Analysis of Chinese Tourists in Japan by Text Mining of a Hotel Portal Site

Elisa Claire Alemán Carreón[1], Hirofumi Nonaka[1], Toru Hiraoka[2]

[1] Nagaoka University of Technology, Nagaoka, Japan
s153400@stn.nagaokaut.ac.jp, nonaka@kjs.nagaokaut.ac.jp
[2] University of Nagasaki, Nagaoka, Japan
hiraoka@sun.ac.jp

Abstract. With an increasingly large number of Chinese tourists in Japan, the hotel industry is in need of an affordable market research tool that does not rely on expensive and time consuming surveys or interviews. Because this problem is real and relevant to the hotel industry in Japan, and otherwise completely unexplored in other studies, we have extracted a list of potential keywords from Chinese reviews of Japanese hotels in the hotel portal site *Ctrip*[1] using a mathematical model to then use them in a sentiment analysis with a machine learning classifier. While most studies that use information collected from the internet use pre-existing data analysis tools, in our study we designed the mathematical model to have the highest possible performing results in classification, while also exploring on the potential business implications these may have.

Keywords: Text Mining • Sentiment Analysis • Support Vector Machine • Machine Learning • Entropy

1   Introduction

In Japan, the population of Chinese tourists has increased over the recent years, with a 107.3% increase from 2014 to 2015 [1] and so on. This change in a customer base for many industries, mainly the hotel industry, brings forward a need for an affordable and reliable method to study this new market; which has not been explored in other studies before. Surveys and interviews, previously used both in business and other studies, present difficulties and an increase in costs and time, besides the fact that these studies [3, 4] do not target Japan or Chinese customers [5].

Using the available text reviews of Japanese hotels by Chinese customers available in the website *Ctrip*, our new method proposal is to use a mathematical model using Shannon's Entropy, a concept which can determine a word's

---

[1]   Ctrip. http://www.ctrip.com/

probabilistic distribution, to determine a list of relevant keywords to be used in the context of classifying positive and negative emotions in a hotel review using a Support Vector Machine.

There are other studies using similar techniques, mostly with English texts [6,7] but not many that apply these knowledges practically, and those that do [8] don't analyze the implications fully. The ones that do focus on Chinese sentiment classification [2] or its business applications [9] are scarce, and use formal words that describe emotions that are unlikely to appear in our reviews. This is why we developed our own sentiment analysis model to provide a practical tool for future marketing choices in our study. In contrast to these studies, in our entropy based keyword extraction we are using words used by a specific group of users that are characteristic of documents written in different emotional states, which has a better capacity for more precise results in the sentiment analysis classification by the Support Vector Machine.

2   Methodology

2.1   Word Segmentation

For a statistical analysis to be made possible for each word, we segmented the collected Chinese texts without spaces into words using the Stanford Word Segmenter [10].

2.2   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 [11]. 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 the figure below.
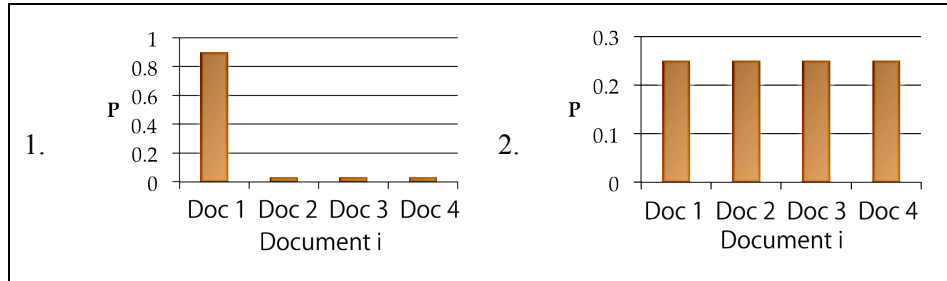
Figure 1. Probabilities of a word *j* being contained in a document *i*
1. Entropy close to zero, 2. High entropy

With this logic in mind, we used a set of documents that were previously tagged as positive or negative by a group of Chinese students. If a word has a higher entropy in positive documents than in negative documents by a factor of alpha (α), then it means its probability distribution is more spread in positive texts, meaning that it is commonly used in positive tagged documents compared to negative ones.

