Sentiment Analysis – Hotels Review

Contributors: Faraz Mirza, Mayank Khare, Sai Manne, Syed Kamil, Vamsi Chinta

MSA 8040 Data Management (Final Project) – Georgia State University

I. Objective

The objective of this document is to confirm that the customer finds a 'good' hotel. In order to make the decision easy we have identified some key elements. In this document, we define how we extracted useful information or knowledge in order to achieve this objective.

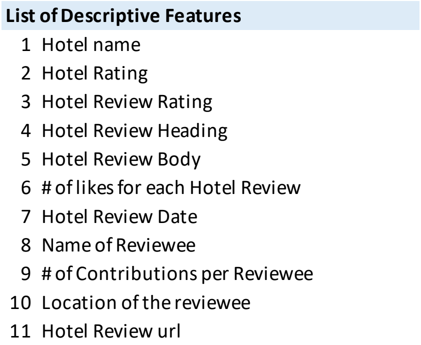
II. Data Procurement

According to 'Forbes'[1], Trip Advisor has the highest number of both hotel's ranking and hotel reviews. With this high amount of volume and market share, we assumed that the hotel reviews taken from the Trip Advisor is a sufficient amount of data to gauge a 'good' hotel regardless of location.

According to the Trip Advisor, Atlanta has 216 hotels total that have an overall hotel rating that range categorically from 1 to 5 with an interval spacing of 0.5. (i.e. 1, 1.5, 2.0, etc.)

Of which, we randomly sampled 3 hotels per rating class level resulting in a sample population size of 23 hotels.

In order to rank the hotels, the following features were selected to be mined:



List of Descriptive Features	
1	Hotel name
2	Hotel Rating
3	Hotel Review Rating
4	Hotel Review Heading
5	Hotel Review Body
6	# of likes for each Hotel Review
7	Hotel Review Date
8	Name of Reviewee
9	# of Contributions per Reviewee
10	Location of the reviewee
11	Hotel Review url

We mined all the reviews TripAdvisor had for the hotels we needed. Each hotel has varying review time ranges but with that said, our sample dataset ranges from 2002 to 2018.

The procedure used for procuring the data stated above is as follows:

We chose to use Scrapy for web crawling because it is optimal for handling redirections and auto-throttling, to avoid overloading the servers[2].

CSS selectors were used to extract the data because all of our queries only required searching for elements in the forward direction [3] . Furthermore, we outputted our queries into MongoDB to serve as our database needs. From here, we used python to extract the needed data into a pandas Dataframe for our data analysis.

III. Descriptive Statistics

Population size (Hotels) = 216 hotels
Sample population size (Hotels) = 23 hotels
Total number of reviews = 20,017 reviews

IV. Prediction Model (sentiment analysis)

A – Metrics of the Sentiments

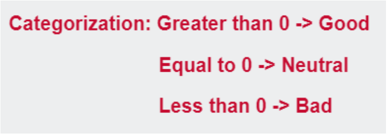
The goal is to provide customer enough information about Hotels and about existing reviews it had received in order for them to make the effective decision of whether or not to choose that hotel.

To accomplish our goal we extracted information from one the reliable third party source Trip Advisor that is broad and available for customers to write and leverage it when choosing the hotel for their needs.

In order to classify our sample dataset, we decided to use lexicon and rule-based sentimental python libraries for natural language processing to perform a sentimental analysis.

We explored and utilize two most effective modules used are 'NLTK's Vader' and 'Textblob'. These two libraries are considered optimal for NLP applications, in fact, GitHub has 'NLTK' ranked number one and 'TextBlob' as second in the field of sentiment analysis. The sentiments were computed as polarity scores between -1 and 1. These scores are numerical based, so in order for customers to make sense of the data and also to evaluate our model, we decided to categorize these values.

Categorization based on the polarity score of text reviews:



Categorization: Greater than 0 -> Good
Equal to 0 -> Neutral
Less than 0 -> Bad

We have also considered customer rating as one of the attribute to compare our text sentiment and make sure it is consistent and provide more insight about the hotel.

