Rating Prediction Based on TripAdvisor Reviews

Dangyi Liu
A53221859
dangyi@ucsd.edu
UC San Diego

Yu Chai
A53213872
yuc385@ucsd.edu
UC San Diego

Chenxi Zheng
A53221997
chz190@ucsd.edu
UC San Diego

Yilun Zhang
A53099979
yiz209@eng.ucsd.edu
UC San Diego

***Abstract*—Rating is becoming increasingly important for customers, which helps them obtain useful information of products more easily and more efficiently. Normally, a customer rates products from different aspects and then makes an overall rating. According to this human thinking process, we are curious about how ratings in different aspects would influence the final overall ratings. Therefore, we make two different prediction models: 1. Use review context to predict overall rates predictions. 2. Use review context to predict rates in different aspects(cleanliness, location, rooms, service, sleep quality and value) and then use these ratings to predict the overall ratings. By comparing MSE of two models, we found out that the first model is more accurate than the second one.**

***Keywords*—linear model, rating predictions**

I. Introduction

In nowadays, more and more people start to search on internet for useful information. In order to help customers finding information efficiently and attracting more customers, many websites starts to ask customers with their reviews based on their experience. In this way, other customers could get useful information easily and choose their products based on those informations. For example, Tripadvisor asks customers' reviews based on their experience on different hotels. Other customers could easily make decisions about hotels, instead of wasting amount of time to do the research. In order to get more review context from customers, agents like Tripadvisor will always give customers some bonus to those who give feedbacks for their products. Also, some customers would like to share their experience with others. Thus, more and more people comment their perspectives on the websites including many aspects like utilities and services. It will not only help other customers to make decisions, but help the hotels to know what aspects that customers care about. However, people rate the hotels based on their objective thoughts. In other words, different people emphasize on various aspects. For example, a customer focuses on cheap price. The customers may still wont be satisfied if the hotel has excellent utilities and service with expensive price. So the hotel may still get low overall review rates from this customer.

Therefore, to give an overall rating, a logical customer will first make rates on different aspects like cleanliness, location, rooms, service, sleep quality or values and so on, based on their experience, then combine them into an overall rating. Therefore, we are interested in how rates in different aspects could influence the overall rates. In other words, we want to make more accurate predictions for overall rates based on rates in different aspects.

We use two models. we first use review context to predict the overall rating. Then we use reivew context to predict the rating in several aspects: cleanliness, location, rooms, service, sleep quality, and value. Then using these rates, we predict the overall ratings. To compare those two models, we identify if the model could predict the overall rating more accurately by using rating in several aspects. A challenge for this task is that, the data set is really large and we didnt get complete data set. So we first need to preprocess the data set then do the predicton.

In this paper, we introduce some related works and we describe the dataset that being used in detail. Then we go into some details about the model. Finally provides our conclusion.

II. Related Work

With the advancement of internet, the Web has become one of the major sources for collecting opinions and suggestions from consumer s. There are now numerous Web sites containing such information and many scientists have done some relative research on it.

In [5], Satoshi pointed out the importance of knowing the reputation of manufacture own product and their competitor's product. Since it was extremely hard to collect and analyze survey data manually, he developed a new framework which could mining product reputations on the Internet. This framework could automatically collect peoples opinions about target products from Web pages, and it uses text mining techniques to obtain the reputations of those products. In this paper, they generated syntactic and linguistic rules to determine whether any given statement is valid, and if so, what is its attitude (positive or negative).

However, only the known attitude in review analysis was not enough for manufactures, potential customers wanted to know more detail about the product. Therefore, later in [6], focused on online customer reviews of products, Liu proposed a framework for analyzing and comparing consumer opinions of competing products, and provided a visual side-by-side and feature-by-feature comparison for potential customers and product manufactures. This paper came up with a new technique based on language pattern mining, which proposed to extract product features from Pros and Cons in a particular type.

