Do tourists have different experiences depending on where they visit? Text mining and sentiment analysis of reviews for Patong and Choeng Thale.

Consumer reviews have fast become the primary source of information used to influence other consumers to buy products, particularly in the tourism industry. Tourism businesses need to use these reviews to provide a service which meets traveller expectations and maintain competitiveness. Due to the sheer number of reviews generated, text analysis and sentiment analysis are important tools to automatically generate insights into what reviews frequently say. This case study used reviews scraped from TripAdvisor for hotels, restaurants, and bars in Patong and Choeng Thale in Thailand to identify what tourists say about the venues in these two areas and the associated sentiment. Quality of food, service, staff, atmosphere, and cost are the most important areas for tourists in both locations, especially concerning either negative or positive reviews. Despite still requiring further development, text mining and sentiment analysis have the ability to improve customer experience as well as business performance in the tourism sector.

1. Introduction

The generation of Web 2.0 has involved a massive growth in user generated content, particularly through social networks, blogs and forums (Valdivia et al., 2017). This has led to a massive increase in consumer-generated opinion, causing a shift in where consumers get their information before making purchase decisions (Barreda and Bilgihan, 2013). Consumer opinion has become increasingly important to potential consumers and so is increasingly important to consider in modern-day business strategies and marketing ideas (Berezina et al., 2016, Barreda and Bilgihan, 2013). This is especially true for the tourism industry, where social networks have caused dramatic change to the tourism sector and play a huge role in the success of this industry (Nicoli and Papadopoulou, 2017). Tourist experience is now one of the most valuable measures of tourism practice today, with negative experiences having the ability to taint the entire view of an area (Aljerf, 2015, Yu and Egger, 2021). There are now dedicated websites for travellers to review their experiences of locations, hotels, food venues and tourist spots. Accessing this customer feedback can be used to inform stakeholder’s strategies for marketing and improvement of services, such as improvement of hotel service quality and management (Dina, 2020). These sites often include rating systems, which can inform both consumers and providers of the overall opinion for a tourism service (positive or negative) (Dina, 2020). However, without reading the reviews, targeting areas for improvement is impossible just using this score. Often the sheer number of reviews generated makes manually reading them a challenge and the unstructured format of review data makes traditional data mining techniques near impossible to apply (Cantallops and Salvi, 2014). One of the new major areas for research in the tourism sector is utilising text mining to detect key terms across reviews and the application of sentiment analysis to detect the predominant emotions attributed to these terms (Alaei et al., 2019). So far text mining and sentiment analysis has been utilised to identify important attributes for travellers using tourism-based services (Cheng and Jin, 2019, Kuhzady and Ghasemi, 2019), to suggest improvements to service-providers (Dina, 2020), to identify management-strategy opportunities (Hu et al., 2022) and to understand the overall associations of an entire tourist destination (Ali et al., 2021, Micera and Crispino, 2017).

This analysis applies text mining techniques and sentiment analysis on reviews of scraped from Trip Advisor with the aim of showcasing the usefulness of the results for stakeholders, such as owners of businesses which cater to tourism and those interested in the reputation of tourist locations. The reviews were gathered from various hotels, bars and food venues in the tourism-heavy province of Phuket in Thailand. Two locations, Patong and Choeng Thale, were selected to compare using a sample of 15 venues from each location. Two software which can be used for texting mining techniques are R and SAS Enterprise Miner (SAS EM). R is a free open-source software for statistical computing, with a variety of add-on packages which provide many built-in functions, including text mining and sentiment analysis (R Core Team, 2021). SAS EM is a software which claims to be able to “streamline the data mining process to develop models quickly” (SAS Institute Inc., 2015). This tool aims to allow non-technical users with limited coding ability to easily generate models and results which are easy to interpret. An additional aim is to compare the results generated by these two platforms.

2. Literature review

Online reviews are now an important part of the tourism sector. Acting as a form of ‘electronic word-of-mouth’, reviews provide honest feedback of a tourist’s experience without the presence of commercial interference (Kaplan and Haenlein, 2010). They often include details of the tourist’s experience, any recommendations and emotions felt (Wilson et al., 2012). The unbiased nature of user reviews offers increased reliability and trust for consumers, so-much-so that reviews by other consumers have now become the most preferred and most influential source of information when making decisions (Li and Hitt, 2008, Barreda and Bilgihan, 2013, Berezina et al., 2016). As destination experiences, such as accommodation and attractions, are difficult to evaluate prior to visiting, many tourists will browse online reviews in order to gain an overview of an area and to reduce the risks associated with booking places unknown (Filieri et al., 2015). It is reported that 97.7% of users of the popular tourism review site TripAdvisor were influenced by reviews and 97% of users revisited the site in order to plan their next trip (Gretzel and Yoo, 2008, Barreda and Bilgihan, 2013). These reviews are not only useful to consumers, however. Online reviews offer a free source of feedback for tourism-based businesses to improve stakeholder decision making and improve consumer experience (Dina, 2020). Reviews have the ability to influence the popularity and performance of businesses, so taking advantage of consumer opinions is necessary if businesses intend to increase their competitive advantage (Xie et al., 2016, Nieto et al., 2014, Simeon et al., 2017).

The issue the tourism sector faces with obtaining meaningful information from textual reviews is the sheer amount generated and the lack of structure to text. Gaining insight into tourist experience and attitude would be time-consuming and difficult for companies to complete manually (Alaei et al., 2019). This has led to increasing use of computer-based algorithms to automatically produce insight from large amounts of textual data. Text mining and sentiment analysis dates back to the 1970s but only recently gained the attention of practitioners in recent time with the advent of big data (Alaei et al., 2019). The process involves using computers to clean and process quantitative data into qualitative data which can then be used to generate some insight. Research into application of this data mining technique in the tourism industry has rapidly increased over the years, with the argument being made that despite sentiment analysis being underdeveloped compared to other data mining techniques, it is an important tool for the tourism industry (Alaei et al., 2019). Using text mining and sentiment analysis in the tourism industry to understand aspects that are important to tourist experience has already begun. Investigation into Airbnb user reviews identified three attributes important for Airbnb users (locations, amenities and host) and found there was an overall positivity bias in Airbnb reviews (Cheng and Jin, 2019). It was suggested that this could be due to users having more of a relationship with hosts compared to staff in hotels, as Airbnb reviews often contained the names of their hosts. Traveller’s destination satisfaction was found to be determined by opinions on location, room quality, staff and restaurants when text mining TripAdvisor hotel reviews (Kuhzady and Ghasemi, 2019). Sentiment analysis conducted on five tourist attractions in Tennessee found that visitor’s reviews tended to have a more positive emotional response (Lee et al., 2020). These investigations can be used to inform business strategy. Dina (2020) suggested that word clouds could be used to improve hotel facilities, such as cooperating with a restaurant to avoid food complaints. Another study used sentiment analysis of reviews to investigate the roles of different staff in providing customer experience and identify which staff were associated with positive emotions (Hu et al., 2022). Beyond tourism businesses, sentiment analysis can also be used to assess and improve the reputation of a destination with tourists. Ali et al. (2021) found that tourists found the city of Marrakech to be dirty and disliked the abuse of animals and behaviour of sellers in the market square when looking into TripAdvisor reviews. Furthermore, Micera and Crispino (2017) showed that Naples has a reputation of being a centre of art, culture and events.

3. Methodology

3.1 Data preparation

The dataset used contained reviews scraped from TripAdvisor.com for various hotels, bars and food venues within the Phuket province of Thailand. Data gathered complies with the UK data protection act 2018 as non-personal details were used, with each reviewer given an ID number of random digits, reviews scraped are publicly available and users who submit reviews agree to the terms and conditions of the TripAdvisor website which complies with GDPR laws. The venues were separated according to the area located in within the Phuket province and the date of the review was also included. There were 49944 individual reviews for 537 venues across 25 locations. A sample of 30 venues was utilised for this analysis (R code details can be found in Appendix A).

In order to select a useful sample of venues, some data exploration and preparation of the dataset needed to be conducted. This section was conducted using R and used the ‘tidyverse’ package (Wickham et al., 2019), due to its flexibility when altering data frames. There was not any missing data, however there was 3700 duplicate rows which were removed. The number of venues per location ranged from 1 to 163. None of the locations contained exactly 30 venues. I then looked into how many reviews each venue had as selecting the venues with the most reviews may provide the best analysis due to potentially containing more information to work with. However, most of the hotels had 100 different reviews each, with only 5 containing more than 100. The maximum number of distinct reviews for a venue was 200.

Researching the reputation of each location for tourists in the Phuket province, I decided to select 15 venues from Patong and 15 venues from Choeng Thale to compare. Patong is a beach city and one of the most popular tourist locations, with an active nightlife (Thai Holiday, 2022b). Choeng Thale on the other hand is known for being quieter and popular for being close to nature and for the hiking-trails (Thai Holiday, 2022a). Comparing the language used and the sentiment for these two very different tourist locations would be interesting. For consistency, all strings were converted to lower case using the functions ‘mutate\_each’ and ‘tolower’. Additionally, the locations contained whitespace at the front of the string so was removed using the function ‘trimws’. Two data frames were created which contained only data for Patong or Choeng Thale. 15 venues were randomly sampled from each data frame and then these two data frames were combined using the function ‘rbind’, which combines data frames with the same column names. This data frame was then used for analysis.

