

Beyond the Star: European Hotel Review Sentiment & Clustering Analysis>

By: Ritchie Paul

Student Number: 500584056

Supervisor: Tamer Abdou

Date of Submission: 05/16/2022

Table of Contents

[Abstract 3](#_Toc452457337)

**Refined Research Question ………………………………………............................…….........3**

**Literature Review ………………………………………………………….…….........………4-7**

**EDA (Exploratory Data Analysis)…………………………………….……............................7-8**

**Pre-Processing and Feature Engineering ………………………………….……...................8-9**

**Model Building and Conclusion ………………………………….……..........,....................9-11**

**Revised Abstract**

With the world getting more digital post-pandemic and the travel industry picking up, understanding your customer’s digital footprint is important in the hotel industry. Reviews play a big part in building trust towards new customers and building brand loyalty among old ones. Beyond a rating system, understanding  the sentiment around good and bad reviews can lead to valuable insights. It can be used in such areas like customer experience, reducing customer churn, upselling opportunities etc. Also, understanding similar characteristics between good and bad hotels can give us insights into hotel segmentation.

The dataset that I have chosen to work on would be the “515K Hotel Reviews Data in Europe” dataset, which can be found here: <https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe?datasetId=2142&language=Python>.

Liu, J. I. A. S. H. E. N. (n.d.). *515K Hotel Reviews Data in Europe*. Kaggle.com. Retrieved May 16, 2022, from https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe?datasetId=2142&language=Python

The theme of my project will be around sentiment analysis and clustering. The following general techniques will be used: Exploratory Data Analysis (EDA), Text Pre-processing, Feature Engineering, Modeling (e.g. Naive Bayes, K-Means, Decision-Tree, TF-IDF) etc.

This project would be able answer the following research questions: Can we predict that a good/bad review is given through the text of the review? To what degree is that the case?**Are certain words typical of either bad or good reviews?** Do good or bad hotels tend to be near each other?

I will be using statistical technique to tokenize the words and semantic analysis as well.

The tools used in this project are: Microsoft Azure Platform, Python and Jupyter Notebook. In terms of packages: Pandas/Matplotlib/Seaborn for EDA, NLTK for the natural language processing tasks and Scikit-Learn for the machine learning tasks.

## Refined Research Question

**What words in the text of a hotel review in Europe indicate whether is it good or bad?** Also, do good and bad hotels tend to congregate with each other?

## Literature Review

Hwang, San-Yih; Lai, Chia-Yu; Jiang, Jia-Jhe; and Chang, Shanlin (2014) "The Identification of Noteworthy Hotel Reviews for Hotel Management," Pacific Asia Journal of the Association for Information Systems: Vol. 6: Iss. 4, Article 1.  
DOI: 10.17705/1pais.06402  
Available at: <https://aisel.aisnet.org/pajais/vol6/iss4/1>

The author of this study wanted to investigate a way to automatically identify customer reviews that would be helpful to hotel management. That could be done for either negative or positive reviews. Most other studies focused on customers while this one focus on help hotel staff. The data used was English reviews from TripAdvisor for the top 10 hotels in each of these cities: Taiwan, Taipei, Kaohsiung, Taichung, New Taipei City, Hualien, Nantou and Illan. The data was collected from 2000 to May 2013. The author uses a two-phased approach to this study. The first phase was to interview two hotel managers to understand the characteristics of worthwhile reviews. The second phase will be building a classification model to identify these newsworthy reviews. After the interview with the hotel managers, the author realized the characteristics of relevant reviews were divided into 3 categories: content, sentiment and quality features. The author figured out that that content features have a much larger impact on precision where we see that sentiment and quality features effect recall more. Three approaches were used to derive the content features: TF-IDF method, semantic-based LDA method and word-based LDA method. The semantic-based LDA method does the best with a higher recall and less features used.

*Impact on my project*

This study gives me a lot of good ideas in terms of what methods I should explore or look into for my own research. This article has allowed to me gain a perspective in methods used and to think more critically about my dataset.

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (01/2016). *Understanding satisfied and dissatisfied hotel customers: Text mining of online hotel reviews* Haworth Press. doi:10.1080/19368623.2015.98363

#### The author of this article wants to examine and understand the common patterns that lead people to leave positive/negative reviews. Text mining is the main method that the author uses to achieve this. The data was collected from TripAdvisor from hotels located at Sarasota, Florida up and until the end of 2013. The text mining approach used was using the “PASW Modeler”. After that a text-link analysis was done to logically link patters or find the connection between concepts. This led to understanding what factors lead to negative or positive reviews. The main categories where we in both positive and negative reviews are around “room for improvement” and customer services. This allows that both tangible and intangible factors are at play when being reviewed online for your hotel.

