CCT College Dublin

Assessment Cover Page

Module Title:	Strategic Thinking
Assessment Title:	Report 2
Lecturer Name:	James Garza
Student Full Name:	Izaias De Oliveira Gomes Junior
Student Number:	2023232
Assessment Due Date:	15th of December of 2023
Date of Submission:	15th of December of 2023

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Google Reviews and Sentiment Analysis

A study case with Machine Learning

Izaias de Oliveira Gomes Junior

James Garza

15 December 2023

Table of Contents

Introduction

This project is a study case of a hotel and camping site in Itu, a small city in Brazil, and has the purpose of analysing the guests' reviews on its page on Google Maps.

The Tourism industry is considered vital for most of the countries being the economic sector that shows its beauty, culture and potential to the rest of the world. Its impact on the GDP goes from jobs in many different areas (i.e. hospitality, agriculture, transportation, communications and others) to the Balance of Payments on bringing foreign exchange. In 2020 the industry represented 10,3% of the global GDP and 7,9% of Brazil's GDP representing almost 7 millions jobs (Colortel, 2022).

It is a sensitivity industry that suffered the most with the Covid-19 Pandemic (Behsudi, 2020) but a resilient one in countries with great potential to be explored like Brazil, where its 2022 GDP increased 2,9% with the Tourism industry being one of the biggest responsible (Maciel, 2023). In a global panorama the number of people travelling to other countries reached 960 million in 2023 almost 63% of the pre-pandemic numbers (UNWTO, 2023).

The hospitality sector is one of the most important sectors of this industry employing around 173 million people worldwide (EHL Insights, 2023). The reviews play an important role on customers' decisions and not all travellers prioritise the same attributes as recent studies have shown. Families and groups of friends can both travel for leisure but they will not look at the same reviews, while the first one will focus on comfort the second one is more interested in the meals offered (Witts, 2015). Understanding how different groups of customers look into the hotel's reviews should be considered as part of any hospitality business marketing strategy, once it will be more personalised facilitating booking sales once the potential customers will look exactly for what they are looking for.

Objectives

As a study case this project aims to help the hotel to improve its marketing strategy through a better understanding of its positives and negatives attributes pointed out by the guests. By exploring the data provided, the investigation will consist in finding those attributes and how they are correlated to the rating given and to the words used in the reviews to express the guests perception of their experience. Sentiments can be extracted from words and from how they are put together.

Problem Definition

One of the main objectives of any company is to optimise its revenue by minimising costs and maximising profits and an efficient marketing strategy can be crucial to reach this goal. The advance of technology and the popularisation of social media has changed the relationship between customer and business. It is much easier and faster to look for potential hotels for our holidays as it is to see what past customers have to say about their experience throughout different platforms.

The reviews are not always good but this does not mean they are not useful. Negative reviews are as important as the positive ones and knowing the reasons why the clients are not satisfied with the service or product is indispensable if the company values improvement and has it as part of its culture. Domino's is a case of success among marketing analysts when it comes to deal with bad reviews properly. In 2009, after analysing thousands of reviews, the company decided to reinvented their pizza turning losses into profits (Duke, 2019).

Scope

CRISP-DM is a logical and natural project management that helps to organise the process of working with data. The first step is Business Understanding which is helpful to draw the initial lines of the project and not waste time with inappropriate techniques. It does not mean that the approach will not change, as CRISP-DM is flexible and cyclical, the project can be adapted more than once throughout the time.

Since the beginning of the semester the idea was to work with a marketing study case. Initially we tried to work with the Food and Beverage sales and its possible correlation with the weather with data from a hotel in Dublin, unfortunately we could not gather all the necessary data forcing us to change the subject. We tried to keep working with a marketing study case in hospitality changing the hotel and the dataset.
Text is the most abundant data format in the world and most commonly in an unstructured format (IBM, 2023). The first step to deal with this type of data is Text Preprocessing which consists in cleaning the data by removing punctuations, stop words and emojis; putting all the words in lower or upper case; removing foreign words if they are not important for the analysis; splitting the text into smaller units; transforming the words in their roots and other steps (Deepanshi, 2021).

