuracy rate is \(P = 80.73\%\), the recall rate is \(R = 64.90\%\), and the F value is \(F = 71.95\%\). Therefore, it concludes that this model is a good evaluation model for additional comment texts.

6. Conclusion

With the rapid development of e-commerce, consumers and businesses urgently need a method to help them find useful information from a large number of disorderly online comments, so as to make correct judgments and choices in a short time. Based on the concept of text analysis, this paper proposes a quality evaluation model of online review, and chooses the additional comment of the MI Bluetooth speaker in Taobao as the test object. The usefulness of the model is verified by experiments.

This model can help consumers judge the quality of online review and provide services for consumers to make purchasing decisions. At the same time, the model can also help merchants analyze and identify reliable review information to accurately grasp the user’s consumer demand. In view of the shortcomings of the model, such as too few commodities as the experimental object, I will continue to improve in the future study.

References

[1] Kim S, Pantel P, Chkolovski T, et al. Automatically assessing review helpfulness [C] //Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2006:423-430.
[2] Hyunmi Baek, JoongHo Ahn, Youngseok Choi. Helpfulness of Online Consumer Reviews: Readers' Objectives and Review Cues [J]. International Journal of Electronic Commerce, 2012, 17(2).

[3] Huang Ting-ting, Zeng Guo-zhen, Xiong Huang-liang. Trusted Sorting Method for Shopping Customer Comments Based on Product Feature Correlation [J]. Journal of Computer Applications, 2014, 34(08): 2232-2327+2.

[4] Tang Xiaobo, Qiu Xin. Study on the topic-oriented high-quality comment mining model [J]. Modern library and information technology, 2015(Z1): 104-112.

[5] Li Zhiyu. Study on the Ranking Model of Online Commodity Review Utility [J]. New Library and Information Technology, 2013(04): 62-68.

[6] Liu Jingjing, Cao Yunbo, Lin Chin-Yew, et al. Low-quality product review detection in opinion summarization [C] //Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics, 2007: 334-342.

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[8] Titov I, McDonald R. Modeling Online Reviews with Multi-grain Topic Models [C]. In: Proceedings of the 17th International Conference on World Wide Web (WWW'08). New York: ACM, 2008: 111-120.

[9] Tang Xiaobo, Qiu Xin. Study on the topic-oriented high-quality comment mining model [J]. Modern library information technology, 2015(Z1): 104-112.

[10] Gross A, Murthy D. Modeling Virtual Organizations with Latent Dirichlet Allocation: A Case for Natural Language Processing [J]. Neural Networks, 2014, 58: 38-49.

[11] Zhou Hong, Li Sujuan. The Influence of Online Appropriate Comments Consistency on Consumer Information Adoption [J]. Journal of Wuhan University of Technology (Information & Management Engineering), 2018, 40(05): 537-543.

[12] Wang Qianqian. A Method for Ranking the Credibility of Online Commodity Comment Information [J]. Journal of Information, 2015, 34(03): 181-185.

[13] Zhang Jing, Jin Hao. Study on Automatic Judgment of Chinese Words' Emotional Tendency [J]. Computer Engineering, 2010, 36(23): 194-196.

[14] Wang Changzheng, He Wei, Wang Kui. Study on the Usefulness Perception of Additional Comments in Internet Word of Mouth [J]. Management Sciences, 2015, 28(03): 102-114.

[15] Zhou Linlin, Li Xiaoqi. Determining Weights Based on Grey Correlation Method and Its Application [C]. //The 8th China Intelligent Computing Conference and International E-Commerce Federation China Chapter 3rd Annual Conference Proceedings. Northeastern University% Northeastern University, 2014: 19-23.