Predicting Hotel Rating Based on User Reviews

Pradeep Kumar Duvvur
Advisor: Dr. Carol Romanowski
# Table of Contents

Abstract............................................................................................................................................. 3
1 Introduction ........................................................................................................................................ 4
2 Related Work .................................................................................................................................. 5
3 Design & Implementation .............................................................................................................. 6
  3.1 Data Extraction .......................................................................................................................... 6
  3.2 Data Pre-Processing ................................................................................................................... 10
    3.2.1 HTML Markups Removal ..................................................................................................... 10
    3.2.2 Special Characters Removal ................................................................................................ 10
    3.2.3 Tokenization ....................................................................................................................... 11
    3.2.4 Stop Words Removal ........................................................................................................... 12
    3.2.5 Stemming ........................................................................................................................... 13
    3.2.6 Spelling Correction ............................................................................................................. 14
  3.3 Implementation .......................................................................................................................... 15
4 Results ............................................................................................................................................... 17
5 Conclusion ....................................................................................................................................... 22
6 Future Work .................................................................................................................................... 23
7 References ....................................................................................................................................... 23
Abstract

Online user reviews are instrumental in helping users to decide which hotel to book for their stay at their choice of location. Online user reviews help potential customers to understand the positive and negative aspects of a hotel. It also allows users to learn about the amenities provided by the hotel. User reviews are not only useful for potential customers, but also help the management to understand the areas at which the hotel needs to improve. This will allow the management to take appropriate steps and ultimately improve the reputation of the hotel. TripAdvisor is a website that contains a number of user reviews for multiple hotels across the globe. Each user leaves a brief review about the hotel, and rates the hotel based on multiple factors like value, location, sleep quality, cleanliness and service. Based on the user reviews and their ratings, each hotel is given an overall rating on a scale of 1 to 5 stars. A number of travelers blog about their experiences at a hotel, while others use social media like Twitter and Facebook to talk about their experiences at hotels. Such reviews do not have a star rating and are not accounted for. Being able to predict ratings for reviews obtained from social media and other sources can additionally help in computing the overall rating of a hotel.
1 INTRODUCTION

The hotels listed on TripAdvisor have a number of user reviews from customers who stayed at the hotel. Each user also gives a rating for the hotel based on his/her experience at the hotel. One of the problems with such ratings is that, two different customers may leave a different rating even if they have a similar opinion regarding the hotel. For example, two customers may leave a review that reads “Great stay!”. However, one customer may leave a 4 star rating, while the other customer may leave a 5 star rating. Thus the ratings are not standardized based on the review text. This difference affects the overall rating of the hotel.

Since each hotel has thousands of reviews, it is difficult for new customers to go through each one of them to understand the positives and negatives of the hotel. Providing review highlights based on all the reviews available will help customers to easily weigh out the pros and cons of the hotel.

Hotel ratings keep changing over time based on multiple factors like the service, the cleanliness of the hotel, the food at the restaurants, the rush during a specific period, etc. It is important for the management to be able to analyze how the ratings have changed over time. If a dip in ratings is seen during a specific time of the year, the management can analyze reviews from that time period and find out what caused the reviews to go down. Based on the reviews, the management can take corrective measures. Figure 1 shows a sample user review on TripAdvisor.
Following are the main goals of this project:

- To predict the rating of each individual review.
- To compute the overall rating of a hotel based on predicted ratings from individual reviews.
- To provide review highlights based on what’s most talked about regarding the hotel.
- To visualize ratings trends over time.

The reviews for hotel “Circus Circus Hotel” in Las Vegas will be used to predict the hotel rating and visualize rating trends.

2 Related Work

Chen et al. [1] propose an approach to predict the star ratings of Yelp reviews provided for restaurants. The other goal was to suggest restaurants to customers based on their reviews. The initial step was to perform a binary classification. Hence, the model predicted if the review sentiment was either positive or negative. This was then extended to predict a star rating on a scale of 1-5. The authors evaluated Naïve Bayes and Support Vector Machine classification algorithms and found that the star rating prediction was best with SVM.

