Relation Analysis between Hotel Review Rating Scores and Sentiment Analysis of Reviews by Chinese Tourists Visiting Japan

Elisa Claire Alemán Carreón[a,*], Hirofumi Nonaka[a], Toru Hiraoka[b]

[a] *Nagaoka University of Technology, Nagaoka, Japan*
[b] *University of Nagasaki, Nagasaki, Japan*

Abstract

In current times, the importance of online hotel review sites has become more and more apparent. Users of these sites reference of reviews strongly influences their purchase behavior and as such, reviews are important to companies and researchers alike. The majority of review sites offer both text reviews and numerical hotel ratings, and both information sources are widely used by researchers as a representation of a customer's sentiment and opinion. However, an opinion is a difficult concept to measure, and as such, depending on the relation these two sources have, it would be apparent whether or not it is safe to consider them equally in research. In this study we utilize an entropy-based Support Vector Machine to classify positive and negative sentiments in hotel reviews from the site *Ctrip*, then calculating the ratio of positive and negative sentiment in each review and examine their correlation with said review's rating score using Spearman and Kendall Correlation coefficients and Maximal Information Coefficient (MIC).

Keywords:

Sentiment Analysis, Tourism, SVM, Machine Learning, Chinese

*Corresponding author

Email addresses: s153400@stn.nagaokaut.ac.jp (Elisa Claire Alemán Carreón), nonaka@kjs.nagaokaut.ac.jp (Hirofumi Nonaka), hiraoka@sun.ac.jp (Toru Hiraoka)

1. Introduction

As the number of Chinese tourists visiting Japan increases, it is important for the hotel industry to conduct market research to analyze the needs of hotel guests. Under these circumstances, grasping needs through questionnaires and interviews has become the center of market research. However, surveys using such questionnaires and interviews have problems in terms of cost and real-time performance. On the other hand, with the spread of the Internet, there are many online reviews. Users will use the opinions of others as reference, and because of the large influence these have on purchase decisions [1, 2], these have also been actively used in the industry. Users' evaluations on many online review sites are numerous, and are divided into comment text as text information and a review score expressed numerically. The review scores are structured numerical information and are easy to use for analysis, so they are being used as evaluation indexes for users' products and services [3, 4, 5]. On the other hand, text reviews are also being analyzed based on natural language processing. For example, research that analyzes sentiment of word-of-mouth using information theory [6] as well as forecast of product sales [7] or the ranking of products based on sentiment analysisis [8] is being conducted.

In customer trend analysis, what is used as an analysis index is extremely important. As mentioned above, in previous studies, customer behavior was analyzed using sentiment analysis only [6, 8] and there are many studies that use only the review score points [3, 4, 5]. Therefore, it is necessary to examine the relationship between these review scores and the sentiment analysis of comments. If the relationship between the review score and the sentiment analysis of the document is high, it would be easy and there would be a great merits to do analysis based only on the numerical information or using the numerical information as training data for future sentiment analysis. On the other hand, if the correlation is low, it is necessary to comprehensively evaluate the reviews by using the score and text information together. Alternatively, it is necessary to select a method that more reflects the user's emotions and opinions and use

it as an evaluation index. In this way, investigating the relationship between the review score and the sentiment analysis of the comment text is extremely important in the analysis of the review.

Therefore, in this study, we investigated the relationship between the review scores and the sentiment analysis of the comment text. In this study, we first collected a large number of review documents written about Japanese hotels and their review score from the online hotel review site *Ctrip* for Chinese tourists. Next, we trained an SVM using the entropy-based feature selection method developed by our researchers, and classified documents that express positive emotions and documents that express negative emotions. Since the feature vector characteristic of this emotion classification was constructed by keyword extraction based on entropy, the classification was performed using a language-independent method based on statistics. Furthermore, based on the emotion classification of each sentence in each review, the ratio of sentences expressing satisfaction and the ratio of dissatisfaction sentences to the full review were calculated in order to quantify emotions. Finally, the interrelationship between the review score and the emotional evaluation of the review was analyzed from Spearman's rank correlation coefficient and Kendall's rank correlation coefficient MIC. This will be described in detail below.

2. Previous Work

In previous research, targeting product reviews is the mainstream. For example, a study that applied Word Cloud to extract words that are often used by consumers [9], and a market analysis using sentiment analysis of product reviews using an emotion dictionary called HowNet [10]. On the other hand, as for quantitative evaluation of the impact of hotel online reviews, there is a study that proves that hotel online reviews have a particular effect on customer motivation [1]. There are other studies that have shown a relationship between sales and the textual information of reviews that are evaluated [11]. In the mentioned studies, the scores were not considered, and the relationship between

the scores and the text information was not focused on.

