Data Mining And Modelling (MGMT4018)

Report: Text And Sentiment Analysis On Vapiano Tower Bridge Restaurant

Prepared by Mirza Razin Alam

101329992

Objective:

For the purposes of this research, we collected data from TripAdvisor of the reviews for Vapiano Tower Bridge restaurant based in London, United Kingdom. I picked this restaurant to carry out this research because it is a place that I would visit quite often during my time in London as a student. This is a chain Italian restaurant that has 4 branches in the city of London and is very popular. By conducting text and sentiment analysis on the reviews scraped from TripAdvisor, I can understand and compare the opinions of other people towards this particular restaurant.

Text analytics and mining process:

Using Python APIs as we were able to scrape the reviews from the URL address of the Vapiano restaurant from TripAdvisor. The tools collected information such as ID, dates, ratings, titles and reviews. This data was then extracted to a csv file. It necessary to remove some additional information that the scraping tool picked up. Once the data cleaned, we were able to load the refined dataset back onto our Python Integrated Development Environment for our analysis. The new dataset contained 84 observations and 4 variables columns.

Outcomes of our analysis:

Our initial step was to generate a wordcloud to get a sense of the most frequent words being used in the reviews (Appendix A). Some of the most occurring words were 'Order', 'Food', 'Pasta', 'Price', 'Staff', 'Time' and so on. Using sentiment analysis APIs within Python we were able to categorize the reviews by assigning polarity scores (-1 being negative, +1 being positive and 0 refers as neutral). In Appendix B we can see a scatter plot that illustrates that most of the observations are green, which tells us that most of the reviews are positive. Further in Appendix C, it shows that out of 84 reviews, almost 70 of them are positive with only around of the reviews as negative. Appendix D shows that all 84 reviews are detected as opinions rather than facts by the API algorithm. In Appendix E there is also a rating distribution where we used the ratings that were already present in TripAdvisor and assigned labels to score of 10 as 'Very Negative'; score of 20 as 'Negative'; score of 30 as 'Neutral'; score of 40 as 'Positive' and score of 50 as 'Very Positive'. Most of the reviews in the rating distribution again shows most of the reviews as 'Postive' and 'Very Positive'.

As part of the analysis, we also assigned each observation to a topic using the tools in Python. We chose 10 topics that are shown in Appendix F with the counts. The topics were chosen by analyzing some of the most occurring words. We can see that most of the reviews were classified as the topics 'Food', 'Pasta' and 'Order'. A sentiment analysis on the topics were also performed. In the stacked bar chart (Appendix G) it clear that most of topics has positive reviews. However, in reviews that classified as 'Food' topic shows that almost 40% of the texts are negative. This is concerning for Vapiano and should addressed as this reveals that there is further room for improvement in the food that is offered.

Recommendations:

