Hotel Rating Prediction via Customer Review Analysis

Huang Chaomin, Huang Yujia, Lin Chenxi, Mao Wenjian, Tao Jingyi

Abstract

This study aims to predict the hotel rating based on hotel reviews and basic information data. The business value of this topic is to provide a solution of reducing the inefficient business decisions on hotel investment. The data source used from Kaggle website includes 1,667 hotel information with 10,000 reviews. Three types of models are conducted in this project: prediction based on sentiment polarity, prediction based on sentiment polarity with hotel information and prediction based on reviews. In this topic, various models are applied: Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Natural Language Processing (NLP) models. In the Natural Languages model, the BERT developed by Google is applied to increase the accuracy of prediction. Eventually, the BERT model leads to highest accuracy of prediction. Besides, keywords generated from the natural language model provided business value to the business operator to better understand customers preference. Furthermore, the recommendation based on the geographical data also provided to hotel investors to investigate the better investment opportunities.

1. Introduction

In this session, background information and motivation are covered to give an overview on the reason why this project is proposed, leading to the problem statement of the project. Some basic facts about the data and the methodology are also included.

1.1 Background Information

Choosing a suitable hotel is one of the most important items on travelers' to-do list, when planning a trip for holidays. How to choose a hotel that matches your preference? The most common way is to go to a hotel booking website, go through hotel rating and reviews, then make decisions based on them.

Reviews are one of the most significant inputs in a customer's buying decision. This sentiment is especially true when it comes to booking hotels. We can quickly tell whether a particular hotel is stay-worthy or should be avoided entirely. More positive reviews for a hotel equate to an increase in business for that hotel.

A recent study from Cornell University's School of Hotel Administration showed that customers are twice as likely to book a hotel with positive reviews as they are a hotel with negative reviews (McCarthy et al., 2010). A second study showed that revenue is strongly correlated with reviews. That is, 1% reputation improvement leads to a 1.42% increase in revenue per available room (RevPAR) (Anderson, 2012).

1.2 Motivation

It is critical for the hotel owner to understand customers' opinions after their stay and then incorporate necessary upgrade and improvement. Investors can tell which hotel facilities or features attract the most business by extracting hot topics in customers' reviews. This information will benefit not only travelers during their hotel selection process, but also hotel owners in hotel development.

Furthermore, hotel booking websites can build up hotel recommendation systems to match travelers' preferences to increase the profit. Secondly, travelers can be classified into different groups based on their top topics of reviews. Hotel marketing groups can customize promotions to different traveler groups. Last but not the least, Machine Learning models help hotels to know what factors (e.g. WIFI, location, cleanness, etc.) are more important for the travelers to give a high rating. New hotels can also refer to such information and make improvements accordingly for reaching a high traveler rating.

1.3 Problem Statement

From the perspective of hotel owners, without reviewing travelers' comments on their hotels, they may lose lots of potential sales, because they do not understand what really matters to the travelers and what are actually chasing them away. Without this knowledge, they may come up with inefficient business decisions in hotel investment, facility maintenance and so on.

With this in mind, we aim to apply Machine Learning techniques including Natural Language Processing (NLP), to extract important business insights from the customers' review comments.

1.4 Dataset

1.4.1 Source of Data

In this project, we use the Hotel Reviews dataset from Kaggle.com. This is a list of 1000 hotels and 10,000 reviews provided by Datafiniti's Business Database. This dataset includes hotel location, name, rating, review data, title, username etc. Link to data source

1.4.2 Data Collection Method

Datafiniti provides instant access to web data. It compiles data from thousands of websites to create standardized databases of business, product, and property information.

1.4.3 Description of the Dataset

The dataset consists of about 1,667 hotels across the United States. There are a total of 10,000 reviews, updated between January 2018 and September 2018. Each review listing includes a range of details, including review date, hotel name & location, review rating etc. The list of variables we are planning to use in the project is shown in *Table 1*.