After the Text Preprocessing, the sentiment analysis will proceed and the Machine Learning Model will be applied. This semester only Random Forest will be used and the choice is based on its easy

applicability, high accuracy and reduced overfitting (E R, 2021). As Random Forest uses a combination of individual models to get the best result it makes a perfect choice at this stage when we are required to use at least one model. Other models will be tested in the next semester with the objective to explore new techniques and compare their results.

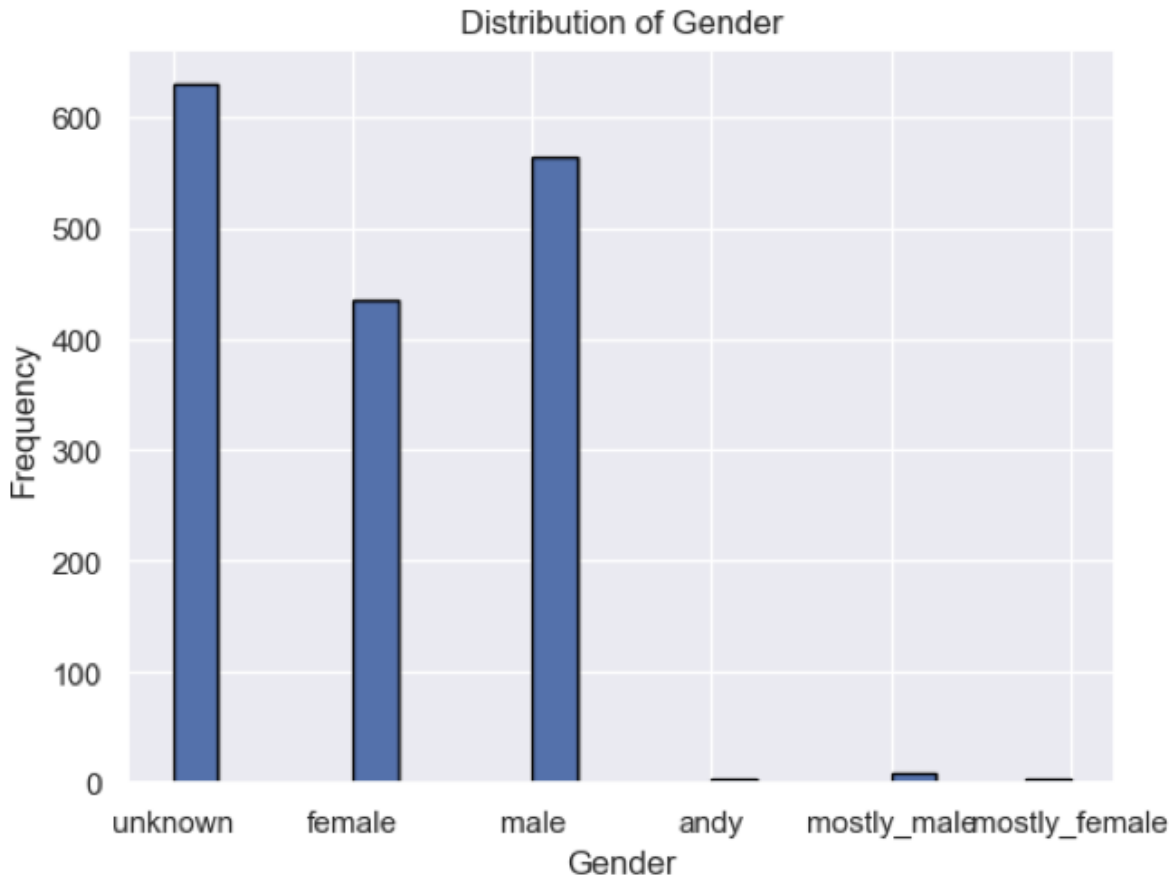
The reviews are written in Portuguese and translated to English by Google. Firstly we will work only with the English text and in the second semester we will work with the original text and compare if there is any change in the sentiment analysis or in the machine learning results for the different languages.