When scientists figured out the method to extract information from reviews, more and more research focused on summarizing users opinions and categorizing reviews. In paper [7], authors suggested that classifying documents not by topic but by overall sentiment. They compared three ML method (Naive Bayes, maximum entropy classification, SVM) with the sentiment classification on traditional topic-based categorization. In our paper, we also consider this factor and will figure out the better one while comparing the overall rating directly from sentence with overall rating derived from aspect ratings and aspect rating derived from sentence.

But, how could we analyze the comparative sentences? We are not the first one to think about it. In [8], the paper prosed the study of identifying comparative sentence. They first analyzed different types of comparative sentences from both the linguistic point of view and the practical usage point of view. Then conclude a rule mining and machine learning approach to identifying such comparative sentences. Those sentences are very useful in many application like bussiness intelligence and so on.

The closet work to our study is [9], which introduced the problem of Latent Aspect Rating Analysis(LARA) and proposed a two-stage approach:in the first stage, keywords specified by users are used in a bootstrapping algorithm to identify the aspects and segment the review content; in the second stage, a generative Latent Rating regression(LRR) model is applied to infer aspect rating and weights in a review.

III. Dataset Description

A. Data Source

We use dataset from DAIS(The Database and Information Systems Laboratory at The University of Illinois at Urbana-Champaign)[2]. This dataset consists of over 1,500,000 reviews of hotels from Tripadvisor and we randomly pick up nearly tenth of it as our dataset.

B. Data Cleaning

Since the dataset is collected from the website and some parts of the review data are missing, we have to first preprocess the datase t and get rid of the incomplete reviews. Normally, each review in the dataset includes eight fields:

1) Author
2) AuthorLocation
3) Content
4) Date
5) HotelID
6) Ratings
7) ReviewID
8) Title

For field $Ratings$, there are seven subfields:

1) Cleanliness
2) Location
3) Overall
4) Rooms
5) Service
6) Sleep Quality
7) Value

After preprocess, we get 64004 pieces of review data, covering 1310 hotels, spanning over 32 months, ranging from Feb. 2010 to Sept. 2012. Then we refine the review data by removing stopword in Content and making content words insensitive to capitalization and shuffle the whole dataset in the end. Then we can conduct some statistics analysis and explore more information about the dataset in order to fulfill our prediction task.

C. Statistiscs of Ratings over Aspects

First, we draw the box plot as Figure.1 over subfields in field Ratings. From the box plot, we can see the medians of each subfield are [4.0, 5.0, 5.0, 4.0, 4.0, 4.0]. The lower quartile(25% percent) of Overall Rating is equal to its median, namely 4.0 and high quartile(75% percent) of Cleanliness and Location are both equal to their medians, namely 5.0. And there are some flier data, or outlier data lie in 1.0 and 2.0 in these three subfields. And for the rest four subfields, their rating distribution are quite equivalent, all with higher quartile of 5.0 and lower quartile of 3.0. And they all get lower whisker at 1.0. From this figure, we have an approximate understand about the distribution of ratings in different aspects.

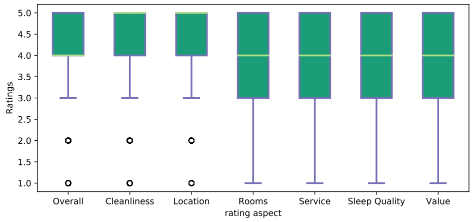

Fig. 1: Box Plot of Ratings over Aspects

D. Average Overall Ratings over Time

Then we explore the relationship between average overall ratings and time. We first calculate the number of reviews over time to see the distribution of review numbers. It is shown in Figure 2 and we can see that there exists several period of time with small amount of reviews which are under 500. Normally, the number of reviews ranges in [1000, 3000], but there also exists period of time with large amount of review over 3000. We then plot the figure to show to relationship between average overall rating and time, and it is shown as Figure 3(there is no specific meaning behind different color). The size of area in Figure 3 represents the amount of review numbers, where bigger circle corresponds to larger amount of reviews. There are some outliers in the upper part of figure which comes from a small amount of reviews. Otherwise, we can see there is no apparent trend over time, the average overall rating normally lies in [3.9, 4.2].

