INDIVIDUAL ASSIGNMENT

MSC IN DATA SCIENCE AND BUSINESS ANALYTICS

APUMF1904DSBA

IN-DEPTH CUSTOMER'S SENTIMENTS IN AIRBNB

Student Name: CHAN KUOK HONG

TP Number: TP034717

Module Code: CT085-5-M

Module Name: CAPSTONE PROJECT – PART 1

Submission Date: 20th DECEMBER 2019

Supervisor Name: PROF. DR. R LOGESWARAN A/L N RAJASVARAN

Second Marker Name: DR. SIVAKUMAR VENGUSAMY

DECEMBER 2019

ABSTRACT

Airbnb has witnessed explosive success in recent years. However, the company must be able to understand its current standing in the constantly evolving market. In fact, the recent entrance of several peer-to-peer accommodation platforms to the market and the acquirement of HomeAway by Expedia have created the competition for greater market share. Although Airbnb may still be the market leader, the constant overlooking of customer experience by the company may eventually lead to the reduction of market share. Particularly, there is a positive bias found in Airbnb’s comment section and this may have limited the company’s awareness about its’ weaknesses or area for improvement. In order to reveal the customer experience, the current study make use of Latent Dirichlet Allocation (LDA) alongside fine-grained sentiment analysis to generate a set of expectations that are unique to the Airbnb users and analyse their type and valence of the sentiments. The current project will help the business gain a better understanding of the customer experience with their brand and help demonstrate a better alternative in conducting sentiment analysis through fine-grained sentiment analysis technique.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

TITLE: IN-DEPTH CUSTOMER'S SENTIMENTS IN AIRBNB

CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION

Back in the days, travelling was perceived as a luxurious experience where tourists spend more time relaxing than venturing out to immerse in any local cultural activities. Recently, however, the industry is witnessing a fundamental shift in travelling trends. The winds of change in travelling is set in motion by the rising number of millennial travellers. Surpassing the Baby Boomers, the millennials will become the largest living generation in human history (Fry, 2018). In terms of the tourism industry, millennials make up approximately 20% of the total market few years ago (Mohn, 2014). However, the number is expected to rise as more millennials join the workforce after their education and the desire to travel continues to grow (Bair and Wright, 2017). Some may argue that the figure discussed above are still insignificant, however, it is important to understand that millennials hold vastly different viewpoint towards travelling. Unlike the previous generations, millennials view traveling as a necessity (Bair and Wright, 2017). Therefore, the millennials may have a smaller budget, but they are also much more willing to spend on travelling and they travel much more often than the previous generations. Considering this, finding out and appealing to the preferences of millennial travellers is essential to ensure the competitiveness of companies within the tourism industry.

One of the most important aspects of travelling experience is the accommodation and Airbnb has become the model example in terms of providing to the accommodation needs of young travellers. Airbnb is a peer to peer sharing platform that allows private individuals to register as "host" and list any property (a room, an entire apartment or house) to be rented by travellers in exchange for money. In just a few years after its initial founding, the company raised $2.3 million in venture and achieved $25 million in valuation. More recently, the revenue of Airbnb during the second quarter of 2019 has reached $1 billion (Lee, 2019).

The success of Airbnb can be attributed to several factors. Firstly, the online platform that operates based on the sharing economy business model kept the operating expenses lean. Furthermore, the transactions that occur primarily in an online setting also appealed to the younger travellers' preference to be contacted through means such as emails or text messages (Chin, 2018). Aside from this, many have pointed out that Airbnb offers greater value for money in terms of accommodation amenities, space, and location when compared to hotels (Sigala, 2017). Furthermore, the company's branding that highlights the authentic cultural experiences is one of the most important contributors to Airbnb's popularity among the younger travellers (Guttentag, 2015). Such experiences appeal to the younger travellers who are more willing to spend on "experiences" (Chin, 2018). Moreover, the company has also emphasized the company's commitment in ensuring the sustainability of travelling by giving back to the local community (Lekach, 2018). Such marketing message is especially well-received by the more environmentally conscious young travellers (Chin, 2018). Combining the competitive advantages of Airbnb, it has been reported that the company that currently does not own a single hotel or lodging property has started to negatively affect the sales of the hotel industry (Sigala, 2017).