Yaonan [2] also proposes an approach to predict star ratings for a given user review. The author chose Yelp dataset for training and testing. The author restricted the data to just Chinese restaurants. The author generates the vocabulary by picking unique words from the user reviews and removing stop words, other words with parts of speech. The author predicts the rating based on the adjectives present in the user review. Yaonan [2] experimented with Naïve Bayes and SVM algorithm and found that SVM provides a better accuracy using Radial Basis Function (RBF) Kernel.
3 DESIGN & IMPLEMENTATION

The architecture of the system shown in Figure 2 describes the overall process. The major steps involved in predicting a hotel’s rating is as follows:

- Extracting user review data from TripAdvisor website.
- Cleaning and pre-processing the extracted data.
- Training the model using SVM.
- Testing the model.
- Predicting the rating for each review in the test set.
- Compute overall hotel rating from the individual ratings predicted.

![Architecture of the system](image)

Figure 2. Architecture of the system

3.1 DATA EXTRACTION

Since there was no hotel reviews dataset readily available to perform mining, the data needed to be extracted from TripAdvisor website by web crawling. Web crawling is the process of implementing a program that is capable of browsing the web in a step by step manner by changing a part of the URL. A Java library called “Jsoup” provides APIs that can be used to extract data from websites. Jsoup was used to extract user reviews of the hotel “Circus Circus Hotel” from TripAdvisor.
Each review on TripAdvisor has the following fields:

- **Review Title** – A brief review of the customer’s experience.
- **Full Review** – Complete user review of the hotel that talks about different aspects of the hotel.
- **Review Date** – The date on which the hotel was reviewed.
- **Star Rating** – The user’s rating on a scale of 1 – 5 stars.
- **Review ID** – The unique id of the review.

Following are the steps to extract user reviews from TripAdvisor using Jsoup:

- A connection is made to the TripAdvisor webpage that holds the first ten user reviews.
- The total number of pages of user reviews is identified by parsing it to a DOM.
- Each page has a maximum of 10 reviews.
- Each review shows only a few lines of text and there is a “See more” button that needs to be clicked to view the rest of the text.
- Jsoup does not have the option to simulate a button click. Hence, the URL of the Full review page must be obtained by parsing the webpage and obtaining the contents of the Anchor `<a>` tag.
- Once Jsoup navigates to the webpage containing the full review, the webpage is parsed and the fields “Review ID”, “Review Title”, “Full Review”, “Review Date”, and “Star Rating” are extracted from the corresponding tags using Jsoup.
- All this information is stored in a “Review” object in Java and inserted in a list.
- This is repeated for all reviews on the page.
- Once all reviews are collected from one page, the program is made to navigate to the next page and all reviews from that page are collected and are added to the list.
- As it can be seen in Figure 3 and Figure 4, the only part of the URL that changes for each page, is highlighted in Red. This part of the URL is changed by the program to iterate through all pages.

```javascript
https://www.tripadvisor.com/Hotel_Review-g45963-d91770-Reviews-or10-Circus_Circus_Hotel_Casino_Las_Vegas-Las_Vegas_Nevada.html#REVIEWS
```

*Figure 3 – URL for page 1*

```javascript
https://www.tripadvisor.com/Hotel_Review-g45963-d91770-Reviews-or20-Circus_Circus_Hotel_Casino_Las_Vegas-Las_Vegas_Nevada.html#REVIEWS
```

*Figure 4 – URL for page 2*
• This process continues for all pages that contain reviews.
• All the reviews in the list are then written to a flat file.

Figure 5 shows the output obtained by extracting data from TripAdvisor.
Figure 6 contains a flowchart that depicts the entire flow of data extraction from TripAdvisor using Jsoup library.