*Impact on my project*

This study gave me a lot of contextual information that I have previously lacked. Understanding what parts of an hotel experience defines your stay and ultimately your rating is a great logical foundation to leap off. Understanding text mining in this industry-specific application helps increase my domain knowledge in the project.

Hu, Y., Chen, Y., & Chou, H. (03/2017). *Opinion mining from online hotel reviews – A text summarization approach* Pergamon Press. doi:10.1016/j.ipm.2016.12.002

This is a novel approach to find out the most informative sentences or information from each review. This is through a text summarization approach, as noted above. The data comes from the Red Roof Inn and Gansevoory Meatpacking Hotel from January 1, 2012 to March 31,2013 on TripAdvisor.com. A lot of pre-processing was done to filter for relevant information such as: stop-word removal, sentence selection and part-of-speech tagging. After that process, a sentence ranking was completed through these factors: the reputations of the review author, the use of the review, recency of the review and the specific language in the sentences. To find the to-n most importance sentences in the study, the n-medoids clustering algorithm was used split sentences into n groups. The medoids from these n-groups were used as the concluding outcome. Other methods were also used to compare to the above proposed method. 20 participants were also picked to contrast the performance between all of the methods used. The unique approach from the author empirically provided more of an exhaustive summary on both good and bad reviews.

*Impact on my project*

#### This is a novel technique that I have found during my literature review. Traditionally, looking through a semantic or statistical lens has been the go-to for any kind of text-processing. Assigning values to things outside of the body of the text and understanding the nuance from a social media perspective has provided insights in how to analyze my data.

Phillips, P., Barnes, S., Zigan, K., & Schegg, R. (2017). Understanding the Impact of Online Reviews on Hotel Performance: An Empirical Analysis. *Journal of Travel Research*, *56*(2), 235–249. https://doi.org/[10.1177/0047287516636481](https://doi-org.ezproxy.lib.ryerson.ca/10.1177/0047287516636481)

This study is taking an empirical approach to understanding if there is a relationship between online reviews for a hotel and the actual performance for a hotel. The data is coming from TrustYou, a private German company. It collected web-based reviews from the whole Swiss hotel industry. The total number of hotels used were 442 that were operating in Switzerland in 2010. The hotel variables that were picked from this study was up to the authors discretion. They are across 3 unique, reasonable area of a hotel. That would be the actual building characters, the quality of the cuisine and customer service. The analysis was done through “PLS-PM in the XLSTAT software package ([XLSTAT 2015](https://journals-sagepub-com.ezproxy.lib.ryerson.ca/doi/10.1177/0047287516636481))”. Generally, the results show that a higher proportion of good reviews associated with good experiences at a hotel have the largest impact on hotel demand and income.

*Impact on my project*

#### While this article does not have a direct impact on my project, it proves a very important point of my project being useful. If there was no correlation between good reviews and a hotel’s bottom line, then my analysis wouldn’t further any discussions or have value to society.

Hu, F. (07/2020). *What makes a hotel review helpful? an information requirement perspective* Haworth Press. doi:10.1080/19368623.2019.1661931

This study seeks to understand what makes a review useful, the attributes that contribute to that and if those attributes differ for different hotels. The data used was from TripAdvisor reviews for Chinese hotels in Beijing and Shanghai between January 1st, 2011 and December 31st, 2015. A total of 483 hotels and 85,963 reviews were used. The data was divided into “helpful” and “non-helpful” reviews. Text-processing was done and 28 attributes related to hotels were picked.

One key finding that was made is showing that the tourism industry isn’t the same at all different classes of hotel. Generally, a conclusion was made higher class travellers tend to value more luxurious, detailed service while lower-class travellers value a good deal and accessibility/ease-of-travel.

*Impact on my project*

The key takeaway from this study is that the travel industry is not just a group of similar people. They all have their own wants and needs at every level. With that knowledge, dividing or separating hotel reviews into their own star category might be a possible approach.

