Springboard: Data Science Career Track Program

Capstone Project 2 Proposal

Abdul Hannan

07/2022

Hotel Reviews Sentiment Analysis Proposal

1) What is the business problem?

- Reading reviews when choosing a hotel may take hours or a considerable amount of time, sentiment analysis eases the procedure, and helps customers learn about the hotel without going through all the reviews, and hotels understand what guests are happy or not happy with. Travel agencies and booking services often provide a score between 1 – 5, however that score is not very helpful for managers as they provide little to no information about what guests considered positive or negative. Extracting that from the textual reviews provided by the guests can provide vital information that may help the managers make better decisions and keep future customers happy.

2) Who are the intended stakeholders, and why is this problem relevant to them?

- The dataset is collected from Booking.com, the analysis of this can help them make a better search engine that provides their customers with the best hotel within their budget, and list a few characteristics extracted from the sentiment analysis done on the previous reviews provided by customers.

3) Where are the datasets available from?

- The dataset is available on Kaggle.

4) What data science approaches do you anticipate you will use to model the business problem as a data science problem?

- For this business problem, we will use an opinion-making/sentiment analysis which is a technique of Natural Language Processing (NLP). Since textual data is not always clean, we will clean it up by removing noise, stop words, normalization, stemming, and lemmatization. We want to represent a review into a sparse array or a collection of text documents into a vector. Since we have two reviews per person, a positive and a negative, we will be retrieving 2 sparse arrays and using those to build classifiers. To do that we will be using CountVectorizer, TfidfVectorizer, HashingVectorizer, and others will be considered during the process. We want to also focus on the terms that do not have much-predicted power and are used in reviews greater than a certain threshold. After that, we will go on and generate models, look for over or under-fitting and find optimal parameters to build the classifier on.

5) How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?

- After establishing the alignment between the business goals and the appropriate performance metrics, models will be ranked concerning these metrics, which will result in an explanation to the managers on how to better the experiences of their future customers.

- Interpretability analyses will hopefully lead to identifying patterns that impact the satisfaction of a customer and make better analysis on recommending hotels based on scores provided by the customers.