3. Methodology

3.1. Preprocessing

At the crawling stage, we used the fact that the URL structure is determined by the hotel ID number, and so each hotel page can be automatically loaded. Next, scraping was performed, and the documents, IDs, review score, etc. of each review were acquired using the structure of the HTML code and saved in our database. Morphological analysis was also performed for the review. As a tool for morphological analysis of Chinese, Stanford Word Segmenter [12] provided by The Stanford NLP Group of Stanford University was used.

3.2. Sentiment Analysis

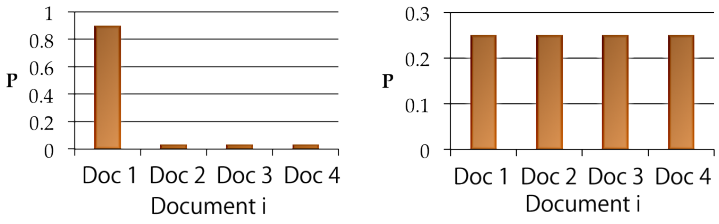
Samples were extracted from the data collected by online hotel reviews for sentiment analysis, and with the cooperation of three Chinese research students, each sentence in each review was tagged according to whether it expressed satisfaction or dissatisfaction. Manual classification of Positive and Negative was performed, and training data was created. Then, the words and emotion classification tags included in the sample review were trained by SVM, and the emotion classification of all the data in the population was performed. A feature vector characteristic of emotion classification by SVM was constructed by keyword extraction based on entropy, which will be described later. The details will be described below.

3.2.1. Entropy Based Keyword Extraction

In this study, we based the extraction of the keywords that are influenced by the users' emotional judgement on the calculation of an entropy value for each word. Speaking in Information Theory terms, Shannon's Entropy is the expected value of the information content in a signal [13]. Applying this knowledge to the study of words allows us to observe the probability distribution of any given word inside the corpus. For example, a word that keeps reappearing

in many different documents will have a high entropy, given that predicting on which document it would appear becomes uncertain. On the contrary, a word that only was used in a single text and not in any other documents in the corpus will be perfectly predictable to only appear in that single document, bearing an entropy of zero. This concept is shown in Fig.1.

Based on the meaning of entropy explained above, keywords that will be considered positive will have a large entropy when they appear in many positive documents, and a smaller entropy in negative documents. The same will occur for negative keywords in the opposite documents. In this study we use the entropy values of keywords to perform a classification. First, we tagged a set of documents as positive or negative. Then, for each word j that appears in each document i, we counted the number of times a word appears in positive comments as N_{ijP}, and the number of times a word appears in negative comments as N_{ijN}. Then, as shown in the formulas below, we calculated the probability of each word appearing in each document shown below as P_{ijP} (1) and P_{ijN} (2).



(a) Entropy close to zero.

(b) High entropy.

Figure 1: Probabilities of a word j being contained in a document i.

$$P_{ijP} = \frac{N_{ijP}}{\sum_{i=1}^{M} N_{ijP}} \tag{1}$$

$$P_{ijN} = \frac{N_{ijN}}{\sum_{i=1}^{M} N_{ijN}} \tag{2}$$

We then substitute these values in the next formula. We calculated the entropy for each word j in relation to positive documents as H_{Pj} (3), and the

entropy for each word j in relation to negative texts as H_{Nj} (5). That is, as is shown in (4) and (6), all instances of the summation when the probabilities P_{ijP} or P_{ijN} are zero and the logarithm of these becomes undefined are substituted as zero into (3) and (5).

$$H_{Pj} = -\sum_{i=1}^{M}[P_{ijP}\log_2 P_{ijP}] \tag{3}$$

$$\lim_{P_{ijP}\to 0+} P_{ijP}\log_2 P_{ijP} = 0 \tag{4}$$

$$H_{Nj} = -\sum_{i=1}^{M}[P_{ijN}\log_2 P_{ijN}] \tag{5}$$

$$\lim_{P_{ijN}\to 0+} P_{ijN}\log_2 P_{ijN} = 0 \tag{6}$$

After calculating the entropies for each word, we adjusted for their α value by testing for the highest F-value. A positive keyword is determined when (7) is true, and likewise, a negative keyword is determined when (8) is true for the best performing α value, using these keywords as elements for training an SVM [14]. The performance was determined using a k-fold cross validation calculating the best F_1 value [15].

$$H_{Pj} > \alpha H_{Nj} \tag{7}$$

$$H_{Nj} > \alpha' H_{Pj} \tag{8}$$

3.3. Correlation Analysis

The following method was experimentally used to measure the correlation of the ratio of positive sentiment obtained by dividing the sentences judged as positive included in each review x to the review scores y.