Melián-González, S., Bulchand-Gidumal, J., & González López-Valcárcel, B. (08/2013). *Online customer reviews of hotels as participation increases, better evaluation is obtained* Sage Pub. doi:10.1177/1938965513481498

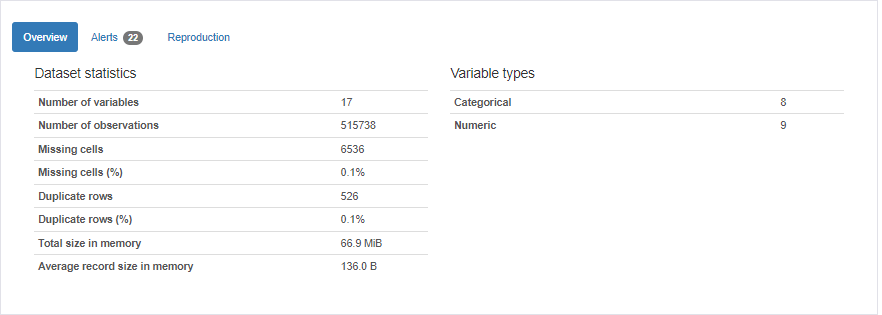
This study seeks to prove that as a hotel gets more reviews, the more positive the score will be until it levels off into their true score. In May 2010, total of 26,439 hotel reviews were used from 200 destinations across Europe. More traditional statistics were used to do the analysis such as regression analysis and using t-tests. We see that and confirm that as the number of a hotel increases, the ratings tend to get better. Also, we see that the negative review tends to be front-loaded at the beginning and peter out as more reviews are added in.

*Impact on my project*

The takeaway for here from me is that my large data set of 550,000 review is useful and not just inflated for no reason. This also confirms my hesitance on using smaller datasets as we see that it is not representative of the true level of the hotel.

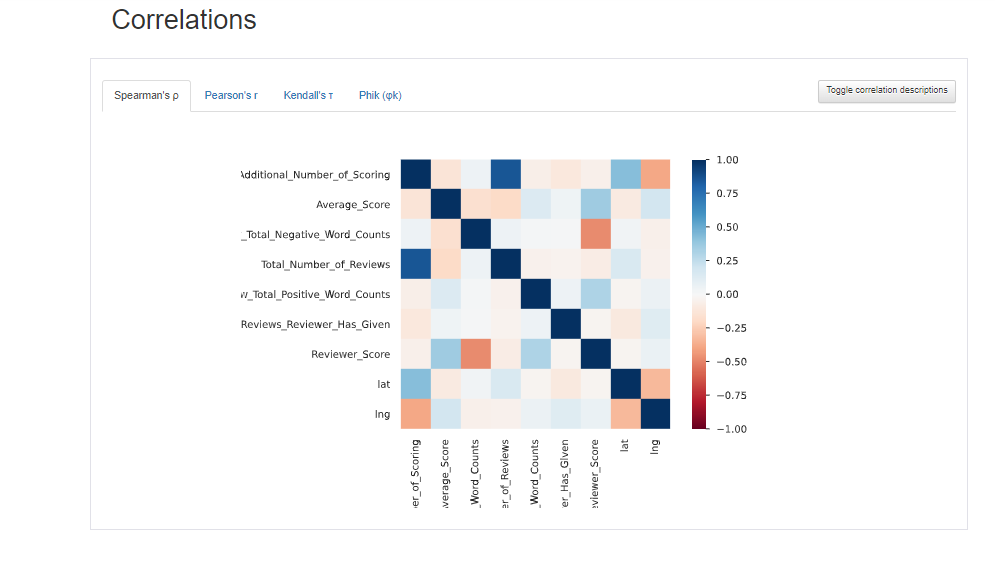
**Exploratory Data Analysis**

In the data-set, there are exactly 515, 738 observations provided with 17 variables. 8 are categorical while 9 are numerical. The *Reviewer\_Score* variable has a minimum rating of 2.5 to a maximum rating of 10, That is across both positive and negative reviews, which shows us that there are no reviews below 2.5, which could be a limitation of data collection. The data-set has 526 duplicate rows (which is approximately 0.1% of the data). Also, we see that there are 6536 observations that have missing cells.



Review\_Total\_Positive\_Word\_ Counts and Review\_Total\_Negative\_Word\_ Counts have a high amount of zeros, 35,946 (7.0%) and 127,890 (24.8%) respectively. This could throw off our analysis but it won’t because the zeroes are intentional. When a negative review is given, the Review\_Total\_Positive\_Word\_ Counts is blank and vice-versa.

According to the below, we see that Additional\_Number\_of\_Scoring & Total\_Number\_of\_Reviews are highly correlated with each other. They should be excluded from the analysis as they won’t have a significant impact.



**Pre-Processing and Feature Engineering**

**Removing Unwanted Symbols**

Unwanted symbols are removed from the text because it would interfere with the analysis doen further in the paper. These symbols might convey some meaning in the text but it won’t be helpful in our analysis. Those symbols are completely removed from the text and not replaced with anything else.

