Are Online reviews important: Exploring the impact of customer review on the hospitality industry

Apurva Singhvi, Pranab Bhadani, Kumar Satyam
asinghvi@iu.edu, pbhadani@umail.iu.edu, ksatyam@indiana.edu

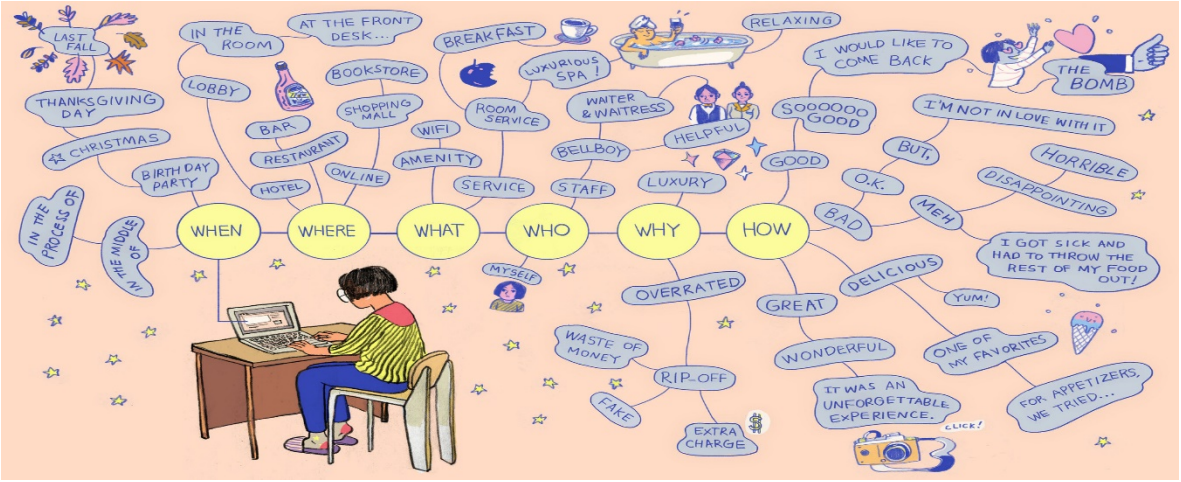


Image Source: nytimes.com

Abstract – The major reason for any hotel to get successful is to understand user needs and wants. The Internet now being a predominant means of travel bookings, people give a lot of thought on other traveler's reviews and opinions before booking a hotel. The main aim of this project was to understand the feedback mechanism of customers like user reviews, rating, and comments; analyze them by employing various techniques like sentiment analysis, Multinomial Logit model and get a better understanding of hotel industry and its features. Another aspect of the analysis was to mine customer reviews and to come up with a list of most important attributes related to a hotel which all the customers are writing about.

Keywords – Online Review, Hotel, text mining, sentiment analysis

1. INTRODUCTION

In this Internet age with the growing consumer adoption of technology in all aspects of life, the hospitality industry has not been left untouched. Hotel Industry has shown a tremendous growth in recent years and they are employing various techniques and ways to attract customers. Various e-travel websites like Expedia, Yelp, and Trip Advisor have given facility to the consumers to provide feedback and share their overall experience. This feedback mechanism like customer reviews, ratings, and comments can be used and analyzed to get a better understanding of the hotel industries and its features. We can employ various techniques like sentiment analysis, Linear Regression model to get a better insight of various features of hotel and other related fields.

2. BACKGROUND

A recent survey shows that more than 61% of vacationers now trust online reviews – they even visit social media sites and check image-sharing platforms to make a conscious decision. In another survey, 79% of the travelers said that a good management response to a bad review reassures them whereas, 78% of the travelers said that a good response to a good review makes them think highly positive of the hotel [2]. With the recent trend of using smartphones in technology driven, culture consumers use their phones for price comparison and reading online reviews before deciding on any hotel.

It has been observed that depth at which the NTOs (National Tourist Organization) uses Social media like Twitter, Facebook or any other blogging platform is tremendous. The research to understand the examination of social media usage in destination marketing had two approaches:

a. Data mining methods and approaches to identify the official presence of NTOs in various social media platforms.

b. Context analysis to understand the depth of the social media utilization by various NTOs [3].