Table 1. List of Variables to be Used

VARIABLE	Type	Description
NAME	String	Name of hotel
REVIEWS.DATE	Date	Date of review
REVIEWS.ID	Int	ID of review
REVIEWS.RATING	Float	Rating of hotel
REVIEWS.TEXT	String	Content of review
REVIEWS.TITLE	String	Title of review
REVIEWS.USERCITY	String	City of user
REVIEWS.USERNAME	String	Username
REVIEWS.USERPROVINCE	String	Province of user

1.5 Methodology

Firstly, we have carried out the data pre-processing to clean up and categorize the raw data for model training, following with the sentiment analysis on all customer reviews to provide some insights on the user feedback. Secondly, we use sentiment polarity and other hotel attributes to fit the classification models (i.e. Logistic Regression and SVM models) and predict the hotel rating. Last but not least, we also use NLP techniques to predict the hotel rating from the customer reviews and compare the results with those generated from the classification models.

1.5.1 Natural Language Processing (NLP) & Deep Learning

We aim to perform text analytics on hotel reviews in order to extract travelers' preference and dislike. We will apply NLP, including the Bag-of-Words and the N-gram model, to break the textual reviews into individual corpus and further convert them into a countVec. Due to the known drawback of both the Bag-of-Words and the N-gram model in capturing semantic meaning of human language, at a later stage of this project, we will also discuss how to utilize Deep Learning to improve our learning outcome.

1.5.2 Sentiment Analysis

The basic task of Sentiment Analysis is to classify the polarity of the text, which is also the label to be used in our supervised models. Each review comment reflects a sentiment level within [-1,1] to indicate its sentiment orientation of positive, neutral or negative. We will explore multiple regression models and select the one that yields the highest accuracy.

1.5.3 Topic Modelling

Topic Modelling is also playing a very important role in our project. Upon Sentiment Analysis, we can nominate the top few keywords from each of the 3 sentiment classes. Focusing on the Positive and Negative classes, we can understand what features of a hotel interest travelers the most, as well as the other way around respectively. We will also be exploring the feasibility of predicting the overall rating of a hotel using the features identified from our Topic Modelling.

2. Data Pre-processing

2.1 Data Cleanup

Before the data can be passed to fit the model, several data cleanup steps need to be done.

As the normal hotel rating range is from 1 to 5, data cleanup should be done to remove ratings that are outside the normal range, especially when rating is 0. A zero rating means the customer did not give a final rating while leaving the review comment for the hotel. As the models need to be trained with the reviews and final rating, there is no point to analyze the data without rating and hence should be removed.

Additionally, empty values are removed as well as they will cause the model to fail.

2.2 Label Discretization

Many of the final hotel ratings (in the range of 1 to 5) contain decimal places (e.g. 4.8, 3.5). In order to train classification models, all ratings are rounded to their nearest integers and then converted from continuous numbers to discrete category variables.

2.3 Feature Engineering

Below are some important features used in our models.

2.3.1 Sentiment

This dataset does not provide any information on the review sentiment. Nevertheless, we think the sentiment information would be a useful feature for rating prediction and we would like to test our hypothesis.

We generated the sentiment information for the customer reviews via the existing sentiment analysis packages. The sentiment information consists of two scores: polarity and subjectivity.

Sentiment polarity has a range of [-1, 1], where 1 means a strongly positive statement, -1 means a strongly negative statement, and 0 indicates neutral.

Sentiment subjectivity has a range of [0, 1]. Subjective sentences (score towards 1) generally refer to personal opinion, emotion or judgment, whereas objective (score towards 0) refers to factual information.

In order to validate the generated sentiment scores, some examples of both positive and negative reviews are examined. For positive reviews, the comments are mainly talking about good things in the hotel, such as “amazing place”, or “the staff is very friendly and helpful”. For negative reviews, comments include something like “heater not working”, or “I’m very disappointed with the hotel services”, etc. In general, the polarity scores here are giving a good estimate on the overall sentiment.

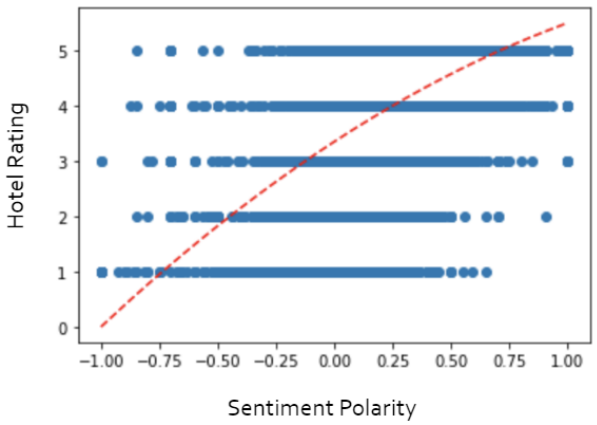

Figure 1: Sentiment Polarity & Hotel Rating

Figure 1 is a scatter plot of hotel ratings against the sentiment polarity. As indicated from the trend line, there is a strong positive correlation between sentiment polarity and final hotel rating. When the reviews are negative, the hotel ratings tend to be low, and when the reviews are positive, the ratings are getting much higher. As a result, we can say that the customer reviews are likely to play an important role in influencing the final hotel rating and the review sentiment could be a useful feature for the rating prediction model.

