Sentiment and Text Analysis of TripAdvisor Reviews

# Introduction

As a consumer it is sometimes difficult to navigate review websites to find meaningful reviews that help you determine if you want to make a purchase or not. Some reviews may be positive in nature but do not provide a lot of insight into why the reviewer had a positive experience.

By classifying reviews as positive or negative we can find features within those reviews that may help businesses drive new approaches to their products.

# Literature Review

In order to prepare for this project, I have reviewed some material regarding sentiment analysis and the caveats that go with opinion mining via review sites, forums and other online sources such as social media. The papers I mentioned below cover similar themes regarding some of the errors involved in drawing conclusions from human generated reviews. One paper in particular discusses the interesting point that sometimes review text does not match a review rating score (human nature to want to be more positive therefore score is more positive than textual review).

**Sentiment Analysis using Product Review Data by Xing Fang and Justin Zhan**:

http://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

This paper covers some of the caveats of opinion mining using online data. For example, sometimes paid reviewers will give “fake” reviews and skew the opinions or results of any type of analysis. Also because people can freely post reviews as they like you cannot always guarantee that they are meaningful.

**Sentiment Analysis of Reviews: Should we analyze writer intentions or reader perceptions? By Isa Maks and Piek Vossen**

http://www.aclweb.org/anthology/R13-1054

This paper discusses weaknesses in using star ratings in opinion mining. They point out discrepancies between the review ratings and the actual review text itself (using some examples where the text was actually more negative than the actual rating score itself). This paper also discusses the effect on a reader when they read the review text versus when they see the ratings.

**Sentiment Analysis for Hotel Reviews by Walter Kasper and Mihaela Vela**

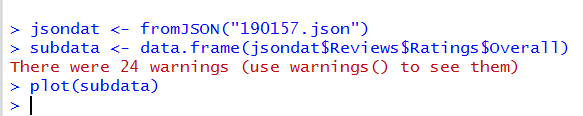
This paper discusses a point that is of particular interest to the data within this project. They cover the topic of considering ratings of a 3-star hotel and a 4-star hotel. Just because the higher end hotel received a lower rating on one attribute does not necessarily mean that the 3-star hotel is inherently better. The textual comments of the reviews would be of more interest to a hotel manager in that case to determine what the customer believed to be objectionable or positive

# Dataset

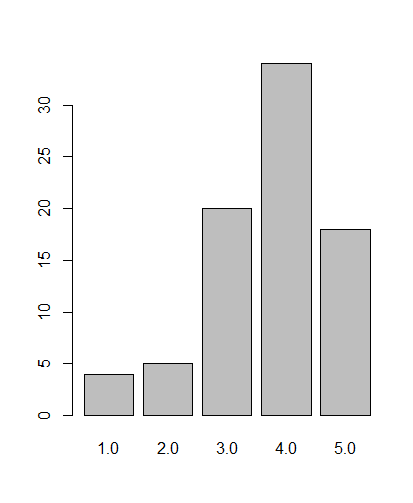
The dataset for this project will be the TripAdvisor reviews curated by Hongning Wang, Chi Wang, ChengXiang Zhai and Jiawei Han for their paper; Learning Online Discussion Structures by Conditional Random Fields. The 34th Annual International ACM SIGIR Conference (SIGIR&#39;2011), P435-444, 2011.

It contains multiple trip reviews grouped by hotel. The hotels are scored on Service, Cleanliness, Overall Value and Location. The review contains the reviewer id, review content and the date the review was posted. The hotels are also rated more in depth for other items that I have decided not to include in this project due to timing; Check-in Rating, Business Services Rating, Overall Room Rating, Sleep Quality of the Room Rating. I have also decided to take a subset of the original 12000 hotels because I would need more time to clean the dataset of that size (each hotel also has approximately 100 reviews each)

Loading 1 Hotel’s reviews into R just to take a preliminary look at the type of data available in the set;



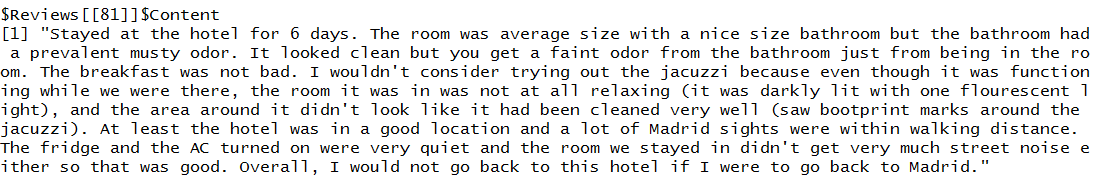
Showing preliminary look at how “Overall” ratings vary for 1 hotel – the “Overall” ranking is a value from 1-5 and the website does allow negative numbers (lowest being -1). I would like to use this rating as the basis for labelling data.



