# SENTIMENTAL ANALYSIS OF TOURIST ACCOMMODATION REVIEWS BY


## DABERECHI CORNELIUS AHANONU

## (UNIVERSITY OF SALFORD)


### Introduction

Sentiment analysis allows persons, organizations, and institutions to know what people think about them or their product. With sentiment analysis, people's opinion can be mined, and the result will be applied for predictive or proactive analysis. For example, sentiment analysis was used to predict the potential winner of the 2016 US election. Therefore, understanding customers feedback/sentiments is necessary to improve customer experience. This is because customer experience has a direct impact on the success of a company. However, it is difficult to gauge customer's feeling and needs. Therefore, the use of natural language processing (NLP) sentiment analysis tool bridges the gap by helping service providers appreciate end users' perspectives and better cater to their needs. Some words might be ambiguous depending on the tone and the context it was used for. But sentiment analysis applies NLP to determine the emotion behind those communication. Positive or negative sentiments can be identified appropriately by the customer service department, with the aim of reinforcing positive sentiments, or to empathically resolve negative ones.

### Datasets

We used a tourism accommodation dataset for this analysis. The dataset contains information about people's opinion towards different hotels/restaurants. The data was collected at different dates and location. The data are analyzed using lexicon-based approach to determine the sentiments of customers. We determined the polarity for the collected texts to help understand customer's opinion for a particular restaurant/hotel. Furthermore, a comparison is made among the restaurants/hotels over the type of sentiment. Also, a word cloud is plotted representing most frequently appearing words in the texts.

### Explanation and preparation of datasets

The dataset contains texts that represents customers opinion which is also the reviews. We also have a column for the different restaurants/hotels. The reviews were given at different dates. The reviews represent customer's feelings about the restaurant/hotel. This is usually their opinion especially with regards to the service they received in each restaurant/hotel.

The dataset is a text, therefore the following steps before was taken to transform the data:

1. **Lowercase the data**

The idea is to convert the input text into the same casing format so that it converts 'DATA', 'Data', 'DaTa', 'DATa' into 'data'. Converting all your data to lowercase helps in the process of preprocessing and in later stages in the NLP application, when you are doing parsing.

2. **Removing Punctuations**

The second most common text processing technique is removing punctuations from the textual data. The punctuation removal process will help to treat each text equally. For example, the word data and data! are treated equally after the process of removal of punctuations.

3. **Stop words Removal**

We also removed stop words from the text. Examples of stop words include so, yet, before etc. These are the most common words in any language (like articles, prepositions, pronouns, conjunctions, etc) and does not add much information to the text. They are termed as 'noise' in the dataset.

4. **Lemmatization**

For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set.

The goal of lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance:

am, are, is $\Rightarrow$ be
car, cars, car's, cars' $\Rightarrow$ car

Lemmatization is a text normalization technique used in natural language processing to bring words to their base forms. For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words. Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.

5. **Sentiment Polarity**

Sentiment polarity is a technique used in natural language processing for dictating the orientation of a sentence. We used the SentimentIntensityAnalyzer module from the python nltk library to identify the polarity of each word for every sentence. Each word is either judged to be positive, negative, or neutral with a polarity score that sums to 1. There is also a compound polarity score that represents the overall polarity score. The compound polarity score ranges between -1 and 1, where a negative value indicates the negative

sentiment and positive value representing positive statement. Also, the closer the values to -1 or =1, the stronger the sentiment.

## Description of the algorithms used

### Sentiment Analysis

There are two approaches for the sentiment analysis. They include the machine learning approach and the lexicon-based approach. The machine learning approach classifies the text using classification algorithm. The machine learning method work in such a way that we apply a trained classification algorithm on the dataset to identify if the text is positive or not. On the other hand, the lexicon-based approach uses sentiment dictionary with opinion words and match them with the text to determine its polarity, A sentence is tokenized, and each token is matched with the available words in the model to find out its context and sentiment (if any). A combining function such as sum, or average is taken to make the final prediction regarding the total text component.

**The application of data-mining techniques to selected datasets that you choose using Python.**

We also applied the classification data mining technique. This is because we want to classify if customer's comment is either positive, negative, or neutral.

**Explanation of the experimental procedure, including the setting and optimisation of model hyperparameters during training, and your approach to validation (for supervised learning tasks).**

We created a column representing polarity scores for positive, negative, neutral and compound.