3.2 Text mining and Sentiment analysis

Text mining uses natural language processing (NLP) to gain insight into unstructured textual data, for example, the reviews used in this instance (Shekhar, 2021). To create structured data which can be interpreted by computer algorithms, individual words (terms) are extracted from the input data. These terms can be used for a multitude of applications, including sentiment analysis, where algorithms can be used to identify positive or negative emotions associated with each term (Shekhar, 2021). The overall opinion of the whole can then be generated.

3.2.1 R Implementation

The full R code for text mining can be found in Appendix B. Once the sample of 30 venues had been loaded into R, the dataset was divided into a Patong dataset and a Choeng Thale dataset, in order to compare the two areas. Additionally, to reduce the number of rows and repeated values (such as the venue name), the reviews for each hotel were collapsed into one row, making sure to include a space between the joined reviews so words can be separated easily. Following this the text within the reviews needed to be cleaned up to allow text analysis and sentiment analysis. The reviews contained digits and leftover HTML/website features features (such as ‘…more” and ‘\n’. These specific features were removed simply by substituting the strings with whitespace. For more specific preparation for text mining, the package ‘tm’ was used (Feinerer, 2013). This package offers a framework for text mining, including preparing the data. Firstly, data frames needed to be converted into a corpus object using the function ‘Corpus’. A corpus object organises the review data into a format suitable for text mining (Feinerer, 2013). Additional features needed to be removed to allow remove terms which are not useful for the analysis, such as removing stop words (common words which have little meaning alone, such as ‘and’, ‘is’), removing punctuation, removing numbers and removing white space. I also opted to remove ‘choeng’, ‘patong’, ‘phuket’ and ‘thailand’ as the reviews are likely to reference the location, however this is not informative. There is also a function to stem the remaining words, which involves removing affixes from terms to reveal the base form (for example ‘disappointed’, ‘disappointment’, ‘disappointing’ are grouped under the base form ‘disappoint’). The two corpus were each converted into a document term matrix. This stores the frequency of each word for each document (review).

The most frequent terms used in reviews for each location could then be visualised using the package ‘ggplot’ (Kahle and Wickham, 2013), allowing comparison [fig.1&2]. Due to the large number of terms, many with small frequencies, a subset was taken containing only words with a frequency of above 90. This retained 48 words for Patong and 47 words for Choeng Thale. A word cloud diagram is a very useful method for visualising frequent terms and facilitating easier comparison between the two locations. These can be produced using the package ‘wordcloud’ (Fellows et al., 2018). A cluster of words is produced, where the size and colour of the word represents its frequency. As the top five most frequent words were the same for both locations (‘food’, ‘good’, ‘restaur’, ‘great’, ‘servic’), these were removed to better identify any potential differences in the most frequently used words between the two locations. As the top five frequent words are the same between locations, it may be useful to look at which words are frequently found with these, as they may be used differently according to location. As each of the venues may have generated different opinions, an individual has the ability to affect the overall frequency of words associated with each location. This is most evident with the different cuisines featuring in the high-frequency words, such as ‘seafood’, ‘pizza’, ‘chicken’ all in high-frequency in reviews for Patong. Therefore, I also looked into the top five frequent terms for each venue in the sample. This was accomplished by using the package ‘corpus’ (Perry, 2021), in particular, the function ‘as\_corpus\_frame’. This function transformed the original corpus objects for the two locations into an object which is similar to a data frame and so could be bound to the original individual location’s data frame. This associated the venue name with its corresponding review, although the version of the review which was prepared for text mining, so the process did not have to be repeated. When looking at the top five most frequent terms for each venue, the terms ‘food’, ‘servic’, restaur’ and ‘staff’ were common across the majority. Alone these terms do not allow for much analysis, particularly when common across venues and it makes sense that these should be the most common given the venues reviewed are establishments providing food and services to tourists. Additionally, some of the venues contained the single letter ‘e’ or ‘u’ as a frequent term, which are likely artefacts leftover from data processing and so needed to be removed. These terms were removed using the function ‘unnest\_tokens’ from the package ‘tidytext’ (Silge and Robinson, 2016), as the cleaned review column contained all terms in one column per venue so to extract individual words, the column needed to be split into individual tokens. This allowed text analysis to extract meaningful information which differentiates the venues.

Chart, bar chart

Description automatically generated

**Figure 1:** Frequency of each word found in Patong corpus (frequency ≥ 90).

Chart

Description automatically generated

**Figure 2:** Frequency of each word found in Choeng Thale corpus (frequency ≥ 90).

For sentiment analysis, the package ‘syuzhet’ was used; a popular open-source tool for extracting the emotions associated from text (Jockers, 2017).The package has access to four sentiment lexicons: ‘Syuzhet’, ‘afinn’, ‘bing’ and ‘nrc’. The lexicon ‘nrc’ was used to extract sentiment regarding Patong and Choeng Thale, as it offers more than just distinguishing ‘positive’ or ‘negative’ association of words, like other sentiment lexicons, but also further separates terms into ‘anger’, ‘fear’, ‘anticipation’, ‘trust’, ‘surprise’, ‘sadness’, ‘joy’ and ‘disgust’. The full R code for this section can be found in Appendix C. Looking at the distribution of these further eight basic emotions may be interesting to compare between the locations. The ‘nrc’-based sentiment for each term was extracted using the function ‘get\_nrc\_sentiment’ and then the frequency of each emotion was determined for Patong and then Choeng Thale. It was also decided to investigate which terms were high in frequency for the positive and negative sentiment using word clouds. Although creating word clouds for each basic emotion may be interesting, in this instance it was decided just to look at positive and negative sentiment, as these present a general sentimental association which would be applicable to all interested stakeholders. The same extraction of emotions was conducted for individual venues, as individual venues may affect overall sentiment, such as if one venue has particularly negative reviews. Additionally, this showcases the ability of sentiment analysis to benefit stakeholders interested in the broader scale of area reputation, as well as on the smaller scale of stakeholders interested in the reputation of individual venues. To simplify the resulting analysis, only negative and positive sentiment were included in the final results per venue.

3.2.1.1 Additional task: Shiny Dashboard

A shiny dashboard to present the results from the text mining and sentiment analysis was produced using the packages ‘shiny’ (Chang et al., 2021) and ‘shinydashboard’ (Chang and Borges Ribeiro, 2021). The full R code for this can be found in Appendix D. As the text mining used in this investigation has the potential to benefit stakeholders interested in the reputation of Patong and Choeng Thale, as well as those interested in the reputation of a particular business, two separate corresponding pages were created called ‘Location’ and ‘Venues’ [fig.3&4]. The ‘Location’ page contained word clouds for each location which presented the frequency of words (related to word size) for Patong and Choeng Thale [fig.3]. Additional, word clouds for each location were included which presented words associated with a particular sentiment. The user can select the desired sentiment from a list containing ‘anger’, ‘fear’, ‘anticipation’, ‘trust’, ‘surprise’, ‘sadness’, ‘joy’ and ‘disgust’. The number of words featured in all word clouds can be altered using the slider, with the maximum number of words set to 50.

Graphical user interface, application

Description automatically generated with medium confidence

**Figure 3:** ‘Location’ page of shiny dashboard.

The ‘Venue’ page contained two bar plots containing the top five most frequent terms and the frequency of terms associated with each sentiment for an individual business [fig.4]. Users can select the location of the business and then select the desired business from the corresponding list of venues.

Chart, bar chart

Description automatically generated

**Figure 4:** ‘Venue’ page of shiny dashboard.

3.2.2 SAS Implementation

Firstly, the data was imported into SAS EM using the file import node, as all textual data was contained in one excel file. The role of ID was assigned to the venue name and the role of text was assigned to the review, as necessary for text analysis in SAS EM. The two separate locations were then separated using the filter node to allow comparison of the terms used to review these separate areas. The filter nodes were renamed corresponding to the location of the data retained, ‘Patong’ or ‘Choeng Thale’. The same text analysis process was then applied to each of the two datasets, as can be seen in figure 5.

A picture containing text, sky, screenshot

Description automatically generated

**Figure 5:** SAS EM workflow for text mining of Patong and Choeng Thale reviews.

The text parsing node was attached next. This node is used to prepare and clean the text. Looking at the term frequency matrix produced can be used to improve preparation, by identifying any odd terms retained, and can be used to direct analysis by looking at the most frequently used terms. Some of the default settings were altered in this node. ‘Detect: Different Parts of Speech’ was set to yes, with ‘Ignore Parts of Speech’ excluding all except abbreviations, adjectives, adverbs, nouns, verbs and verb-adjectives. These types of terms are likely to provide the most semantically interesting results due to them being the most major forms in sentences. Additionally, ‘Find Entitities’ was set to ‘Standard’ and ‘Ignore Types of Entities’ excluding all entity types, except ‘Company’, ‘Organisation’, ‘Person’, ‘Product’ and ‘Prop\_misc’, as it is possible reviews may mention particular venues, people and products provided, such as the food provided and particular staff/venues which could be linked to specific opinion-terms. ‘Ignore Types of Attributes’ involved ‘Num’ and ‘Punct’ so both numbers and punction could be excluded. Looking at the terms retained by this step, it appears further cleaning is required, as website-related features appear to have been retained (‘x000d’). Location-related terms, such as ‘phuket’, ‘choeng’ and ‘thailand’ have also been retained, which provide little semantic value in the text analysis.