Below contains a list of the attributes found within the dataset;

|  |  |
| --- | --- |
| **Attribute** | **Data Type** |
| Service Rating | Numerical |
| Cleanliness Rating | Numerical |
| Overall Score | Numerical |
| Location Score | Numerical |
| Review Title | Text |
| Author | Text |
| Content | Text |
| Date | Date |
| HotelID | Numerical |
| HotelName | Text |

Below is a sample of the review content:



# Approach

Below is an outline of the steps I will take to complete this project.

**Step 1: Clean and Explore Data**

The dataset is quite large so I have extracted a subset of 2000 reviews for initial work. These reviews are from a cross section of 30 unique hotels from the original TripAdvisor dataset. The useful data was parsed from the original json format of the dataset. I extracted the HotelIDs, ReviewIDs, Review Content and relevant rating scores.

Cleaning steps taken to this point – The positive, negative and neutral text corpora were cleaned in R. Whitespaces were removed, stopwords were removed, the length of words used to determine term frequencies were limited to over 4 letters (to reduce words that are used frequently but that do not affect the analysis of the review significantly – ‘a’, ‘by’, ‘the’, ‘we’, etc.).

During the classification steps taken so far the text was reduced to a few sentences since many of the reviews are multi-paragraphed and processing time was too high. I will include information about the training data used so far in coming sections. Punctuation was also removed from the training corpora in order to ensure that the classifier could read the text (some errors occurring on hyphens and html characters that were not cleaned in the original data).

**Step 2: Label Data based on known scored sentiment dictionary**

2000 reviews from 30 hotels were used in the original labelling process. I have uploaded these into Hive tables. I joined the tokenized review text and a dictionary table where the words were the same. Then calculated a sum for each review based on the number of words that were included in the review from the dictionary that was used. (The dictionary was an openly available dataset of scored positive and negative sentiment words from the following sentiment analysis studies; Opinion Extraction and Summarization on the Web Minqing Hu and Bing Liu Department of Computer Science University of Illinois at Chicago 851 South Morgan Street, Chicago, IL 60607-7053 {mhu1, [liub}@cs.uic.edu](mailto:liub%7d@cs.uic.edu)

The scored dictionary included a -1 for negative sentiment words and a +1 for positive words. The hive analysis concluded positive, negative or neutral for each review in the training set based on the sum of 1s found in the review (in some cases a -1 and +1 would obviously cancel each other out therefore leading to a neutral sentiment overall).

This labelled data was then extracted to a text file and then moved to R for further analysis. The steps taken to move the data from hive to R will be included in the links to code further in this document.

**Step 3: Extract Review text for Analysis**

After getting the initial review text from the hive step above, I created Positive, Negative and Neutral review text corpora. From here I was able to find the frequent terms used in the positive, negative and neutral reviews to get a better sense of the dataset.

Analysis of the frequently occurring terms in the Positive, Negative and Neutral training data will be provided in the Results section of this document.

Word clouds of frequently used words in the TripAdvisor reviews will be included in the Results section of the document.

**Step 4: Build Classifier and Use Subset of Data to Train**

The following tutorial was used to model the classifier portion of this study;

<http://datascienceplus.com/sentiment-analysis-with-machine-learning-in-r/>

The initial training set for the positive review text was taken from 5 reviews that were labelled as positive from the dictionary comparison of positive words that were found in those reviews. And likewise for the training set for negative and neutral reviews – I also chose 5 reviews for each of those training sets as well. The training data was not sufficient enough to get an accuracy of higher than 50% at this time. Prior to the final due date of this project I would like to extend the work done so far and see if I can improve the accuracy of sentiment predictions.

# Initial Results

The dataset was imported into Hive initially. From there the data was labelled as positive, negative or neutral based on the polarity score calculated from a sum of words found in the reviews that are also included in a known dictionary.

