Intelligent Text Scoring System Based on Deep Learning

Qi Lin *

International School, Beijing University of Posts and Telecommunications, Beijing
100876, China

*Corresponding author e-mail: linwel@bupt.edu.cn

Abstract. In view of the lack of unified standards for emerging local service industries and the impact on users' reasonable choices, this thesis takes the field of B&B as an example to design an analysis and scoring system based on massive online data. The system first obtains massive online data, including text-type data, through web crawlers. Then analyze and pre-process the captured information to build a corpus; use Chinese word segmentation and label clustering to mine the keywords in the listing introduction and comments. The sentiment polarity of the tenant reviews was analyzed based on the support vector machine (SVM) and sentiment lexicon. Finally, an optimized scoring system was obtained.

1. Introduction

In recent years, with the rapid development of the economy and the tension of social resources, a new economic model -- the sharing economy has effectively promoted the flow of social wealth, and improved the efficiency of the use of social resources. Some new services for daily life such as online car-hailing and homestay services came into being. Among them, as a more convenient accommodation service model, the homestay is not only to change the short-term lease of the house from offline to online, but also to share the unused space resources to travelers who need them worse, therefore, it is widely welcomed. However, compared with traditional hotels, the new daily-life service industry represented by the homestay also lacks uniform standards and cannot be strictly controlled, which impacts on tenants' accommodation experience and may even cause problems and security risks that threaten the lives and property of the tenants.

Some current scoring systems lack certain objectivity. This paper takes the field of B&B as an example. Through the operation of crawler construction, information processing and scoring system, a local service analysis and scoring system based on massive online data is established to help users to be more targeted under massive data. Choose the services you need. The system first uses the network crawler to obtain the data available for the listing on the network, including text data, such as the description of the room, comments, and the image data of the landlord uploaded by the landlord, and establishes a large amount of online data stored in the database. For textual data, we used a house description and a guest review to manually build a corpus of homestays. Through experiments, this paper explores and finally selects the emotional polarity analysis method based on SVM and sentiment dictionary, calculates the emotional value of the tenant's comment, and obtains the comment score of the room.

The main innovations and contributions of this paper are:

1. Build a home dataset in the field of homestay, and transfer the pre-trained model of the NIMA network to make the model more suitable for specific fields, so that image aesthetic evaluation can be used not only for aesthetic purposes, but also more practical applications;
2. Integrate natural language processing, computer vision, etc. together, and provide a more objective basis for scoring, so that the product score displayed online depends not only on the user's rating. This avoids the click farming behavior of the merchant which reduces the authenticity of the score;

2. Related work
In terms of text data mining and the recommendation system, Liu et al. [1] designed a film recommendation system based on movie reviews in order to solve the problem that the film rating of the existing recommendation system is inaccurate, affecting the recommendation accuracy. This system uses the web crawler technology first to obtain film review data, then analyzes and pre-processes it, mines the keywords in film review and classifies the positive and negative words, and then calculates the favorable rate of the film review. This film recommendation system based on favorable rate can make the recommendation result more objective and true. Experiments show that the recommendation by using favorable rate improves the accuracy of movie recommendation, and can effectively guarantee the quality of the movie recommendation system. However, this system only judges the emotional polarity through the keywords extracted from the audience's film review, and does not make full use of other valuable text information such as the movie introduction. Referring to this system and making our system more adaptable to specific areas, this paper adopts a more appropriate method of label extraction and label clustering, and uses different emotional polarity analysis methods. It can be better used for recommendations.

For label clustering, Luo et al. [2] proposed a collaborative filtering recommendation algorithm based on label clustering. By constructing a user-tag correlation matrix, the user's hobbies were obtained. Then the K-means clustering algorithm was improved to obtain the user clusters with the same hobbies. Finally, they found the nearest neighbor set in the user cluster that matches the target user, and generated a recommendation. Experimental results show, compared with the user-based collaborative filtering algorithm, the improved collaborative filtering recommendation algorithm which has better quality of recommendation results. However, experiments have shown that for the K-means clustering algorithm more suitable for English, it takes too long to use Chinese vocabulary training and testing, and the dictionary universality is slightly worse. Therefore, this paper selects the HIT IR-Lab Tongyici Cilin, which is more suitable for Chinese, to carry out label clustering. The experiment proves that its effect and efficiency are in line with expectations.