Nevertheless, continuous improvement is required to ensure the company's success can be maintained. This is especially relevant to Airbnb due to the entrance of few competitors into the market. For instance, the recent acquirement of HomeAway, a similar peer to peer accommodation platform by Expedia has created a sizable competitor to Airbnb. As Expedia has established itself as the giant in tourism industry for many years, Airbnb must devise ways to retain its' customers in the face of a competitor with much greater financial backing and recognition within the industry. Aside from the challenges posed by competitors, the recent accusations of the company in driving up the housing prices has caused several locations to limit the amount of Airbnb allowed (Barron, Kung, and Proserpio, 2019). To make matters worse, recent scandals regarding the Airbnb hosts involvement in sexual and physical assault has definitely tarnished the reputation of the company (Hohman, 2018; Levin, 2017).

Therefore, it is vital for Airbnb to understand the company's current standpoint within the industry in order to maintain its competitiveness. One of the most effective ways to confirm the company's standpoint is through customer reviews. Although Airbnb has provided a comment section for both Airbnb hosts and guests to rate each other after the stay, numerous studies have highlighted that the comments are overwhelmingly positive (Fradkin, Grewal, and Holtz, 2018; Bridges and Vasquez, 2017). As the hosts are also allowed to rate the guests, the overwhelming positive comments are due to the fear of retaliation should the guests left any negative comments. Should the hosts retaliate by rating the guest negatively after receiving a low rating, it may affect the guests' ability to make another reservation on Airbnb in the future. Consequently, guests that have negative experiences chose to voice their opinions in other independent sites. This negatively affects the company's ability to identify areas for improvement and growth as the comments on the site may not represent actual experiences of the guests.

Considering the importance of obtaining an accurate representation of the guests' experience, it may be better for the company to venture beyond the company sites' review section. Therefore, the current study will conduct a fine-grained sentiment analysis on Airbnb guests' comments collected in both Airbnb platform and from another independent site. By including the reviews from Airbnb into the analysis, the current project will also help to verify if the recent changes made by the company to overcome the positive bias in the review section has resulted in any improvement. In addition to that, unlike the general sentiment analysis, fine-grained sentiment analysis do not force the sentiments into a dichotomous category but offers a much wider range of human sentiments. In doing so, the company will gain better understanding about the customers' sentiments towards the company. In addition to the fine-grained sentiment analysis, the sentiments will also be performed at an aspect level. In other words, the findings of the current research will reveal the various types and levels of customer sentiments in regard to the various aspects of Airbnb experience from reservations to the customer service.

1.2 PROBLEM STATEMENT

The lack of accurate understanding about Airbnb guests' experiences severely limits the company's ability to identify its' current standpoint and areas for growth. Specifically, the fear of retaliation from the hosts has contributed to the serious imbalance in the amount of positive and negative comments. Although the company has recently made changes to the commenting system whereby the comments will only be revealed once both the host and guest have completed their comments, studies conducted few years ago have revealed that the positive bias remains (Porges, 2017). Therefore, the changes made to rectify the positive bias remains unknown. In addition to the positive bias, the current sentiment analysis techniques may not provide sufficient depth in capturing the customers' experience. Specifically, the overly simplistic dichotomous or trichotomous sentiment analysis do not offer sufficient information about the customers' experience with Airbnb.

1.3 RESEARCH OBJECTIVES

The aim of the current project is to perform an in-depth analysis on the customer reviews of Airbnb customers. In order to achieve aim stated above, the objectives listed below must be fulfilled:

1. To identify the main aspects of Airbnb guests' comments.
2. To perform fine-grained sentiment analysis on each aspect identified in objective 1.

1.4 RESEARCH SIGNIFICANCE

As mentioned above, vast majority of the customer reviews left on the platform of Airbnb are positive. Consequently, the overwhelmingly positive reviews can seriously hinder the company's ability to establish an accurate understanding of the customers' experience. Furthermore, the company may never be able to rectify any negative experiences that customers may have had without receiving their negative reviews. In fact, recent response from Airbnb regarding the assault cases of Airbnb hosts have indirectly confirmed that the company is unaware of the actual customers experience (Levin, 2017; Hohman, 2018). Particularly, the company's representative has expressed that such negative experiences are extremely rare while mentioning that the comments on the company's platform have shown that most customers have had positive experiences with the company's service.