2.3.2 Review Length

The length of customer reviews are extracted as a separate column, which can be used as an additional feature and passed to the model together with other features in the dataset.

2.3.3 Hotel State

As this dataset only contains hotels in the US, we include the state of the hotel as the model fitting feature. We hope this information can give us some insights on whether hotels in some particular states are generally getting a higher rating than the other states.

2.3.4 Hotel Longitude and Latitude

The hotels’ longitude and latitude values are also passed to the model as hotel features. While the state information provides some insights on location preference, the longitude and latitude information could tell us additional information on whether the customers prefer hotels more towards south rather than north, or whether they prefer the hotels on the east or west shorelines.

3. Rating Prediction Models

In order to use the same dataset to validate our model, we split the original dataset into 70% training data and 30% testing data. Models are built on training dataset and validated on testing dataset.

Three models are built to predict hotel review ratings.

3.1 Logistic Regression - Polarity

The first model is the Logistic Regression model. Firstly, we use review text sentiment result "Polarity" as the only independent variable and review ratings as dependent variable. After fitting the model on training data, we applied the model on testing data to make predictions. accuracy_score method is used to compare predicted value with real value, the accuracy score is around 0.51.

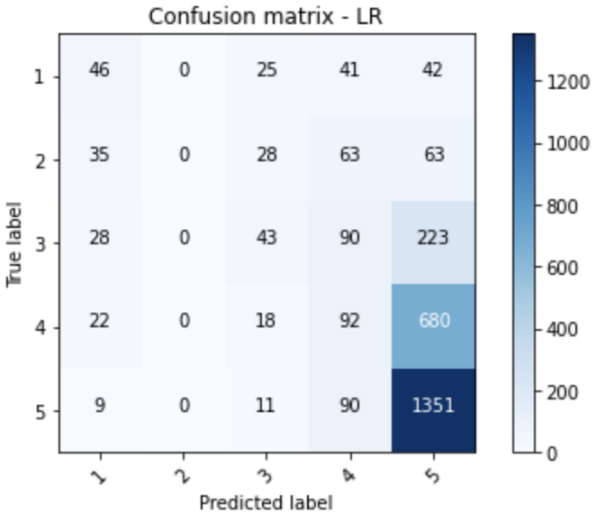

Figure 2: Confusion Matrix - LR polarity

3.2 Support Vector Machine (SVM) - Polarity

The second model is the Support Vector Machine model. Similarly, "Polarity" is used as a single independent variable to build the model. Model accuracy score on the testing data is around 0.50, which is slightly lower than Logistic Regression model.

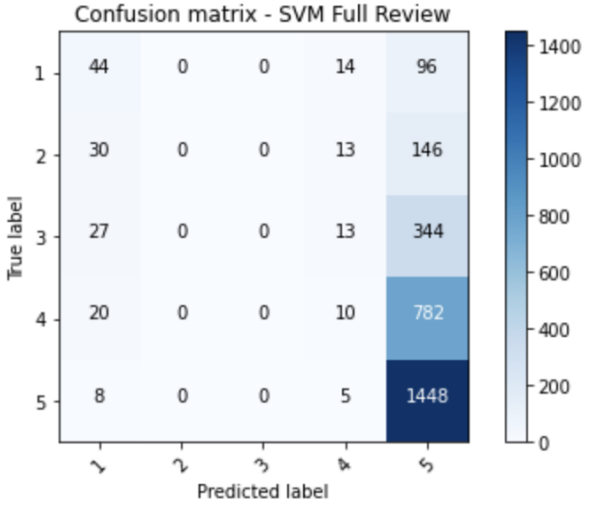

Figure 3: Confusion Matrix - SVM polarity

3.3 Logistic Regression - More Variables

Will adding more independent variables to models give better prediction results? To test our hypothesis, besides "Polarity", we add subjectivity, review length as well as latitude and longitude of the hotel as independent variables. The accuracy score of the Logistic Regression model using more variables on testing data is around 0.462, which is lower than the same model but only using polarity as the independent variable.