**Visualisation of the results.**

**Distribution of compound, positive and negative scores**

Let us visually examine the distribution of the overall polarity scores. We observed that we have more positive sentiments by looking at the compound score graph.

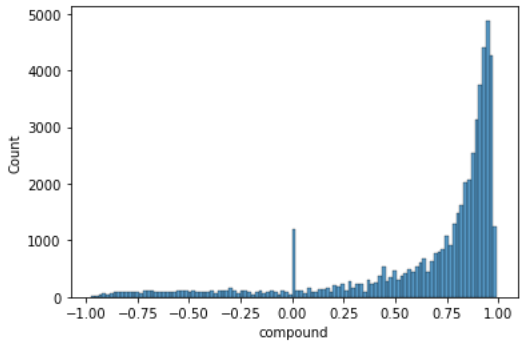


**Fig 1**: Compound Scores

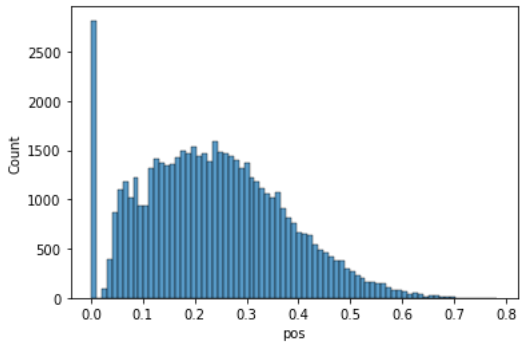


**Fig 2**: Positive Scores

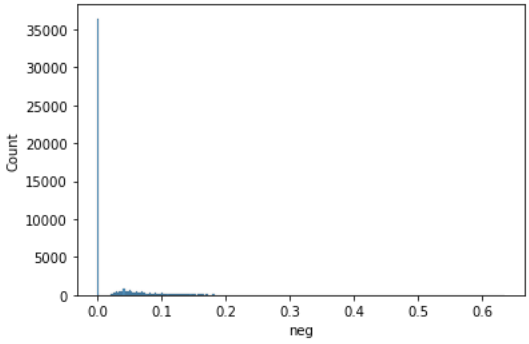


**Fig 3**: Negative Scores

By comparing the negative and positive score graphs, we observe that there are more positive words than negative words.

**Results analysis and discussion**

**Table 1:** Summary statistics for the polarity scores

| | Compound | Negative | Neutral | Positive |
|---|---|---|---|---|
| count | 53644 | 53644 | 53644 | 53644 |
| mean | 0.659426 | 0.026919 | 0.737566 | 0.235517 |
| Std | 0.423817 | 0.051132 | 0.127370 | 0.136031 |
| Min | -0.975700 | 0.000000 | 0.217000 | 0.000000 |
| 25% | 0.585900 | 0.000000 | 0.654000 | 0.133000 |
| 50% | 0.844200 | 0.000000 | 0.744000 | 0.228000 |
| 75% | 0.927100 | 0.042000 | 0.830000 | 0.327000 |
| Max | 0.993200 | 0.635000 | 1.000000 | 0.783000 |

Looking at the summary statistics, it is obvious that we have more positive sentiments. In fact, we can see that the median compound score is 0.84 – which means that over 50% of the reviews have a compound score of more than 0.84, which suggests strong positive sentiment.

**Table 2:** Frequency of the top 10 negative reviews per restaurant/hotel

| Restaurant/hotel | Frequency |
|---|---|
| Da Mario | 49 |
| Outdoor Restaurant | 48 |
| La Casa | 45 |
| Dada Yura Restaurant | 44 |
| Pizza Hut – Jungceylon | 40 |
| Ali Baba Restaurant | 39 |
| Mama Restaurant – Karom Beach | 37 |
| Restaurant La Croisette | 37 |
| Khan Baba Phuket | 36 |
| Food Market Restaurant | 34 |

Table 2 shows the frequency of the top ten negative reviews per restaurant/hotel. We observed that Da Mario, Outdoor Restaurant and La Casa has more negative sentiments.

