Cômodos mais relevantes em um Airbnb: Uma análise quantitativa com 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. Introdução

Airbnb é uma das soluções que funciona como se fosse uma rede social, no qual o objetivo dessa rede é unir as pessoas que querem partilhar sua hospedagem àquelas que querem algum alojamento em mente, unindo anfitriões e viajantes que tem como objetivo reservar alojamentos únicos em qualquer lugar do mundo. [1]

Tendo em vista a importância da plataforma em levar as pessoas a qualquer lugar do mundo tendo sempre um local para alugar, ter uma política de avaliação das hospedagens de torna necessário para que elas consigam mensurar a qualidade do ambiente.

Além disso, devido a plataforma não explicitar algumas informações, se torna necessário entender quais são os esforços que os anfitriões devem ter ao agradar da melhor forma possível os seus clientes e, pensando nisso, optamos por construir essa solução que facilita a vida da pessoa que está ofertando o alojamento em descobrir qual o cômodo da casa trará uma melhor avaliação do viajante tornando a instalação melhor avaliada e, consequentemente, aumentando a sua busca.

Para isso, utilizamos um dataset contendo cerca de 9 milhões de reviews com o objetivo de classificar-los em reviews bons ou ruins e então, extrair quais são os cômodos contidos nesses comentários.

Já foram feitas pesquisas com a intenção de analisar comentários de Airbnb de forma qualitativa [2]. Entretanto nossa trabalho pretende extrair informações das reviews por meio de uma análise majoritariamente quantitativa, apoiando-se em ferramentas disponíveis em bibliotecas da linguagem python.

II. Metodologia

Tendo em vista o tamanho do dataset, inicialmente foi optado por reduzir o seu tamanho a fim de testar todo o procedimento que será melhor abordado nas subseções a seguir.

A. Compreensão do dataset

O dataset utilizado, como comentado previamente, é um conjunto de 9 milhões de reviews em hospedagens do Airbnb, dentre esses comentários, é importante ressaltar que a linguagem que eles foram escritas variam numa série de mais de 10 línguas distintas. Deste modo, para este trabalho, foi optado por trabalhar somente com as linguagens em inglês, assim não seria necessário um processo prévio de tradução desses textos, evitando de trazer todo o contexto e dificultando uma análise de sentimentos futura.

Além disso, existiam comentários que possuíam somente numerais como informação, estes dados também foram removidos do dataset pois não trariam valor à busca.

Este conjunto de dados possui 6 colunas, sendo elas o id da hospedagem, a data deste comentário, quem fez o comentário e o comentário efetivamente, no qual, para este projeto, o dado mais importante e o único utilizado foi o comentário executado.

É importante ressaltar também que em um mesmo comentário, existem múltiplas frases que o usuário pode escrever. Houveram casos em que um só comentário possuía mais de 3 sentenças. Assim, foi optado por separar cada um desses reviews por frases a fim de ter uma melhor análise por conteúdo, dado que um mesmo comentário pode possuir frases com sentido positivo e negativo.

B. Tratamento

Tendo em vista essa grande quantidade de dados, optamos por normalizar todos os comentários feitos pelos usuários, tendo em vista o ponto anterior que um usuário, em uma mesma hospedagem, pode fazer mais de um comentário.

Assim, utilizamos a ferramenta simple_preprocess, da biblioteca gensim, que tem como o objetivo converter um documento em uma lista de tokens, tendo passado como parâmetros o tamanho mínimo dos tokens que permanecerão nessa lista, em nosso caso mantivemos termos maior ou igual a três.

Ademais, tendo essa lista de tokens, fizemos a remoção de stop words do comentário, a fim de trazer somente informações relevantes para futuras análises e optamos por lematizar, a fim de unir os tokens em seus significados.

Acrescido a isso, foram encontrados alguns casos em que os comentários ficaram vazios, devido ao fato de algum erro de

escrita pelo próprio usuário em seu comentário, desta forma, foi optado por remover estes comentários.

Adicionalmente, foi executado o Porter stemmer a fim de reduzir todas as palavras modificadas (inflexões, derivações) para a sua forma radicalar, visando ter uma melhoria na análise final.

De forma complementar, após todos os processos prévios de tratamento, utilizando os comentários stemerizados, foi criado um processo responsável por filtrar todos os cômodos contidos em um comentário específico.

Para este ponto, utilizamos de uma lista pré-criada de cômodos na linguagem escolhida, em nosso caso inglês, a fim de englobar o máximo de possíveis acomodações e não perder informação. Além disso, utilizamos do synset da biblioteca wordnet, a fim de obter todos os possíveis sinônimos de cada uma das palavras elencadas, visando abranger ainda mais a listagem de sinônimos.

Ao final de todo este processo, obtivemos um dataframe, no qual possuíamos as informações dos comentários (uma sentença por linha), seus comentários tratados, lematizados, stemerizados e uma coluna no qual contém a lista de cômodos citados neste comentário [3].

C. Análise de Sentimentos

Este tópico é um tópico crucial para a análise, tendo em vista que o dataset não possui explicito nenhum tipo de sentimento a respeito dos comentários e, para trazer maior valor de negócio, é necessário que a gente consiga dividir quais os cômodos são mais comentados (positivamente ou negativamente). Deste modo, como não é o objetivo final do projeto, utilizamos a biblioteca textblob que é capaz de mensurar a polaridade da mensagem, tal como a sua subjetividade [4], deste modo é possível realizar um mapeamento do sentimento desse comentário.

