Most relevant rooms in an Airbnb accommodation: A quantitative analyses on reviews

Bruno D. de Paiva
Escola Politécnica - UFRJ
Rio de Janeiro, Cidade Universitária
bruno.dantas@poli.ufrj.br

João P. Leite Pinho
Escola Politécnica - UFRJ
Rio de Janeiro, Cidade Universitária
jpinheirolp@poli.ufrj.br

***Abstract*—This paper describes a project that gets every Airbnb review and clusters it to get the rooms more commented on, positively or negatively. The idea is that the person who announces its houses can focus on improving these rooms aiming for the best reviews of the platform.**

***Index Terms*—airbnb, text analysis, reviews**

I. Introduction

Airbnb is a application that works as if it were a social network, in which the objective of this network is to unite people who want to share their accommodation with those who want some accommodation to stay, uniting hosts and travelers who aim to book unique accommodations anywhere. place in the world. [1]

Bearing in mind the importance of the platform in taking people anywhere in the world, always having a place to rent, having a lodging evaluation policy makes it necessary for them to be able to measure the quality of the environment.

In addition, because the platform does not specify some information, it is necessary to understand what efforts the hosts must have to please their customers in the best possible way and, with that in mind, we chose to build this solution that facilitates the life of the person who is offering accommodation in discovering which room in the house will bring a better evaluation of the traveler, making the facility better evaluated and, consequently, increasing your search.

For this, we use a dataset containing about 9 million reviews in order to classify them into good or bad reviews and then extract which rooms are contained in these comments.

Research has already been done with the intention of analyzing Airbnb reviews qualitatively [2]. However, our work intends to extract information from reviews through a mostly quantitative analysis, based on tools available in python language libraries.

II. Methodology

Considering the size of the dataset, it was initially decided to reduce its size in order to test the entire procedure, which will be better addressed in the following subsections.

A. Understanding the dataset

The dataset used, as previously mentioned, is a set of 9 million reviews on Airbnb's accommodations, among these comments, it is important to notice that the language they were written in varies in a total of more than 10 different languages. Thus, for this work, it was chosen to work only with the languages comments in English, so a previous process of translating these texts would not be necessary, avoiding bringing all the context and making a future sentiment analysis a harder process.

In addition, there were comments that had only numerals as information, these data were also removed from the dataset as they would not add value to the search.

This dataset has 6 columns, which are the hosting id, the date of this comment, who made the comment and the actual comment, in which, for this project, the most important data and the only one used was the comment executed.

It is also important to point out that in the same comment, there are multiple sentences that the user can write. There were cases where a single comment had more than 3 sentences. Thus, it was chosen to separate each of these reviews by phrases in order to have a better analysis by content, since the same comment can have phrases with positive and negative meanings.

B. Data Treatment

Having a large amount of data, we chose to normalize all comments made by users. Bearing in mind the previous point, that a user in the same acomoddation can make more than one comment.

Thus, we use the simple_preprocess tool, from the gensim library, which aims to convert a document into a list of tokens, having passed as parameters the minimum size of the tokens that will remain in this list, in our case we kept terms greater than or equal to three.

Furthermore, having this list of tokens, we removed stop words from the comment, in order to bring only relevant information for future analysis. We than chose to lemmatize, in order to unite the tokens in their meanings.

In addition, some cases were found in which the comments were empty, due to the fact that the user himself wrote a mistake in his comment, thus, it was chosen to remove these comments.

Additionally, Porter stemmer was run in order to reduce all modified words (inflections, derivations) to their root form, aiming to improve the final analysis.

In a complementary way, after all the previous treatment processes, using the stemerized comments, a process was created responsible for filtering all the rooms contained in a specific comment.

For this point, a list of rooms created by the authors in the chosen language was used, in this case English, in order to encompass as many accommodations as possible and not lose information. In addition, the synset funcyion of the wordnet library was used in order to obtain all possible synonyms of each of the words listed, aiming to cover even more the list of synonyms.

At the end of this entire process, we obtained a dataframe, in which we had the information from the comments (one sentence per line), their comments treated, lemmatized, stemmerized and a column containing the list of rooms mentioned in this comment [3].

C. Sentiment Analyses

This topic is a crucial topic for the analysis, given that the dataset does not have any kind of explicit feeling about the comments and, in order to bring greater business value, it is necessary for us to be able to divide which rooms are most commented on (positively or negatively). In this way, as it is not the final objective of the project, we use the textblob library that is capable of measuring the polarity of the message, as well as its subjectivity [4], in this way it is possible to map the sentiment of this comment.

For this, we chose to define the polarity levels in 7 possible feelings. Knowing that this polarity varies from -1 to 1, and that 0 is considered neutral sentiment, we made the following mapping.

TABLE I
SENTIMENT MAPPING TABLE

Sentiment	Interval
horrible	[-1, -0.75[
very negative	[-0.75, -0.5[
negative	[-0.5, 0[
neutral	0
positive	[0.01, 0.5[
very positive	[0.5, 0.75[
amazing	[0.75, 1]

Given the sentiments above, for each polarity measured in a document, it was verified whether the value obtained was contained within some defined range, thus, a new column was created in this dataset that aims to indicate the sentiment of this comment.