Ethical Considerations

The reviews are written by guests using their personal Google Account in the Google Business Account of the company. The name of the reviewer is displayed according to what they set, some examples of sets are: full name, first name, business name, nickname and others unidentified names. Even though the reviews are public and everyone on the internet can check them it is not ethical to expose them in a research with their personal identifiable information without their consent for it.

One idea that was raised during the Exploratory Data Analysis was to use the first name to predict the gender using the library gender-guesser from PyPI. The intention was to analyse which gender is more prone to give a positive review and a negative one. Unfortunately this will not be possible at this stage because the number of “unknown” corresponds to almost 30% of the dataset, which would result in a significant loss of information. An alternative would be to create our own name dictionary using the most frequent names which would be more accurate but also more time consuming. This idea will be on hold at the moment for further analysis in the next semester.

Figure 1 - Distribution of Gender



Data Understanding and Preparation

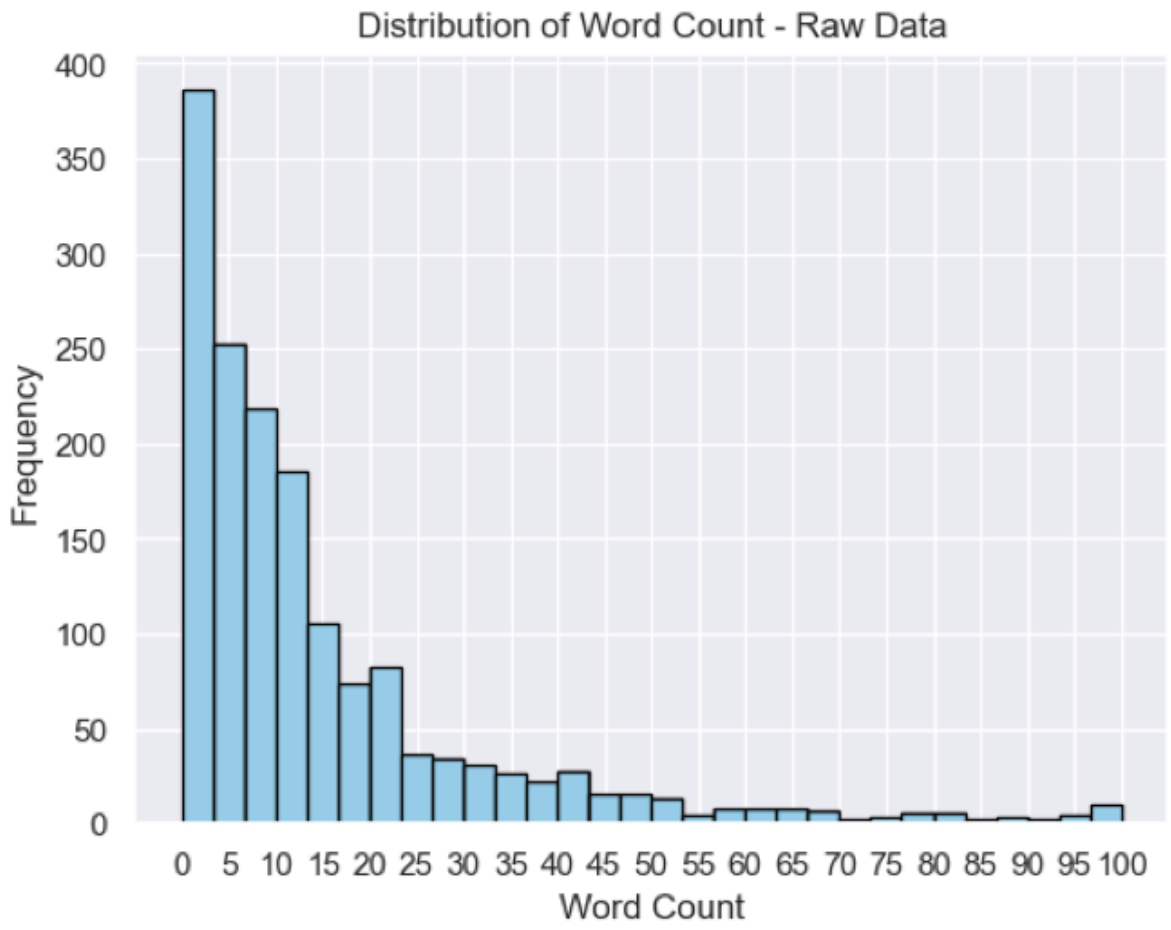
The reviews were extracted from the hotel's Google My Business page through the tool Google Takeout which allows the download of all your data in different Google applications. As I am the Marketing Manager from Camping Chapéu de Sol I have access to this data.