Figure 6 – Flowchart for data extraction from TripAdvisor
3.2 DATA PRE-PROCESSING

Data pre-processing phase is one of the most important phases of text mining. The data that is extracted from any source is usually not clean and contains a lot of special characters, emoticons, and other text that is not necessary for mining. It is important that such unnecessary data is removed from the dataset before any further processing.

The CSV file that is obtained by extracting user reviews from TripAdvisor, is read by a Python program. It is important to specify the encoding format while reading the CSV. The encoding format used is “latin-1”.

The CSV file which is used as input, contains the following fields:

- Review ID
- Review Date
- Star Rating
- Review Title
- Full Review

3.2.1 HTML Markups Removal

Since the data is extracted from a webpage, some of the reviews contained HTML Markups that needed to be removed. Some of the HTML markups are:

- “&lt” for “<” (Lesser than symbol)
- “&gt” for “>” (Greater than symbol)
- “&amp” for “&” (Ampersand symbol) etc.

Since these markups do not have any significance in the mining process, they were removed. This was done using a Python library. The output generated by removing markups is then passed on to the next stage for further pre-processing.

3.2.2 Special Characters Removal

The reviews also contain a lot of punctuation, special characters and emoticons like “.”, “!” “:), “:(” etc. Special characters like these need to be removed before mining is performed. Hence, these were removed using Regular Expressions (Regex) in Python.
Figure 7 shows a sample review that contains special characters, punctuation and emoticons that need to be removed.

Figure 7 – Sample review containing special characters.

Figure 8 shows the output obtained after the special characters, emoticons and punctuations are removed. This output is then used as input for further pre-processing.

Figure 8 – Review with all special characters removed.

Once the special characters are removed, the text is converted to Lower Case since python does not remove stop words that are capitalized. Stop words removal will be discussed in the upcoming sections.

3.2.3 Tokenization

Tokenization is one of the steps performed in pre-processing which involves cutting down sentences into separate words. These words are called tokens. In this step, the text is split into tokens based on white spaces between words. The process of breaking text into tokens is called Tokenization. The tokenization step is performed in Python using “NLTK” library.
3.2.4 Stop Words Removal

Words that are frequently used and do not have any value in text mining need to be removed. Such words are called “Stop Words”. Some examples of stop words are – “an”, “and”, “is”, “the”, etc. If such words are removed, the dataset will contain only words that are important. This will help in improving the accuracy and it betters the process of mining. Stop words are removed from the text in a Python program using a Python library called “NLTK”.
Figure 11 shows an image of what the text looked like before the stop words were removed.

Figure 11 – Text before stop words removed.

Figure 12 shows the text after stop words are removed.

Figure 12 – Text after stop words were removed.

3.2.5 Stemming

Stemming is done as part of pre-processing to reduce a word to its root form. Once a word is reduced to its root form, the number of unique words in the dataset reduces and the training becomes more efficient. This ultimately results in better accuracy. Porter Stemmer algorithm is used to perform stemming. This is done in Python using “NLTK” library.
3.2.6 Spelling Correction

Since the dataset is obtained from TripAdvisor, there are a number of reviews that contain words spelt incorrectly. This causes the same two words to be treated as different words during the training phase and leads to a decrease in accuracy. To improve the accuracy, spellings of some words that are incorrectly spelt, need to be corrected. This is done in Python using TextBlob.
Figure 15 shows an example of an incorrectly spelt word.

**Intermediate Review**

*since* stayed last time ok nothing special *recently* stayed nights surprise husband check done within min even tho que lovely

**Corrected Spellings**

*since* stayed last time ok nothing special *recently* stayed nights surprise husband check done within min even

Figure 15 - Incorrect spellings in review

Figure 16 - Spellings corrected

Figure 17 shows the final output after the spellings are corrected.