Once we are able to classify a review as negative, positive or neutral we can then dig deeper into those reviews and extract words and topics that may be valid in determining points for improving the product or services that the review is covering. In this case, we are looking to find out what people are saying about Hotels on TripAdvisor. What is it about the hotel that makes the experience for them positive, negative or neutral?

Below is a word cloud showing frequently used words in the positive reviews that were used more than 10 times per document.



**Frequent Topics Covered Positive Reviews**

The following words were used more than 20 times in the positive reviews that were studied in this project. They cover multiple topics listed in the chart below;

|  |  |
| --- | --- |
| **Topic** | **Frequent Words In Positive Reviews** |
| *Business Services Available* | Wi-fi, Internet |
| *Food and Drink* | Breakfast, Buffet, Champagne, Coffee, Dinner, Fruit, Meal, Menu, Restaurant |
| *Amenities* | Fitness, Attractions, Atmosphere, Bars, Areas, Location, Beach, Facilities, Market |
| *Room* | View, Bathroom, Bathrooms, Toilet, Toiletries, Sink, Sleep, Pillows, Shower, Bed, Balcony, Clean, Furniture, Décor, Sheets, Phone |
| *Staff* | Doormen, Housekeeping |
| *Costs* | Cheap, Cheaper, Price, euro, euros |

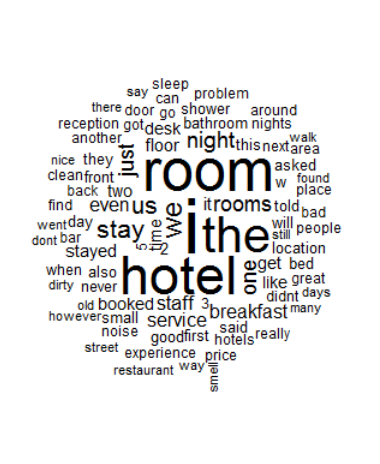
Below is a word cloud showing frequently used words in the negative reviews that were used more than 10 times per document.



**Frequent Topics Covered Negative Reviews**

|  |  |
| --- | --- |
| **Topic** | **Frequent Words In Negative Reviews** |
| *Business Services Available* | Internet |
| *Food and Drink* | Food, Breakfast |
| *Amenities* | Lobby, Reception, Parking, Pool |
| *Room* | Bathroom, Desk, Dirty, Smell, Floor, Door, Noise, Experience, Booking, Shower, Toilet, Towels, View |
| *Staff* | Manager, Rude, Service, Valet, Staff |
| *Costs* | Expensive, Charge, Price |

Below is a word cloud showing frequently used words in the neutral reviews that were used more than 10 times per document.