**Table 3:** Frequency of the top 10 positive reviews per restaurant/hotel

| Restaurant/hotel | Frequency |
|---|---|
| Da Mario | 230 |
| No. 6 Restaurant | 182 |
| The Family Restaurant | 180 |
| Sabai Sabai | 177 |
| The Pizza Company | 153 |
| Outdoor Restaurant | 145 |
| Rock Salt | 99 |
| Curry Delight Indian Restaurant | 99 |
| Sam's Steaks and Grill | 99 |
| Thong Dee The Kathu Brasserie | 98 |

Table 3 shows that Da Mario, No. 6 Restaurant and The Family Restaurant has the top three (3) positive sentiments.

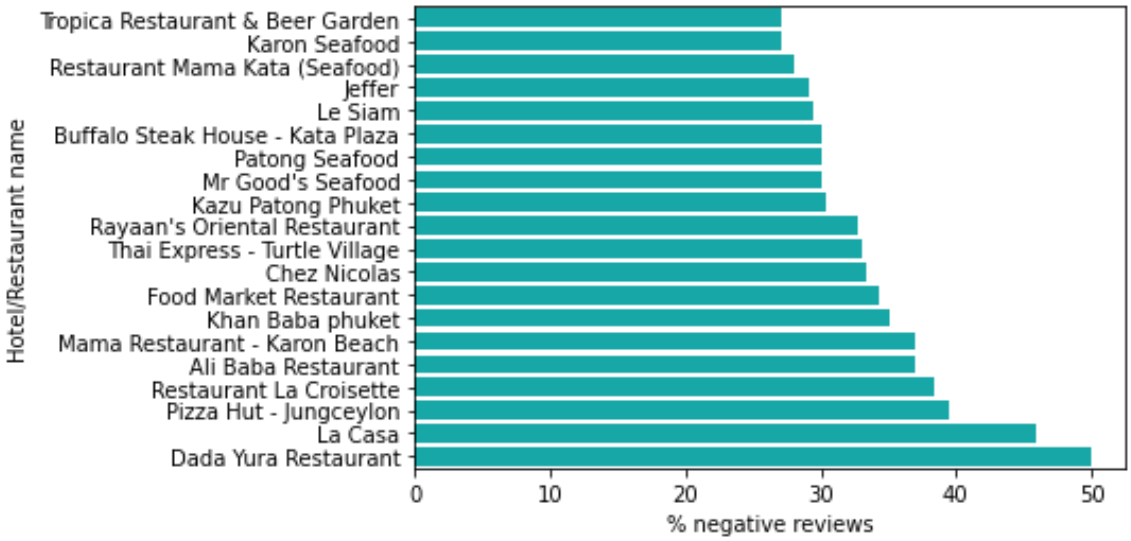


**Fig 4:** Bar Plot of negative sentiments

The bar plot above shows the top twenty (20) negative sentiments per hotel/restaurant. It is apparent that Dada Yura Restaurant, La Casa and Pizza Hut – Jungceylon has more negative sentiments.

**Word Cloud**

The Word Cloud gives a graphical representation of the most frequently appearing words in the text. The most occurring word will have noticeable appearance in the Word Cloud.


**Fig 5**: Wordcloud of words from negative reviews for Dada Yura Restaurant.

The wordcloud above represents the most frequent words mentioned by the customers. We can see that ‘untidy’, ‘service’, ‘staff’, ‘terrific’ are the most used words. This represents potential concerns and hence deserve further actions.


**Fig 6**: Wordcloud of words from positive reviews for Dada Yura Restaurant

The wordcloud above shows the most frequent positive words used to describe the Dada Yura restaurant. We can see that ‘food’, ‘friend’, ‘service’, ‘russian’ are the most used words used to describe the Dada Yura restaurant.