**Review Updated**

*since* stayed last time ok nothing special *recently* stayed nights surprise husband check done within min even

Figure 17 - Final output after correcting spellings

### 3.3 IMPLEMENTATION

Once the data pre-processing is completed and the data is in the desired format, the model needs to be trained and tested for accuracy. For this purpose, the original dataset is divided into a “Test Set” and a “Training Set”. The model is trained using the “Training Set”. 70% of the original data is used for creating the “Training Set”. To test the model created, the “Test Set” is used. 30% of the original data is used to create the “Test Set”. To split the dataset into test and training set, a python library called “scikit-learn” was used.

There are totally 5036 records of user reviews in the original dataset. Since the dataset is split into a training set and a test set with a 70:30 ratio, the training set contains 3526 records and the test set contains 1510 records.

Since data pre-processing is completed at this stage, the dataset is devoid of stop words and since stemming is performed, all the words are in their root form. A “Bag of Words” model was created using the pre-processed data. The dataset totally contained around 9100 features after the pre-processing stage. The dataset had a number of features that had very high frequency and some features that had very low frequency. Usually, the high frequency words are the most commonly used words and the low frequency words are
the most rarely used words. The common words are not significant and it confounds the analysis if they are retained in the text. The rare words do not occur often and do not contribute significantly to the analysis. Hence they were removed. This was achieved by tweaking the “max_df” and “min_df” parameters in CountVectorizer. The threshold values were set based on data exploratory analysis. The parameter “max_df” specifies the threshold for frequency. If there is any term with a frequency greater than the threshold, that term will be neglected. Similarly, the parameter “min_df” specifies the threshold for frequency. If a term has a frequency lower than the threshold, that term will be neglected. The total number of features was reduced to around 1100 once the max_df and min_df parameters were changed. This helped improve accuracy since there were lesser but more important features to train on. In order to further help improve accuracy, “bi-gram” is used. Bi-grams consider an adjacent pair of words at a time, i.e. two adjacent words are considered together for the feature space. Using bi-grams further helps improve accuracy.

The training is performed with Support Vector Machine (SVM) using the training dataset. SVM is a classification algorithm that classifies data using labeled training data. SVM comes under supervised learning. Chen et al. [1] have concluded that SVM produces the highest accuracy for 5-star rating prediction. The training data has labels about the “Rating” which is the class that needs to be predicted. Once the model is trained using the training set, the model is tested to measure accuracy. The model predicts a rating on a scale of 1 – 5 stars for all the 1150 records in the test data set. The accuracy of star rating prediction is found to be 91.27% using a confusion matrix.

As a next step, the data was visualized to understand the average hotel rating trends and to analyze the changing rating trends for individual rating score. A word cloud was generated to understand what was most talked about regarding the hotel. All the visualizations were performed using “Tableau”. Tableau cannot generate word clouds if the input given is a stream of text. It requires a table containing two columns - the word and its corresponding frequency. A java program was written to generate this table. The table was then given as input to generate the word cloud. The visualizations are described in more detail in the Results section.
4 Results

The results were obtained after testing the model with the test set which had 1150 records. The accuracy with which the rating score for each review is predicted, was found to be 91.27%.

A confusion matrix depicts how well a model performs in terms of classifying the records, by comparing the predicted class label with the actual class label. Table 1 shows the confusion matrix that is generated.

The labels 1,2,3,4 and 5 indicate the star rating predicted by the model.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>91</td>
<td>10</td>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>131</td>
<td>3</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>367</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>497</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>34</td>
<td>431</td>
</tr>
</tbody>
</table>

*Table 1 - Confusion Matrix – Where 1,2,3,4,5 denote class labels (Star Rating)*
Figure 18 shows the original rating of the “Circus Circus Hotel” based on user reviews on TripAdvisor website. This hotel has an overall rating of **3.5**

![Figure 18 - Original Rating of Hotel - 3.5](image1)

Figure 19 shows the overall rating predicted by the model. The overall rating is computed by averaging the individual rating values predicted by the model for each review. The rating predicted by the model is **3.68**

```
>>> runfile('C:/Users/PradeepKumar/.spyder2/Code')
The predicted hotel rating is: 3.6853188929
```