3. Method

3.1. Web Crawler and Database
The system is built on the analysis of massive online data, so the acquisition and storage of massive data is the basis of all work. To this end, we select an industry-leading homestay website as a data source, and analyze the company's website structure. We use the regular expression to get information about each stay, including the title, the description, the latitude and longitude, the room type, amenities, ratings from guest, host scores, reviews, and URLs of home images with captions, and store information in the database established in advance. Then, we access the URL of the room image stored in the database, and images can be crawled and stored locally. In this way, we finally crawled and stored more than 400,000 guest reviews and about 150,000 real shot images of about 5,000 stays. In the follow-up study, our database is constantly getting new data and adding new information.

3.2. Information Processing
For the large amount of listing information obtained by the web crawler, how to process these texts into a basis for scoring and recommendation is one of the most important issues of our system. So, this requires us to carry out detailed information processing.

The text data we need to process is mainly the host's introduction to the stay and the guest's review of the stay. By studying characteristics of Chinese, we find that many characters without real meaning interfere with text processing, so in the data pre-processing stage, we use the stop word dictionary to
remove the emoji, some modal and degree adverbs in the introduction and reviews, and build the corpus used to assist us in the selection of word segmentation tools. Word segmentation is the process of recombining consecutive word sequences into word sequences according to certain specifications. Chinese is just a simple delimitation of characters, sentences, and paragraphs through explicit delimiters. There is no formal delimiter for words. The computer can't read the whole sentence directly like humans, so we choose the Chinese word segmentation tool jieba [3], which is the most suitable for our task, to divide the whole sentence into an independent word accurately and efficiently.

We need to extract valid information from the reviews and generate labels indicating the guest’s rating on the stay. We explored two ways: term frequency–inverse document frequency (TF-IDF) and dependency parsing. TF-IDF is a method of information extraction that evaluates how important a word is to a file set or to one file in a corpus. The importance of a word increases in proportion to the number of times it appears in the file, but at the same time it will decrease inversely with the frequency it appears in the corpus [4]. For example, "traffic" is mentioned many times in a review, but it is rarely mentioned in other reviews. So, TF-IDF algorithm considers "traffic" to be important for this review. However, the label we need is a phrase, which is subject-verb relation generally composed of "noun + adjective" instead of a separate keyword. Therefore, dependency parsing analysis [5], that analyze the dependencies between the feature words and the evaluation words in the statement to obtain the combination of label, is more suitable for the label acquisition process of the system. In order to improve the effect of dependency parsing analysis, we combine the syntax analysis with domain knowledge. In the previous participle about jieba, we use the domain dictionary covering common vocabulary and fix collocation in the field of the homestay such as "staying experience" and "facilities" to avoid mistakes in word segmentation results, which cause the sentence template to extract the wrong content. After relying on the dependency parsing analysis to obtain the dependency relationship, we screen out the combination tagged with "SBV" (subject-verb) whose first half is in the domain dictionary, and the second half is adjective, such as "SBV (host, enthusiasm)", "SBV (room, neat)", etc..