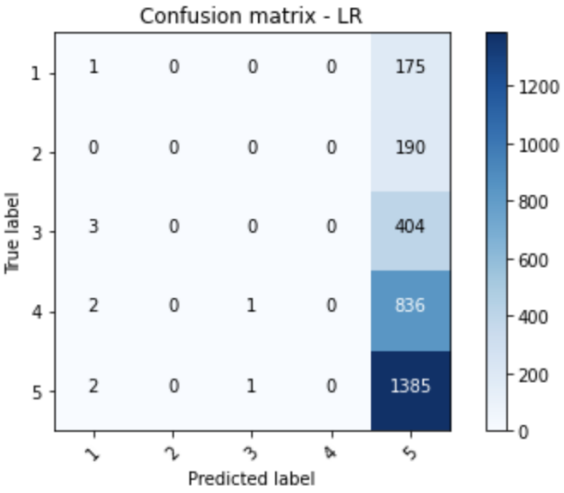

Figure 4: Confusion Matrix - LR more variables

3.4 Support Vector Machine (SVM) - More Variables

Similar to the Logistic Regression model, we add more independent variables to train the SVM model, the accuracy score on testing data is around 0.482, still lower

than the same model that uses only polarity as the independent variable.

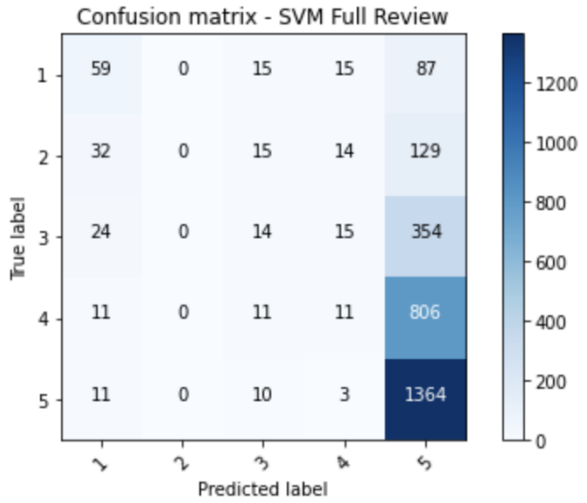

Figure 5: Confusion Matrix - SVM more variables

Here is the summary table of prediction accuracy of machine learning models:

Table 2. Prediction accuracies for Logistic Regression and Support Vector Machine.

MODELS	ACCURACY	BETTER?
LR - POLARITY	0.51	✓
SVM - POLARITY	0.50	
LR - MORE VARIABLES	0.462	
SVM - MORE VARIABLES	0.482	

From the above table, we find out that using only "Polarity" will have better prediction results, which indicates that review text impacts customer review ratings more than other factors.

3.5 Natural Language Processing (NLP)

Natural language processing, short as NLP, is a field of artificial intelligence in which machines are able to read, even understand and derive meaningful information from various languages. Machines cannot understand the human language words; instead, machines can understand numbers. Multiple-classification models can be developed based on review directly. The NLP model applied the same datasets but only 'reviews.text' and 'reviews.rating' columns.

Tokenization is the process of slicing the words into pieces. The name of the unit is token, which can represent one word, multiple words or sentences. The Bag-of-Words model is the simplest tokenization model. It categorizes each word to a unit token. The words need to be encoded as integers or floating point values. The simplest method of features extraction is Counter Vectorizer. It treats each token as a vector and counts the total appearance times. We applied a tokenization method on the data preprocessing. Counter Vectorizer is used and generated to list all the words, and convert them to a numerical value. Therefore, the mathematical methods to build up our models.

Term Frequency - Inverse Document Frequency(TF-IDF) is a technique computing a weight to each word signifying the importance of the word in the document and corpus. Thousands of reviews are used in our NLP model. So, the TF-IDF is also used in our model. After this operation, the weights for all words are generated and ready to pass into the model building.

The baseline model is built by Naive Bayes method. For the baseline model, the training accuracy is at 0.503 while the test accuracy at 0.428. After this, the logistic regression and k nearest neighbor. The best results we got is around 0.536. As for a multiple classification task, the accuracy is fairly good.