Para isso, optamos por definir os níveis de polaridade em 7 possíveis sentimentos. Sabendo que essa polaridade varia de -1 até 1, e que o 0 é considerado sentimento neutro, fizemos o seguinte mapeamento.

TABLE I
TABELA DE MAPEAMENTO DE SENTIMENTOS

Sentimento	Intervalo
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]

Dado os sentimentos acima, para cada polaridade medida dado um documento, foi verificado se o valor obtido estava contido dentro de algum intervalo definido, assim, foi-se criada uma nova coluna neste dataset que tem como o objetivo indicar qual o sentimento deste comentário.

Além disso, tendo em vista que queremos somente verificar os sentimentos negativos e positivos, optamos por remover todos os comentários que trariam uma identificação neutra, mesmo tendo em vista que a qualidade desta análise de sentimentos não tem uma acurácia e precisão de 100%.

D. Word Cloud

Antes da criação da wordcloud, basicamente medimos qual a frequência de cada um dos cômodos que apareceram na coluna rooms, criada durante o processo de tratamento. Para tal foi criada um dicionário com a acomodação sendo a chave e a frequência que ela apareceu nessa coluna o valor para que, desta forma, o wordcloud pudesse ser gerado a partir das frequências previamente calculadas.

Visto que o objetivo desse trabalho é auxiliar aqueles que ofertam seus alojamentos no airbnb, se faz necessário expor as informações extraídas dos comentários num formato de fácil compreensão. Para tal fizemos uso da biblioteca wordcloud que constrói um esquema gráfico onde o tamanha das palavras corresponde a frequência que aparecem no texto, ressaltando as palavras mais relevantes [5]. Foram construídas duas wordclouds, uma com os comentários positivos, Fig. 1, e uma com os comentários negativos, Fig. 2, como é possível observar abaixo.

Fig. 1. Cômodos mais falados em comentários positivos.

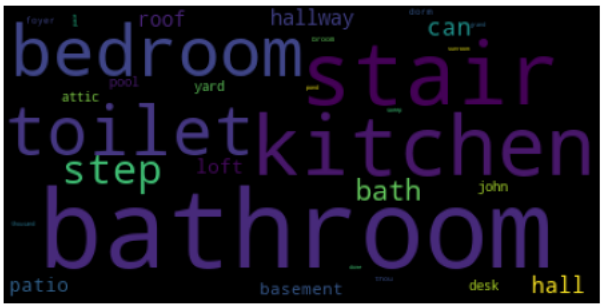

Fig. 2. Cômodos mais falados em comentários negativos.

III. RESULTADOS

Ao analisar ambas as wordclouds geradas, é possível perceber que existem destaques facilmente observados, indicando quais acomodações os anfitriões da plataforma devem concentrar seus esforços. Imageticamente, a partir do resultado da wordcloud, é possível notar que ”kitchen”, ”bedroom”, ”stair”,

”bathroom” e seu sinônimo ”toilet” são as mais evidentes em ambos wordclouds. Sendo que tanto nos comentários positivos quanto nos negativos palavra mais frequente é ”bathroom”, seguida por ”kitchen”. Além disso, é notório que, por conta da utilização do wordnet, que palavras como ”Jhon”, que é um sinônimo para banheiro em inglês britânico e ”can” como um sinônimo informal para banheiro, em inglês norte americano aparecem como possíveis palavras desse wordcloud, impactando negativamente na obtenção do resultado. Uma possibilidade era basicamente juntar todos esses possíveis sinônimos em uma única palavra, a fim de obter um melhor resultado em qual o cômodo foi mais comentado, seja negativamente ou positivamente, porém tais esforços foram deixados para futuras implementações desse projeto.

Outro ponto observado também é que ambas as wordclouds possuem cômodos bem similares, a visualização imagética da wordcloud não facilita para a obtenção dessa informação, porém, como é possível observar nas tabelas abaixo, tem-se a seguinte relação de cômodo e frequência de aparecimento. É importante frisar, além disso, que para um número de 50000 comentários, após todos os filtros, aqueles que relacionam aposentos e são comentários positivos ou negativos são cerca de 10000, no qual 20% são comentários negativos e os outros 80% são positivos.

TABLE II
TABELA DE PALAVRA VS FREQUÊNCIA (SENTIMENTOS POSITIVOS)

Palavra	Frequência
bathroom	2066
kitchen	1998
bedroom	1104
stair	537
step	384

TABLE III
TABELA DE PALAVRA VS FREQUÊNCIA (SENTIMENTOS NEGATIVOS)

Palavra	Frequência
bathroom	730
kitchen	256
stair	242
bedroom	240
step	110

Assim, selecionando as 5 palavras mais comentadas positivamente e negativamente, tem-se a possibilidade de indicar melhor ao dono do estabelecimento qual deverá ser o maior cuidado que ele deve possuir com relação ao seu ambiente visando receber uma melhor avaliação dentro da plataforma e também trazer um maior agrado aos seus clientes.

IV. CONCLUSÃO

Observando os resultados obtidos é possível concluir que de fato banheiro, cozinha, quarto e escada são os cômodos com os quais anfitrião de Airbnb deve prestar mais atenção. É recomendável sobretudo se ater a qualquer problema que possa haver no banheiro, visto que este aparece um número de vezes muito maior do que os outros cômodos nos comentários negativos. Se faz importante ressaltar que este estudo foi feito em um Dataset que continha comentários de airbnbs localizados nas cidade de Los Angeles e Nova Iorque, nos Estados Unidos. Sendo assim os resultados obtidos devem ser pensados dentro deste contexto. Além do mais como foram analisados apenas cometários em inglês a análise está restrita a os hospedes falantes deste idioma.

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.