E. Length of Review and Overall Rating

Intuitively, the length of review has some connection with overall rating. We also explore the relationship between average overall ratings and length of content in reviews. We first count the number of reviews for different length of content and show the result in Figure 4. We can see normally the length of reviews lies in [6, 200]. There are only few number of reviews with content length over 500. The relationship between length of review and corresponding overall review can be shown in Figure 5. We can see for those with review length under 200, their average overall rating almost intensively lies in [3.5, 4.5]. However, for those with larger review length, the average overall ratings varies from [1.0, 5.0] dispersively.

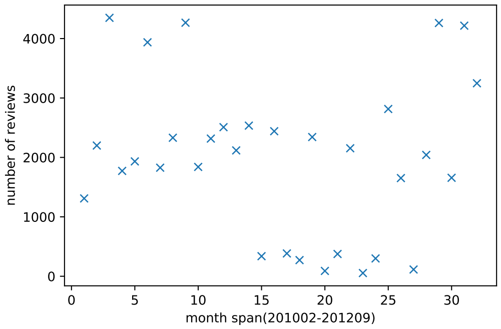

Fig. 2: Box Plot of Ratings over Aspects

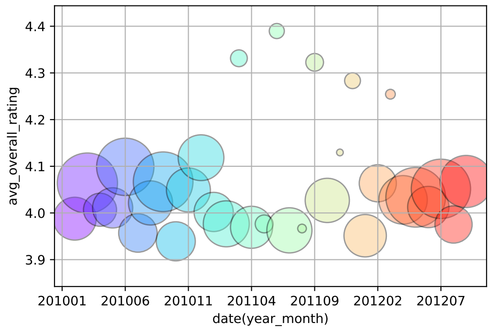

Fig. 3: Changes of Overall Rating over Time

IV. Prediction Task

Our task is to predict overall rating using two different models

- (Direct method) predict overall rating directly based on review text.
- (Indirect method) predict ratings for individual aspects first, then use individual ratings to predict overall rating.

We want to know which model produces a better result. The metric we use to evaluate models is mean

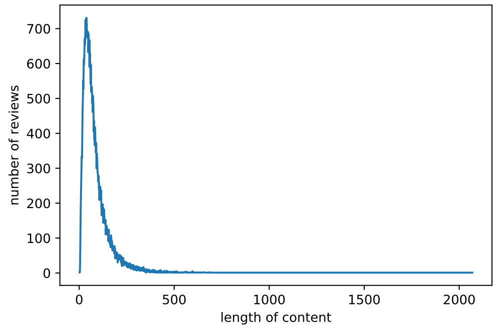

Fig. 4: Number of Reviews over Content Length

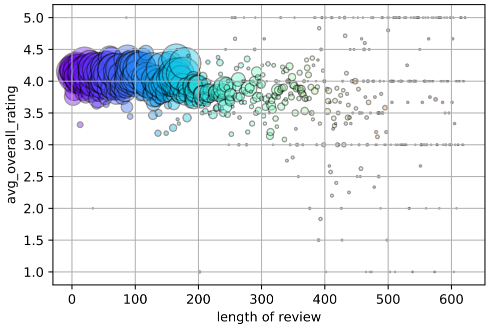

Fig. 5: Length of Review and Corresponding Overall Rating

square error,

$$\text{MSE} = \| \, y - \text{prediction} \, \|_2^2$$

We take 10000 reviews out of all 64004 reviews as the test set, the MSE is calculated on test set.

V. Models Description

A. Preprocess of review text

Both methods require us to convert the review content into word vector. We're taking the following steps.

1) Split text into words:

1) Remove non-roman characters. These includes digits,
2) punctuations, greek letters, etc. To perserve the seperator between words, we replace these characters with spaces.
3) Convert letters into lowercase.
4) Split the text into words. It's possible that we mistakenly split the contractions, e.g. split "we'll" into "we" and "ll".
5) Remove stopwords such as "a", "they", "because", "between", "is". It's also necessary for us to remove the fragments of contractions such as "ll", "ve", "t", "s".