Table 3. NLP statistical models with accuracy

MODEL NAME	TRAINING ACCURACY	TEST ACCURACY
NAIVE BAYES	0.503	0.428
LOGISTIC REGRESSION	0.697	0.536
K-NEAREST NEIGHBOR	0.524	0.306

Rather than the statistical model, the machine learning model also applied to this NLP classification task. Bidirectional Encoder representations from Transformers (BERT) are used in this project. Bert uses the transformers, it contains two essential mechanisms: encoder and decoder. In this project, only the encoder used to do the classification job. The decoder is designed for prediction for the next sentence. Compared to directional models which read words from left to right, or right to left, Transformer encoders read the entire sequence of the words. The model allows it to learn from the surroundings of words as well.

In this project, the 'ktrain' package is used to build up the bert model. Besides, to reduce the workload of training, the pretrained model 'bert-base' is also used.

For the neural network model, it is important to find the proper learning rate. Neither too high nor too low will lead to the bad results. In Figure 6, the optimized learning rate of the BERT is 1e-4.

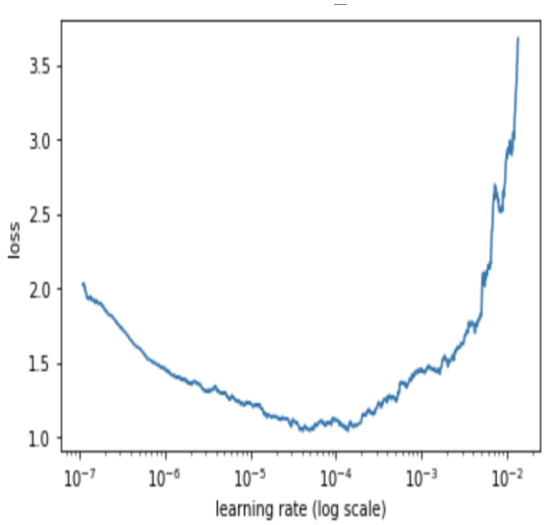

Figure 6: BERT learning rate

The final BERT model is built up based on the learning rate at 1e-4. The model is trained for 5 epochs. The final training accuracy is 0.901 while test accuracy is 0.600. In Figure 7, the confusion matrix of BERT is displayed. The accuracy scores show that BERT works better than the statistical models.

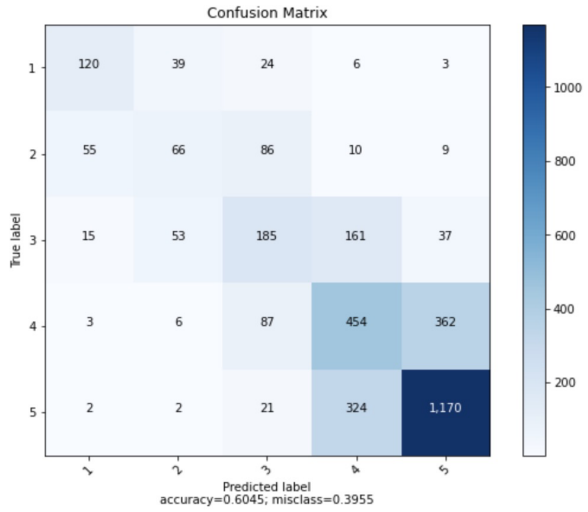

Figure 7: The confusion matrix of BERT model

In conclusion, the hotel rating can be predicted from the geographically data with reviews or only reviews. The business value of the model is that the investor can apply the model to figure out whether the location is valuable for investment. And, the review NLP model can tell investors to better understand the hotel which does not have a rating system.

4. Recommendations

4.1 Key Hotel Facilities

After exploring machine learning techniques to predict hotel rating, we are aware that the overall rating is mainly influenced by text reviews. Next, we'll see, from the above analysis, what business recommendations we can make to the hotel owners as well as the investors.

Based on our NLP, we further extract the hot topics mentioned by the customers, regardless they're part of a positive or negative comment. The topics are ranked according to their term frequency (TF-IDF) in the review comments. Topics with high term frequency give us the hint that when people assess a hotel, what key factors they take into account and what really matters from the customers' perspective.

Figure 8: Hotel Reviews Word Map

Seeing from the above word map, several words can be easily identified - hotel, room, front desk, pool, and so on.