To calculate the entropy in a set of documents, 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).

$$P_{ijP} = \frac{N_{ijP}}{\sum_{i=1}^{M} N_{ijP}} \qquad (1)$$

$$P_{ijN} = \frac{N_{ijN}}{\sum_{i=1}^{M} N_{ijN}} \qquad (2)$$

We then substitute these values in the formula that defines Shannon's Entropy. 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).

$$H_{Pj} = -\sum_{i=1}^{M} \left[P_{ijP} \log_2 \left(P_{ijP} \right) \right] \qquad (3)$$

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

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

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

After calculating the positive and negative entropies for each word, we measured their proportion using the mutually independent coefficients α for positive keywords and α' for negative keywords, for which we applied several values experimentally. A positive keyword is determined when (7) is true, and likewise, a negative keyword is determined when (8) is true.

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

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

2.3   Sentiment Analysis Using Support Vector Machine

In machine learning, Support Vector Machines are supervised learning models commonly used for statistical classification or regression [12]. Using already classified and labeled data with certain features and characteristics, an SVM learns to classify new unlabeled data by drawing the separating (p-1)-dimensional hyperplane in a p-dimensional space. Each dimensional plane is represented by one of the features that a data point holds. Then each data point holds a position in this multi-dimensional space depending on its features. The separating hyperplane and the supporting vectors divide the multi-dimensional space by minimizing the error of classification. A two-dimensional example is shown below.
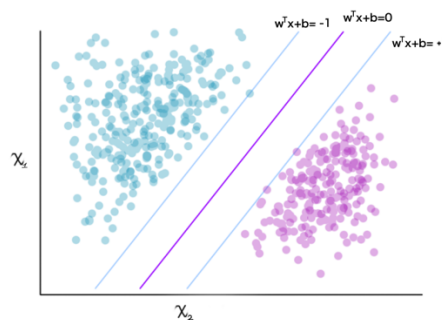

Figure 2. 2D example of the Linear SVM classification

The linear kernel for the SVM classification is defined by the formula (9) below. The influence that each point of the training data inputs into the vector is defined by their weight w_n, included in the Weight Vector w. The bias coefficient b determines the position of the hyperplane.

$$f(x) = w^T x + b \qquad (9)$$

Then, the conditions shown in (10) are applied when classifying new data.

$$f(x) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases} \qquad (10)$$

Now, the initial condition is set as w = 0. Then each possible separating vector is tested, and when a classification for x_i fails, the value for w is changed as follows in (11) by a value of α.

$$w \leftarrow w + \alpha sign(f(x_i))x_i \qquad (11)$$

This process is repeated until all points of the training data are classified correctly. The resulting formula for classification (12) is as follows.

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i (x_i^T x) + b \qquad (12)$$

In Natural Language Processing, when it comes to statistically analyzing documents, each possible word in that corpus is a feature, or a dimensional plane. Then, the value of that feature will be marked as the number of times a word j is contained in a document i.

We implemented this theory in Python using the Support Vector Classifier (SVC) included in the library *scikit-learn*[2]. To vectorize the documents into word vector spaces we used the method *CountVectorizer* included in the same library. We then managed these vectors using the mathematics library *numpy*[3].

To evaluate each of our trained machines, we used the K-fold Cross Validation method, which has been proven to provide good results [13]. In each test, we calculated then the Precision, Recall, F_1 [14] and Accuracy values for our predictions.

---

[2]   Scikit-Learn http://scikit-learn.org/
[3]   Numpy. http://numpy.org/

3   Experiments

3.1   Dataset

We crawled a total of 1,541,424 html files, from which 5,938 were unique Japanese hotels. From these pages, we extracted a total of 44,912 reviews, which turned into 286,109 separate sentences. In our corpus, there were 23,443 different words used, from which 2,802 were noise characters.

3.2   Sentiment Analysis Experiments

Before performing the sentiment analysis, we had a group of Chinese student collaborators tag a sample of our collected data, (159 documents, 117 tagged positive and 42 tagged negative) and then calculated the Shannon's entropy values for each word in positive and negative tags.

We then altered the comparison coefficient α and α' to different values, from which we extracted different lists of positive and negative keywords to be tested in an SVM. We changed the values of α and α' in 0.25 intervals from 1.0 to 3.75.

In our study, we investigated different parameters to decide the kernel and model of the learning machine. The best performing kernel for the SVM was the Linear one, with a C=3.0.

Based on our K-Fold Cross Validation (k=10) results, we chose the best performing positive and negative lists. We decided to make a combined list of these two previous lists and train the SVM with this. With this Combined list, the Accuracy was of 0.90 ± 0.09 and the F_1 measure was 0.93 ± 0.05, both excellent results for classification.