Similar to Text Rating we have rank customer ratings so it is diverse and tell the complete story:

[1] Forbes Website
[2] Scrapy Tutorial
[3] CSS vs Xpath reference

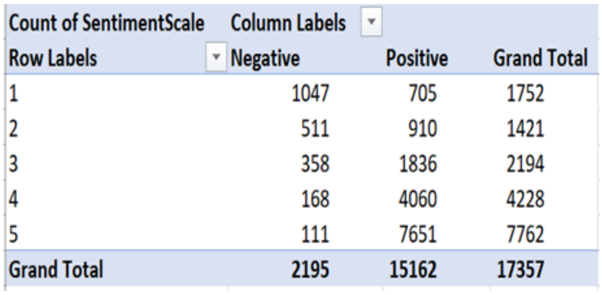
- **Greater than and Equal to 3 -> Good**

  Equal to 2 -> Noise

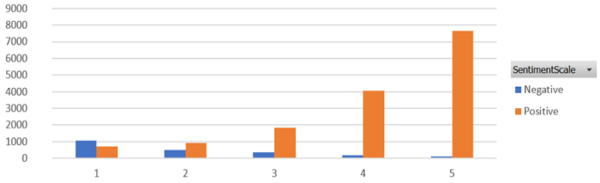
  Equal to 1 -> Bad

Note: Main reason we are using customer rating in our analysis, so we can identify the credibility of the review.

B – Comparing the Credibility of Customer Rating with the text review sentiment scale:



Count of SentimentScale	Column Labels ▾		
Row Labels ▾	Negative	Positive	Grand Total
1	1047	705	1752
2	511	910	1421
3	358	1836	2194
4	168	4060	4228
5	111	7651	7762
Grand Total	2195	15162	17357



C – In order to understand the scalability of the data and interpret the best results we have provided further analysis and compare the accuracy and the precision of the results.

Accuracy = No. of Correct Predictions/Total Predictions

Precision = True Positives/(True Positives + False Positives)

Recall = True Positives/(True Positives + False Negatives)

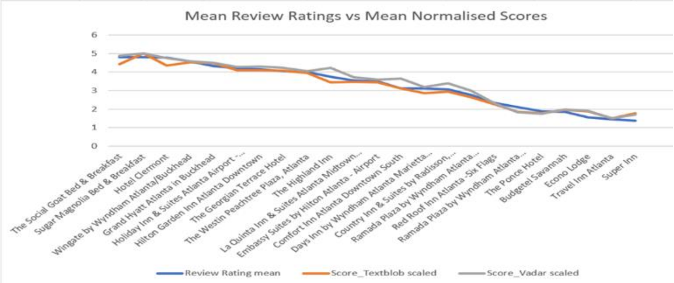
F1 Score = 2 * ((Precision * Recall) / (Precision + Recall))

D – Results between the two libraries with respect to customer rating and the text review:

Vader:	Textblob:
Accuracy: 88.3%	Accuracy: 84%
Precision: 0.96	Precision: 0.9535
Recall: 0.945	Recall: 0.9566
F1 Score: 0.953	F1 Score: 0.955

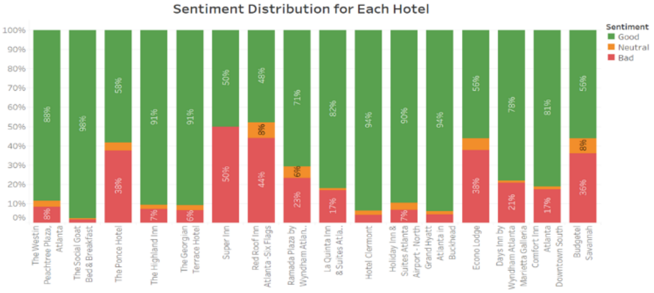
E – Consistency Element:

This is the further drill down on each hotel, in this section we have identified how customer rating have taken place when comparing with the sentiment scale. In this graph, we have shown each hotel average review rating along with average normalize rating. Furthermore, we are comparing the average rating by the classifier so it can bring about the more precise comparison.



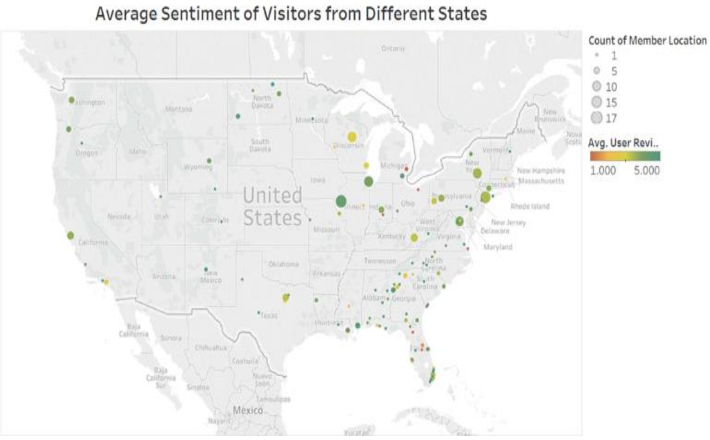
F – Distribution Element:

Above we have shown how the classifier's predicted the sentiment score compare to the actual rating, but in the image below we are showing what percentages it is assigned to each category (Bad, Neutral, Good). As you can see it provides the scale in percentages on each hotel, this is to further enhance our point of view of how rating have been received and also it is distributed when reviewing all rating at once.