In another study, the authors [4] tried to imitate the planning of random tourists using an online search engine i.e. Google when searching for information about destinations. It also talked about how the social media has made the life of the tourists easy by getting the reviews and recommendation about the hotels which they are planning to stay during a tour. For the analysis purpose, the authors of the paper chose a set of predefined keywords which are the most relevant to

the tourism industry and then applied data mining techniques to come up with relevant outcomes which not only help tourists to choose best options as destination but also help businesses and other organizations benefit which are directly related to tourism industry.
An another research was carried out to study the impact of electronic word-of-mouth(eWOM) on the consumer decision-making process and tourist expectations and accordingly hotel industry should develop specific marketing strategies to consider the synergy among social media. [5].

These researches influenced us to analyze how tourism sector is getting affected by social media usage. We used Customer Feedback Systems(CDS) from Trip Advisor Website (one of the leading e-business company in the tourism industry) for judging the actual performance of a hotel by doing sentiment analysis and also considering the other rating parameters given by the customers on the website to build a model, and to build a hypothesis whether these parameters are having any impact on the hotel's overall ratings. [6]

3. ARTICLE BODY

The entire execution of the project can be depicted as shown below (Fig 1) which started with Data extraction, Data preprocessing, performing text mining and sentiment analysis, applying the statistical model and eventually depicting the results in the format which are easy to understand.

Figure 1: Flow Diagram of Project Execution

3.1 Data Extraction:

The data was made available by the university of Illinois at Urbana-Campaign on their website http://times.cs.uiuc.edu/~wang296/Data/ under the section **TripAdvisor Data set.** We collected data sets for all the formats that are JSON, Text and, Processed.
The data consists of reviews from various users for around 1850 hotels. The raw data consist of three formats: 'Text', 'Processed', and 'JSON'. Initially, each one of us picked one format of mentioned formats for data transformation. We transformed the semi-structure to structure format for further analysis. After analyzing the structure of the datasets of these three formats, we ended up using the 'Processed' format files for our initial conversion process which converts all the 'Processed' file into a .csv file.
We studied the reviews and other ratings given by the customers for a subset of 1850 hotels on trip advisor.

Since we used the 'Processed' format file for our further analysis, following are the features for the 'Processed' format files:
a. ***Hotel ID***: It a unique identification number given to every hotel.
b. ***Author***: It is the customer name who has given the review.
c. ***Content:*** The review given by the customer. It was basically a text feedback which was either positive, negative or neutral.
d. ***Date***: Date when the review was posted.
e. ***Overall Rating***: The rating given by the customer on the overall experience of a hotel.
f. ***Value Rating***: It is the rating for the value for money factor.
g. ***Rooms Rating***: It is the rating given by the customer based on the condition of the hotel rooms.
h. ***Location Rating***: It is the rating given by the customer based on the location of the hotel. The important places in the nearby town.
i. ***Cleanliness Rating:*** It is the rating given by the customer based cleanliness of the hotel including rooms, bathrooms, lobby etc.
j. ***Check in /Front Desk Rating***: It is rating which measures customer experience with respect to the front desk.
k. ***Service and Business Service Rating***: It is the rating given by customers if the hotel has conference rooms or provide services like shuttle services or is apt for conducting any business conferences.

3.2 Data Parsing and Cleaning:

Data parsing is a very important step to convert unstructured data to a structured format. Firstly, we tried to parse the JSON format of data where we used java codes to parse the JSON formatted data into CSV tabular formatted data. But because of the large number of unwanted encoded characters in the dataset, the java code was not able to load the data into java object and hence we were not able to parse the JSON data properly.

Next, we tried to convert the "parsed-text" formatted data set into csv tabular formatted form using python codes. Even though the input data contained the unwanted encoded characters, we were able to remove them using special capabilities of python and then loaded the data into python objects. As the input data was following a pattern/ format, we used python parsing techniques to fetch valuable data frames out of it and programmatically load the data frames into a CSV file.

This CSV data had some missing values and some unwanted encoded character which we removed using MS excel tool. In totality, we had around 250,000 records after parsing and

after removing encoded characters we eliminated around 52,000 rows. We also had values for attributes as -1 which indicated that they were missing in the original file. After eliminating all such records, we had around 72,404 valid records on which sentiment analysis was performed.