Table 1. Results of the K-fold Cross Validation performance tests[4]

Keyword List	Accuracy Average	Accuracy Std. Dev.	F1 Average	F1 Std. Dev.
Positive (α=2.75)	0.58	0.16	0.69	0.16
Negative (α=3.75)	0.88	0.07	0.92	0.05
Combined	**0.90**	**0.09**	**0.93**	**0.05**

Following this observation, we predicted whether a sentence was positive or not from the remaining unlabeled data using this model. We then made an-

---

[4]   Not all the lists are displayed. Only the ones with the highest performance rates are shown.

other prediction of whether a sentence was negative or not using the Negative (α=3.75 in formula (8)) keyword list trained linear SVM. Having both of these predictions, we decided to classify the documents based on a consensus of the two, having the categories "positive", "neutral" and "negative".

4 Results and Discussion

4.1 About the Methodology

We have shown that the entropy based keyword extraction method has many advantages for this field of study. Not only this, but we have obtained a result that exceeds a 0.9 performance (accuracy and F_1), which can be considered as high.

4.2 About the Results

Now, these are the relevant keywords extracted from the entropy calculations that were used to train the SVM for positive and negative classifying. Because the keyword list for positive words is larger than 100 words, we will only show the top words that are better suited for analysis, such as nouns and adjectives.

Table 2. Keywords worthy of note

Word	Translation	Word	Translation	Word	Translation
园林	garden	烤肉	Barbecue	人员	personnel
花园	flower garden	中文	Chinese text	华人	Chinese person

5 Conclusion and Future Work

In our study, with the objective to understand what Chinese customers of Japanese hotels demand and feel we extracted keywords from a Chinese portal site with reviews of Japanese hotels. In the evaluation process of our machine learning experiments, we obtained our highest performance (F1=0.93±0.05 and Accuracy=0.90±0.09) using both positive and negative keywords extracted using Shannon's entropy as a base to train a linear kernel SVM.

In regards to the needs of Chinese customers of Japanese hotels we found that the Chinese customers in our database were less concerned with price and

more particularly with food, gardens or parks near the hotel, good service and transport, as well as the use of Chinese language to cater to them specifically.

It could be thought that some of the keywords being used were not about the hotels themselves, but for the surrounding area, which is a subject we will leave for future works. Another subject to investigate is how these keywords could be rated to determine the most important of the demands of Chinese customers of Japanese hotels.

Acknowledgements

References

1. Japan National Tourism Organization,
   `http://www.jnto.go.jp/jpn/statistics/visitor_trends/`
2. Lee, Y.M.: A linguistic approach to emotion detection and classification. In: The Hong Kong Polytechnic University (2010)
3. Chang, C.Y., Kivela, J., Mak, H.N.: Food preferences of Chinese tourists. In: Annals of Tourism Research, Vol. 37, No. 4, pp.989-1011. (2010)
4. Thuy-Huong Truong, Brian King. 2009. An evaluation of satisfaction levels among Chinese tourists in Vietnam. In: International Journal of Tourism Research. Vol. 11. No. 6. pp.521-535. (2009)
5. Ma, Y.F., Zheng, P., Bai, K.: Study on the Discrepancies of Decision-making Behavior of Tourists to China——Take Tourists from Japan and America as a Case Study. In: College of Tourism and Environment of Shaanxi Normal University, Xi`an 710062, China (2008)
6. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. In: Journal of Computational Science Vol. 2, No. 1, March 2011, pp.1-8. (2011)
7. O'Connor, B., Balasubramanyan, R., Routledge, B., Smith, N.: From tweets to polls: Linking text sentiment to public opinion time series. In: ICWSM 11, No. 122-129 (2010): 1-2. (2010)
8. He, W., Zha, S., Li, L.: Social media competitive analysis and text mining: A case study in the pizza industry. In: International Journal of Information Management. Vol. 22, No. 3, June 2013, pp.464-472 (2013)
9. Zhang, H., Yu, Z., Xu, M., Shi, Y.: Feature-level sentiment analysis for Chinese product reviews. In: 2011 3rd International Conference on Computer Research and Development (ICCRD) Vol. 2. pp.135-140 (2011)
10. Chang, P.C., Galley, M., Manning, C.: Optimizing Chinese Word Segmentation for Machine Translation Performance. In WMT. (2008)
11. Shannon, C.: "A Mathematical Theory of Communication". In: Bell System Technical Journal. Vol. 27, No. 3, pp. 379 - 423 (1948)
12. Cortes, C., Vapnik, V.: Support-vector networks. In: Machine Learning. Vol. 20, No. 3, pp. 273-297 (1995)
13. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*. Vol. 14. No. 2 (1995)
14. Powers, D.: Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, Vol. 2, No. 1, pp. 37-63 (2011)