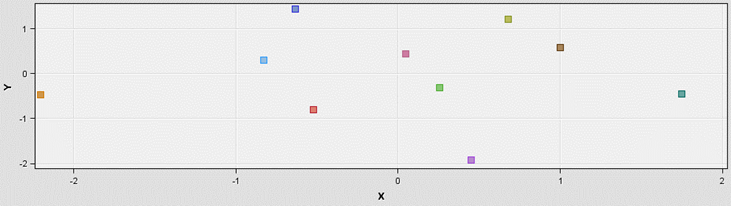
Graphical user interface, table

Description automatically generated

**Figure 6:** Interactive ‘Filter Reviewer’ with terms ordered by frequency.

The text filtering node allows more specific cleaning of the terms using the interactive ‘Filter Reviewer’ [fig.6]. Here terms can be manually removed, or multiple words can be grouped as synonyms and so will be represented by one word. Grouping similar terms was already attempted by SAS EM (indicated by a ‘+’ next to terms), however sometimes this is not perfect at catching all. With ‘Check Spelling’ set to ‘Yes’, any potential misspelled words can be detected and associated with the correctly spelled term. During this stage multiple terms were removed from both the Patong data and the Choeng Thale data, including website-related terms (‘x000d\_i’, ‘f0’, ‘009f’), location related terms (‘thailand’, ‘phuket’) and time related terms (‘year’, ‘month’, ‘minute’). Additionally, some terms required adding to synonym groups. Some terms were linked to website-related features, such as ‘\_x000d\_food’ and ‘\_x000d\_staff’ and so were not able to be linked to their corresponding terms (‘food’ and ‘staff’ respectively). If spotted these terms were allocated as being synonymous with their base term. With the number of terms provided in the ‘Filter Viewer’ it was not possible to manually group all items and retain all potentially missed terms, so there may have been some missed. The resulting concept links (which terms associated with selected terms) were investigated and recorded in the results analysis section.

The final stage is the text mining stage. One of the nodes used for this stage is ‘Text’ Cluster’, which clusters the documents into sets and then produces the descriptive terms for that cluster. The maximum number of clusters was set to 15 and the maximum number of terms per cluster set at 8. These values were selected to avoid overcomplicated results. The Patong dataset produced 10 clusters which appeared to be relatively even sized and well-spaced apart [fig.7A&B], however, the Choeng Thale clusters produced did not share these attributes [fig.8A&B]. The maximum number of clusters was lowered for the Choeng Thale data to 12, which produced 11 clusters that appeared to be more acceptable [fig. 9A&B].



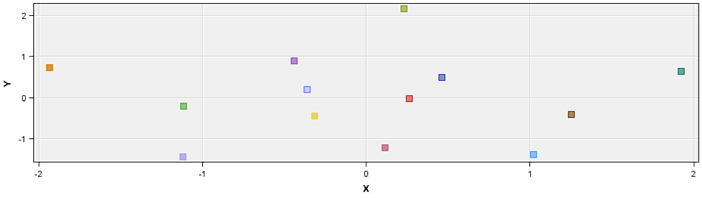
**A**

**B**

**Figure 7:** Maximum clusters set to 10 for Patong. **A)** Pie chart with size of segment indicating frequency. **B)** Distance between clusters.

**A**

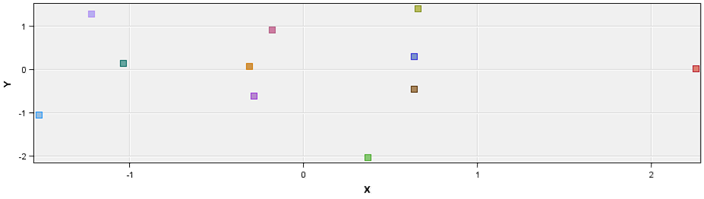
**B**



**Figure 8:** Maximum clusters set to 10 for Choeng Thale. **A)** Pie chart with size of segment indicating frequency. **B)** Distance between clusters.

**A**

**B**



**Figure 9:** Maximum clusters set to 12 for Choeng Thale. **A)** Pie chart with size of segment indicating frequency. **B)** Distance between clusters.

The other node used was the ‘Text Topic’ node which associates terms and documents to topics, which are collections of terms that fit a general theme. Documents and terms can associate with more than one topic or no topics at all. Users have the option to define their own topics, if there are any the user is particularly interested in, otherwise the algorithm will produce its own topics. As this investigation was for exploratory purposes, the algorithm was used to produce its own topics. No single-term topics were set to be created due to the high likelihood of specific terms, such as ‘food’, which have already been identified as high in frequency being used. The maximum number of multi-term topics was set to 15, as this is only for exploration purposes and so do not want too many results. This number of topics appears to generate well separated results, with terms that are strongly associated for one topic not also being highly associated with another topic [fig.10&11].

Graphical user interface, diagram, application

Description automatically generated

**Figure 10:** Matrix showing interactions between pairs of topics for Patong by plotting terms describing topics. Rank of topics indicates frequency (1 = most frequent).

Graphical user interface, application

Description automatically generated

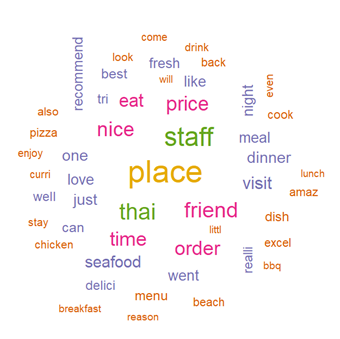
**Figure 11:** Matrix showing interactions between pairs of topics for Choeng Thale by plotting terms describing topics. Rank of topics indicates frequency (1 = most frequent).

4. Results and Discussion

The word clouds generated for Patong and Choeng Thale contain very similar frequent words, with ‘staff’, ‘place’, ‘friend’ and ‘nice’ being most frequent in both clouds [fig.12A&B]. From this it can be suggested that experience with staff and friendliness are important for reviewers due to how frequently this is mentioned. Further to this, ‘good service’ is associated with the word ‘good’ [fig.15A&B]. For Patong, ‘service’ is associated with ‘friendly service’ and ‘waitress’ and ‘rude’ were associated with ‘bad’ [fig.15A]. As food was actually the most frequent term, frequent associations with this term were investigated [fig.12A&B]. ‘Food’ was associated with opinions of ‘good’ and ‘great’ [fig.13A&B]. The term ‘service’ was also associated with ‘food’ [fig.16A&B]. Other associations appeared to be related to specific cuisines, such as Italian for Patong [fig.13A] and Indian food for Choeng Thale [fig.13B]. These cuisines are likely related to the specific venues sampled for each area. Reference to Thai cuisine is also present for both areas [fig.13A&B]. For Choeng Thale, in association with ‘bad’ the food ‘chicken’ appeared [fig.17B]. Some interesting terms associated with ‘restaurant’ were ‘beach’ for Patong [fig.14A] and ‘air’ and ‘design’ for Choeng Thale [fig.14B]. These potentially suggest the importance of aesthetics and desire to be close to the outdoors for customers. Other additional atmosphere-related terms included ‘music’ associated with ‘good’ [fig.15A] and ‘busy’ associated with ‘bad’ in Choeng Thale [fig.17A]. Pricing is also a topic which appears frequent among reviews, with price being strongly associated with ‘good’ [fig.15A&B] and food being linked to ‘pricey’ in Patong [fig.13A] and for Choeng Thale ‘overprice’ being associated with ‘bad’ [fig.17B]. The predominant takeaway from investigating word frequencies is that food and service quality are the two most important factors for consumers. Consumers appear to value friendly service provided by staff and desire that a service is of good quality for a reasonable price. Although Thai cuisine is popular, tourists will also frequent other cuisines, so are not only interested in the destination country’s cuisine. Food is considered one of the most important aspects of tourist experience and another study on tourist reviews for Phuket also found that food was the most frequent term used by tourists (Sangkaew and Zhu, 2022). The frequent terms associated with Patong and Choeng Thale are similar to those found in another study, where ‘staff’ and ‘restaurant’ were frequently used in positive reviews (Kuhzady and Ghasemi, 2019). Any businesses looking to improve should focus on providing a quality experience, with friendly, attentive staff, a good atmosphere, quality food and good overall service, which should justify any pricing.



**B**

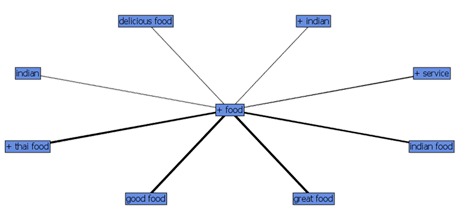
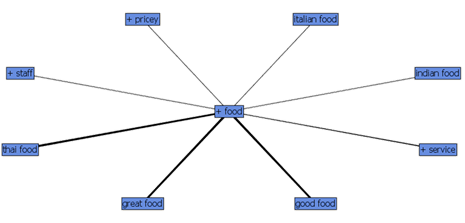


**A**

**Figure 12:** Word clouds with size and colour indicating relative frequency of terms. **A)** Patong. **B)** Choeng Thale.