3.3 Sentiment Analysis:

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service [7]. The primary purpose of performing sentiment analysis of the reviews given by the visitors to the hotel is to generate data for the categorical feature "sentiment" in our Multinomial Logit model. The first task in performing sentiment analysis is to read the reviews from the CSV files into a data structure that can be easily manipulated and accessed. Training data was prepared manually by doing sentiment analysis of around 7000 user reviews. We then used Panda data-frame for storing all the data coming from trip advisor website. Our goal was to classify the reviews into three broad categories i.e., +1 for the positive sentiments, -1 for negative sentiments and 0 for any neutral comment. Next, we have used "TfidfVectorizer" function from the NLTK package to fetch all the important features from our training dataset. Once we have created the feature vector from our training dataset we needed to train our model on this data. The model we used here is SVM to predict the sentiment. The main reason of using SVM classifier is because of its robustness from noise and outliers. On fitting the model and finding the accuracy, we got nearly 91% accuracy. Considering the performance of our model, we were satisfied enough to pass all the test data and get the predicted result i.e., the sentiment of the reviews from them. Please find some sample of user reviews from our dataset based on their polarity.

Positive Sentiment Reviews

1. *Fine hotel in a great location with friendly staff I stayed here as part of a conference, and found it to be a fine, pleasant property with a good mix of local character and modern amenities. The staff were all helpful, friendly, proficient in English, and efficient. I found the rooms to be spacious and quiet, and the morning breakfast was very good. Although it didn't vary much from day to day, there was enough to choose from to make each morning different. My room on floor 3 ended up being too warm for my taste, but the windows didn't have screens, so I didn't feel comfortable leaving the windows open for better ventilation (plus the windows shut out street noise, which can be quite noticeable in the city). It would have helped to have a TV remote, a clock in the room, and a bit more explanation in the room (how to dial using the telephone, how to use the room control system, internet connectivity). But what the room lacked in explanation, the staff made up for in helpfulness (happy to give directions, provide wake-up call).I'm not sure if internet works this way in all the other hotels, but internet service was purchased in advance by the hour, and could be spread out over several days. So, your time on and off the internet was measured (be sure to disconnect when you're not using it), and I ended up spreading 10 hours of purchased internet over 5 days (but I didn't figure out if there was a way to monitor how much time you had left).The hotel was a great location for my conference in the Fortessa di Basso, and is very close to the train station. It was right next to a street market, which was interesting to browse through, and close enough for a nice walking tour of all the sights of Firenze. Overall, a very pleasant stay and I would recommend it for visitors to Florence.*
2. *Highly recommended Stayed here for three nights with my mother just before Easter, and this was the perfect base for our visit - beautifully decorated, immaculately clean, and in an ideal location in a fascinating part of Venice, an easy 15 minute walk from the main bus station but away from the crowds. The staff are charming and go out of their way to be helpful - obviously they take pride in what they do. And we paid just 355 Euros for two people for three nights - an excellent deal given how expensive I had been led to believe Venice would be. Highly recommended.*

Negative Sentiment Reviews

1. *Horrible Area, Very rude and unhelpful staff My mum and I stayed at Cova this July.I will say the rooms were lovely, brand new and very clean!Be warned about the area, Ellis St is nicknamed Crackville. We saw Dozens of people partaking in unsavoury activities day and night on the next block down from the hotel. It is very noisy and there are all sorts of goings on outside the hotel at all hours. Do not walk down Ellis St after dark, whatever you do!!The Staff do not speak very good English and are not helpful at all!When I checked out they charged me $20 for a phone call I didnt make, I had picked up the phone, dialed a number and it was engaged. They insisted on charging me for the call anyway. I was grateful we only had 4 nights there, it was more than enough!!!*
2. *Horrible Customer Service! My husband and I stayed at the Silver Cloud because we were in town*

for a show. We had problems with the staff from the very beginning. I travel often with my work all over the country and over seas. I have never experienced such horrible treatment. Upon checking in the front counter staff asked for a credit card for deposit on the room as most hotels do. I informed them that I would be doing a cash deposit as was my custom due to past issues with identity theft and fraud. I have never had an issue with this at any other hotel. Often hotels will ask for a larger deposit if you do cash, but they always except it. In fact it is against the law in our country to refuse cash because as it states on any bill this note is legal tender for a debts public and private. Despite my explaining all of this the staff continued to argue with me. I explained to them all the various hotels I have stayed at and have never experienced any problem and in fact I even do a cash deposit for rental cars. The manager finally came out and was even more offensive. In light of the time and our need to get to our show, I told them I would consent to them charging a $25 dollar deposit on my credit card but no more and asked that they turn of my phone and pay channels because I did not want to be charged for any services by mistake. The next morning upon checkout the matter only got worse. As I was checking out I needed to pay for valet parking and expected them to charge against the previously authorized $25. They told me I would need to pay an additional $20 because they did not know how to do that. As a business owner I have a merchant account and know if you authorize a credit card you can then go in and use the pre-authorized amount for the actual charge. In fact that is what an authorization is for. Then the girl asked me if I would prefer to pay CASH for the parking! I was irate and said that I had tried to pay cash last night, but they wouldn't accept it! I refused to pay and insisted she charge against the already authorized amount and if she did not know how to do that she could call her merchant account provider and find out how to. The whole experience was absolutely ridiculous. I suggest the entire staff take a course in proper customer service skills and front desk procedure.