2) Generate Word Set: In our review context, we only care about unigrams because most of informative words are adjective such as dirty and terribles. Also, we take 1000 most common unigrams to make the predictions, in order to avoid rare frequency and meaningless words.

3) Convert Word Bag into Word Vector: Given the context review from customers, we need to formalize how important of each word in the review context. The importance of words increased proportionally to the frequency of the word appears in the document. However, there are some words are not important but appears a lot of times like ”the”, ”is”. So in order to see the importance of the word in the review context, we need to balance the frequency of the word appears in current review context and other reivew context. Therefore, we choose to use tf-idf (term frequency-inverse document frequency) to filter the context in different fields like cleaning and so on. The tf-idf is defined as follows.

$$\text{idf}(t) = \log \frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}$$

$$\text{tf}(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}$$

$$\text{tf-idf}(t) = \text{tf}(t) \times \text{idf}(t)$$

B. Direct Method

In direct method, we predict overall rating directly based on word vector extracted from review context using the following linear model.

$$\text{overall rating} = \alpha + \theta_0 \text{tf-idf}(t_1) + \theta_1 \text{tf-idf}(t_2) + \dots$$

To prevent overfitting, instead of minimizing MSE, we add L2 regularization. We'll also need to seperate a validation set from training set in order to train the hyperparameter λ. We take 10000 reviews out of the training set as the validation set.

The result shows that, this method could achieve an MSE of 0.6598 on test data, when $\lambda = 1000$.

C. Indirect Method

In constract, when using indirect method, we take the following steps to make prediction.

1) Train a model to predict overall rating based on individual ratings
2) For each aspect, train a model to predict the rating based on word vector.

The general idea is similar with direct method though we use the aspect rating as mediator. The results of MSE is 0.6632.

1) From Individual to Overall: We use a linear model to fit between overall rating and individual ratings, with the following form.

$$\begin{aligned}\text{overall rating} = \alpha &+ \theta_0 \text{ cleanliness } + \theta_1 \text{ location}\\ &+ \theta_2 \text{ rooms } + \theta_3 \text{ service}\\ &+ \theta_4 \text{ sleep quality } + \theta_5 \text{ value}\end{aligned}$$

Since we only have 7 parameters in this model, we could use linear regression without regularization.

2) From Word Vector to Individual Rating: For each aspect, we use linear regression with L2 regularization to predict the individual rating, with the following form.

$$\text{individual rating} = \alpha + \theta_0 \text{tf-idf}(t_1) + \theta_1 \text{tf-idf}(t_2) + \dots$$

3) Result: For the first step, we calculate the combination of aspect ratings in order to predict the overall rating. Here is the result of the coefficients for each aspect:

Aspect	Coefficient
Cleanliness	0.13864
Location	0.08302
Rooms	0.25638
Service	0.27541
Sleep Quality	0.12467
Value	0.18121

which achieves an MSE of 0.2118 on test data. We could see that the aspect people care most is service, which has a heavy impact on overall rating, and location has the least impact on overall rating.

For the second step, we take $\lambda = 1000$. The MSE for each aspect are

Aspect	MSE
Cleanliness	0.6566
Location	0.6504
Rooms	0.7569
Service	0.7811
Sleep Quality	0.8298
Value	0.8218

We could draw a wordcloud for each aspect, showing words with highest and lowest coeffients.