3.3.1. Pearson correlation coefficient r

In order to measure the correlation of the sentiment ratio x, obtained by dividing the number of sentences judged as positive included in each review i_P by the total number of sentences i_T, to the score y, one of the methods used was Pearson's correlation coefficient r (**??**). The formula is shown below. The value of the formula (**??**) is substituted into the formula (**??**). When calculating the negative rate, substitute the number of sentences judged to be negative i_N into the formula (**??**).

$$x = \frac{i_P}{i_T} \tag{9}$$

$$r = \frac{\sum_{i=1}^{M}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{M}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{10}$$

3.3.2. Spearman's rank correlation coefficient ρ

In order to measure the correlation of the sentiment ratio x, obtained by dividing the number of sentences judged as positive included in each review i_P by the total number of sentences i_T, to the score y, since twe consider the score to be a ranked variable, we used Spearman's ranked correlation coefficient ρ (**??**) which is also based on Pearson's correlation coefficient. The formula is shown below. Substitute the value of the formula (**??**) into the formula (**??**).

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{11}$$

3.3.3. Kendal's rank correlation coefficient τ

Like Spearman's rank correlation coefficient, Kendal's rank correlation coefficient is used to investigate the relationship between the values that represent rank. The formula (**??**) is shown below. However, substitute the expression (**??**) and the expression (**??**) into the expression (**??**). Substitute the expression (**??**) for each.

$$\tau = (K - L)/\binom{n}{2} \tag{12}$$

$$L = \#\Big\{\{i,j\} \in \binom{[n]}{2} \mid \neg(x_i \lesseqgtr x_j, y_i \lesseqgtr y_j)\Big\} \tag{13}$$

$$K = \#\Big\{\{i,j\} \in \binom{[n]}{2} \mid (x_i \lesseqgtr x_j, y_i \lesseqgtr y_j)\Big\} \tag{14}$$

3.3.4. MIC

Pearson's correlation coefficient can only extract linear relationships. On the other hand, there is MIC (Maximal Information Coefficient) as a method to analyze the relationship between two variables including non-linearity [16]. It is an index for analyzing the relationship between variables including non-linearity based on the amount of mutual information, considering the two variables for which you want to analyze the relationship as random variables. Fig. **??** shows several examples of comparing the coefficients of MIC and Pearson. Pearson's coefficient is a value from 0 to 1 in a linear relationship, and it is determined whether it is a positive value or a negative value depending on the direction of inclination. On the other hand, in the case of MIC, even if it is a non-linear relationship, it can express the correlation using the value from 0 to 1 as long as there is a relationship. Examples of this are shown in Fig.2.

The procedure for calculating MIC will be described. First, for the variables X and Y to be analyzed, after plotting the two variables on the coordinate space, the space is divided by $a * b$ (Split the X direction into a parts, and the Y direction into b parts). Then, for each of the two variables, the cell existence probability can be calculated by dividing the number of sample points belonging to each cell by the total number of samples. That is, X and Y are regarded as random variables based on the existence probability in the cell. This makes it possible to calculate the Mutual Information for X and Y.

At this time, even if the original two variables have a non-linear relationship as well as a linear relationship, the dependency between the random variables becomes strong, so the mutual information takes a large value. Therefore, unlike Pearson's correlation coefficient, it is possible to extract a non-linear relationship. Now, with the MIC, the width and length of each cell are unequally spaced,

so there are innumerable division methods (note that each cell has a maximum resolution). For the purpose of extracting non-linearity, it is necessary to find a dividing grid that maximizes the amount of mutual information as much as possible. In MIC, it is assumed that X is now divided into arbitrary a pieces and Y is divided into arbitrary b pieces. At this time, we find a division that maximizes the amount of mutual information in the division of $a * b$ by brute force. The formula (15) for mutual information is shown below.

$$I(X;Y) = \int_Y \int_X p(x,y) \log(\frac{p(x,y)}{p(x)p(y)})\, dx\, dy \tag{15}$$

In this study, the MIC was calculated using the Python library minepy [17]. The Fig. **??**, which compares Pearson's correlation coefficient and MIC, was obtained from the minepy API site.

4. Experiment Results

The following describes what was specifically done in this study using the method described earlier.