Neutral Sentiment Reviews

1. *Good hotel... for the right price. Overall we were pleased with the stay. Parking was abundant and free, internet was fast and free. The hotel restaurant was good. Indoor/outdoor pool was nice (but could use a little leaf cleaning). The room iteself was OK... but they have a bit of a mold problem (minor, but you could see some, and it smelled a little musky... although Febreeze seemed to make it better). Overall a nice place, but needs to go after the mold issue and the rooms need updated.*
2. *OK for business travellers. The room I got was OK. It is standard suite from an Embassy Suites. The hotel is generally clean except the underground garage, which is a little too smelly and dirty for me. It is located at the intersection of major freeways so the noise is a given. But I slept OK.Some of things inside need major improvement. Like their gym, pool, furniture. They are just TOO old.We had to venture out to get a good meal. They are quite a few good restaurants around. The one inside the hotel is just mediocre.In general it is all right. But don't expect it to Wow you.*

3.4 Text mining for finding important attributes of a hotel:

Our next step in the project, after finding the sentiments of the reviews was to look for all the important attributes of the hotel for example staff services, cleanliness, etc. which are mostly talked about. By analyzing this we can find out what all attributes visitor of the hotel generally cares about. To achieve this goal, we have extensively used Python and NLTK package. For preprocessing the user reviews, we used tokenization, Stop Word, Stemming and pos_tagger steps. Firstly, we started with tokenizing all the words in the reviews and then removed all the stopping words from it. Next, we applied pos_tagger to tag all the words with related parts of speech. Then we fetched all the noun words from the list and calculated its frequency. The idea or notion behind performing this analysis is that we can safely assume that the noun that has appeared most number of time is the target features in the hotel that have been talked or written most by the customer. We then sorted the words in descending order of frequency to find the top 10 things that customer cares about.

3.5 Statistical modeling using Multinomial logit model

Our final task was to build a statistical model that would help us predict the overall rating of the hotel based on the various independent features. The independent features included in the model are value, rooms, location, cleanliness, check-in front desk, services and business services. We also kept sentiment as one of the categorical variables to check whether sentiment itself has an impact on deciding the overall rating of the hotel or not. We have used the multinomial logit model to analyze the relationship between overall rating and the other features mentioned earlier. Multinomial logit model gives the log-odd ratio probabilistic value of dependent features based on the given independent features. The estimation techniques used in the model is maximum likelihood estimation. Based on the result shown in the SAS output result (Fig.4), keeping the

reference for overall rating as 4 for clear comparison, we can see that ‘value’ (Beta = -1.349) has the greatest impact on the overall rating. On the other hand, ‘check-in’ value of the reviews although turns out to be significant but it doesn’t impact much on the overall rating of the hotel (beta = -0.3987). Thus, we did a detailed comparative analysis of all the important features related to a hotel in predicting the overall rating of the hotel. We also found that the t-statistics for all the features turned out to be significant at 0.01 confidence level. This justifies our selection of all the features in our overall analysis.

4. RESULTS & OBSERVATIONS

4.1 Data Visualization and Results

The ratings and sentiment visualizations were prepared by using Matplotlib library in Python. We randomly took data of 5 sample hotels and compared each of the hotels with respect to the 8 features including the sentiments. The features include value, rooms, location, cleanliness, Check-in front desk, services and business services.