**B**

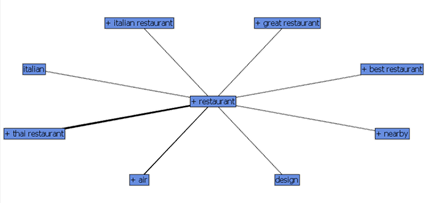
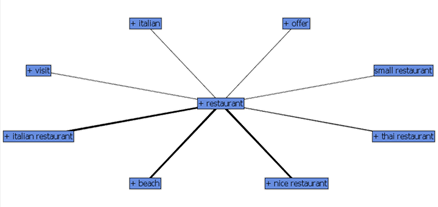
**A**



**Figure 13:** Terms associated with food. **A)** Patong. **B)** Choeng Thale.

**B**

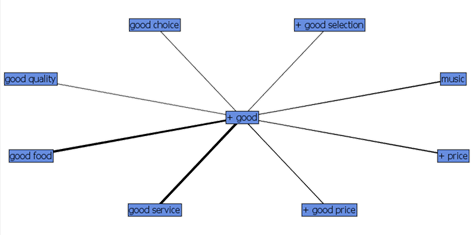
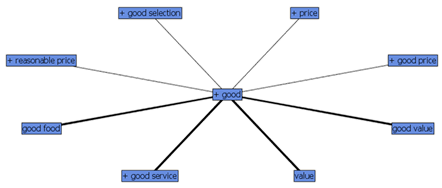
**A**



**Figure 14:** Terms associated with restaurant. **A)** Patong. **B)** Choeng Thale.

**B**

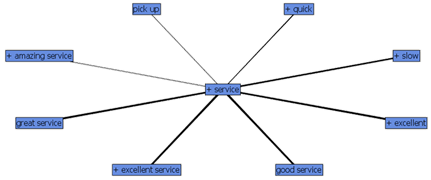
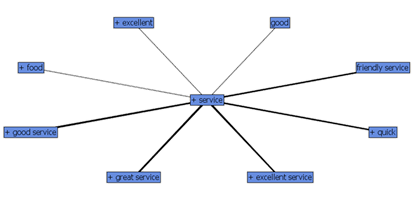
**A**



**Figure 15:** Terms associated with good. **A)** Patong. **B)** Choeng Thale.

**B**

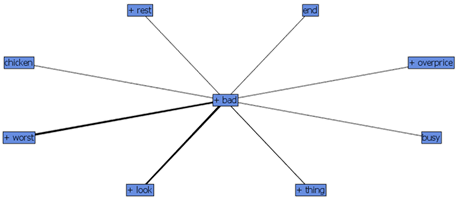
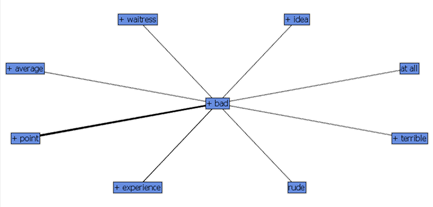
**A**



**Figure 16:** Terms associated with service. **A)** Patong. **B)** Choeng Thale.

**B**

**A**



**Figure 17:** Terms associated with bad. **A)** Patong. **B)** Choeng Thale.

Looking at the clusters and topics produced SAS EM for Patong, there appears to be some similarity with terms associated [table 1&2]. Most of the clusters and topics produced appear to be related to specific venues, such as the restaurant White Box, Don’s bar bq, the Holiday Inn resort and the Sea Hag. These are not very useful when interested in general tourist views on Patong. If interested in individual venues, the extra analysis which plotted the top 5 words for each restaurant in the Patong sample was more useful [fig.18]. A few interesting findings included the live music at the Port and the views associated with the White Box being particularly popular among reviewers [table 1&2]. The clusters with those most associated terms and the topics with the most reviews associated appear to emphasise the importance of reasonable pricing, quality food and friendly staff for customers and that customer’s do not like waiting [table 1&2]. Another interesting cluster with frequent terms and topic with many reviews associated suggests that tourists in Patong enjoy music and drinking in the evening [table 1&2]. This would make sense with the nightlife of Patong being particularly popular and well-associated with the area as a holiday destination (Thai Holiday, 2022b). Additionally, once more the atmosphere appears to be important to reviewers [table 1&2]

**Table 1:** Clusters produced by SAS EM for Patong with the frequency of each cluster and the corresponding percentage.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster ID** | **Descriptive Terms** | **Frequency** | **Percentage** |
| 1 | +Italian, +bbq, +don, +pizza, +meat, +pasta, +salad, +eat | 127 | 0.09065 |
| 2 | +beach, +view, +white, box, +road, +dinner, +recommend, +restaurant | 127 | 0.09065 |
| 3 | Sea, hag, 'sea hag', +disappoint, +day, 'thai food', +thai, +want | 48 | 0.034261 |
| 4 | Chicken, rice, fried, +'pad thai', +pad, +salad, +fish, +curry | 136 | 0.097074 |
| 5 | +holiday, inn, +stay, resort, 'holiday inn resort', charm, +Indian, +thai | 58 | 0.041399 |
| 6 | +seafood, fresh, +prawn, +crab, +dish, +fish, +cook, +curry | 157 | 0.112063 |
| 7 | +great, music, +band, +atmosphere, +drink, +visit, +night, +time | 218 | 0.155603 |
| 8 | +pizza, +table, +wait, +order, +bad, +meal, +menu, +road | 223 | 0.159172 |
| 9 | +breakfast, café, +coffee, +egg, bacon, siam, +owner, +real | 78 | 0.055675 |
| 10 | +good, +price, reasonable, +friendly, quality, +nice, +food, +staff | 229 | 0.163455 |

**Table 2:** Topics produced by SAS EM for Patong with the number of terms and number of documents associated.

|  |  |  |  |
| --- | --- | --- | --- |
| **Topic ID** | **Topic** | **Number of Terms** | **Number of Docs** |
| 1 | +view, box, +white, +white box, +dinner | 54 | 83 |
| 2 | fried, rice, +curry, +order, +green | 67 | 160 |
| 3 | +bbq, +pork, +don, +meat, +salad | 43 | 97 |
| 4 | +holiday, inn, resort, +stay, holiday inn resort | 38 | 83 |
| 5 | +breakfast, cafe, +coffee, +owner, bacon | 41 | 93 |
| 6 | +pizza, +italian, +pasta, +good, +wood | 46 | 112 |
| 7 | +seafood, fresh, +prawn, +cook, +fish | 45 | 163 |
| 8 | +visit, +time, +menu, +restaurant, +last | 67 | 198 |
| 9 | +nice, music, +night, +band, +drink | 75 | 179 |
| 10 | +road, +beach, +restaurant, +street, main | 53 | 137 |
| 11 | +great, great food, +atmosphere, +service, +recommend | 41 | 162 |
| 12 | +thai, thai food, +pad, +pad thai, +good | 52 | 163 |
| 13 | +staff, +friendly, +friendly staff, +nice, +recommend | 65 | 187 |
| 14 | +order, +place, +wait, +bad, +service | 88 | 192 |
| 15 | +price, reasonable, +good, +good, +reasonable price | 65 | 191 |

Chart, bar chart

Description automatically generated

**Figure 18:** Frequency of the top five most frequent term for each venue in the Patong sample.

The clusters and topics produced for Choeng Thale by SAS EM were also very similar and most appeared to be related to specific venues. However, upon closer inspection, some potential overall differences between experiences in Patong and experiences in Choeng Thale exist. The terms used in Choeng Thale appear to have connotations of being a high standard, such as ‘chef’, ‘wine’, ‘steak’ and ‘cocktail’ [table 3&4]. This could be due to Choeng Thale being a smaller tourist destination and so contains fewer venues of a higher quality (Thai Holiday, 2022a). However, similar to Patong, reviewers valued good food, a nice atmosphere, reasonable prices and good service [table 1&2]. There was not any reference to the term ‘bad’, which could imply that reviewers frequently had a more positive experience at Choeng Thale than those who visited Patong. When looking at the top five terms for each venue, there was mostly reference to the type of food served [fig.19]. At the bodega bar & grill, the chef was specifically mentioned. Identifying staff mentioned frequently and their impact on tourist experience can allow better allocation of awarded bonuses and so improve staff motivation (Khoo-Lattimore and Ekiz, 2014). Additionally, it appears that tourists preferred to visit the 360⁰ bar at sunset because of the available views [fig.19].