4.1. Preprocessing

First, we collected 1,541,424 HTML files from May 2016 to September 2016 from *Ctrip*. Of the 1,541,424 HTML files, we collected data on 5,938 hotels in Japan. Among them, he scraped the comment text of 44,912 reviews. 286,109 sentences divided into sentence units are analyzed. In addition, the scores of the reviews were also collected.

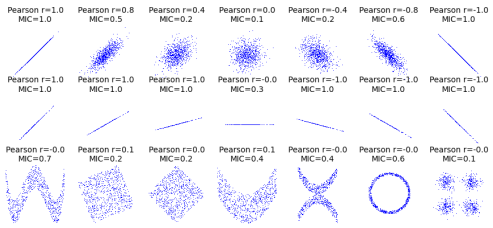


Figure 2: Comparison of minepy MIC and Pearson r in various cases

4.2. Sentiment analysis performance

In order to create training data, a sample of reviews was randomly extracted from all the data, divided into sentences by expressing satisfaction or dissatisfaction, tagging work of 159 sentences was performed manually. The entropy was calculated from the sentences that become the training data.

After calculating the entropy for the training data, the features with the maximum F_1 value were selected by evaluating α from 1.0 to 3.75 with a step size of 0.25 in order to obtain its optimum value. After training the SVM, the keyword list was selected based on the α which led to the maximum F_1 value as a result of 5-Fold Cross Validation ($k = 5$) for the evaluation data. Furthermore, both lists were combined to create a new list, and 5-Fold Cross Validation was performed in the same manner. The evaluation results are shown in Table 1. The feature that combines both has the highest F_1 value, and F_1 classifies all data with an SVM with a high accuracy of 0.95. Positive keywords are, for example, "情" and "景色" (indicating "friendly" and "(good) scenery" respectively), and in the case of Negative keywords, there was an example of "价格" or "price" (indicating dissatisfaction because it is high).

Table 1: 5-fold Cross Validation performance results

Keyword list	C	$F_1\mu$	$F_1\sigma$
Positive keywords ($\alpha = 2.75$)	2.5	0.91	0.01
Negative keywords ($\alpha' = 3.75$)	0.5	0.67	0.11
Combination	**0.5**	**0.95**	**0.01**

Table 2: Correlation of sentiment analysis and score results

Sentence ratio	**Spearman's ρ**	**Kendall's τ**	**MIC**
Positive ratio	0.161	0.125	0.049
Negative ratio	-0.149	-0.122	0.0447

After learning with the optimal model, sentiment analysis was performed on unknown data, and the above-mentioned "Positive ratio" and "Negative ratio"

were calculated. Since these values are the ratio of sentences expressing satisfaction and sentences expressing dissatisfaction, they were treated as numerical coefficients expressing emotions in each sentence, and correlation analysis was performed.

4.3. Correlation analysis

The positive and negative rates of each review in all the data and the review score of the reviews were analyzed using Spearman's rank correlation coefficient, Kendall's rank correlation coefficient, and MIC. The results are shown in Table 2.

5. Discussion

It was shown that both the Positive ratios and the Negative ratios were very low in relation to the review score in all the indicators.

Therefore, since the relationship between the result of sentiment analysis in the comment text and the review score is low, when analyzing the user's opinion from the review, considering both the content of the text and the numerical review score and their differences is very important.

In the past, although this relationship has not been shown, only review scores are often used as indicators of satisfaction and emotional evaluation. For example, the studies by Xie et al. And Burchand-Gidumal et al. used the scores for the hotel as a proxy for satisfaction[3, 4], and Zhou et al. investigated the factors of satisfaction using a multivariate analysis to do this, but the dependent variable was the review score [5]. Based on the results of this study, it was suggested that it is not appropriate to use only the review score or sentiment analysis as an index to measure the satisfaction of tourists.

6. Conclusion

In this study, we investigated the relationship between the results of sentiment analysis in the text text of online hotel reviews and the evaluation points.

We constructed a method with high classification performance ($F_1 = 0.95$) for sentiment analysis, and calculated the ratio of sentences classified as Positive and the ratio of sentences classified as Negative for each review. For the relationship, Spearman's rank correlation coefficient, Kendall's rank correlation coefficient, and MIC were used. As a result, all showed low values. Therefore, it was clarified that the relationship between the result of sentiment analysis in the comment text and the evaluation point, which is numerical information, is low. Therefore, it was considered that a more comprehensive evaluation was important. In the future, based on these results, we will investigate whether the result of sentiment analysis or the evaluation score expresses the user's opinion more after comparing with the analysis using multilingual information, and comprehensively reviewing. We will proceed with the development of analysis methods.

Acknowledgements

We would like to thank Mr. Liangyuan Zhou and Ms. Eerdengqiqige for their support in creating teacher data in Chinese. In addition, this paper was supported by the "Japan Construction Information Center".