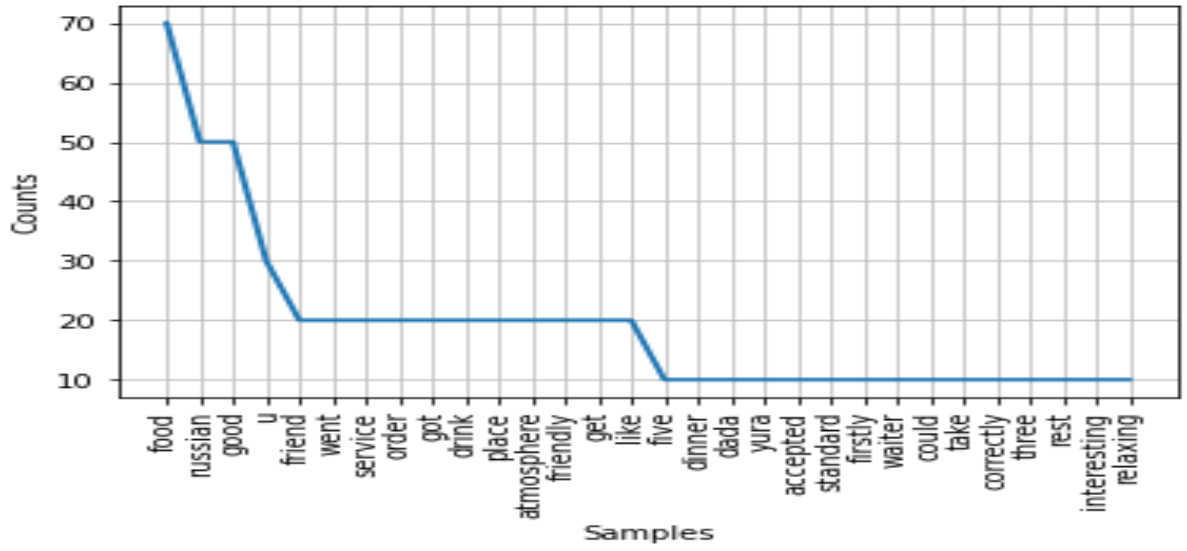


**Fig 7:** Frequency of positive words for Dada Yura Restaurant.

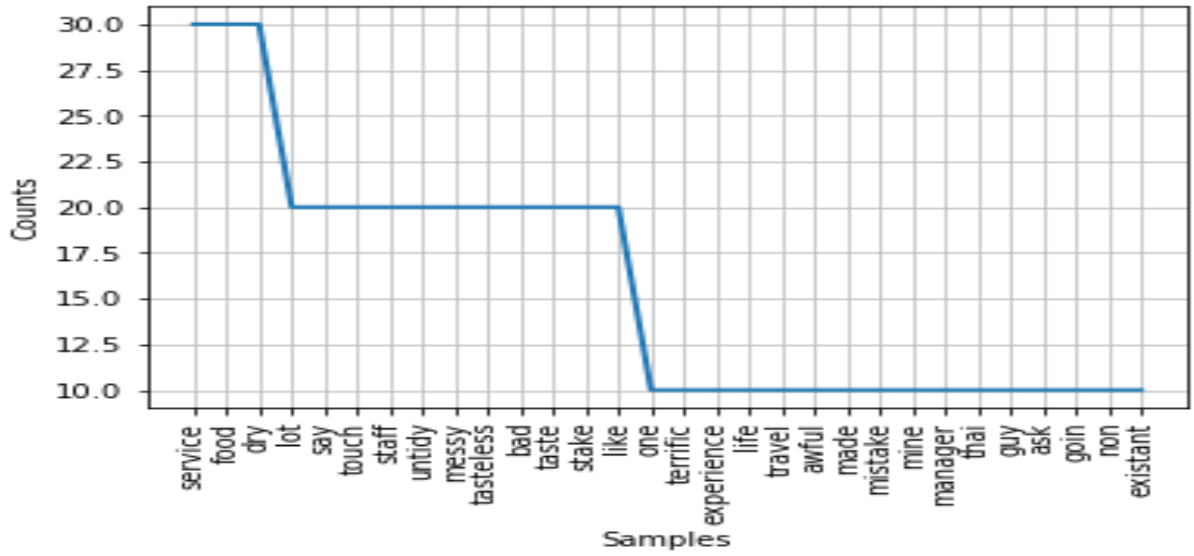


**Fig 8:** Frequency of negative words for Dada Yura Restaurant.

The last two plots above show the counts of the positive and negative words.

**Conclusion**

In this work, we applied the lexicon-based sentiment analyzer which classifies the reviews based on its sentiment value. The text considered are customers opinion for different hotel/restaurant. The sentiment classification is done based on polarity measures. These measures signify the positive, negative, or neutral attitude of users towards a particular restaurant/hotel, thereby enabling us to present the comparison between the top accommodations.

**References**

Rastogi, K. (2022). Text Cleaning Methods in NLP. Available at: www. https://www.analyticsvidhya.com. (Accessed 12 November 2022).

F. Nausheen and S. H. Begum, "Sentiment analysis to predict election results using Python," *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 1259-1262, doi: 10.1109/ICISC.2018.8399007.