**Table 3:** Clusters produced by SAS EM for Choeng Thale with the frequency of each cluster and the corresponding percentage.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster ID** | **Descriptive Terms** | **Frequency** | **Percentage** |
| 1 | +chef, +salad, experience, +dish, +serve, +meal, +fantastic, +taste | 123 | 0.089325 |
| 2 | +Indian, 'indian food', paradise, +owner, +chicken, +order, +thai +cook | 51 | 0.037037 |
| 3 | +wine, +metzo, +laguna, +selection, +steak, +stay, +dinner, +excellent | 161 | 0.116921 |
| 4 | +thai, +'thai food', +curry, +cook, +dish, good, +hotel, +chicken | 184 | 0.133624 |
| 5 | +place, +eat, +love, +main, +bar, +drink, +stay, +visit | 121 | 0.087872 |
| 6 | Amazing, +dee, plee, +Anantara, 'dee plee', +friendly, +staff, +thai | 67 | 0.048656 |
| 7 | +view, +cocktail, +sunset, +lake, +tapa, +bar, +drink, lovely | 92 | 0.066812 |
| 8 | +price, +reasonable, music, 'good food', +good, +atmosphere, +nice, +service | 188 | 0.136529 |
| 9 | +time, +visit, +last, +table, +staff, +night, little, +friendly | 246 | 0.178649 |
| 10 | +beach, surin, +sea, 'surin beach', +breeze, +'sea breeze', +burger, +location | 58 | 0.042121 |
| 11 | +order, +pizza, Italian, +pasta, +wait, +arrive, +main, +starter | 86 | 0.062455 |

**Table 4:** Topics produced by SAS EM for Choeng Thale with the number of terms and number of documents associated.

|  |  |  |  |
| --- | --- | --- | --- |
| **Topic ID** | **Topic** | **Number of Terms** | **Number of Docs** |
| 1 | +resort, +dee, plee, +thai, +anantara | 67 | 78 |
| 2 | +beach, surin, surin beach, +day, little | 52 | 124 |
| 3 | +indian, indian food, +owner, +recommend, +eat | 77 | 142 |
| 4 | +view, +cocktail, +sunset, +tapa, +drink | 62 | 148 |
| 5 | +price, +good, +nice, +reasonable, +atmosphere | 57 | 181 |
| 6 | +sea, +breeze, +sea breeze, +burger, +beach | 71 | 89 |
| 7 | +order, +thai, +curry, +starter, +salad | 89 | 167 |
| 8 | +staff, +friendly, +friendly staff, little, +meal | 63 | 201 |
| 9 | +pizza, italian, +pasta, toto, +nice | 66 | 129 |
| 10 | +dinner, +chef, bodega, +year, experience | 90 | 197 |
| 11 | +excellent, +wine, +steak, +menu, food | 66 | 207 |
| 12 | +dinner, +lake, +nice, +lunch, +club | 80 | 144 |
| 13 | +laguna, +thai, +thai food, +area, +stay | 74 | 186 |
| 14 | +visit, +night, +highly, +recommend, +year | 80 | 139 |
| 15 | +table, +meal, +minute, +last, +wait | 94 | 190 |

Chart, bar chart

Description automatically generated

**Figure 19:** Frequency of the top five most frequent term for each venue in the Choeng Thale sample.

Finally, the sentiment for each location was gathered. Both Patong and Choeng Thale received mostly positive opinions from visitors [fig20&21]. Choeng Thale appeared to have a higher frequency of terms with a positive emotion associated, so travellers could have an overall more positive experience visiting Choeng Thale [fig.21]. The positive term wordclouds associated with both Choeng Thale and Patong serve to reinforce the idea that customers value the food and the atmosphere of the location [fig.22B&23B]. It could be argued that negative sentiment is more important to investigate as potential customers are more likely to use these to avoid wasting money and time on bad experiences. For Patong, frequent negatively associated terms included waiting, blandness, cheapness and coldness [fig.22A]. For Choeng Thale, frequent negatively associated words also referenced cheapness, blandness waiting and coldness, although there were much fewer negative terms associated with this area [fig.23A]. Overall positivity and negativity associated with reviews for individual venues allows the identification of venues which have similar amounts of each sentiment (such as the Ali Baba restaurant) and those which are overwhelmingly more positive (such as the Bodega & Grill) [fig.24&25]. The sampled venues for Choeng Thale appear to have more positive associations compared to Patong, which has more venues with a smaller difference between the frequency of negative and positive terms [fig.24&25]. This could further suggest that venues in Choeng Thale produce better experiences for travellers compared to Patong.

Chart, bar chart

Description automatically generated

**Figure 20:** Frequency of words associated with each emotion for Patong.

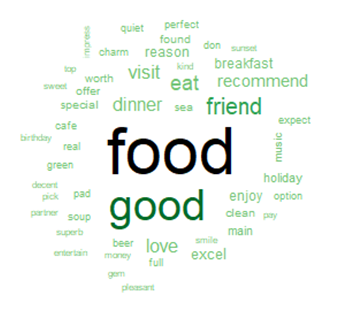
Chart, bar chart

Description automatically generated

**Figure 21:** Frequency of words associated with each emotion for Choeng Thale.

**B**

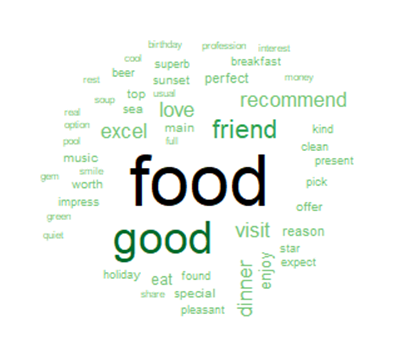
**A**



**Figure 22:** Word clouds for Patong with size and colour indicating relative frequency of terms. **A)** Negative sentiment. **B)** Positive sentiment.

**B**

**A**



**Figure 23:** Word clouds for Choeng Thale with size and colour indicating relative frequency of terms. **A)** Negative sentiment. **B)** Positive sentiment.

Chart, bar chart

Description automatically generated

**Figure 24:** Frequency of the positive and negative terms for each venue in the Patong sample.

Bar chart

Description automatically generated with medium confidence

**Figure 25:** Frequency of the positive and negative terms for each venue in the Choeng Thale sample.

5. Conclusion

Using text mining and sentiment analysis, this investigation identified that most tourists primarily expected tasty food, quality and quick service, friendly staff, and reasonable pricing. Any negative associations were often caused by the opposite of these being provided. Tourists also appreciated a good atmosphere, with nice views and they liked the live music played at particular venues. Although both locations had similar results, Choeng Thale appeared to have more positive terms associated with the sample venues than Patong. This is an example of how tourist venues can improve their reputation by focusing on quality of food, service, staff, and pricing. However, these tools are not perfect. The clusters and topics produced were primarily grouped according to the corresponding venue. This is likely because the sample is too small, and reviews associated with the same venue will likely contain the same phrases. Increasing the number of venues investigated could significantly improve the groups of terms formed, particularly for Patong which had the most venues in the original dataset. Additionally, the single terms provided as associations with other terms or in word clouds could sometimes be difficult to interpret without context. For example, ‘die’ is present in in a word cloud [fig.23A]. Without context it is difficult to understand what this term could refer to. Overall, text mining and sentiment analysis has a lot of potential but the methodology still requires some improvements.

6. References

ALAEI, A. R., BECKEN, S. & STANTIC, B. 2019. Sentiment analysis in tourism: capitalizing on big data. *Journal of Travel Research,* 58**,** 175-191.

ALI, T., MARC, B., OMAR, B., SOULAIMANE, K. & LARBI, S. 2021. Exploring destination's negative e-reputation using aspect based sentiment analysis approach: Case of Marrakech destination on TripAdvisor. *Tourism Management Perspectives,* 40**,** 100892.

ALJERF, L. 2015. Change theories drift conventional tourism into ecotourism. *Acta Technica Corviniensis-Bulletin of Engineering,* 8**,** 101.

BARREDA, A. & BILGIHAN, A. 2013. An analysis of user‐generated content for hotel experiences. *Journal of Hospitality and Tourism Technology*.

BEREZINA, K., BILGIHAN, A., COBANOGLU, C. & OKUMUS, F. 2016. Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. *Journal of Hospitality Marketing & Management,* 25**,** 1-24.

CANTALLOPS, A. S. & SALVI, F. 2014. New consumer behavior: A review of research on eWOM and hotels. *International Journal of Hospitality Management,* 36**,** 41-51.

CHANG, W. & BORGES RIBEIRO, B. 2021. shinydashboard: Create Dashboards with 'Shiny'.

CHANG, W., CHENG, J., ALLAIRE, J., SIEVERT, C., SCHLOERKE, B., XIE, Y., ALLEN, J., MCPHERSON, J., DIPERT, A. & BORGES, B. 2021. shiny: Web Application Framework for R.

CHENG, M. & JIN, X. 2019. What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management,* 76**,** 58-70.

DINA, N. Z. 2020. Tourist sentiment analysis on TripAdvisor using text mining: A case study using hotels in Ubud, Bali. *African Journal of Hospitality, Tourism and Leisure,* 9**,** 1-10.

FEINERER, I. 2013. Introduction to the tm Package Text Mining in R. *Accessible en ligne: http://cran. r-project. org/web/packages/tm/vignettes/tm. pdf*.

FELLOWS, I., FELLOWS, M. I., RCPP, L. & RCPP, L. 2018. Package ‘wordcloud’. *R package version,* 2**,** 331.