References

[1] I. E. Vermeulen, D. Seegers, Tried and tested: The impact of online hotel reviews on consumer consideration, Tourism Management 30 (1) (2009) 123–127. doi:10.1016/j.tourman.2008.04.008.
URL https://www.sciencedirect.com/science/article/pii/S0261517708000824

[2] B. A. Sparks, V. Browning, The impact of online reviews on hotel booking intentions and perception of trust, Tourism Management 32 (6) (2011) 1310–1323. doi:10.1016/j.tourman.2010.12.011.
URL https://www.sciencedirect.com/science/article/pii/S0261517711000033

[3] K. L. Xie, Z. Zhang, Z. Zhang, The business value of online consumer reviews and management response to hotel performance, International Journal of Hospitality Management 43 (2014) 1–12. doi:10.1016/j.ijhm.2014.07.007.
URL https://www.sciencedirect.com/science/article/pii/S027843191400125X

[4] J. Bulchand-Gidumal, S. Melián-González, B. González Lopez-Valcarcel, A social media analysis of the contribution of destinations to client satisfaction with hotels, International Journal of Hospitality Management 35 (2013) 44–47. doi:10.1016/j.ijhm.2013.05.003.
URL https://www.sciencedirect.com/science/article/pii/S0278431913000728

[5] L. Zhou, S. Ye, P. L. Pearce, M.-Y. Wu, Refreshing hotel satisfaction studies by reconfiguring customer review data, International Journal of Hospitality Management 38 (2014) 1–10. doi:10.1016/j.ijhm.2013.12.004.
URL https://www.sciencedirect.com/science/article/pii/S0278431913001801

[6] R. K. Amplayo, M. Song, An adaptable fine-grained sentiment analysis for summarization of multiple short online reviews, Data & Knowledge Engineering 110 (2017) 54–67. doi:10.1016/j.datak.2017.03.009.
URL https://www.sciencedirect.com/science/article/pii/S0169023X16301525

[7] Z.-P. Fan, Y.-J. Che, Z.-Y. Chen, Product sales forecasting using online reviews and historical sales data: A method combining the bass model and sentiment analysis, Journal of Business Research 74 (2017) 90–100. doi:10.1016/j.jbusres.2017.01.010.
URL https://www.sciencedirect.com/science/article/pii/S0148296317300231

[8] Y. Liu, J.-W. Bi, Z.-P. Fan, Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory, Information Fusion 36 (2017) 149–161. doi:10.1016/j.inffus.2016.11.012.
URL https://www.sciencedirect.com/science/article/pii/S1566253516301580

[9] C. Hargreaves, Analysis of hotel guest satisfaction ratings and reviews: An application in Singapore, American Journal Of Marketing Research 1 (4) (2015) 208–214.

[10] H. Zhang, Z. Yu, M. Xu, Y. Shi, Feature-level sentiment analysis for chinese product reviews, in: 2011 3rd International Conference on Computer Research and Development, Vol. 2, IEEE, 2011, pp. 135–140. doi:10.1109/ICCRD.2011.5764099.

[11] S. Basuroy, S. Chatterjee, S. Ravid, How critical are critical reviews? the box office effects of film critics, star power, and budgets, Journal Of Marketing 67 (4) (2003) 103–117. doi:10.1509/jmkg.67.4.103.18692.

[12] P. Chang, M. Galley, C. Manning, Optimizing Chinese word segmentation

for machine translation performance, in: Proceedings of the Third Workshop On Statistical Machine (Statmt '08), Columbus,Ohio,USA, 2008, pp. 224–232.
URL http://nlp.stanford.edu/pubs/acl-wmt08-cws.pdf

[13] C. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (3) (1948) 279–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

[14] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297. doi:10.1007/bf00994018.

[15] D. Powers, Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation, Journal Of Machine Learning Technologies 2 (1) (2011) 37–63.
URL http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf

[16] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, P. C. Sabeti, Detecting novel associations in large data sets, Science 334 (6062) (2011) 1518–1524. arXiv:https://www.science.org/doi/pdf/10.1126/science.1205438, doi:10.1126/science.1205438.
URL https://www.science.org/doi/abs/10.1126/science.1205438

[17] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, C. Furlanello, minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers, Bioinformatics 29 (3) (2012) 407–408. arXiv:https://academic.oup.com/bioinformatics/article-pdf/29/3/407/17105876/bts707.pdf, doi:10.1093/bioinformatics/bts707.
URL https://doi.org/10.1093/bioinformatics/bts707