Aside from hindering the company's ability to improve, the lack of negative or neutral comments have also indirectly led to the lack of research conducted in this regard. So far, majority of the studies conducted have focused primarily at the positive reviews (Hu, Zhang, Pavlou, 2009). Although a pioneering study on negative comments of Airbnb was conducted by Sthapit and Bjork (2019), the researchers have chosen the qualitative methods where only major themes were extracted from the negative comments collected. Although the study has provided some guidance in regard to the dissatisfaction experienced by customers of Airbnb, the study did not indicate the extent of the customers dissatisfaction on various aspects of Airbnb's service.

By performing an in-depth analysis on the comments of the Airbnb guests, the current project may provide contribution in both the research community and the business context. Specifically, the findings may demonstrate the advantages of using data analytics techniques in understanding customers experience over the use of purely qualitative methods. Furthermore, the fine-grained sentiment analysis technique chosen will also provide greater nuances in terms of indicating the extent of customers' emotion polarity. Additionally, the aspects and sentiments extracted will also be unique to Airbnb as they will be derived purely from actual customer reviews. By making use of the findings, the company will be able to understand the true experiences of the customers and rectify aspects that the company is lacking.

1.5 RESEARCH SCOPE

In order to capture the customer experience of Airbnb guests, the study will collect comments left by previous Airbnb guests collected from Trustpilot.com and Airbnb website. Although Airbnb has included a comment section in its platforms, vast majority of the comments were found to be positive, as mentioned in the introduction. Although efforts have been made to rectify the positive bias within the comment section, the actual improvement remain unknown. In this case, Trustpilot provides a large amount of objective comments with a mixture of positive and negative reviews. Therefore, the comments from Trustpilot.com may be a better representation of the customer experiences while making the comments from Airbnb website as a control. Furthermore, a comparison of the customer sentiments may help to gauge the effectiveness of the company's solution as well as to gain a better understanding of the customers' experience. In addition to the source, the current project will only focus the analysis on English comments only. Furthermore, as the analysis and the findings are limited to the defined scope above, it may not be suitable to be used as an actual reference of the customer experience of Airbnb.

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

In the following sections, the research conducted on customer reviews in online context will be critically discussed. Furthermore, the importance of customer reviews on the success of companies that operates primarily in an online setting will be discussed. Additionally, the methods used by these studies will be reviewed. Following the discussion of studies on customer reviews, research performing sentiment analysis on customer reviews will also be discussed.

2.2 CUSTOMER REVIEWS IN ONLINE SETTINGS

Ever since the access to Internet increase, business models that operate purely in an online setting have started to proliferate. One of the forerunners in this field is the e-commerce industry. Following the success of e-commerce, rapid digitalization of various services has given rise to the emergence of sharing economy while Airbnb has become the model example of this industry. Although the lack of physical store and lower overhead cost have kept the operating cost lean, both industries are plagued by the intangibility that both industries have worked hard to overcome. Due to the intangibility of these companies, customer reviews have become the primary source of reference for potential customers regarding to the quality of products or service. Over time, customer reviews have become the norm in any online platform.

Following the rapid development of customer reviews, the research communities have also started a lively conversation in regard to the study of customer reviews, especially in the field of tourism and hospitality (Schukert el al., 2015). In general, this field of research can be divided into three main types of studies. The first type of research, which is also the majority in this field, is the study on customer textual reviews and categorization of the texts to either positive or negative sentiment. As expected, topic modelling techniques and sentiment analysis are fairly common in many of such studies. Take the study conducted by Zhou et al. (2014) for example, the researchers have analysed over 4000 customer comments from agoda.com. Following the analysis, the researchers categorized the reviews into four categories, which are positive, negative, mixture of both positive and negative, and neutral sentiment. Another example of such study is a more recent research conducted by Gao et al. (2018) where the customer reviews of various restaurants were analysed. Interestingly, the researchers have used customer reviews to not only provide the sentiment polarity but also identify potential competitors within the industry. A few more similar lines research is by Dickinger et al. (2017), and Berezina et al. (2015).