Furthermore, given that we only want to check the negative and positive feelings, we chose to remove all comments that would bring a neutral identification, even considering that the quality of this sentiment analysis does not have an accuracy and precision of 100%.

D. Word Cloud

Before creating the wordcloud, we basically measured the frequency of each of the rooms that appeared in the rooms column, created during the treatment process. For this, a dictionary was created with the accommodation being the key and the frequency that it appeared in this column the value so that, in this way, the wordcloud could be generated from the frequencies previously calculated.

Since the objective of this work is to help those who offer their accommodation on airbnb, it is necessary to expose the information extracted from the comments in an easy to understand format. For this we made use of the wordcloud library that builds a graphic scheme where the size of the words corresponds to the frequency they appear in the text, highlighting the most relevant words [5]. Two wordclouds were built, one with positive comments, Fig. 1, and one with the negative comments, Fig. 2, as you can see below.

Fig. 1. Types of rooms that appear the most in positive comments.

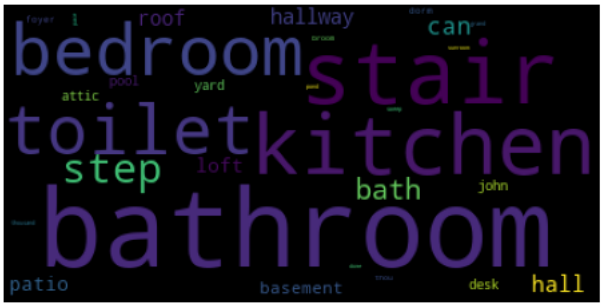

Fig. 2. Types of rooms that appear the most in negative comments.

III. Results

When analyzing both generated wordclouds, it is possible to see that there are easily observed highlights, indicating which accommodations the platform's hosts should focus their efforts on. Imagetically, from the result of the wordcloud, it is possible to notice that "kitchen", "bedroom", "stair", "bathroom" and its synonym "toilet" are the most evident in both wordclouds. In both positive and negative reviews, the most frequent word is "bathroom", followed by "kitchen". Furthermore, it is clear that, due to the use of wordnet, words like "Jhon", which is a synonym for bathroom in British English and "can" as an informal synonym for bathroom, in North American English appear as possible words for this wordcloud, negatively impacting the achievement of the result.

One possibility was basically to combine all these possible synonyms in a single word, in order to obtain a better result in which the room was most commented, either negatively or positively, but such efforts were left for future implementations of this project.

Another point observed is that both wordclouds have very similar rooms, the image visualization of the wordcloud does not make it easier to obtain this information, however, as can be seen in the tables below, there is the following relationship between room and frequency of appearance. It is also important to note that for a number of 50000 comments, after all the filters, those that list rooms and are positive or negative comments are about 10000, in which 20% are negative comments and the other 80% are positive.

TABLE II
TABLE OF WORDS VS FREQUENCIES (POSITIVE COMMENTS)

Words	Frequencies
bathroom	2066
kitchen	1998
bedroom	1104
stair	537
step	384

TABLE III
TABLE OF WORDS VS FREQUENCIES (NEGATIVE COMMENTS)

Words	Frequencies
bathroom	730
kitchen	256
stair	242
bedroom	240
step	110

Thus, by selecting the 5 Words most commented positively and negatively, it is possible to better indicate to the owner of the establishment what should be the greatest care that he must have in relation to his environment in order to receive a better evaluation within the platform and also to bring greater customer satisfaction.

IV. CONCLUSION

Observing the results obtained, it is possible to conclude that, in fact, the bathroom, kitchen, bedroom and stairs are the rooms with which an Airbnb host should pay more attention. It is recommended, above all, to stick to any problem that there may be in the bathroom, as this appears a much higher number of times than the other rooms in the negative comments. It is important to notice that this study was carried out in a Dataset that contained comments from airbnbs located in the cities of Los Angeles and New York, in the United States. Therefore, the results obtained must be considered within this context. Furthermore, as only English comments were analyzed, the analysis is restricted to English speaking guests.

REFERENCES

[1] Airbnb (2022). Airbnb, O que é a Airbnb e como funciona.

[2] Me senti em casa: análise das revisões de experiências de hospedagem colaborativa no site Airbnb sob o prisma da confiança.

[3] Chris D. Paice, "Another Stemmer".

[4] S. Ahuja and G. Dubey, "Clustering and sentiment analysis on Twitter data," 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), 2017, pp. 1-5, doi: 10.1109/TEL-NET.2017.8343568.

[5] M. A. Hearst, E. Pedersen, L. Patil, E. Lee, P. Laskowski and S. Franconeri, "An Evaluation of Semantically Grouped Word Cloud Designs," in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 9, pp. 2748-2761, 1 Sept. 2020, doi: 10.1109/TVCG.2019.2904683.