Research Plan

a. Rationale

After the Singapore government relaxed COVID–19 travel restrictions in recent years, the number of tourists travelling in and out of Singapore has increased. As such, hotels have seen a rise in the number of tourists to be housed, and this may lead to an increase in the number of reviews hotels receive.

Today, it is common to use social networks and review websites to receive data from customer opinions. This is especially true for hotels, where occupants may evaluate the hotel based on several factors through their reviews, e. g. cleanliness, facilities, location and convenience, etc. Reviews come in two broad forms: quantitative reviews, based on stars, diamonds, hearts, etc., and qualitatives review through short, continuous text.

Quantitative reviews do not always paint the full picture of customers' opinions towards a certain hotel. Though it is helpful to have a more objective rating system using numerical scores, e. g. the Department of Tourism grading system in the Philippines, or the European Hotelstars Union system, these do not reflect the underlying reasons for giving such a rating. There is also evidence of manipulation of ratings by hotel management itself, where hotels may be compelled to forge positive or negative ratings to skew the overall rating to serve ulterior motives (TripAdvisor, 2019). We propose using sentiment analysis to extract customers' opinions on hotels from qualitative reviews as an alternative approach.

b. Research Questions and Objectives

b.i. Research Questions

1. How could we quantify the sentiments of individual words on a numerical scale?
2. How could we quantify the sentiments of paragraphs on a numerical scale?
3. How could we use tokens in hotel reviews to predict the overall sentiment of the review?

b.ii. Objectives

1. To run sentiment analysis on individual words and quantify them on a numerical scale
2. To run sentiment analysis on paragraphs and quantify them on a numerical scale
3. To use sentiment analysis on hotel reviews to determine consumers' overall opinions of hotels

b.iii. Data Sources

The hotel reviews used in this project were sourced from DataInfiniti's Business Database (https://datafiniti.co/products/business-data/) via the open-source database sharing site Kaggle (https://www.kaggle.com/datasets/datafiniti/hotel-reviews?resource=download).

c. Literature Review

c.i. Sentiment Analysis

Sentiment analysis is a field of study which utilises computational methods to analyse text, and then categorises the text, usually into polarities—positive, neutral and negative. It has broad applications which range from determining consumers' opinion in sales and product analysis, to competitor research in marketing, and even detecting public opinion in social media monitoring. Sentiment analysis will be used in this project to analyse hotel occupants' reviews, and also determine the most significant upsides and downsides of each hotel, without interference from human bias.

The stock market opinion of StockTwits communities of expert investors was predicted. Sentiment analysis was used in a Deep Learning model to extract sentiment from Big Data. A Pearson Correlation Coefficient combined the linear correlation between users' sentiment and future stock prices, which proved the accuracy of user sentiment to be 53 %. It was concluded that convolutional neural networks (CNN), a type of Deep Learning artificial neural network, was able to predict stock market movement based on sentiment (Sohangir et al., 2018).

Using the social networking site Twitter, Filipinos' sentiments towards the Philippine government's efforts at tackling COVID–19 through vaccination programmes were determined

(Villavicencio et al., 2021). Sentiment analysis was used to extract sentiment from text in English and multiple Filipino languages, which was then used to train a Naïve–Bayes model. The model classified sentiments into positive, neutral and negative categories. It was concluded that the model had a high accuracy of 81.77 %, even helping the Philippine government better conduct budget planning and coordinate COVID–19 efforts.

Tourism quality in Spain was analysed (Borrajo-Millán et al., 2021) by extracting sentiment from reviews by Chinese tourists on the tourism social networking sites Baidu Travel, Ctrip, Mafengwo, and Qunar. Two sentiment analysis methods, lexicon-matching and corpus-based machine learning methods, were used. These methods allow the processing of unstructured text of comparatively longer lengths. Clustered data visualisation categorised aspects of Spanish tourism into positive and negative groups, with the majority holding a positive sentiment. It was concluded that sentiment analysis could be used to improve tourism quality and sustainability decision-making.

SentiStrength, a tool for lexical sentiment analysis was used to study emotions expressed in GitHub commit comments of different open-source projects (Guzman et al., 2014). Their method involved assigning scores to each word, then calculating the net score for each comment. SentiStrength splits each comment into snippets, assigns each a score by computing the maximum and minimum scores of the sentences it contains. Following which, the average of the positive and negative scores is taken as the sentiment score of the entire commit. This study showed that Java projects warranted more negative comments, and projects which had more distributed teams tended to be received more positively.

c.ii. Conclusion

In conclusion, the literature reviewed showed many possible applications of sentiment analysis in quantifying the underlying emotion of feedback on online platforms. Lexicon-based sentiment analysis, which assigns each word a sentiment, then calculates a sentence's total sentiment score, can be used, due to its simplicity in implementation, and the availability of many open-source sentiment lexicons. In addition, this strategy makes accurate predictions upwards of 70 % of the time (Khoo and Johnkhan, 2017).

SentiStrength would also be useful for detecting sentiment from hotel reviews which are usually short in length quickly and efficiently. Using SentiStrength for sentiment generation is also rather accurate, generating both positive and negative sentiments with more than 60 % accuracy (Thelwall et al., 2010). Therefore, the strategies listed above could be adopted or

emulated on a smaller scale for this project.

d. Methodology

d.i. Research Question 1

How could we quantify the sentiments of individual words on a numerical scale?

1. Source for a lexicon, as well as a dataset of international hotel reviews in English
2. Split each review into individual tokens and remove stop words
3. Assign each token a single sentiment score

d.ii. Research Question 2

How could we quantify the sentiments of paragraphs on a numerical scale?

1. Make use of SentiStrengthh to evaluate positive and negative sentiments
2. Determine the overall polarity (positive, negative or neutral) of each review

d.iii. Research Question 3

How could we use tokens in hotel reviews to predict the overall sentiment of a review?

1. Determine relative token frequency in review dataset
2. Build logistic regression and random forest classifier machine learning models fed with dataset obtained above
3. Evaluate accuracy of models by using receiver operating characteristic curve and precision-recall graph

e. References

Borrajo-Millán, F., Alonso-Almeida, M.-d.-M., Escat-Cortes, M., & Yi, L. (2021). Sentiment analysis to measure quality and build sustainability in tourism destinations. *Sustainability*, *13*(11). https://doi.org/10.3390/su13116015

Guzman, E., Azócar, D., & Li, Y. (2014). Sentiment analysis of commit comments in GitHub: An empirical study. https://www.researchgate.net/profile/Emitza_Guzman/publication/266657943_Sentiment_analysis_of_commit_comments_in_GitHub_An_empirical_study/links/5b8305ba4585151fd134f10c/Sentiment-analysis-of-commit-comments-in-GitHub-An-empirical-study.pdf

Khoo, C. S. G., & Johnkhan, S. B. (2017). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*. https://hdl.handle.net/10356/83570

Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep learning for financial sentiment analysis. *Journal of Big Data*, *5*(3). https://doi.org/10.1186/s40537-017-0111-6

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*.

TripAdvisor. (2019). *2019 TripAdvisor Review Transparency Report*. https://www.tripadvisor.co.id/TripAdvisorInsights/wp-content/uploads/2019/09/2147_PR_Content_Transparency_Report_6SEP19_US.pdf

Villavicencio, C., Macrohon, J. J., Inbaraj, X. A., Jeng, J.-H., & Hsieh, J.-G. (2021). Twitter sentiment analysis towards COVID–19 vaccines in the Philippines using Naïve–Bayes. *Information*, *12*(5), 204. https://doi.org/10.3390/info12050204