We got 143 files in json format with approximately 20 reviews each totalising 2793 raw data. Each review was stored in a json string with: "displayName", "starRating", "comment", "createTime", "updateTime" and "name". The first step was to merge all those files into a unique csv file using a code to transform the files into Pandas DataFrame, concatenating them into one main DataFrame and transforming it into a csv file. The json strings were stored in just one column, and to separate the different information in new columns we used the json library and the lambda function when the strings were evaluated assuming that their content is dictionary-like.

The “comment” contains the translated review by Google in English, the original review in Portuguese and the reply from the hotel. Some techniques were experimented to separate that information into new columns. The first one was using a code which intended to separate the texts according to the language detected but unfortunately this code did not perform as expected and a different approach was used. As the column comment had the same structure - “(Translated by Google)” text in English “\n\n(Original)\n” and text in Portuguese” - we separated the text before and after the “\n\n(Original)\n” item.

The reviews with only Star Rating and no comment were removed from the dataset and a total of 1648 reviews were left. The majority of the reviews have a number of words between 1 and 25. The outliers are concentrated in the reviews between 50 and 500 words totalising 130 occurrences. Among the 13 biggest ones - with more than 200 words - 55% have more than 3 as a star rating, and the biggest review - with more than 500 words - has 5 as a star rating.

Figure 2 - Distribution of Word Count Raw Data



The hotel is well-reviewed with 73.9% of the reviews with 5 star rating; however when we compared with the sentiment analysis the number of positive reviews are 53.0%, a decrease of 20.9 percentual points. Some hypothesis can be drawn:

1. The reviews with 5 stars might contain some negative words as the guests may highlight some aspects of the experience that could have been better but did not affect the overall sentiment. The example is the largest review with 500 words, 5 as star rating and sentiment score of 0.053516;
2. The sentiment analysis range is between -1 and 1 and the threshold used to separate them into negative (less than 0.00), neutral (between 0.0 and 0.5) and positive (more than 0.5) is not the most accurate one and has to be tuned.

The most 3 common trigrams are in accordance with the high amount of 5-star ratings. These trigrams, and their respective frequencies, are as follows: “highly recommend it” (18), “place spend weekend” (15) and “great place spend” (12).