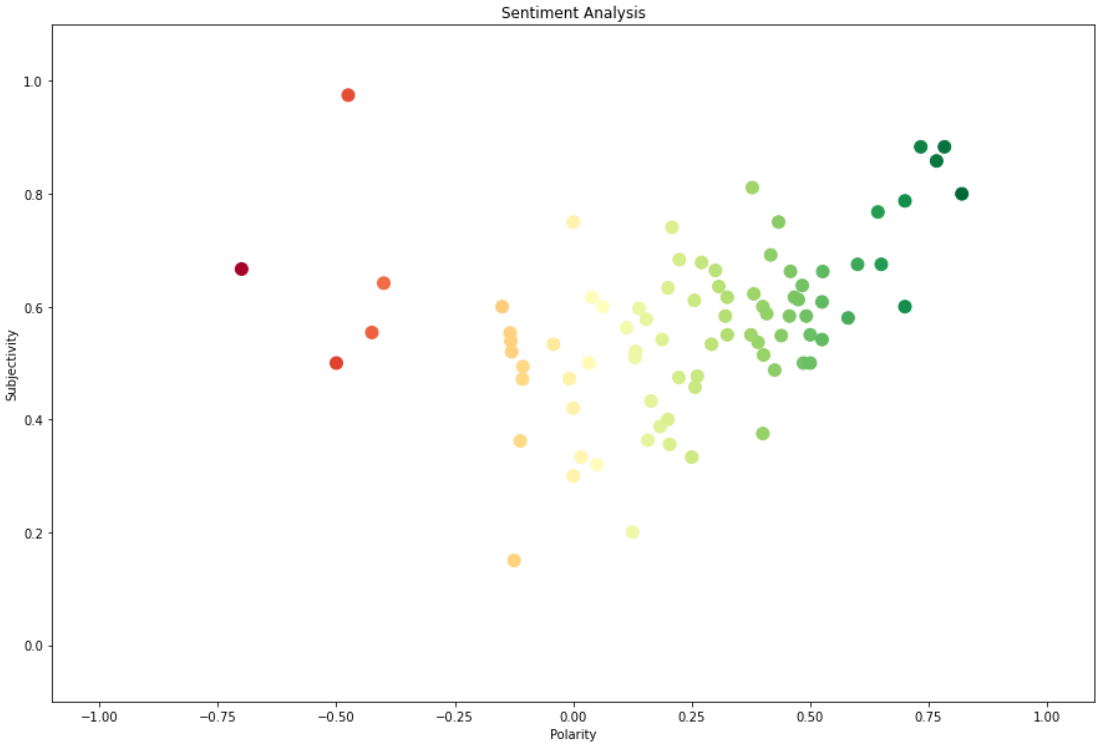
Although majority of the reviews are shown as positive, the restaurant can try to improve by analyzing the negative reviews. Most of the negative comments were complains regarding food. There Vapiano can further take actions to maintain a higher food standard which could improve their ratings on websites such

as TripAdvisor even further. There are also comments regarding poor behavior from the staff of Vapiano. We can recall from our analysis that the word ‘staff’ had high occurrence. Therefore, the manager of Vapiano should investigate such matters and make sure such events do not happen again.