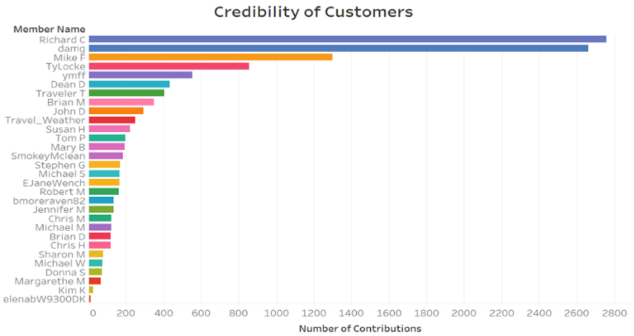
G – Based on the geographical location of frequent visitors:

Since not all the customers that have visited the hotel and provided the feedback are locals, and we want to point that out by categorizing customers input results based on the states they belong to, this is another way to validate the credibility of all customers and to echo that our analysis is based on the actual authentic users. Also, it provides more insight to the user if they are visiting from one those states.



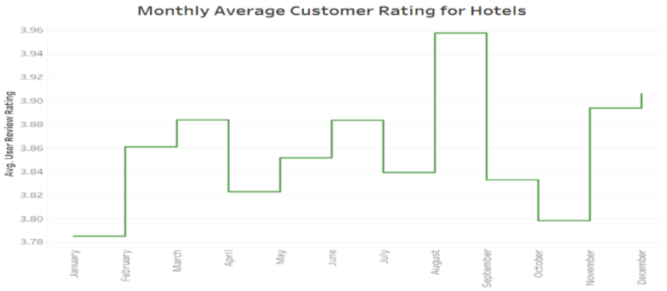
H – Credibility:

Throughout in our findings, we have made sure to extra care for the credibility of the data, and that's why we have scanned our data thoroughly and make sure each review received by the customer is clean and authentic. We have scan each users and make sure to only include reviews from the customers that are genuine and have also provided reviews on more than one hotel. This exercise validates our accuracy results in granular level.



I – Hotel reviews on the monthly level:

We understand that how important it is to keep seasonality in mind, and how hotel make adjustments to accommodate customer needs based on the season; hence, our reason to scale hotel by month. This way new customers would know which month the star month for the hotel and which months they are not so good, earlier I have mentioned how credibility and customer needs are important for us and since we are providing recommendations of which hotel is the best hotel, we have made sure to provide monthly average results by hotel. This way customer would know which month to visit the hotel and when not to visit the hotel.



V. CONCLUSION:

This analysis is strictly based on the hotels that are located in the Atlanta area, and the reviews are strictly based on the users that have visited the hotel. To provide more precise recommendations, we have analyzed customer's reviews along with their ratings and created our recommendation. By no means this list can be used for marketing strategy. Our group have extracted this data through one of the reputable website (www.tripadvisor.com) because in it you will find all types of reviews and rating based on the customer's experience. Furthermore, our recommendations are strictly driven from the customers reviews and their ratings. We have performed a sentiment analysis for 23 unique hotels and around 20K reviews. Sentiment analysis converts the review received in text into a known scale: for example (-1 to 1), where -1 means review is bad and 1 means review is good. To encompass and enhancing the visibility we have provided accuracy between two classifiers that were used for the sentiment analysis. Through this project our group have utilize Python modules such as: Panda, Numpy, Pymongo, and Matplotlib, and in to store our information we have used the MongoDb while MS excel and Tableau to explore for graphic needs.

VI. OPPORTUNNITIES:

Potential opportunities for improvement are:

1) Trip Advisor does not verify whether the review is made by someone who has stayed at the hotel. Due to Trip advisor's potential dominance in the market of hotel review sites. This enables our source of data to contain X% of fradulant claims. If given more time, we would have considered running our analysis on multiple hotel review sites.

2) There were customer reviews in the dateset that were conflicting with their ratings. A negative review with a good rating or a positive review with a bad rating. However, we can't remove all these reviews progmatically based on our sentiment scores because a) we can't be so sure if our scores correct and b) it would actually result in overfitting. So a manual removal is needed which would require a lot of time but for achieving better accuracy, it is a valid point.

References

1. https://www.forbes.com/sites/christopherelliott/2018/09/09/these-are-the-best-hotel-review-sites-in-the-world/#5dac30867fbd
2. https://doc.scrapy.org/en/latest/intro/tutorial.html
3. http://www.way2automation.com/selenium-webdriver-xpath-vs-css.php