FILIERI, R., ALGUEZAUI, S. & MCLEAY, F. 2015. Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth. *Tourism management,* 51**,** 174-185.

GRETZEL, U. & YOO, K. H. 2008. Use and impact of online travel reviews. *Information and communication technologies in tourism 2008***,** 35-46.

HU, F., TRIVEDI, R. & TEICHERT, T. 2022. Using hotel reviews to assess hotel frontline employees’ roles and performances. *International Journal of Contemporary Hospitality Management*.

JOCKERS, M. 2017. Package ‘syuzhet’. *URL: https://cran. r-project. org/web/packages/syuzhet*.

KAHLE, D. & WICKHAM, H. 2013. ggmap: Spatial Visualization with ggplot2. *The R Journal,* 5**,** 144-161.

KAPLAN, A. M. & HAENLEIN, M. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons,* 53**,** 59-68.

KHOO-LATTIMORE, C. & EKIZ, E. H. 2014. Power in praise: Exploring online compliments on luxury hotels in Malaysia. *Tourism and Hospitality Research,* 14**,** 152-159.

KUHZADY, S. & GHASEMI, V. 2019. Factors influencing customers' satisfaction and dissatisfaction with hotels: A text-mining approach. *Tourism Analysis,* 24**,** 69-79.

LEE, J., BENJAMIN, S. & CHILDS, M. 2020. Unpacking the emotions behind TripAdvisor travel reviews: The case study of Gatlinburg, Tennessee. *International Journal of Hospitality & Tourism Administration***,** 1-18.

LI, X. & HITT, L. M. 2008. Self-selection and information role of online product reviews. *Information Systems Research,* 19**,** 456-474.

MICERA, R. & CRISPINO, R. 2017. Destination web reputation as “smart tool” for image building: the case analysis of Naples city-destination. *International Journal of Tourism Cities*.

NICOLI, N. & PAPADOPOULOU, E. 2017. TripAdvisor and reputation: a case study of the hotel industry in Cyprus. *EuroMed Journal of Business*.

NIETO, J., HERNÁNDEZ-MAESTRO, R. M. & MUÑOZ-GALLEGO, P. A. 2014. Marketing decisions, customer reviews, and business performance: The use of the Toprural website by Spanish rural lodging establishments. *Tourism Management,* 45**,** 115-123.

PERRY, P. O. 2021. corpus: Text Corpus Analysis.

R CORE TEAM 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

SANGKAEW, N. & ZHU, H. 2022. Understanding tourists’ experiences at local markets in Phuket: An analysis of TripAdvisor reviews. *Journal of Quality Assurance in Hospitality & Tourism,* 23**,** 89-114.

SAS INSTITUTE INC. 2015. SAS® Enterprise Miner™ 14.1: Administration and

Configuration. Cary, NC: SAS Institute Inc.

SHEKHAR, S. 2021. Text Mining and Sentiment Analysis.

SILGE, J. & ROBINSON, D. 2016. tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software,* 1**,** 37.

SIMEON, M. I., BUONINCONTRI, P., CINQUEGRANI, F. & MARTONE, A. 2017. Exploring tourists’ cultural experiences in Naples through online reviews. *Journal of Hospitality and Tourism Technology*.

THAI HOLIDAY. 2022a. *Choeng Thale* [Online]. Available: https://thaiholiday.com/choeng-thale [Accessed 2022].

THAI HOLIDAY. 2022b. *Patong Town* [Online]. Available: https://thaiholiday.com/patong-town [Accessed 2022].

VALDIVIA, A., LUZÓN, M. V. & HERRERA, F. 2017. Sentiment analysis in tripadvisor. *IEEE Intelligent Systems,* 32**,** 72-77.

WICKHAM, H., AVERICK, M., BRYAN, J., CHANG, W., D'AGOSTINO MCGOWAN , L., FRANÇOIS, R., GROLEMUND, G., HAYES, A., HENRY, L., HESTER, J., KUHN, M., LIN PEDERSEN, T., MILLER, E., MILTON BACHE, S., MÜLLER, K., OOMS, J., DAVID, R., SEIDEL, D. P., SPINU, V., TAKAHASHI, K., VAUGHAN, D., WILKE, C., WOO, K. & YUTANI, H. 2019. Welcome to the {tidyverse}. *Journal of Open Source Software,* 4**,** 1686.

WILSON, A., MURPHY, H. & FIERRO, J. C. 2012. Hospitality and travel: The nature and implications of user-generated content. *Cornell hospitality quarterly,* 53**,** 220-228.

XIE, K. L., CHEN, C. & WU, S. 2016. Online consumer review factors affecting offline hotel popularity: evidence from tripadvisor. *Journal of Travel & Tourism Marketing,* 33**,** 211-223.

YU, J. & EGGER, R. 2021. Tourist experiences at overcrowded attractions: a text analytics approach. *Information and Communication Technologies in Tourism 2021.* Springer.

7. Appendix

**Appendix A**

R code used for dataset exploration and preparation.

library(tidyverse)

##DATASET PREPARATION

rm(list=ls())

#set working directory

setwd("C:\\Users\\smann\\OneDrive\\University\\Data Science\\ASDM\\Project\\Sentiment analysis")

#import dataset

df <- read.csv("tourist\_accommodation\_reviews.csv")

head(df)

tail(df)

str(df)

names(df)

colSums(is.na(df))

#no missing data

#Data exploration

df %>% count(Location)

n\_distinct(df$Location)

n\_distinct(df$ID)

n\_distinct(df$Hotel.Restaurant.name)

#check for duplicated data

table(duplicated(df))

table(duplicated(df$ID))

df <- df %>% distinct()

table(duplicated(df))

#remove duplicate rows

df %>% distinct(Hotel.Restaurant.name, .keep\_all=TRUE) %>%

select(Hotel.Restaurant.name, Location) %>% count(Location)

#work out number of rows per hotel

prop\_df <- df %>% count(Hotel.Restaurant.name) %>% as.data.frame()

prop\_df <- merge(prop\_df, df, by="Hotel.Restaurant.name")

prop\_df <- prop\_df %>% distinct(Hotel.Restaurant.name, .keep\_all=TRUE) %>% select(Hotel.Restaurant.name, n, Location)

prop\_df <- prop\_df[order(-prop\_df$n), ]

#convert all text to lowercase for consistency

df <- mutate\_each(df, funs(tolower))

#location has whitespace at beginning

df$Location <- trimws(df$Location)

# PATONG VS CHOENG

#randomly sample 15 Patong venues

set.seed(123)

patong\_unique <- df %>% filter(Location=="patong") %>% distinct(Hotel.Restaurant.name) %>%

as.data.frame() %>% sample\_n(15)

patong\_sample <- merge(patong\_unique, df, by="Hotel.Restaurant.name", all=FALSE)

#check for duplicates

table(duplicated(patong\_sample))

n\_distinct(patong\_sample$Hotel.Restaurant.name)

n\_distinct(patong\_sample$Location)

#randomly sample 15 Choeng Thale venues

set.seed(234)

choeng\_unique <- df %>% filter(Location=="choeng thale") %>% distinct(Hotel.Restaurant.name) %>%

as.data.frame() %>% sample\_n(15)

choeng\_sample <- merge(choeng\_unique, df, by="Hotel.Restaurant.name", all=FALSE)

#check for duplicates

table(duplicated(choeng\_sample))

n\_distinct(choeng\_sample$Hotel.Restaurant.name)

n\_distinct(choeng\_sample$Location)

#combine to create sample 30 hotels dataset

sample\_df <- rbind(patong\_sample, choeng\_sample)

n\_distinct(sample\_df$Hotel.Restaurant.name)

table(duplicated(sample\_df))

#output dataset

write.csv(sample\_df, "hotel-sample.csv",row.names = FALSE)

**Appendix B**

R code used for text analysis and sentiment analysis.

###TEXT ANALYSIS

library(tm)

library(wordcloud)

library(ggplot2)

library(tidyverse)

#set working directory

setwd("C:\\Users\\smann\\OneDrive\\University\\Data Science\\ASDM\\Project\\Sentiment analysis")

rm(list=ls())

#import sample dataset

df <- read.csv("hotel-sample.csv")

#subset by location

patong <- subset(df, Location=="patong")

choeng <- subset(df, Location=="choeng thale")

#combine reviews for hotels into one row

patong <- patong %>% group\_by(Hotel.Restaurant.name) %>%

summarise(Review = str\_c(Review, collapse=' ')) %>%

ungroup

head(patong$Review)

choeng <- choeng %>% group\_by(Hotel.Restaurant.name) %>%

summarise(Review = str\_c(Review, collapse=' ')) %>%

ungroup

head(choeng$Review)

#text cleaning

#PATONG

patong\_reviews <- patong$Review

patong\_reviews <- gsub("...more", " ", patong\_reviews)

patong\_reviews <- gsub("\n", " ", patong\_reviews)

patong\_reviews <- gsub("[[:digit:]]", " ", patong\_reviews)

#patong\_reviews <- gsub("...", " ", patong\_reviews)

head(patong\_reviews)

#CHOENG THALE

choeng\_reviews <- choeng$Review

choeng\_reviews <- gsub("...more", " ", choeng\_reviews)

choeng\_reviews <- gsub("\n", " ", choeng\_reviews)

choeng\_reviews <- gsub("[[:digit:]]", " ", choeng\_reviews)