It seems that most of the topics are related to the hotel's facilities. Let's see the actual ranking against their frequency below:

Table 4. Keywords Ranking

NO.	Key Words	Term Frequency
1	hotel	6753
2	room	6626
3	staff	4029
4	stay	3769
5	great	3519
6	good	3516

7	clean	3370
8	breakfast	2936
9	nice	2815
10	rooms	2304

The top 10 list validates our findings that more than 50% of the frequently mentioned topics are the facilities and the services that hotels offer.

Therefore, we would want to make a suggestion to the hostel owners that they should emphasize more in these fields when managing the hotel operations. Hotels can conduct systematic staff training and to lead a customer-centric hotel principle. The training should not be a one-time onboarding exercise, but throughout the employment. Hotels will also need to consider ways to upgrade their breakfast set. Hotels may want to gather customer feedback on the meals offered and evaluate if they need to introduce additional cuisine or focus on a few popular dishes. Most importantly, from the result, it tells that customers mentioned a lot about the "room" and the "stay". Hotel environment has to be kept clean and tidy all the time. Room condition needs to be retained at a high quality to make the customers feel home and comfortable.

4.2 State Ranking per Average Hotel Rating

In addition, we draw a simple yet important insight for the investors to understand where to invest. We rank the states as per their average hotel rating ("A Comparison of Hotel Reviews by State", 2016), and we realize some expected as well as some shocking results here.

Table 5. Top 3 States with the Highest Average Hotel Ratings

NO.	State Name	Average Hotel Rating
1	Rhode Island	4.60
2	Mississippi	4.50
3	Hawaii	4.36

Table 6. Top 3 States with the Lowest Average Hotel Ratings

NO.	State Name	Average Hotel Rating
1	Delaware	3.32
2	Alaska	3.33
3	New Jersey	3.51

Rhode Island, not very well known for its scenery, surprisingly leads the ranking with the highest 4.60 average hotel rating. Mississippi and Hawaii, the 2 traditional tourism states, secure the 2nd and the 3rd places respectively.

On the other hand, New Jersey, Alaska and Delaware enter at the bottom of the list with pretty depressing average ratings of around or below 3.50. These low average ratings may imply that something is not doing right with the hotel industry in these places. From a different point of view, it also hints free-up market opportunities. If hotels are not meeting customer expectations, there's a spare market open to new businesses. In particular, large hotel chains/franchises may want to consider what they can do to increase their competition in those markets.

Another interesting insight is that most states' average ratings are between 3.80 and 4.28. If we regard 3 to be the average on a 5-star scale, then this suggests either of the two possibilities.

1. Most states and hotels are factually outperforming customers' expectations, as even those losing states can achieve a rating of around 3.50;
2. Most reviewers have a bias toward giving a high-score rating.

In any other circumstances, receiving a rating of 3 out of 5 may be a good and average thing; nonetheless, based on our hotel review data, receiving a 3 is actually pretty bad.

As such, hotels may need to re-evaluate their market position and competition power when taking the ratings into consideration. Being able to reach an average rating does not necessarily represent average performance in the competition.

5. Conclusion

To summarize, in order to uncover the relationship between hotel review rating and customer text comments, we apply multiple machine learning techniques: data pre-processing with feature engineering, model selection from

logistic regression, Support Vector Machine, Bag-of-Words and BERT, based on the test accuracy scores.

Having explored these models, we arrive at interesting insights on how hotels can improve their ratings. By identifying keywords in review texts, we learn the most likely reasons for which customer provides a good or bad review. These are the areas that we would like to propose to the hotel owners for more attention.

Overall, in the US hotel market, hotels with ratings above 3.50 are considered of average quality only, which gives us a sense of intensive competitions in this industry. New hotels, entering the US, can assess the facilities and services and come up with an anticipated hotel rating. The ratings will assist hotel owners to reliably develop a long-term strategic marketing approach for their hotels.

6. Link to Github Code Repository

https://github.com/BellaTao19/BT5153Group06

7. Reference

A Comparison of Hotel Reviews by State. (2016, March 9). Retrieved April 23, 2020, from https://datafiniti.co/state-state-comparison-hotel-reviews/

Anderson, C. (2012). The impact of social media on lodging performance [Electronic article]. Cornell Hospitality Report, 12(15), 6-11.

McCarthy, L., Stock, D., & Verma, R. (2010). How travelers use online and social media channels to make hotel-choice decisions [Electronic article]. Cornell Hospitality Report, 10(18), 6-18.