Figure 3 - Distribution of Rating Stars

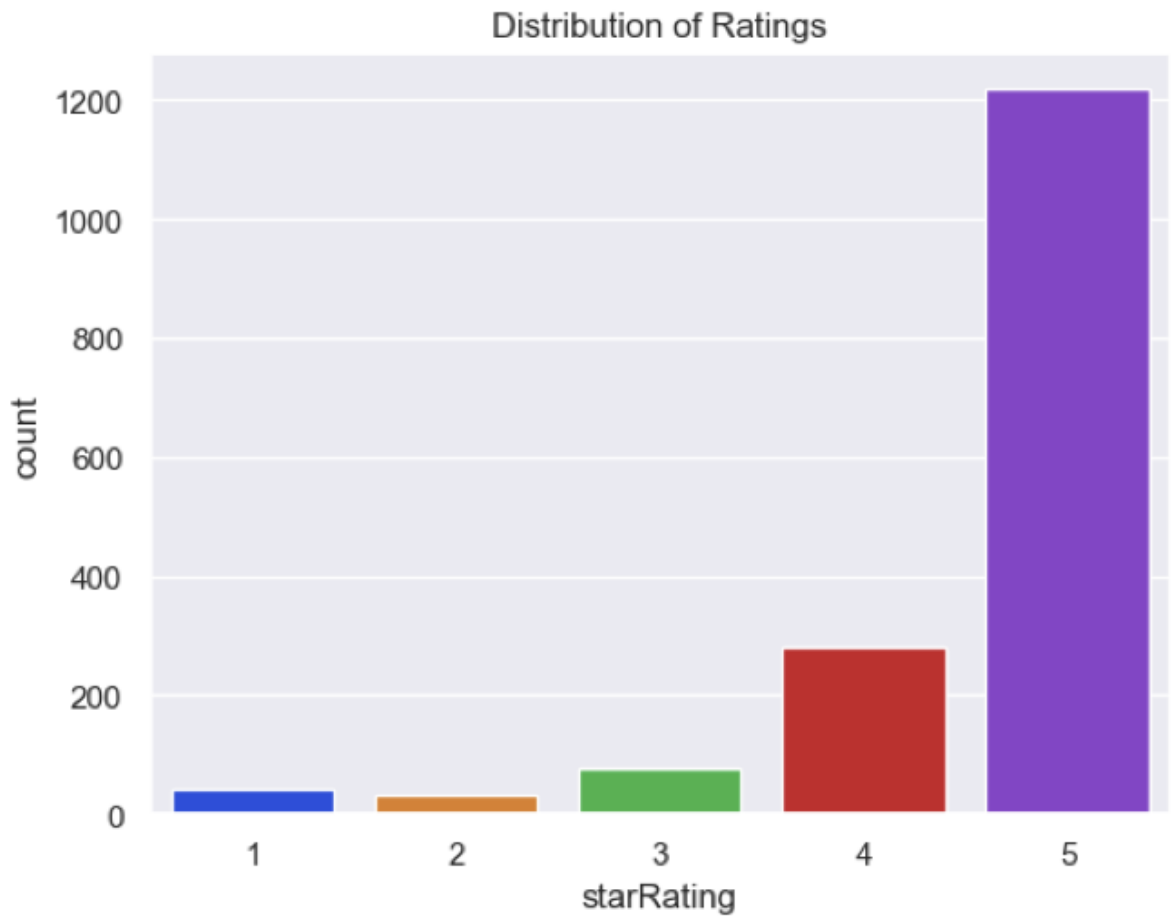


Figure 4 - Distribution of Sentiment Score Cleaned Data

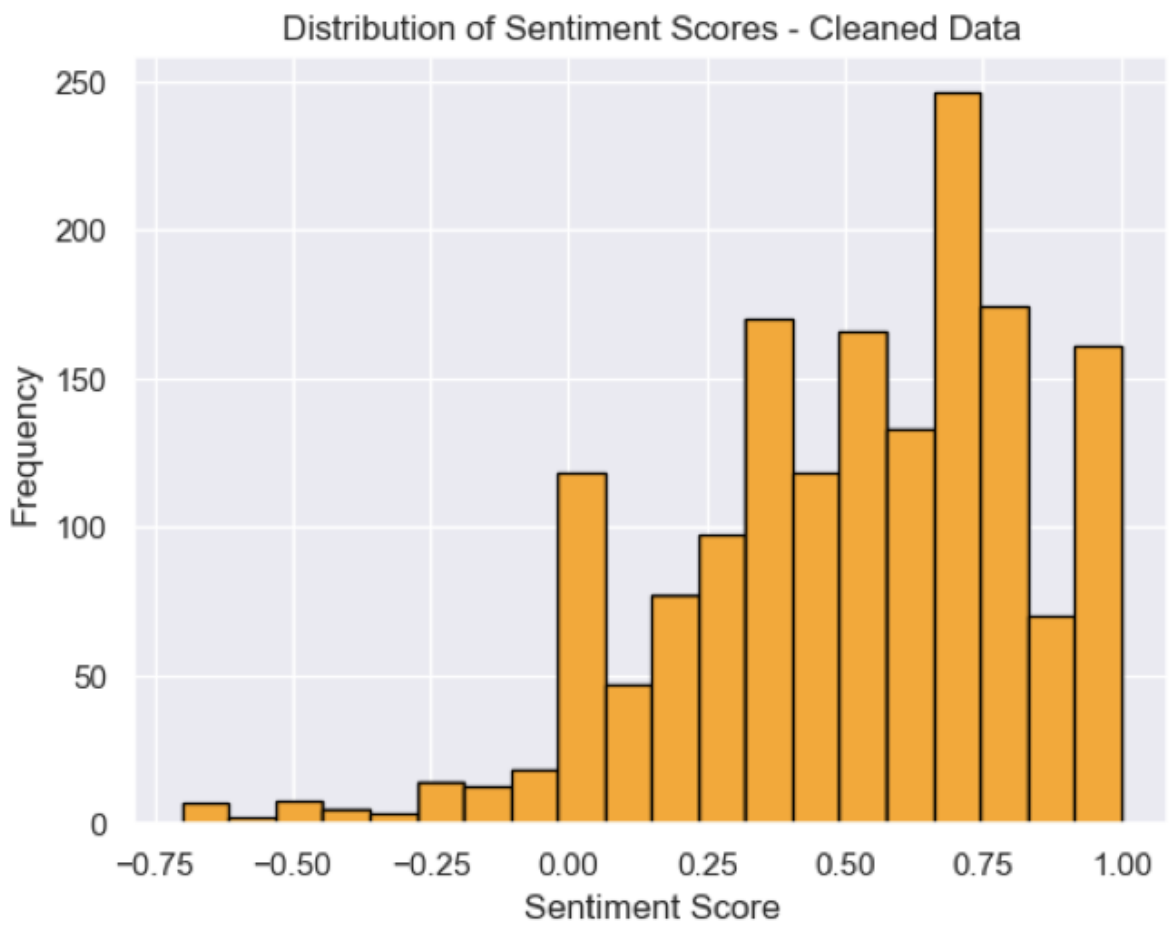
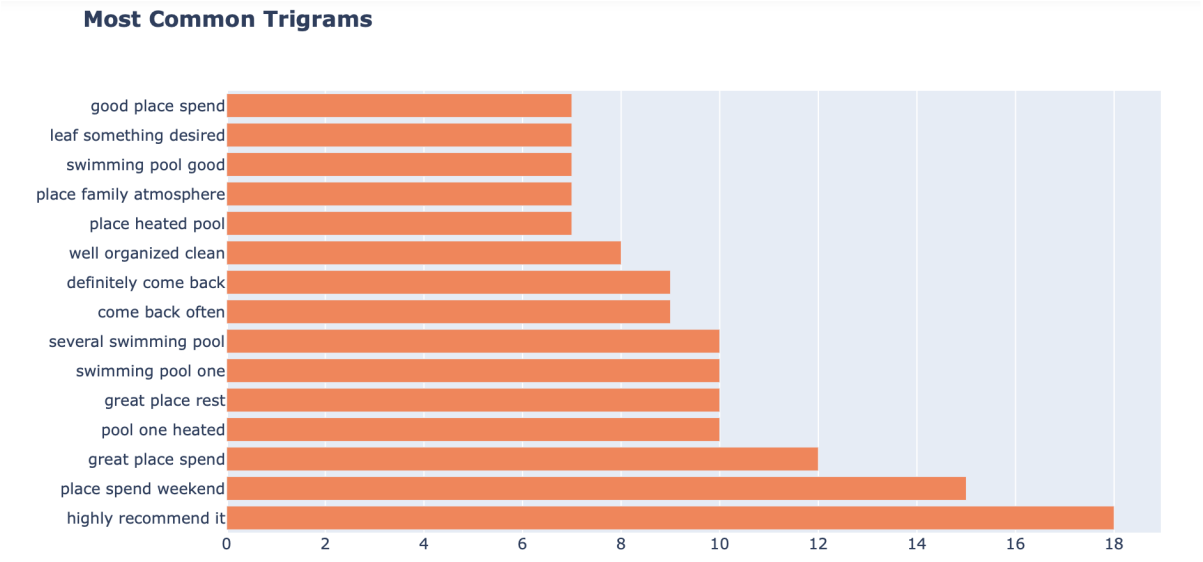


Figure 5 - Most Common Trigrams



Machine Learning

In the first algorithm we mapped the star ratings into: Positive (4 and 5), Neutral (3) and Negative (1 and 2). The best results were obtained with a test size of 40% and max features of 1000 resulting in an accuracy of 0.9181. The model is a little bit overfitting (Accuracy on Training Set = 0.992 and Accuracy on Test Set = 0.918) as the amount of positive reviews are superior to the negative ones. For the Neutral category the precision and recall were 0.00 due the small amount of it. A code with the Grid Search was used to find the best parameters but the results were almost the same with the first algorithm when the tuning was made manually.