The second type of research focuses primarily on the numerical ratings. For instance, the study conducted by Radojevic et al. (2017) have analysed numerical ratings of customer reviews collected from Tripadvisor. The website provides an overall numerical rating alongside with several aspects rating such as value, price, and location. In this case, the researchers have analysed the relationship between overall rating and the individual aspects rating provided by websites using multilevel analysis. Another similar study is conducted by Schukert et al. (2015) where the difference in rating behaviour is compared between English speaking and non-

English-speaking customers. Furthermore, the study also compared the difference in terms of overall rating and aspects rating behaviour of the two groups.

The third type of research is the combination of the previous two types of studies whereby the numerical ratings and textual reviews are included in the analysis at the same time. For instance, the study by Zhao et al. (2019) analysed the impact of textual reviews on the overall rating score. In the research, the textual reviews are represented as a vector of features such as readability, subjectivity, and diversity. Meanwhile, Zhang et al. (2016) examined the potential influence of experts' reviews on the rating behaviour of customers.

As a summary of the current state of the research in this field, the numerical ratings and textual reviews are often studied as separate entities. Recent works, however, have shown attempts to combine both numerical ratings and textual reviews to get a better understanding about customers' experience. However, the focus of the current project is not the combination of the numerical and textual data but to address the potential issues posed by using the pre-defined aspects by many online review systems. So far, most if not all of the studies that have studied customers' experience in various aspects of the accommodation have obtained the data from Tripadvisor and Agoda. However, only a limited number of platforms provide aspects rating. Furthermore, for accommodation, the aspects listed such as location, cleanliness, value for money, and price can differ for different platforms. For instance, Trivago currently provides aspects such as location, value, sleep quality, rooms, and cleanliness. On the other hand, hotels.com provides service, location, cleanliness, and comfort. In addition to that, the constant upgrade of many platforms and websites also resulted in the changes in aspects within the same platform. As a result, the comparison of customers reviews within and across platforms is made increasingly difficult. Consequently, the comparison of findings between studies is also a challenging task due to the lack of standardization.

Aside from the difficulties in comparing, researchers have also pointed out that some of the aspects listed can be vague and users may not always understand the intended meaning (McAuley and Leskovec, 2013). For example, lovehomeswap.com has listed one aspect as "accuracy". However, there has not been any explanation if the "accuracy" was referring to the accuracy of the descriptions or the accuracy of the service provided. More importantly, many of the listed aspects are inconsistent with what the customers are mentioning in their textual reviews. In general, aside from the overall rating, most platforms provide 5 to 8 aspects for customers to provide their rating. However, an extensive study conducted by Dolnicar and Otter

(2003) has identified a total number of 173 aspects that are relevant to the condition of the accommodation by conducting a meta-analysis on studies published between 1984 to 2000. Besides that, the researchers also pointed out that the relevance of the aspects depends on the type of accommodation. In other words, customers of hotels and peer-to-peer accommodation have different sets of expectations for the different accommodation and tend to evaluate them differently. Therefore, based on the findings above, the current aspects listed by most websites or platforms may not reflect actual customers' experience. Despite the findings by Dolnicar and Otter (2003), most studies have continued to use the pre-define aspects listed by most websites as a "one-size fits all" solution.

Looking back at the focus of the current study, which is Airbnb, no aspects have been listen in the review section yet. Currently, customers only provide an overall numerical rating and a textual review. Given the large amount of information in the textual reviews, the lack of aspects may make the understanding of customer experience more challenging. Furthermore, the current aspects listed in most other platforms cannot be an accurate reference to gauge the customer experience of Airbnb, as pointed out in the research by Dolnicar and Otter (2003). However, several studies have improved the prediction of customer experience by complementing the analysis with textual reviews (Ganu et al., 2009; Long et al., 2014). Therefore, the richness of information in textual reviews may be used to generate the aspects unique to the customers of Airbnb and gain better understanding of their experience with the company's service. In doing so, the generated aspects will provide structure in terms of analysing the customers' experience. In other words, the company can then focus on extracting relevant comments on certain aspects that the company would like to improve. Furthermore, by using the aspects extracted, a fine-grained sentiment analysis performed on any individual aspect can provide an in-depth understanding of the customers experience with Airbnb.