**Removing Unwanted Symbols**

Stop words are common words such as ‘a’, ‘and’, ‘but’, and etc. In NLP, these are words which are ignored/removed. I will be using the English stopwords from the NLTK (Natural Language Toolkit) to remove the stop words.

**Feature Engineering**

Text-mining techniques will be applied to figure out what word/categories are most cited in the reviews. Therefore, only the *Positive Reviews* and *Negative Review* variables will be used.

**Model Building and Conclusion**

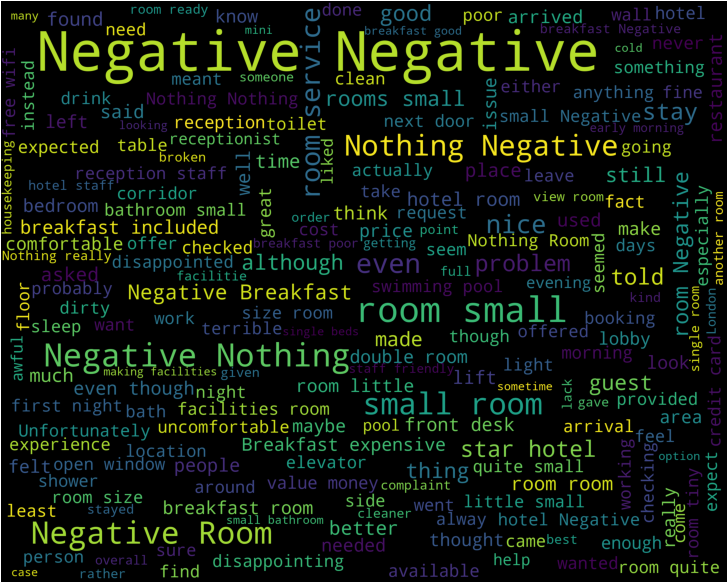
After preparation in the previous sections are done, a word cloud will be built and analyzed to understand what categories are impacting both types of reviews.

**Positive Reviews**



Two categories stick out when analyzing this world cloud: location and service. The largest words simply refer to having a “great” or “good” location, which gives us a general sense of the word but we can dig deeper. Other words that pop-up are: “minute walk”, “tube station”, “train station”, “location near” , “location friendly” etc. This implies that amenities that are close to the hotel and ease of transportation play a big part in the stay of someone at their hotel. Similarly, we see a large positive sentiment around the staff with words like “friendly” and “helpful”. That should come as no surprise as service is a large part of the hospitality industry. To observe some of the more specific reasons, we see that “good breakfast”, “hotel clean”, “clean comfortable” indicate what people look for in terms of service. A comfortable, clean room that is well taken care of and having the option for a complementary breakfast seem to be on the top of the list of what customers want.

**Negative Review**



For the negative review, an emphasis on the service, cost/price and the room/physical amenities itself can be seen in the word cloud. Many references too the room being small or inadequate are prominently on display (e.g. “room small”small room”,”room little”,”bathroom small”). This suggests that the expectation of what the room is or looks like is not being well represented by the hotel. On the service side, similar to the positive reviews mentions of breakfasts are being made but not in a positive light (e.g. “Negative Breakfast”, “breakfast expensive” etc.). Also, many mentions of the staff are made through the word cloud (e.g. “receptionist”,”reception”,”hotel staff”,”housekeeping”). No specifics can be gleaned but it as these are negative reviews, I would assume that they would not be praising them. Words surrounding the concept of cost/ price are peppered throughout the negative review (e.g. “value”,”money”,”price point”,”cost”). Nothing definitive can be made out of that but people that leave negative reviews are more price-conscious.

**Conclusion**

The main purpose of this project narrowed down to understand what words make up a good review vs. a bad review. This was to better understand what categories/factors contribute and ultimately have an impact on a guest’s experience at a hotel. Whether that is a negative or positive experience. The common category between both types of reviews are related to services.That shows that the staff, in how to interact with the clients and keeping the hotel running at tip-top shape is imperative to the positive or negative experience of a guest. For positive review, location is the other standout category for a great experience. Hotel guests tend to want to be near the city centre, public amenities and just accessibility to as many things as possible. This is something that guests and hotels have to pay attention to ensure a good experience. For negative reviews, cost/price seemed to be the other category that really stood out. An accurate representation of what kind of hotel room the guest is receiving is imperative for the hotel staff to communicate. This gap of value for the guest had lead to negative experience and is something that hotels need to work on. Overall, these categories can help hotels understand what they should emphasis and work on to create the best experience for their guests.