However, there are some semantically similar words in these combinations. In order to avoid the redundancy of label results and the confusion of importance, we also need to cluster the labels. We have tried several commonly used label clustering methods. Word2vec converts words into vectors and calculates the distance in vector space [11]. For a given sample set, k-means clustering algorithm divides the samples into k clusters according to the distance between samples. It lets the points in the cluster be as close as possible, and makes the distance between the clusters as large as possible [12]. They have a good effect on English, but there are a lot of Chinese words, and it's difficult for K-means to calculate the similarity of word vector. After testing, it takes a long time to train the model with Chinese vocabulary. The effect is also not good. In addition, we also tried Latent Dirichlet Allocation (LDA), but the method is biased towards the generalization of the article, which can extract the main content of an article [13]. Some of the reviews are relatively short, so the test results are not good. Therefore, we use HIT IR-Lab Tongyici Cilin [5], which has a good understanding of the Chinese, to perform clustering of labels, and store many sets of synonyms in the thesaurus. Finally, we add the combination of "nouns + adjectives" that are not similar to those already in the list into the tag library after the label clustering judgment. Through this technology, we get the five most important tags for each stay.

In order to more deeply explore the likes and dislikes of the tenants' experience, the system also conducted an emotional analysis of reviews from tenants. We divide emotions into positive and negative polarities to ensure the efficiency and accuracy of the analysis. We test the effects of five sentiment analysis methods. Finally, emotional polarity analysis methods based on SVM [6] and sentiment dictionary are selected to calculate the emotional value and obtain the average score. A score close to 1 represents positive emotion, while a score close to 0 represents negative emotion. The result of the sentiment analysis is part of the final scoring system.

3.3. Scoring System
For the traditional hotel industry, the regulatory authorities conduct a standardized star rating of the hotel from various aspects, which can effectively reflect the service level of the hotel. With reference to
the hotel's star rating standards [7] and the relevant regulations of the property industry [8], combined with the data obtained from previous web crawlers and information processing, we have established a complete homestay rating scale model (as shown in Table 1). Each homestay has a score of 120 points and consists of hardware facilities, space environment, geographical location, service guarantee, truthful description and user experience with 20 points each. For convenience, we map all the scores obtained from the previous work to between 1-20 to unify the dimensions. The facilities are the basis of the quality of the homestay. We look at the amenity points and the type of stay (entire apartment / private room / shared room), each with 10 points. The survey shows that the clean and beautiful environment is the premise for most users to choose a home. At the same price, the closer the location to the city center, transportation facilities, attractions, etc., the more popular with guests. The score about the average distance to the important locations such as the city center in the map calculated by making use of the latitude and longitude of the place and the district to which it belongs account for 50% of the location score, and the other half depends on the "location" rating from the guest. Service is an important part of services for daily life such as homestay. The host is the provider of the service, so the host score account 30% for the service score. The guest is the person who enjoy the service, so the "communication" score and "check-in" score rated by the guest account for the remaining 70%. The accurate description is the key to the authenticity of the listing information. The similarity between the "Show and tell" results and the description written by the host can objectively reflect this point, accounting for 60% of the description score, and the other 40% depends on the "accuracy" rating. The user experience best reflects the guest's satisfaction with the stay. The emotional score of the guest's reviews accounts for 60% of the experience score and the other 40% depends on the "value" rating. Multiply the scores of all the above details by the corresponding weights and sum these products, and we will get the comprehensive score of the stay.

Table 1: Homestay rating model

| Rating items          | Scoring rubric                                      | Maximum score |
|-----------------------|-----------------------------------------------------|---------------|
| Facilities            |                                                     |               |
| amenity points        | 1.5 points per amenity (no more than 10 points)     | 10            |
| type of stay          | entire apartment: 10 points                         | 10            |
|                       | private room: 6 points                              |               |
|                       | shared room: 3 points                               |               |
| Environment           |                                                     |               |
| aesthetic score of the images | 1-20 points * 60%                                | 12            |
| "cleanliness" score  | 1-20 points * 40%                                  | 8             |
| Location              |                                                     |               |
| average distance to the important locations | 1-20 points * 50%                                | 10            |
| "location" score     | 1-20 points * 50%                                  | 10            |
| Service               |                                                     |               |
| "communication" score| 1-20 points * 35%                                  | 7             |
| "check-in" score      | 1-20 points * 35%                                  | 7             |
| host score            | 1-20 points * 30%                                  | 6             |
| Description           |                                                     |               |
| similarity between the "Show and tell" results and the description | 1-20 points * 60%                                | 12            |
| "accuracy" score      | 1-20 points * 40%                                  | 8             |
| Experience            |                                                     |               |
| emotional score of reviews | 1-20 points * 60%                                | 12            |
| "value" score         | 1-20 points * 40%                                  | 8             |
4. Experiments