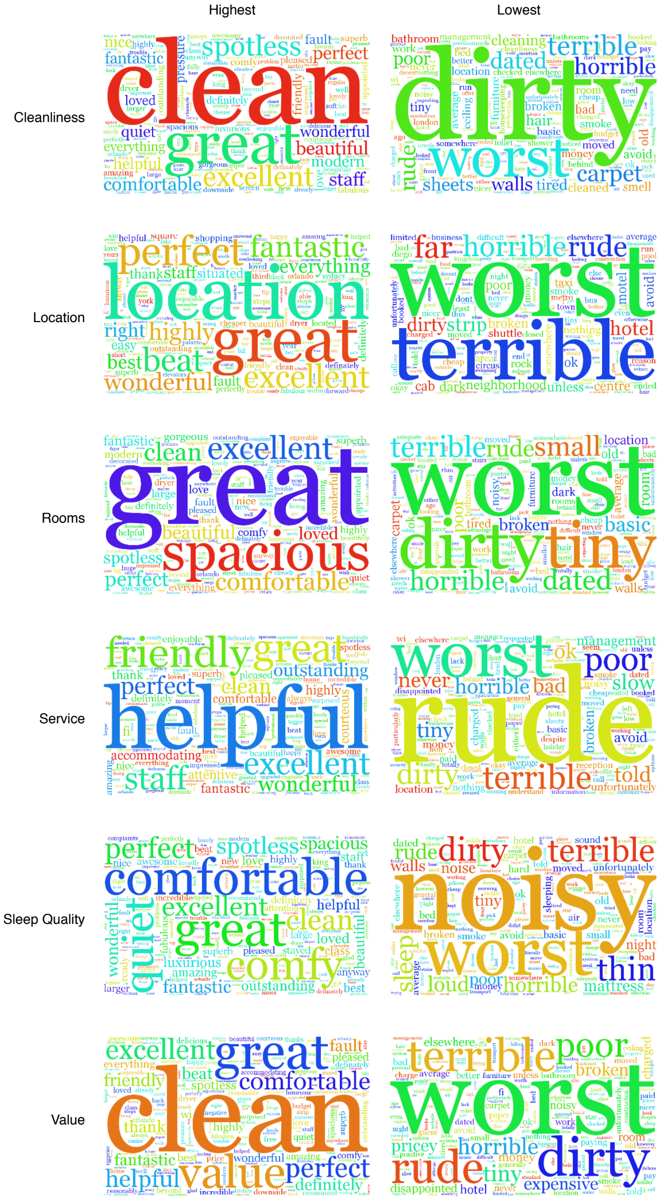

Fig. 6: Wordcloud for Each Aspect

VI. Conclusion

The paper discusses two ways to predict the overall rating from user review context. In the first method, we apply the indirect method. The first step is that we predict the aspect rating from review context; the second step is that we predict the overall rating from the weights of aspect ratings. While in the another method, we calculate the overall rating directly from the review context. Both of them are applied with linear model and derive the model and result according to the minimization of MSE. In our result, we found the MSE of the indirect method is 0.6632 and the MSE of direct model is 0.6598. Based on the result, we could conclude the following things:

- Contrary to our original assumption, direct method performs better than indirect method. As we known, when any person goes through a review text, s/he would label the key word without conscious. For example, when one notices the word "rude" in a review, s/he would label it as service attitude and then reduce their overall rating to this hotel according to the different importance of service attitude when they choose hotels. But, in contrast, the machine learning does not care about it. It does not care about how the word labels but how it would influence overall rating.
- Though the indirect model use the mediate keys, aspect rating, they are still derived the final result by linear model, which is same as the direct model. In addition, since we derive the model according to MSE, the direct model would provide the result with the minimal MSE, while indirect model could only guarantee the local minimal.
- However, the difference between those two methods is very small which means both methods could be accepted in usage.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] Hongning Wang, Chi Wang, ChengXiang Zhai and Jiawei Han. Learning Online Discussion Structures by Conditional Random Fields. The 34th Annual International ACM SIGIR Conference (SIGIR'2011), P435-444, 2011.

[3] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classication of product reviews. In WWW '03, pages 519 - 528, 2003.

[4] M. Hu and B. Liu. Mining and summarizing customer reviews. In W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, editors, KDD, pages 168- 177. ACM, 2004.

[5] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In KDD '02, pages 341-349, 2002.

[6] Liu, Bing, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web." Proceedings of the 14th international conference on World Wide Web. ACM, 2005.

[7] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[8] Jindal, Nitin, and Bing Liu. "Identifying comparative sentences in text documents." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.

[9] Wang, Hongning, Yue Lu, and Chengxiang Zhai. "Latent aspect rating analysis on review text data: a rating regression approach." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.