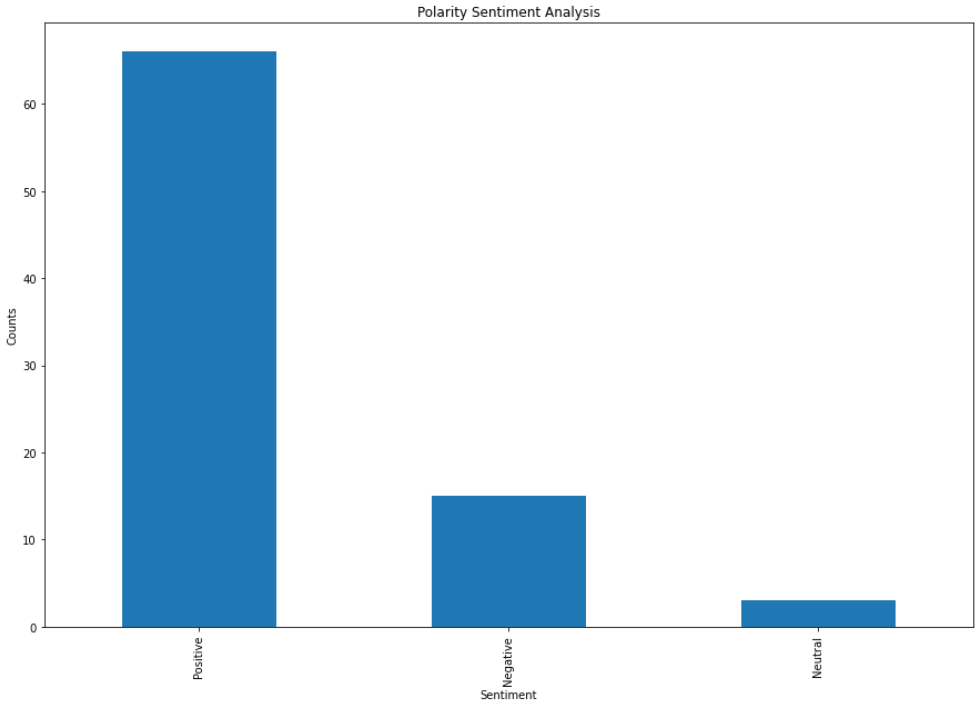
Appendix

Appendix A: Wordcloud

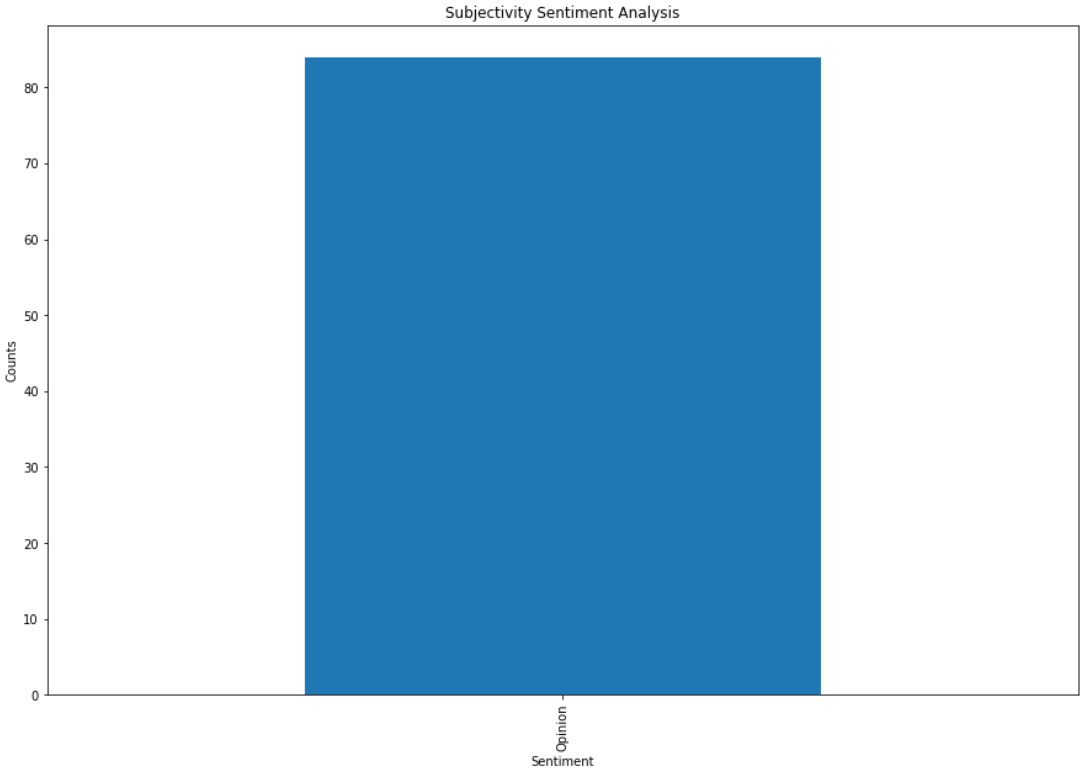
Appendix B: Sentiment Score Scatterplot



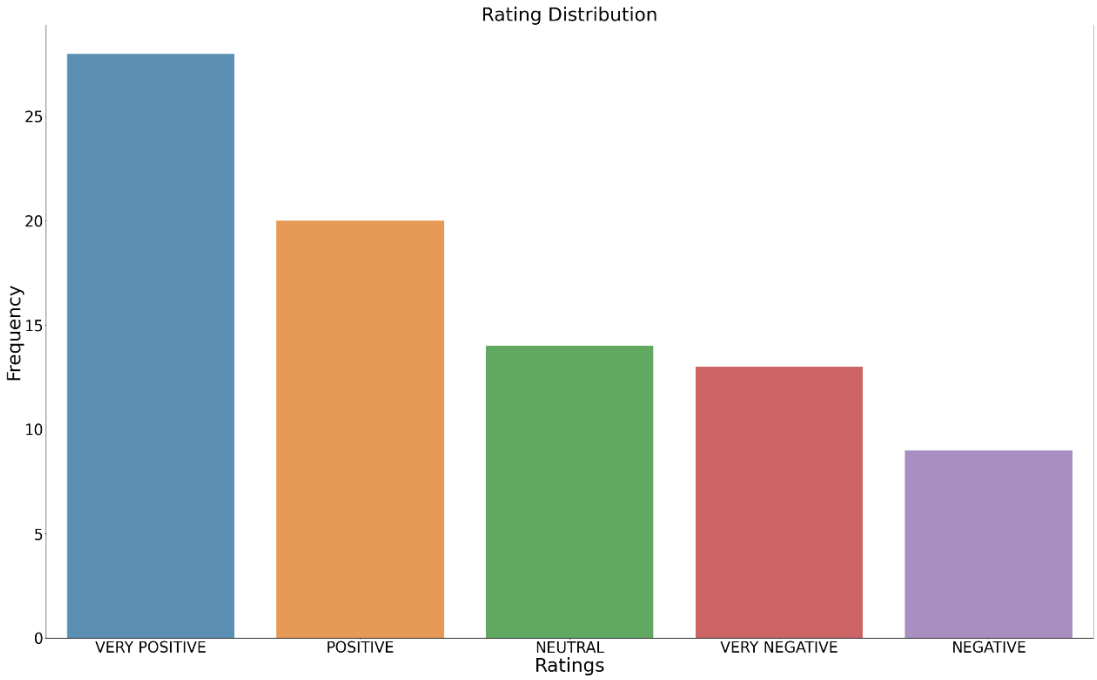
Appendix C: Polarity Sentiment Analysis