#choeng\_reviews <- gsub("...", " ", choeng\_reviews)

head(choeng\_reviews)

#create corpus

patong\_corpus <- Corpus(VectorSource(patong\_reviews))

choeng\_corpus <- Corpus(VectorSource(choeng\_reviews))

#further cleaning

patong\_corpus <- tm\_map(patong\_corpus,removeNumbers)

patong\_corpus <- tm\_map(patong\_corpus,removePunctuation)

patong\_corpus <- tm\_map(patong\_corpus,removeWords,stopwords("english"))

patong\_corpus <- tm\_map(patong\_corpus,removeWords, c("patong", "phuket","thailand"))

patong\_corpus <- tm\_map(patong\_corpus,stripWhitespace)

patong\_corpus <- tm\_map(patong\_corpus,stemDocument) #--> cuts words up

choeng\_corpus <- tm\_map(choeng\_corpus,removeNumbers)

choeng\_corpus <- tm\_map(choeng\_corpus,removePunctuation)

choeng\_corpus <- tm\_map(choeng\_corpus,removeWords,stopwords("english"))

choeng\_corpus <- tm\_map(choeng\_corpus,removeWords, c("choeng", "phuket", "thailand"))

choeng\_corpus <- tm\_map(choeng\_corpus,stripWhitespace)

choeng\_corpus <- tm\_map(choeng\_corpus,stemDocument) #--> cuts words up

#create word matrix

dtm\_patong <- TermDocumentMatrix(patong\_corpus,

control = list(minWordLength=c(1,Inf)))

dtm\_choeng <- TermDocumentMatrix(choeng\_corpus,

control = list(minWordLength=c(1,Inf)))

#find top most frequent words (chose over 50 frequency due to

#large number of words)

#PATONG

termFreq\_patong <- rowSums(as.matrix(dtm\_patong))

wordFreq\_patong <- termFreq\_patong %>%

subset(termFreq\_patong>=90) %>%

sort(decreasing=TRUE) %>%

as.data.frame() %>% rownames\_to\_column

#plot word frequency

ggplot(wordFreq\_patong) +

aes(x=reorder(rowname,(-.)), weight = .) +

geom\_bar(fill = "#112446") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1)) +

xlab("Word") +

ylab("Frequency")

#CHOENG THALE

termFreq\_choeng <- rowSums(as.matrix(dtm\_choeng))

wordFreq\_choeng <- termFreq\_choeng %>%

subset(termFreq\_choeng>=90) %>%

sort(decreasing=TRUE) %>%

as.data.frame() %>% rownames\_to\_column

#plot word frequency

ggplot(wordFreq\_choeng) +

aes(x=reorder(rowname,(-.)), weight = .) +

geom\_bar(fill = "#112446") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1)) +

xlab("Word") +

ylab("Frequency")

#word clouds

#PATONG

wordcloud\_patong <- termFreq\_patong %>%

as.data.frame() %>% rownames\_to\_column %>%

subset(rowname!='food' & rowname!='good'& rowname!='restaur' &

rowname!='great' & rowname!='servic')

wordcloud(words = wordcloud\_patong$rowname,

freq=wordcloud\_patong$.,max.words=50,

min.freq = 5,

scale = c(2,0.25),

size = 0.05,

random.order = F,

colors = brewer.pal(6,"Dark2"))

#CHOENG THALE

wordcloud\_choeng <- termFreq\_choeng %>%

as.data.frame() %>% rownames\_to\_column %>%

subset(rowname!='food' & rowname!='good'& rowname!='restaur' &

rowname!='great' & rowname!='servic')

wordcloud(words = wordcloud\_choeng$rowname,

freq=wordcloud\_choeng$.,max.words=50,

min.freq = 5,

scale = c(2,0.1),

size = 0.05,

random.order = F,

colors = brewer.pal(6,"Dark2"))

#words strongly associated with top 4 words for each

findAssocs(dtm\_patong, terms = c("food","good","restaur", "great", "servic"), corlimit = 0.7)

findAssocs(dtm\_choeng, terms = c("food","good","restaur", "great", "servic"), corlimit = 0.7)

#top 5 words per hotel

library(corpus)

library(tidytext)

#PATONG

corpusdf\_patong <- as\_corpus\_frame(patong\_corpus)

corpusdf\_patong <- cbind(patong,corpusdf\_patong)

corpusdf\_patong$Hotel.Restaurant.name <- as.factor(corpusdf\_patong$Hotel.Restaurant.name)

corpusdf\_patong$text <- as.character(corpusdf\_patong$text)

head(corpusdf\_patong)

str(corpusdf\_patong)

corpus\_patong <- corpusdf\_patong %>% select(-Review) %>%

unnest\_tokens(words,text) %>%

subset(words!='food' & words!='e' &words!='u' &

words!='restaur'&words!='servic'& words!='staff') %>%

group\_by(Hotel.Restaurant.name) %>%

count(words, sort=TRUE) %>%

top\_n(5) %>% ungroup()

ggplot(corpus\_patong) +

aes(x= reorder(words, -n), n, fill = Hotel.Restaurant.name,) + # use data from each novel for plot and distinguish novel by color

geom\_bar(stat = "identity") +

facet\_wrap(~ Hotel.Restaurant.name, scales = "free\_y") + # separates plots by novel, "free\_y" shares scales across the y-axis

labs(x = "", y = "Frequency") + # set x and y-axis labels

coord\_flip() + # Flip cartesian coordinates so horizontal becomes vertical vice versa.

theme(legend.position="none")

names(corpus\_patong)

str(corpus\_patong)

#CHOENG THALE

corpusdf\_choeng <- as\_corpus\_frame(choeng\_corpus)

corpusdf\_choeng <- cbind(choeng,corpusdf\_choeng)

corpusdf\_choeng$Hotel.Restaurant.name <- as.factor(corpusdf\_choeng$Hotel.Restaurant.name)

corpusdf\_choeng$text <- as.character(corpusdf\_choeng$text)

head(corpusdf\_choeng)

str(corpusdf\_choeng)

corpus\_choeng <- corpusdf\_choeng %>% select(-Review) %>%

unnest\_tokens(words,text) %>%

subset(words!='food' & words!='e' &words!='u'&

words!='restaur' &

words!='servic' & words!='staff') %>%

group\_by(Hotel.Restaurant.name) %>%

count(words, sort=TRUE) %>%

top\_n(5) %>% ungroup()

ggplot(corpus\_choeng) +

aes(x= reorder(words, -n), n, fill = Hotel.Restaurant.name,) +

geom\_bar(stat = "identity") +

facet\_wrap(~ Hotel.Restaurant.name, scales = "free\_y") +

labs(x = "", y = "Frequency") +

coord\_flip() +

theme(legend.position="none")

names(corpus\_patong)

str(corpus\_patong)

**Appendix C**

R code for sentiment analysis implementation.

##SENTIMENT ANALYSIS

library(syuzhet)

#whole patong area

emotions\_patong <- get\_nrc\_sentiment(patong$Review)

emotionsDF\_patong <- emotions\_patong %>% colSums() %>% as.data.frame() %>%

rownames\_to\_column("emotions")

head(emotionsDF\_patong)

emotionsDF\_patong <- rename(emotionsDF\_patong,freq = .)

#whole choeng area

emotions\_choeng <- get\_nrc\_sentiment(choeng$Review)

emotionsDF\_choeng <- emotions\_choeng %>% colSums() %>% as.data.frame() %>%

rownames\_to\_column("emotions")

head(emotionsDF\_choeng)

emotionsDF\_choeng <- rename(emotionsDF\_choeng,freq = .)

#plot emotion frequency

#PATONG

ggplot(emotionsDF\_patong) +

aes(x = reorder(emotions, -freq), weight = freq) +

geom\_bar(fill = "#112446") +

labs(x = "Emotions", y = "Word Frequency") +

theme\_minimal()

#CHOENG THALE

ggplot(emotionsDF\_choeng) +

aes(x = reorder(emotions, -freq), weight = freq) +

geom\_bar(fill = "#112446") +

labs(x = "Emotions", y = "Word Frequency") +

theme\_minimal()

#Create word clouds for negative and positive sentiment

#PATONG

c\_patong <- corpusdf\_patong %>%

select(-c(Review, Hotel.Restaurant.name)) %>%

unnest\_tokens(word,text) %>%

subset(word!='u' & word!='e') %>%

group\_by(word) %>%

count(word, sort=TRUE)%>%

left\_join(get\_sentiments("nrc")) %>%

subset(sentiment=="positive" | sentiment=="negative") %>%

ungroup()

neg\_patong <- c\_patong %>% subset(sentiment=="negative") %>%

select(-sentiment)

wordcloud(words = neg\_patong$word,

freq=neg\_patong$n,

max.words=50,

min.freq = 5,

scale = c(2,0.25),

size = 0.1,

random.order = F,

colors = brewer.pal(6, "OrRd")[3:6])

pos\_patong <- c\_patong %>% subset(sentiment=="positive") %>%

select(-sentiment)

wordcloud(words = pos\_patong$word,

freq=pos\_patong$n,

max.words=50,

min.freq = 5,

scale = c(2,0.25),

size = 0.1,

random.order = F,

colors = brewer.pal(6,"Greens")[4:7])

