Project Increment 1

**Topic**

Hotel rating based on customer review and classification & sentimental analysis of review. The implementation will allow hotel owners/managers decide what areas they can improve or adjust to their environment, and leverage the data for insights into a 360 view of consumers and their sentiment.

**Team Members**

Mdakbar Sarkar

Robert Fajardo

Rajesh Pahari

Platini Dacheu

**GitHub Link**

[**https://github.com/RobFaj/NLP\_Project2022**](https://github.com/RobFaj/NLP_Project2022)

**Video Link**

[**https://www.youtube.com/watch?v=PcKUa6bOjk0**](https://www.youtube.com/watch?v=PcKUa6bOjk0)

**Motivation**

The goal is to find a good hotel and having an enjoyable vacation is everyone’s very basic expectation which will not be feasible without having customer feedback or review on the hotel. As we are living in a world where data plays an important role in the decision-making process for both customers and service/accommodation providers. We need a developed system based on customer reviews to make an unbiased rating of the hotel that not only helps the customer to choose the best one but also opens the opportunity to improve for the service provider. We need to have an effective system in place, we need to do the classification of review as well as sentimental analysis.

**Objectives**

The main objective of this project is to build a data centric application that not only helps the customer but also identifies key areas to improve from the provider’s perspective. After completion of the project, we will have a solid understanding of NLP and how we use the concept in the practical field.

**Significance**

You can’t manage what you can’t measure. With the advancement of technology and the rise of the internet, now the world becomes a global village. The recent explosion of digital data opens the door to a new way of thinking, working, and living. Information is the key but managing information is important as making a critical business decision. For example, if anyone wants to travel to a new place and doesn’t have any knowledge about the vicinity, then the plan for vacation may not materialize. In this case, customer reviews will play a big role to make the marketing strategy successful. At the same time, if we have millions of reviews without labeling and sentimental analysis, no one will go for the review of millions of customers. So, managing customer reviews in an efficient manner are necessary to promote, improve, and flourish unseen opportunities into fruition.

**Features**

We will use the following NLP features: tagging, summarization, and sentiment analysis. Within hotel ratings, we mainly focus on the sentimental analysis of customers - positive, negative, or neutral feedback which can be measured on a scale of 1 to 5 where 1 is low and 5 is excellent.

We will tag and classify the review with labels. We will use features such as cost, crime history, amenities, flexibility, parking, indoor facility, room space, air-conditioning, and neighborhood to help classify the hotels.

On hotel recommendation, we will use labeled classified data based on sentimental analysis.

**Diagram

Description automatically generated**

**Background**

Research or project topics are usually not formulated for the first time. Thus, the often build on previous works either in the same domain or some close domain. They can even be inspired by a totally different branch of science. Our topic does not escape this generality. To understand the state of the art of the problem, we have researched previous works accomplished in the area. This helped us more precisely fix the boundaries of our topics. Therefore, areas like sentiment analysis, product rating and review have been deeply explored to help us accomplish our task.

Sentiment analysis is a subfield of NLP that draws on approaches from information retrieval and computational linguistics to identify opinions expressed in text. It is considered a specific type of text mining (Han et al., 2011), and it has been called opinion mining. The main goal of sentiment analysis is to identify positive or negative overall attitudes or opinions toward a brand, product or service based on text comments (Liu, 2010). While the terms appraisal extraction or review mining have also been applied, they are not always completely accurate (Pang and Lee, 2008). Several machine learning and data mining algorithms have been used to detect sentiment (Khoo et al., 2012), and sentiment strength (Thelwall et al., 2010). Pang and Lee (2005) predicted star ratings of movie reviews based on a five-point sentiment scale instead of merely classifying the reviews as positive or negative. They employed a novel similarity measure with a meta-algorithm based on metric labeling and performed. several comparisons of pairs of reviews to identify when the first review was less positive than, more positive than or as positive as the second review. Even simple algorithms have been shown to work well with large data sets, as in the case of the naïve Bayesian approach (Wu and Kumar, 2008). Rutilo et al (2015) rated the hotel by the transformation of the positive percentage of its comments.

Our current project aims to utilize rules based approach in order to generate the hotel rating based on customer review and classification & sentimental analysis of review. Rule based approach is used by defining various rules for getting the opinion, created by tokenizing each sentence in every document and then testing each token, or word, for its presence.

**Dataset**

The dataset is titled “Trip Advisor Hotel Reviews” based on hotels from online reviews of hotels around the world and collected in the TripAdvisor.com database and queried in 2018 as part of research for a paper (Bansal, 2018). The dataset contains two columns “Review” and “Rating” and has 20,491 unique rows of text and ratings scaled 1-5. As seen in Figure 1, we have a visual of the descriptive analysis which provides the ratings have a mean of 3.9 on the scales where we have a minimum of 1 to a maximum of 5 in our ratings data.

**Fig 1:** Description of ratings and row counts

Table

Description automatically generated

Next, we remove stop words, emojis and special characters and lemmaize the text. run the dataframe column “Review” through the class “Clean” and remove emojis and special characters (Figure 3). Then, the program tokenizes the text and splits the text and counts the number of tokens in the “Review” column.