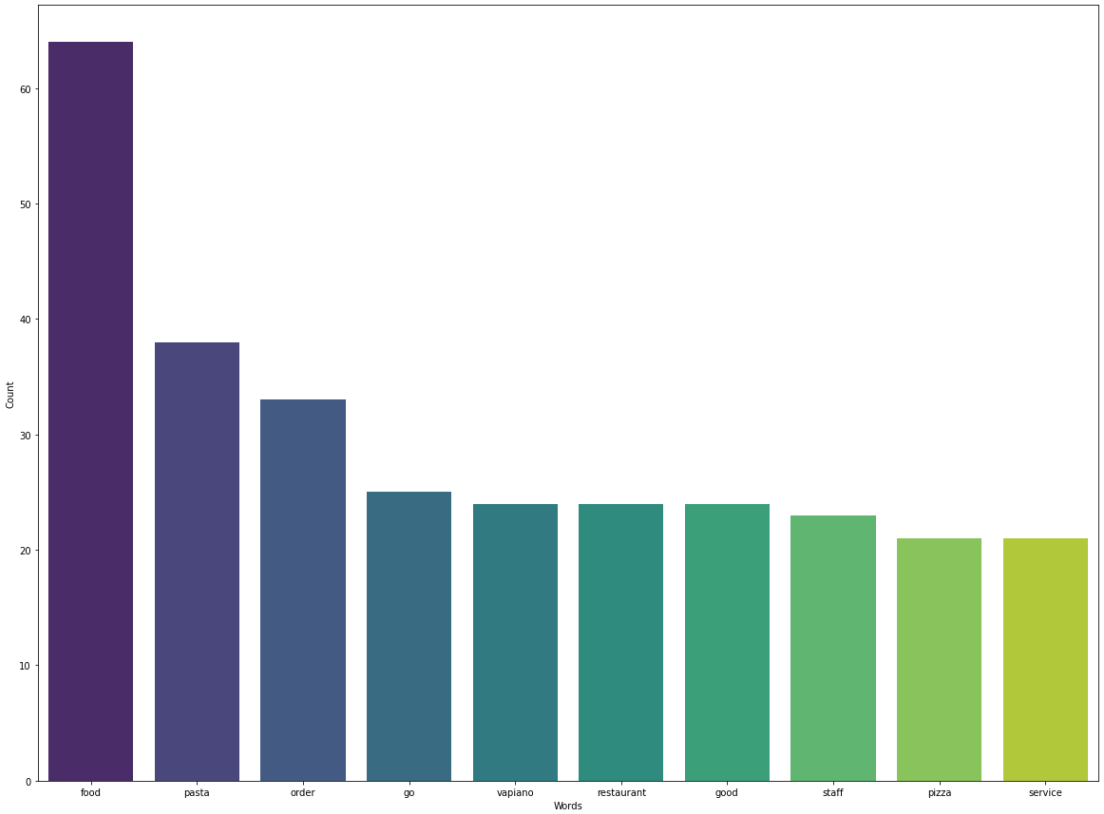
Appendix D: Subjectivity Sentiment Analysis



Appendix E: Rating Distribution



Appendix F: Frequency of the Topics



Appendix G: Sentiment Analysis on the Topics