#CHOENG THALE

c\_choeng <- corpusdf\_choeng %>%

select(-c(Review, Hotel.Restaurant.name)) %>%

unnest\_tokens(word,text) %>%

subset(word!='u' & word!='e') %>%

group\_by(word) %>%

count(word, sort=TRUE)%>%

left\_join(get\_sentiments("nrc")) %>%

subset(sentiment=="positive" | sentiment=="negative") %>%

ungroup()

neg\_choeng <- c\_choeng %>% subset(sentiment=="negative") %>%

select(-sentiment)

wordcloud(words = neg\_choeng$word,

freq=neg\_choeng$n,

max.words=50,

min.freq = 5,

scale=c(2,0.25),

size = 0.1,

random.order = F,

colors = brewer.pal(6, "OrRd")[3:6])

pos\_choeng <- c\_choeng %>% subset(sentiment=="positive") %>%

select(-sentiment)

wordcloud(words = pos\_choeng$word,

freq=pos\_choeng$n,

max.words=50,

min.freq = 5,

scale = c(2,0.25),

size = 0.1,

random.order = F,

colors = brewer.pal(6,"Greens")[4:7])

#comparing hotels

#PATONG

hotel\_emot\_df\_patong <- as.data.frame(emotions\_patong)

hotel\_emot\_df\_patong <- cbind(patong, hotel\_emot\_df\_patong)

hotel\_emot\_df\_patong <- hotel\_emot\_df\_patong %>% select(-Review) %>%

pivot\_longer(c(anger, anticipation, disgust, fear, joy, sadness,

surprise, trust, negative, positive))

#all emotions

ggplot(hotel\_emot\_df\_patong) +

aes(x= reorder(name, -value), value, fill = Hotel.Restaurant.name,) +

geom\_bar(stat = "identity") +

facet\_wrap(~ Hotel.Restaurant.name, scales = "free\_y") +

labs(x = "", y = "Frequency") +

coord\_flip() +

theme(legend.position="none")

#CHOENG

hotel\_emot\_df\_choeng <- as.data.frame(emotions\_choeng)

hotel\_emot\_df\_choeng <- cbind(choeng, hotel\_emot\_df\_choeng)

hotel\_emot\_df\_choeng <- hotel\_emot\_df\_choeng %>% select(-Review) %>%

pivot\_longer(c(anger, anticipation, disgust, fear, joy, sadness,

surprise, trust, negative, positive))

#all emotions

ggplot(hotel\_emot\_df\_choeng) +

aes(x= reorder(name, -value), value, fill = Hotel.Restaurant.name,) +

geom\_bar(stat = "identity") +

facet\_wrap(~ Hotel.Restaurant.name, scales = "free\_y") +

labs(x = "", y = "Frequency") +

coord\_flip() +

theme(legend.position="none")

#just positive and negative

#PATONG

hotel\_posneg\_df\_patong <- hotel\_emot\_df\_patong %>%

subset(name=="positive" | name=="negative")%>% as.data.frame()

ggplot(hotel\_posneg\_df\_patong) +

aes(x= reorder(name, -value), value, fill = Hotel.Restaurant.name,) +

geom\_bar(stat = "identity") +

facet\_wrap(~ Hotel.Restaurant.name, scales = "free\_y") +

labs(x = "", y = "Frequency") +

coord\_flip() +

theme(legend.position="none")

#CHOENG

hotel\_posneg\_df\_choeng <- hotel\_emot\_df\_choeng %>%

subset(name=="positive" | name=="negative")%>% as.data.frame()

ggplot(hotel\_posneg\_df\_choeng) +

aes(x= reorder(name, -value), value, fill = Hotel.Restaurant.name,) +

geom\_bar(stat = "identity") +

facet\_wrap(~ Hotel.Restaurant.name, scales = "free\_y") +

labs(x = "", y = "Frequency") +

coord\_flip() +

theme(legend.position="none")

**Appendix D**

R code for shiny dashboard.

#####R SHINY DASHBOARD

library(shiny)

library(shinydashboard)

ui <- dashboardPage(

dashboardHeader(title = "Comparing reviews of tourist venues

in Patong and Choeng Thale", titleWidth = 300),

dashboardSidebar(

sidebarMenu(

menuItem("Location",

tabName = "location\_tab",

icon = icon("mountain-city")

),

menuItem("Venues",

tabName = "venue\_tab",

icon = icon("hotel")

))

),

dashboardBody(

tabItems(

tabItem(

tabName="location\_tab",

fluidRow(box(sliderInput(inputId = "no\_words",label = "Number of Words",min = 1, max = 50,

value = 5)),box(uiOutput("emotion\_select"))),

fluidRow(box(title="Patong","Frequency-based Wordcloud", plotOutput("patong\_wordcloud","325px", "325px")),

box(title="Patong","Sentiment-based Wordcloud",plotOutput("patong\_emotion","325px", "325px"))),

fluidRow(box(title= "Choeng Thale","Frequency-based Wordcloud",plotOutput("choeng\_wordcloud","325px", "325px")),

box(title= "Choeng Thale","Sentiment-based Wordcloud",plotOutput("choeng\_emotion","325px", "325px"))),

),

tabItem(tabName = "venue\_tab",

fluidRow(box(uiOutput("location\_select")), box(uiOutput("hotel\_select"))),

fluidRow(box(title="Top 5 Frequent Terms",plotOutput("hotel\_word")),

box(title="Sentiment Frequency",plotOutput("hotel\_posneg"))),

)

)

)

)

server <- function(input,output) {

output$emotion\_select <- renderUI({

selectInput(inputId = "v\_emotion\_select",

label = "Sentiment",

choices = c("positive", "negative", "anger", "fear", "anticipation", "trust",

"surprise","sadness", "joy", "disgust"))

})

output$patong\_wordcloud <- renderPlot({

wordcloud(words = wordcloud\_patong$rowname,

freq=wordcloud\_patong$.,max.words=input$no\_words,

scale = c(2,0.25),

random.order = F,

colors = brewer.pal(6,"Dark2"))

})

output$choeng\_wordcloud <- renderPlot({

wordcloud(words = wordcloud\_choeng$rowname,

freq=wordcloud\_choeng$.,max.words=input$no\_words,

scale = c(2,0.25),

random.order = F,

colors = brewer.pal(6,"Dark2"))

})

output$patong\_emotion <- renderPlot({

sent\_patong <- corpusdf\_patong %>%

select(-c(Review, Hotel.Restaurant.name)) %>%

unnest\_tokens(word,text) %>%

subset(word!='u' & word!='e') %>%

group\_by(word) %>%

count(word, sort=TRUE)%>%

left\_join(get\_sentiments("nrc")) %>% subset(sentiment==input$v\_emotion\_select)

wordcloud(words = sent\_patong$word,

freq=sent\_patong$n,

max.words=input$no\_words,

scale = c(4,0.5),

random.order = F,

colors = brewer.pal(6, "Set1"))

})

output$choeng\_emotion <- renderPlot({

sent\_choeng <- corpusdf\_choeng %>%

select(-c(Review, Hotel.Restaurant.name)) %>%

unnest\_tokens(word,text) %>%

subset(word!='u' & word!='e') %>%

group\_by(word) %>%

count(word, sort=TRUE)%>%

left\_join(get\_sentiments("nrc")) %>% subset(sentiment==input$v\_emotion\_select)

wordcloud(words = sent\_choeng$word,

freq=sent\_choeng$n,

max.words=input$no\_words,

scale = c(4,0.5),

random.order = F,

colors = brewer.pal(6, "Set1"))

})

output$location\_select <- renderUI({

selectInput(inputId = "v\_location\_select",

label = "Location",

choices = df %>% select(Location) %>% distinct())

})

output$hotel\_select <- renderUI({

selectInput(inputId = "v\_hotel\_select",

label = "Venue",

choices = df %>% subset(Location==input$v\_location\_select) %>%

select(Hotel.Restaurant.name) %>% distinct())

})

output$hotel\_word <- renderPlot({

hw <- rbind(corpus\_choeng, corpus\_patong)

ggplot(data = hw %>% filter(Hotel.Restaurant.name==input$v\_hotel\_select)) +

aes(x= reorder(words, -n), n, fill = words) +

geom\_bar(stat = "identity") +

scale\_fill\_brewer(palette="Set3") +

labs(x = "", y = "Frequency") +

coord\_flip() +

theme(legend.position="none")

})

output$hotel\_posneg <- renderPlot({

he <- rbind(hotel\_emot\_df\_choeng, hotel\_emot\_df\_patong)

ggplot(data = he %>% filter(Hotel.Restaurant.name==input$v\_hotel\_select)) +

aes(x= reorder(name, -value), value, fill = name) +

geom\_bar(stat = "identity") +

scale\_fill\_brewer(palette="Set3") +

labs(x = "", y = "Frequency") +

coord\_flip() +

theme(legend.position="none")

})

}

shinyApp(ui= ui, server = server)