1) Choice of word segmentation method

Due to the particularity of Chinese, we need Chinese word segmentation to recombine successive character sequences into word sequences according to certain specifications. Through investigation and research, we find the leading four Chinese word segmentation methods in the industry: jieba [3], SnowNLP [9], pyltp [10], THULAC (THU Lexical Analyzer for Chinese) [11]. By calculating the accuracy, the recall rate, the F-Measure and the running time of segmenting the same corpus, the Chinese word segmentation results of the four methods are tested, so that the optimal segmentation method is selected for our system.

First, we use python to connect to the database, randomly get the ID list of 50 stays, and get the corresponding 8000 reviews. After removing the characters that affect the Chinese word segmentation results such as English text and emojis, we save them as the test corpus of this experiment. By manually segmenting the corpus and inspecting it over again, we get the ground-truth. Then, we use these word segmentation tools to segment the test corpus, and record the running time. After obtaining the results, we count the number of words in the word segmentation results of using each word segmentation tool, the number of words in the ground-truth and the number of words overlapping the first two, that is, the number of words segmented correctly in the results of using each word segmentation tool. Use the number of words in the segmentation results to divide the number of words segmented correctly in the results of using each word segmentation tool, and get the accuracy rate. Use the number of words in the ground-truth to divide the number of words segmented correctly in the results of using each word segmentation tool, and get the recall rate. Then we can calculate the F-Measure. Finally, the evaluation data of these word segmentation methods are shown in Table 2.

Table 2: The evaluation data of the four word segmentation methods

| Method   | Accuracy | Recall rate | F-Measure | Running time |
|----------|----------|-------------|-----------|--------------|
| jieba    | 0.96     | 0.6         | 0.74      | 1.47         |
| SnowNLP  | 0.74     | 0.5         | 0.6       | 4.86         |
| Pelt     | 0.81     | 0.51        | 0.63      | 0.7          |
| THULAC   | 0.76     | 0.49        | 0.6       | 2.4          |

Experiments show that the accuracy, the recall rate and F-Measure of jieba are higher than other three word segmentation methods, and the running speed also has better performance. Therefore, we use jieba as the word segmentation method in this system.

2) Results of sentiment analysis

The sentiment analysis is a classification task of judging that the emotion of the input sentence is positive or negative by itself. In order to get the emotional attitude in the reviews, we test the sentiment-dictionary-based, K_NN-based, Bayes-based, maximum-entropy-based, SVM-based [12] emotional polarity analysis method by analyzing the same corpus. We calculate the accuracy, the recall rate, the F-Measure and the running time to compare the advantages and disadvantages of these five methods, and then choose the sentiment analysis method applicable to our system.

First, we selected 5,000 reviews with strong tendency to likes and dislikes as our test corpus, and according to the attitude of the reviews, we manually mark: 1 represents positive polarity; 0 represents negative polarity, as the ground-truth. We use these five methods to calculate the sentiment values for each review in the test corpus and record the program running time. Then we compare the test results and the ground-truth, and count the number of positive reviews correctly analyzed in the analysis results of each method, the total number of positive reviews in the analysis results, and the number of positive reviews in the ground-truth. The accuracy is obtained by dividing the number of positive reviews correctly analyzed in the analysis results by the total number of positive reviews in the results of each method. The recall rate is obtained by dividing the number of positive reviews correctly analyzed in the
analysis results by the number of positive reviews in the ground-truth. The final experimental result is shown in Table 3.