2.3 SENTIMENT ANALYSIS ON CUSTOMER REVIEWS

Sentiment analysis is a common technique used by numerous studies in the field of customer experience (Yu et al., 2013). Generally, sentiment analysis can either be conducted on a document level, sentence level, or on a aspect level (Vinodhini and Chandrasekaran, 2012). As suggested, sentiment analysis can analyse an entire document, a complete sentence, or an entity aspect. Considering the studies from previous section, aspect level sentiment analysis is fairly popular in the field of research on hospitality and business (Pan et al., 2007).

For instance, the study conducted by Nakayama and Wan (2018), customers from different cultural backgrounds were compared in terms of their preference for restaurant entrée items based on the customer reviews. On the other hand, the study by Guo et al. (2017) analysed the customer reviews on hotels and identified the various factors that may contribute to customers' satisfaction with their accommodation. Furthermore, the customers are also divided based on their demographic factors and the findings revealed that customers from different demographic groups show different levels of emphasis on various aspects of their accommodation.

However, majority of the studies that have made use of sentiment analysis have either classified the sentiments as only two or three classes, which are positive, negative, or neutral (Yan et al., 2018; Geetha et al., 2017; Kirilenko et al., 2018; Gitto and Mancuso, 2017). Regardless of the methods, vast majority of these studies shared one common shortcoming. A large number of studies in this field have assumed that human emotions can be forced into a one-dimensional scale. In other words, most of these studies have forced the studies into a dichotomous or trichotomous categories of emotions. For instance, most studies have categorised the sentiments into negative, neutral or positive sentiments. However, numerous studies on human emotions have stated that human emotions do not always fall neatly into either one category (Lichtenstein and Slovic, 2006; Tversky and Thaler, 1990). Furthermore, by examining at several customer textual reviews, the crude division of the various emotions such as surprise, disgust, disappointment, and joy into either positive or negative dimensions lack nuances and cannot provide an accurate representation of customers' feelings. In fact, it is also perfectly possible for customers to have a mixture of positive and negative feelings about the service. Moreover, the overly simplistic division of sentiments may fall short in reviews with mixed sentiments such as: "This place is so convenient with plenty of restaurants and bars but the amount of noise at night is terrible." Based on the example, sentiments expressed do not conform to the dichotomous division as it expressed different sentiments for different aspects of the accommodation. Although emotions can be better understood as a multi-dimensional continuum, the categorization of emotions is still required given the current limitations in computation capabilities. Furthermore, several researchers have suggested that the representation of emotions in sentiment analysis should increase in terms of dimensionality to provide a more accurate representation of human emotions (Inkpen et al., 2010; Esmin et al., 2012).

2.4 SENTIMENT ANALYSIS VIA PLUTCHIK'S EMOTIONAL WHEEL

In the field of psychology, numerous theories have been proposed in terms of categorizing the wide range of human emotions. The most widely studied theories include the 6 fundamental emotions by Ekman (1992) and the emotional wheel by Plutchik (1994). Ekman (1992) proposed 6 emotional states that include fear, sadness, anger, joy, disgust, and surprise. Building on the theory proposed by Ekman, Plutchik (1994) proposed the emotional wheel and added two more emotional states, which are anticipation and trust. Aside from adding two more emotional states, the author also categorized the emotional states into 4 opposing pairs of emotional states, which are fear-anger, disgust-trust, surprise-anticipation, and sadness-joy. The figure below is an adaptation of the emotional wheel proposed by Plutchik.