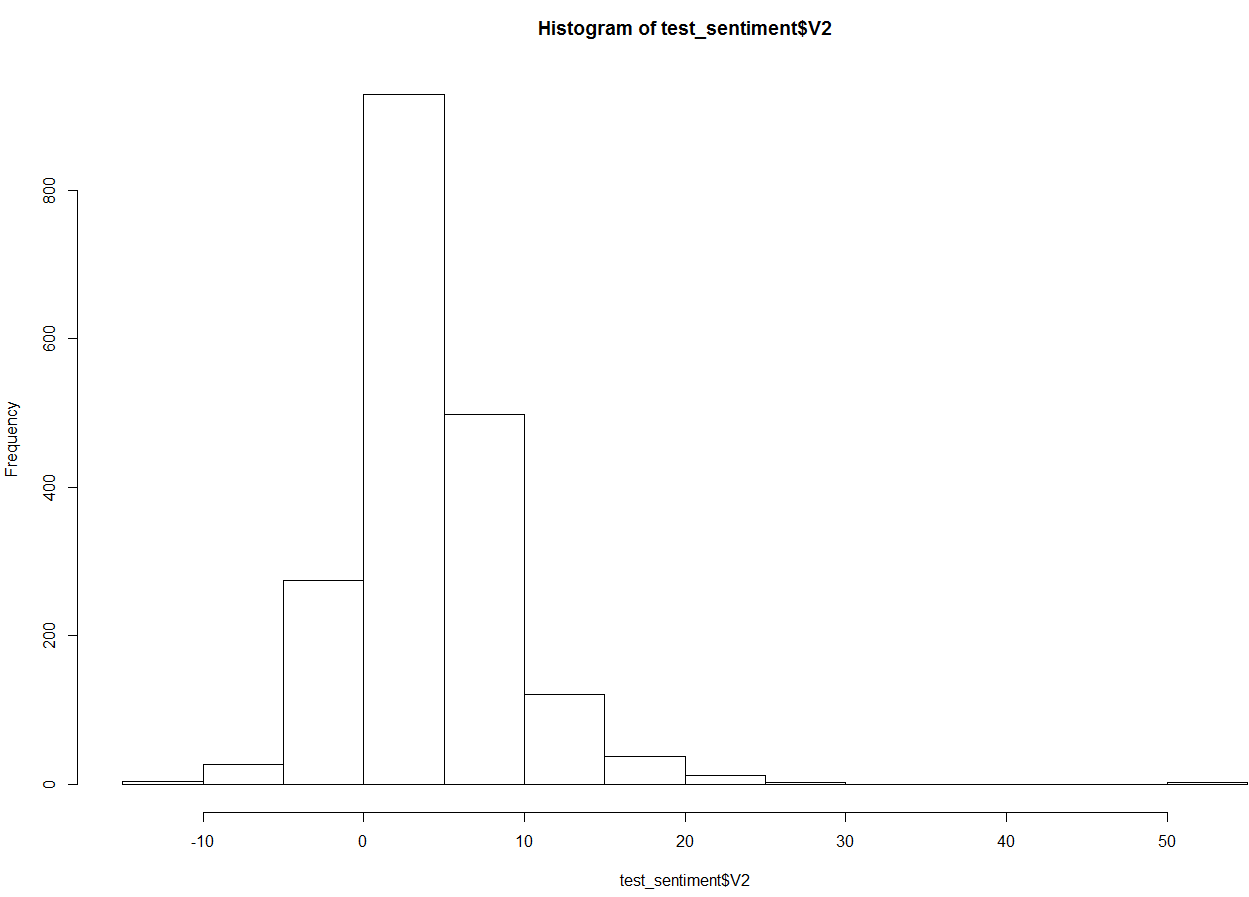
**Frequent Topics Covered Neutral Reviews**

|  |  |
| --- | --- |
| **Topic** | **Frequent Words In Neutral Reviews** |
| *Business Services Available* | Internet |
| *Food and Drink* | Breakfast, Restaurant |
| *Amenities* | Pool, Lobby, Parking |
| *Room* | View, Access, Shower, Toilet, Cold, Experience, Beds, Desk, Noise, Location, Small |
| *Staff* | Friendly, Staff, Service, Poor |

We can determine that similar topics are covered in all reviews even if the content is regarded as having a different sentiment. It would seem that the neutral reviews saw both pros and cons to their stay or maybe overall had a good experience but ran into a problem while they were there. The topics discussed above outline in what business areas those issues might’ve happened in (lack of service, poor quality food, few amenities or issues with their room).

From all three test sets we can see that there are key factors that consumers look at when deciding to book a hotel stay; Food and nearby bar options, Amenities, Quality of the Hotel Room, the quality of the staff and the availability of business services such as internet access.

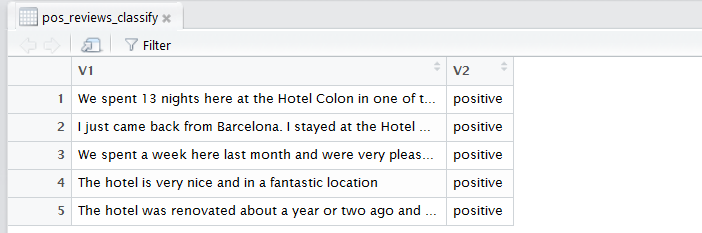
**A distribution of the Sentiment Scores for the test sample**;