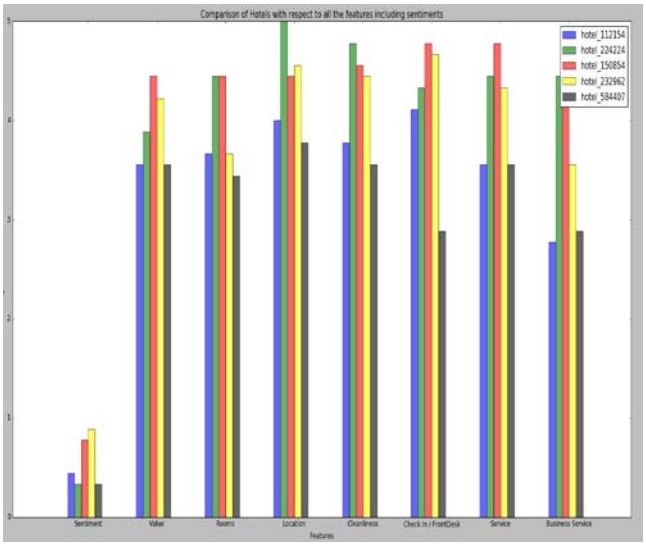


Figure 2: Comparative Study of 5 sample hotels with respect to all the features including sentiment

On analyzing the graph, we can say that hotel_224224 is in the best location. Talking about business service features, the hotel_584407 does not have good feedback from customers. In a similar way, we can compare the sample hotel w.r.t other features. We can also extend these features for all the list of hotels that are available in the data set and can identify the best features of a particular hotel and help a hotel identify their weakness and its strengths, so that the hotel can attract the customers via their talked about features and can improve on their negative features.

The next graph was developed graph by doing a feature based sentiment analysis on all the reviews given by the customers for a sample hotel. The motivation behind this analysis is to mine customer reviews to come up with a conclusion that says about the most important aspects related to a hotel which all the customers are writing about. The graph (Fig 3) shows the importance of different attributes of the hotels.

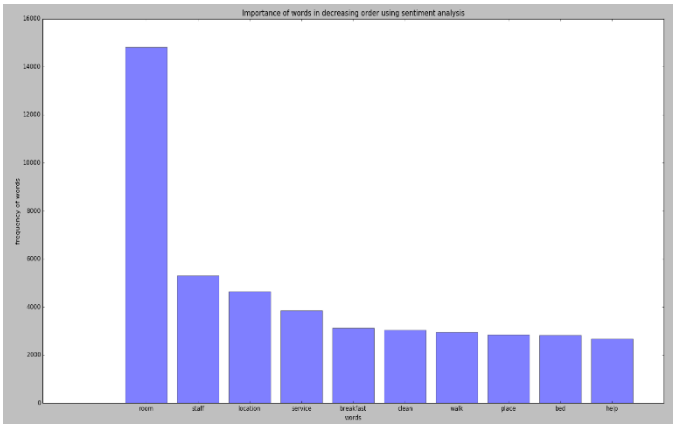
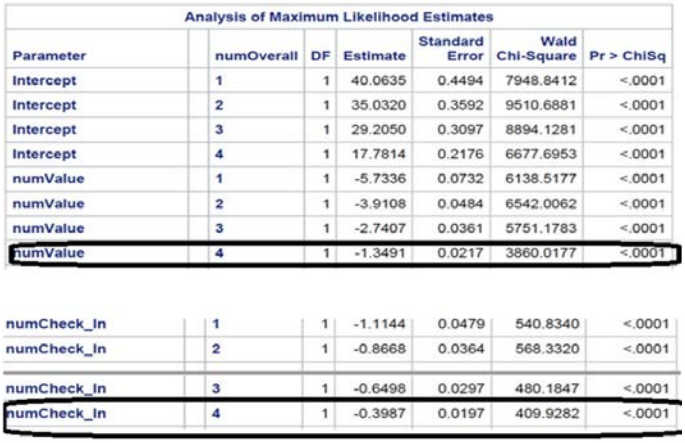


Figure 3: Importance of words in decreasing order using sentiment analysis

The figure 3 indicates that the guests of the hotel are mostly concerned with the condition of the ‘room’. For eg. neatness of the rooms or ventilation of the rooms. The ‘room’ is followed by ‘staff’. For example, ‘staff’ behavior or courtesy towards the customer. Whereas ‘help’ is the least important word as it is being used the least number of time.

The next graph (Fig 5) demonstrates our SAS output. As in our case the overall rating can range from 1 to 5. So it's an example of multiclass multivariate regression equation. Since it is a multiclass example we have decided to use multinomial logit model instead of normal logit model. The model gives us the relationship between log odd ratio between dependent variable (overall rating in this case) and other dependent variables. To estimate we have used maximum likelihood estimation technique for estimating or Beta estimates.

Based on the result we got from SAS output, we found that the feature "Value" has the highest impact overall rating whereas on the other side feature 'check-in' has the least. Thus our model keeping the reference level '5' in the SAS code gave us the detail comparative study of the various features.