**Fig 2:** Clean TextGraphical user interface, text, application

Description automatically generated

After, cleaning the text we see that our maximum token length is 2,052 tokens in row 7072, and we have a minimum token count of 7 in row 1501.

**Fig 3:** Token count

Graphical user interface, text, application, email

Description automatically generated

Next, we perform a word count on the “Top 20 words in Hotel Review” and display it in the chart below. This knowledge will help us gauge review quality a point of reference in terms of validating the applications of this data and assessing the value of the reviews.

**Fig 4: Top 20 Words**

Chart

Description automatically generated

In Figure 5, we get a mean count of 109 tokens for the dataset. We need to make sure that we are gathering enough tokens to derive a business use case for understanding our consumer market and their sentiment towards hotels in general. It is crucial to drive our sentiment analysis, topic modeling, and text classification model with significant token counts per row. In figure 5, it also shows that 50% of the token counts are at or near 80 tokens which is close to our average token count.

**Fig. 5:** Token length statistics

**Text

Description automatically generated**

**2. Exploratory Data Analysis**

All data science projects have this one fuel in common: data. For our project, we are using the TripAdvisor Hotel dataset. Every data-based work starts by understanding the data at hand. We have therefore thoroughly “searched” our data to get the more sense of it. To that end, we have massively used visualization as the mean to uncover our data. Suitably chosen charts have made that gives us big to detailed pictures of the data we will be using and will guide us through the next steps of our work.

We gathered the count of all ratings in the dataset. The largest Rating count by far is a Rating of “5” at over 9,000 rows. We have approximately 1,200 ratings of “1”, 1,800 ratings of “2”, and approximately 6,000 ratings of “4”. Generally, the population seems to share positive experiences.

**Fig 6**: Histogram

Chart

Description automatically generated

As a whole we decided to capture the most frequent words through out all of the ratings. These words such as room, crime, police, pool, breakfast, and drink all seem to fit our project features. These features included cost, crime history, amenities, flexibility, parking, indoor facility, room space, air-conditioning, and neighborhood.

**Fig 7:** Word Cloud size based on frequency

Text, application, chat or text message

Description automatically generated

The word frequency within each rating is computed and segregating Ratings from 1 to 5. In the lowest rating in figure 8, we see the top three tokens are police, crime, and stay. In ratings at 2, we see a mixture of police, crime, and good, which could indicate a shift in concerns. Ratings of 3 place good in the top category based on counts, the complimentary positive words have more counts. In the highest rating of 5, crime is nearly at the bottom and surpassed by staff and great. However, we do see that drugs are also in the count. This will require a deeper understanding of the content surrounding the word drug and the sentiment within the review.

**Fig. 8**: Ratings 1 to 5 (starting from top left to bottom )

Chart, bar chart

Description automatically generated

Next, we gathered the most common Bigrams and split them by rating class from 1 to 5.

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Table

Description automatically generated

These bigrams and unigrams will be the root of our topic tagging as we decide which bigrams will be most valuable to tagging our reviews.

The primary concern here is confirming our analysis is complete by checking the quality and cleanliness of our NLP model. Even after text cleaning there are some tokens that managed to escape the regex filtering and stemming process. We will fine tune our text cleaning, and perform an additional cleaning step involving token removal less than two characters, and decide which NTLK method is best for removing the least valuable tokens.