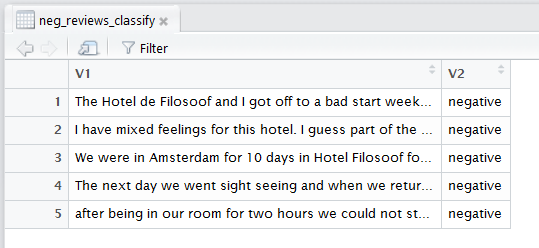


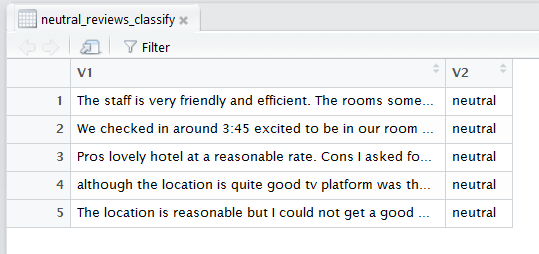
Based on the above you can see that the dictionary labelling step found on average 10 words of polarity within the review itself. Very few reviews in the test set were at either extreme of polarity/sentiment score. Although there were a few negative reviews scored well below 0 and several positive scores that were well above 20 – those seemed to be outliers or perhaps indicate that there just wasn’t enough data used to test with at this point.

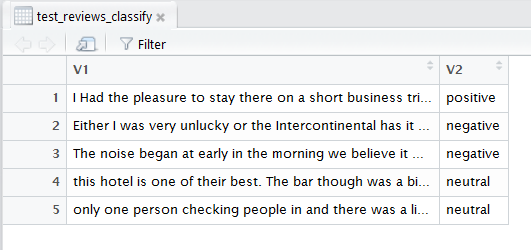
**Classification Results**

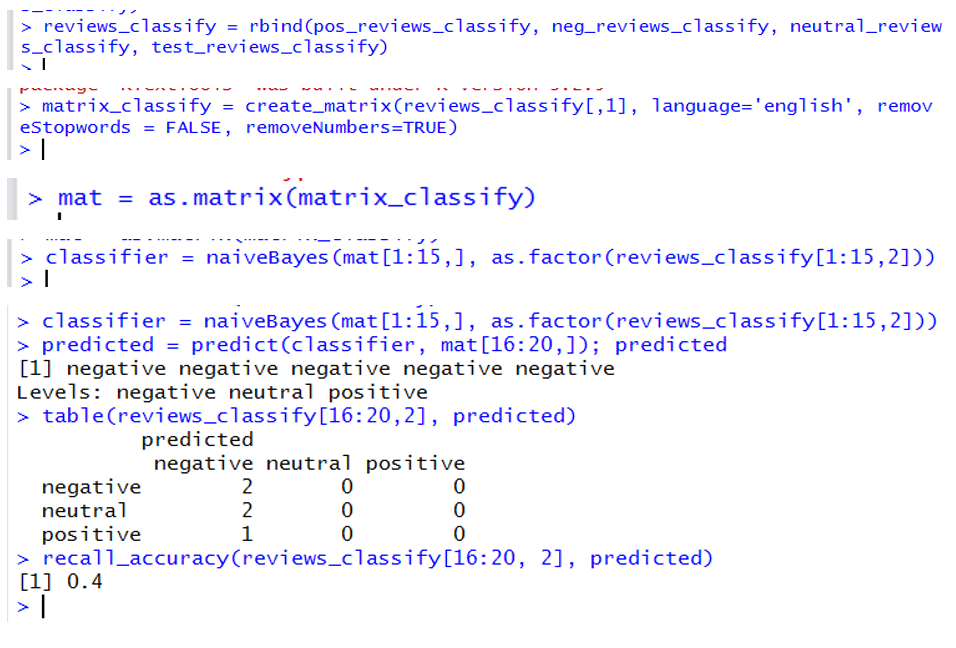
The following test data was used in building a classifier in R and using Naïve Bayes to classify test reviews. In order to run these tests in R I removed sentences from the Hive classified review set to make a more manageable dataset to work with.











As you can see above the recall accuracy is quite low due to the small data set that was used to test the classifier at this point. There is opportunity to improve this accuracy score by using more in-depth training.

**Next Steps/Areas for Improvement**

Continuing to train the classification model further with additional reviews should improve the recall accuracy score achieved so far. However, based on some of the text analysis there is already a lot of information to go on to choose areas to explore further – such as topic modelling for example. Stretching this further into modelling topics within these reviews would help focus business intelligence even more, as we would then be able to extract topical areas for business improvement or we could see in which areas we already excel in as a Hotel Manager or Hotel Corporation.

The term frequency analysis already shows some areas where reviewers have pointed out concerns and where they believe certain hotels are already excelling.

Another possible area for improvement within this project may be to find clusters of hotels where there are mostly negative reviews or mostly positive reviews and determine if there are any location correlations.