Table 3: The evaluation data of the five sentiment analysis methods

| Method                | Accuracy | Recall rate | F-Measure | Running time |
|-----------------------|----------|-------------|-----------|--------------|
| Sentiment-dictionary-based | 0.89     | 0.95        | 0.92      | 5.45         |
| K_NN-based            | 0.94     | 0.69        | 0.79      | 685.59       |
| Bayes-based           | 0.91     | 0.92        | 0.92      | 24.87        |
| Maximum-entropy-based | 0.92     | 0.95        | 0.93      | 405.43       |
| SVM-based             | 0.97     | 0.88        | 0.92      | 109.62       |

Experiments show that the accuracy and F-Measure of SVM-based methods are the highest among the five methods. The comprehensive evaluation index – F-Measure based on K_NN method is worst. The method based on maximum entropy takes too long to train the model; The accuracy of Bayes-based method is not high, and it does not meet the practical requirements. The method based on sentiment dictionary does not involve machine learning, so its running speed is the fastest. With the increase of manual workload, the accuracy will be further improved. Therefore, the SVM-based method with outstanding performance, and the simple and efficient sentiment-dictionary-based method, are most suitable for the text sentiment analysis in the B&B field of this system.

5. Conclusion
This paper used an approach to natural language processing and computer vision in the field of deep learning to create an intelligent text scoring system. From the establishment of the corresponding database of reptiles to the establishment of information processing and scoring system, this paper discussed the process in detail. This paper used the folklore scoring system as an example to verify the experiment. Compared with the listing score displayed by the website itself, the evaluation standard of our scoring system was more scientific, objective and comprehensive, and the result was more differentiated than the original score. More reference value. In the future, the application of this system can be extended to more fields, and more extensive and in-depth exploration and analysis of massive online data, providing comprehensive technical support for the intelligent development of the sharing economy.

References
[1] Liu Hui, Li Fengyin, Qi Jiguo, Cui Wei, & Ge Rui. (2018). Design and Implementation of Film Recommendation System Based on Film Criticism Mining. Electronic Technology, 47(12), 89-92.
[2] Luo Zhengqing, & Zheng Tao. (0). Collaborative filtering recommendation algorithm based on label clustering. The 13th (2018) China Management Annual Conference.
[3] Sun Junyi. jieba. Retrieved August 20, 2019, from github.com/fxsjy/jieba
[4] Yu Yu, & Wang Hongyan. (2018). Text information extraction based on tf-idf algorithm. Science and Technology Vision, No.238(16), 122-123.
[5] Nie Hui, & Du Jiazhong. (2014). Research on Product Feature Label Extraction under Dependency Syntax Template. Data Analysis and Knowledge Discovery, 30(12), 44-50.
[6] Qi, X., Wang, T., & Liu, J. (2017, December). Comparison of Support Vector Machine and Softmax Classifiers in Computer Vision. In 2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCCE) (pp. 151-155). IEEE.
[7] Cui Yu. (2014). Research on Hotel Star Rating System in China——Based on Comparative Analysis of Hotel Rating System in the United States and the United Kingdom. China Market (46), 146-150.
[8] Zhong Ying. (2018). How to keep up with the management of the “B&B Hot” in the Magic City. Procurator Fengyun, 569(21), 46-47.

[9] isnowfy. snowlp 0.12.3. Retrieved August 20, 2019, from pypi.org/project/snowlp/

[10] endyul, liu946. pyltp. Retrieved August 20, 2019, from pyltp.rtfd.io

[11] Maosong Sun, Xinxiang Chen, Kaixu Zhang, Zhipeng Guo, Junhua Ma, Zhiyuan Liu. THULAC: An Efficient Chinese Lexical Analysis Toolkit. Retrieved August 20, 2019, from thulac.thunlp.org

[12] Chang Dan, & Wang Yuzhen. (2019). Research on user analysis sentiment analysis based on svm. Journal of Zaozhuang University, 36(02), 78-83.

[13] Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.