Analysis of Maximum Likelihood Estimates

Parameter		numOverall	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1	40.0635	0.4494	7948.8412	<.0001
Intercept		2	1	35.0320	0.3592	9510.6881	<.0001
Intercept		3	1	29.2050	0.3097	8894.1281	<.0001
Intercept		4	1	17.7814	0.2176	6677.6953	<.0001
numValue		1	1	-5.7336	0.0732	6138.5177	<.0001
numValue		2	1	-3.9108	0.0484	6542.0062	<.0001
numValue		3	1	-2.7407	0.0361	5751.1783	<.0001
numValue		4	1	-1.3491	0.0217	3860.0177	<.0001
numCheck_In		1	1	-1.1144	0.0479	540.8340	<.0001
numCheck_In		2	1	-0.8668	0.0364	568.3320	<.0001
numCheck_In		3	1	-0.6498	0.0297	480.1847	<.0001
numCheck_In		4	1	-0.3987	0.0197	409.9282	<.0001

Figure 4: SAS Output

This model can also be used for predicting the overall rating of the hotel, which will be very helpful for hotels to identify their strengths and weakness.

For detail output of SAS please refer: https://github.com/satyamsah/social-media-mining

5. CONCLUSION

After analysis, it was found out that the online reviews posted by people play an important role in decision making for selecting a hotel. Firstly, we found that there is a significant relationship between the reviews given by the user and the overall rating of the hotel. Secondly, on mining the reviews we found the guests have mostly discussed the rooms and the staffs of the hotel.

After conducting the sentiment analysis using NLTK and SVM classifier, we achieved an accuracy of 91% in classifying the sentiments. The overall rating predicted by multinomial logit model is in accordance with the overall ratings given by the users.

6. CHALLENGES & FUTURE OPPORTUNITIES

While doing the data pre-processing many reviews were neglected because of the presence of foreign language, as we had limited the scope to English language only. Some of the reviews were removed as the ratings were -1 for our six attributes indicating that they were missing in the original file.

The training data set for performing the sentiment analysis was prepared manually by going through the reviews and identifying them as positive, negative or neutral. As a future project, a continuous dynamic lexicon must be used, instead of the manually created training dataset. Also, the sentiment analysis should be carried with noises in it.

ACKNOWLEDGEMENTS

We would like to thank our professor Vincent Malik for his support and guidance throughout the project execution.

REFERENCES

[1] "The Art of the Amateur Online Review." 2014. *Michael Erard*. Nov 29. https://www.nytimes.com/2014/11/30/business/the-art-of-the-amateur-online-review.html.

[2] "Why Online Reviews are important for Hospitality Industry." 2015. *Nimesh Dinubhai*. Feb 10. https://www.linkedin.com/pulse/why-online-reviews-important-hospitality-industry-nimesh-dinubhai

[3] "An examination of use of social media in destination marketing" 2014. Abbas Alizadeh, and Rosmah Mat Isa. Aug 03. http://globalbizresearch.org/Singapore_Conference/pdf/pdf/S466.pdf

[4] "The Use of Social Media and Internet Data-Mining for the Tourist Industry". 2016 Krask, B. and Kyela, K. Feb 25. https://www.omicsgroup.org/journals/the-use-of-social-media-and-internet-datamining-for-the-tourist-industry-2167-0269-1000197.php?aid=69470

[5] "Web Reviews Influence on Expectations and Purchasing Intentions of Hotel Potential Customers" 2013. Aurelio G. Mauri and Roberta Minazzi. May 07. https://www.researchgate.net/publication/257118239_Web_Reviews_Influence_on_Expectations_and_Purchasing_Intentions_of_Hotel_Potential_Customers

[6] Retrieved from https://www.usatoday.com/travel

[7] Sentiment Analysis Wikipedia. (2016). https://en.wikipedia.org/wiki/Sentiment_analysis

Contributions of Team Members

1. **Apurva Singhvi**: Data cleansing and pre-processing. Discussion on what models can be applied on the data. Preparing training dataset for sentiment analysis and drawing out analysis of the sentiments and writing report.
2. **Pranab Bhadani**: Writing python code for data conversion using processed files, Discussion on what models can be applied on the data. Text mining and applying multinomial logit model.
3. **Kumar Satyam**: Writing Java code for Data conversion using JSON files, Discussion on what models can be applied on the data, writing code in Python for Data visualization and writing report.