The Random Forest was performed with the Sentiment Score as well and the accuracy was lower (0.816), more overfitting (Accuracy on Training Set = 0.997 and Accuracy on Test Set = 0.816) but the Neutral category had a precision of 0.50 and recall of 0.05. In a separated Jupyter Notebook both experiments were tested merging the Neutral Category to the Negative one but the results were similar to the original.

Conclusion

As written above, the project was changed from the first task to the current one, keeping only the idea of working with a study case in marketing. After defining the business understanding of the new project, almost two weeks were spent in the process of getting the new dataset and cleaning it. The most difficult part was to transform all the json files into an unique readable csv file with all the information separated in different columns. After this the work flowed naturally and some adjustments were done in previous steps such as: data visualisation, outliers and the cleaning process.

For next semester the agenda is to improve data visualisation, to apply other Machine Learning models and to get more reviews for the dataset. One of the limitations that I could not overcome in this project was to plot a graphic from the WordCloud library with the most frequent words in the reviews. As for the ML models, the Convolutional Neural Networks and Support Vector Machine are being considered to be added. There was no clear correlation found between the sentiment score and the rating star, and the creation of a name dictionary is being studied to identify the gender to test a possible correlation between it and those two variables.

Two more experiments are scheduled for the following steps: the analysis with the original text (in Portuguese) and the possibility to get the reviews from the competition to compare the difference

between the sentiment score. It was challenging to do everything again with less time but it was important to learn how to prioritise and make better decisions.

Reference list

Behsudi, A. (2020). Impact of the pandemic on tourism – IMF F&D. [online] IMF. Available at: https://www.imf.org/en/Publications/fandd/issues/2020/12/impact-of-the-pandemic-on-tourism-behsudi.

Colortel (2022). Setor Hoteleiro: grande gerador de emprego e renda. [online] Colortel. Available at: https://colortel.com.br/setor-hoteleiro-grande-gerador-de-emprego-e-renda-2/ [Accessed 29 Nov. 2023].

Deepanshi (2021). Text Preprocessing in NLP with Python Codes. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/# [Accessed 8 Dec. 2023].

Duke, C. (2019). How 3 Smart Brands Turned Negative Reviews into Opportunities: A Quick Case Study. [online] Mondovo Blog. Available at: https://www.mondovo.com/blog/how-3-smart-brands-turned-negative-reviews-into-opportunities-a-quick-case-study/.

E R, S. (2021). Random Forest | Introduction to Random Forest Algorithm. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.

EHL Insights (2023). Hospitality Industry: All Your Questions Answered. [online] hospitalityinsights.ehl.edu. Available at: https://hospitalityinsights.ehl.edu/hospitality-industry.

IBM (2023). What is Text Mining? | IBM. [online] www.ibm.com. Available at: https://www.ibm.com/topics/text-mining.

Maciel, V. (2023). IBGE confirma atividade turística como importante indutora da economia brasileira. [online] Ministério do Turismo. Available at: https://www.gov.br/turismo/pt-br/assuntos/noticias/ibge-confirma-atividade-turistica-como-importante-indutora-da-economia-brasileira# [Accessed 29 Nov. 2023].

UNWTO (2023). Tourism on Track for Full Recovery as New Data Shows Strong Start to 2023. [online] www.unwto.org. Available at: https://www.unwto.org/news/tourism-on-track-for-full-recovery-as-new-data-shows-strong-start-to-2023#:~:text=Overall%2C%20international%20arrivals%20reached%2080.

Witts, S. (2015). What do consumers look for when booking hotels? [online] restaurantonline.co.uk. Available at:

https://www.restaurantonline.co.uk/Article/2015/07/03/What-do-consumers-look-for-when-booking-hotels [Accessed 29 Nov. 2023].