Confusion matrix:
```
[[ 91 10  1  6  0]
 [ 5 131  3 18  5]
 [ 1  2 367 28  0]
 [ 1  7  7 497 15]
 [ 0  0  2 34 431]]
```

Accuracy is:
```
0.912755716005
```

![Figure 19 - Rating predicted by model - 3.68](image2)

The average rating trend of the hotel was visualized using Tableau. The resulting graph is shown in Figure 20. The hotel has shown an overall positive trend in the rating since January 2015.

The hotel’s average rating was at its **highest** in March 2013. The average rating of the hotel was **4.13** at that point.
The hotel’s average rating hit its **lowest** in January 2015. The average rating of the hotel was **3.23** at that point.

Such details help the management to look into what went wrong when the rating was at its lowest and fix them. It also helps the hotel to understand what it needs to continue doing, to provide customers with a pleasant experience.
The rating trends of individual rating scores were also visualized and the results are shown in Figure 21.

![Figure 21 - Individual rating trend over time](image)

It can be seen that the number of 5 star ratings were 70 during the month of April in 2013. This is the best in the hotel's history. If the management can analyze what made the customers happy, they can continue to provide similar services to keep improving the hotel’s rating.

Also, it can be seen that the hotel had the maximum number of 1 star ratings in September 2015. There were 21 user reviews with a rating of 1 star. This is a cause of concern for the management and if the cause for the poor ratings is investigated, the hotel can improve on the service / facilities that they were lacking in.
Figure 22 shows a pie chart of the star ratings. It can be seen that the majority of the user ratings are between 3-5 stars. This is an indicator that the hotel is doing good overall based on ratings.

It’s important for potential customers to be able to know what other customers feel about the hotel and what’s most talked about regarding the hotel. For this purpose, a word cloud was created from the all the reviews provided by users. The word cloud can be seen in Figure 23.
The size of a word in the word cloud depends on the number of times the words was used in user reviews. Hence, if the size of a word is big, it indicates that the word occurred frequently in reviews.

From this word cloud, it can be seen that the word “Circus” has been used the most number of times. This is because, the hotel offers live circus shows every day and people talk about their experiences. It can also be seen that a number of users are talking about the “Adventuredome”, which provides indoor entertainment. The word cloud gives an idea for potential customers to know what to lookout for at the hotel.

5 Conclusion

Based on the results, it can be concluded that SVM is a good machine learning algorithm to use for predicting star ratings based on user reviews. The accuracy of the model is 91.27%. It can also be understood that the dataset being chosen should have reviews distributed across all star ratings. This will help the model to train better. For example, if a dataset is chosen in such way that a big majority of the user reviews are 5 stars, and
there are hardly any reviews between 1-4 stars, then the model is likely to misclassify most reviews as 5 star ratings. This will bring down the accuracy.

It can be seen that the model is capable of predicting a hotel’s rating based on the user reviews with a good accuracy. The original rating of the hotel was 3.5 and the model predicted a rating of 3.68.

Based on the visualization, it can be seen that hotel has garnered positive reviews since January 2015 and it is likely that the positive trend will continue. The word cloud helps new customers to get an overview of the hotel and learn what’s most talked about regarding the hotel.

6 Future Work

Each user review has an overall rating on a scale of 1 - 5 stars and sub ratings for specific aspects like “Cleanliness”, “Service”, etc. The specific sub ratings can also be considered in the future during the training phase so that the rating for specific aspects of the hotel can also be predicted.

In some user reviews, customers mention other hotels and talk about how the other hotels are better in terms of price, location, etc. If the names of other hotels can be identified in the reviews, the hotel can learn about its competitors.

TripAdvisor had user reviews in languages other than English but they were not extracted in the data collection phase. In future, reviews in other languages can also be included.

7 References