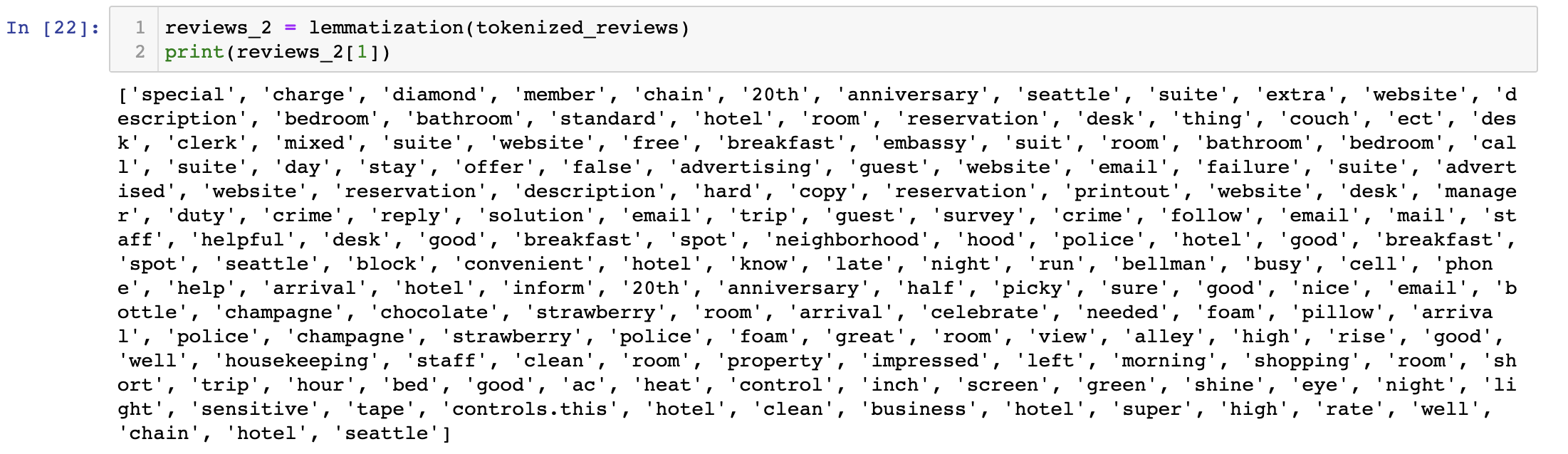
**3. Feature Extraction**

Successful machine learning model rely on data. Data need to be in good quantity as well as quality. Unfortunately, raw data is almost never appropriate to train a model. In order to make it suitable for model training, data needs to go through a series of steps called globally data processing. In this process, data is cleaned and prepare for model building. One important thing in these steps is the extraction of features. Data are defined by a lot of features or attributes, but not all of them are important to build a specific model. Each model has its requirements and specifications. When it comes to text data as the one we are using, there is a huge amount of words in a language vocabulary. However, not all of them are relevant in the transmission of information. The idea here is to get rid of unnecessary words and keep only those that are meaningful in transmitting information. That is what this section have been focusing on.

**Topic Modeling**

After cleaning and exploring our data we can begin extracting our features to achieve our goal to be build a better model for hotel managers and owners. After loading and cleaning the

In this token sample we are able to tokenize the reviews into valuable text for modeling.



**Text Classification**

In this area, were are starting to lemmatize the reviews and capture text sequences.

Text

Description automatically generated

**Sentiment Analysis**

We classify reviews and their sentiment into the buckets “neautral”, “positive”, and “negative”. Here are our preliminary results for all 3 classes and their counts in our dataset.

Chart

Description automatically generated

We can see that we have a large portion of positive reviews. Customers may be more enticed to leave positive reviews. First, we must analyze the text to ensure we are getting quality responses.

Here are character counts which we can **A picture containing text

Description automatically generated**use to help ensure that our review data are generating a balanced insight into consumer experience.

**A picture containing table

Description automatically generated**

Graphical user interface

Description automatically generated with low confidence

Next we have the word count per review.

Chart

Description automatically generated with medium confidence

Based on these histograms above on “Character Counts” and “Word Counts” per sentiment class, it is safe to assume that reviewers are leaving enough data within all 3 classifications.

In the final portion of sentiment so far we have classified Unigrams, Bi-Grams and Tri-Grams.

**Fig :** Most Common Unigrams

Chart, bar chart

Description automatically generated

**Fig. :** Most Common Bigrams

Chart, bar chart, histogram

Description automatically generated

A picture containing table

Description automatically generatedChart, bar chart

Description automatically generated

**Fig. :** Most Common Trigrams

Chart, bar chart

Description automatically generated  
Chart, bar chart

Description automatically generatedA picture containing table

Description automatically generated

We are on a good track to complete our ngram classification and begin modeling our review classification system. We plan to deploy the model within the next two weeks, and begin sharing with our clients soon.

**References**

1. Sentimental Analysis: <https://www.kaggle.com/code/emirkocak/in-depth-series-sentiment-analysis-w-transformers>
2. Hotel Rating: <https://www.kaggle.com/code/nidhaypancholi/predicting-rating-from-reviews-trip-advisor>
3. Sentiment Analysis in Hotel Reviews: <https://www.altexsoft.com/blog/sentiment-analysis-hotel-reviews/>
4. Han, J., Kamber, M. and Pei, J. (2011), Data Mining: Concepts and Techniques, Morgan Kaufmann, Waltham, MA.
5. Pang, B. and Lee, L. (2008), “Opinion mining and sentiment analysis”, Foundations and Trends in Information Retrieval, Vol. 2 Nos 1-2, pp. 1-135.
6. Liu, B. (2010), “Sentiment analysis and subjectivity”, in Indurkhya, N., Damerau, F.J. (Eds), Handbook of Natural Language Processing, Vol. 2, Taylor and Francis Group, Boca Raton, FL, pp. 627-666
7. Khoo, C.S.G., Nourbakhsh, A. and Na, J.C. (2012), “Sentiment analysis of online news text: a case study of appraisal theory”, Online Information Review, Vol. 36 No. 6, pp. 858-878
8. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010), “Sentiment strength detection in short informal text”, Journal of the American Society for Information Science and Technology, Vol. 61 No. 12, pp. 2544-2558.
9. Wu, X. and Kumar, V. (2008), “Top 10 algorithms in data mining”, Knowledge and Information Systems, Vol. 14 No. 1, pp. 1-37.
10. Pang, B. and Lee, L. (2005), “Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales”, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 115-124.
11. Bansal, B. (2018). *TripAdvisor Hotel Review Dataset* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.1219899>