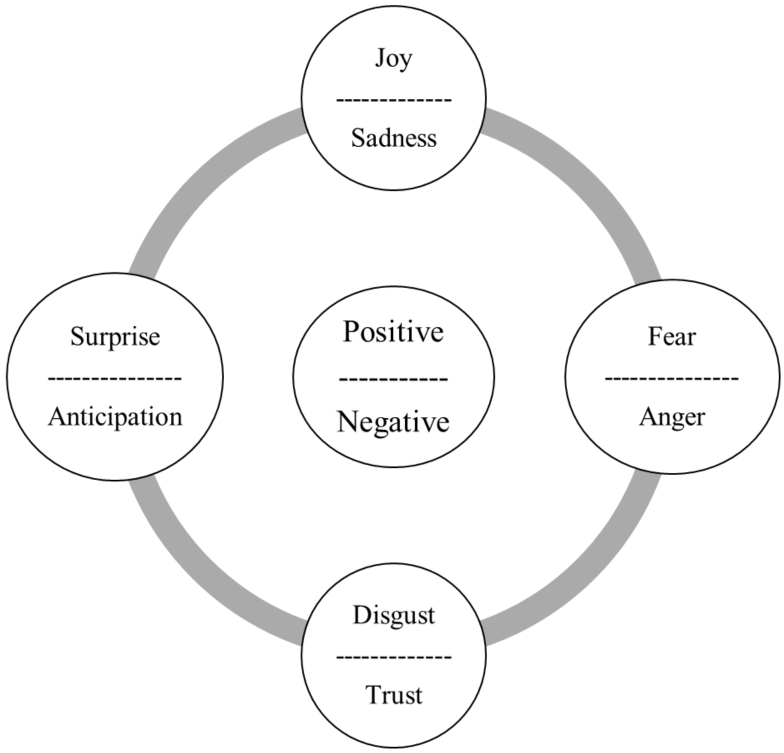


Figure 2.4.1 Adaptation of Plutchik's wheel of emotions

The Plutchik's emotional wheel is deemed suitable as a better alternative to the dichotomous categories of emotions as the emotional wheel is well supported by numerous psychological researches and has started being incorporated into a few studies using sentiment analysis (Borth, Chen, Ji, and Chang, 2013; Chafale and Pimpalkar, 2014; Terada, Yamauchi, and Ito, 2012). Furthermore, in comparison with the original theory proposed by Ekman (1992), Plutchik's model has a more balanced amount of positive and negative emotions. Therefore, the emotional states proposed in Plutchik's emotional wheel may be a better alternative to the dichotomous division of sentiments in sentiment analysis.

As a summary, the current state of the research on customers experience and reviews are still very much reliant on the readily available information from websites and online platforms. As mentioned above, most studies have taken the aspects listed from the platforms and have used those aspects to categorize customers' sentiments. However, such methods may not provide an accurate understanding of the customers' experience because the listed aspects are not reflective of what the customers are discussing in their textual reviews, as highlighted by the meta-analysis by Dolnicar and Otter (2003).

Therefore, in order to further understand the customer experience of Airbnb, the listed aspects provided by other independent sites cannot be used. In fact, to the best of the researcher's knowledge, there is no research conducted on the customer reviews on Airbnb from sources other than Airbnb own's platform. As pointed out by Dolnicar and Otter (2003), customers of each type accommodations tend to evaluate different types of accommodations with different standards. In other words, there may be a set of important aspects that are unique and important to Airbnb. So far, no study has been conducted yet. Consequently, the overwhelmingly positive reviews on Airbnb also limits the company's understanding of the actual customer experience as customers are still reluctant to share their complaints on the site (Fradkin, Grewal, and Holtz, 2018; Bridges and Vasquez, 2017).

Hence, by going beyond the company's own website to extract the textual reviews from another independent site may provide a comparison in terms of the customers' experience in different platforms. Furthermore, the extraction of aspects from Airbnb textual reviews without relying on pre-existing aspects will also provide results that are unique to the Airbnb customers. Specifically, the findings may reveal the set of expectations that customers tend to have when they are using the services of Airbnb. Following the extraction of aspects, sentiment analysis that is based on the emotional wheel will provide a nuanced picture of the emotions experienced by the customers when interacting with the company's service. From this finding, the company will understand the current performance in meeting the customers' expectations.

CHAPTER 3: METHODOLOGY

3.1 INTRODUCTION

As shown in the flowchart below, the current study will begin with the collection of data from Airbnb.com and Trustpilot.com. After the collection of data, the collected data will be stored in separate files for subsequent processing. In order to prepare the dataset, a series of screening procedures will be performed to ensure the data quality is appropriate for subsequent analysis. Furthermore, some exploration will also be conducted to gain an overview of the trends within the data. Following the initial inspection, a series of pre-processing procedures will be carried out to convert the dataset into formats that are suitable for the main analyses. The steps taken to optimize each model will be discussed and results will be reviewed. All the analyses will be performed through Python programming due to the wide range of libraries available that are suitable for the analyses involved in the current project. Before the explanation of the data analyses, the algorithms and techniques that will be used are first introduced.

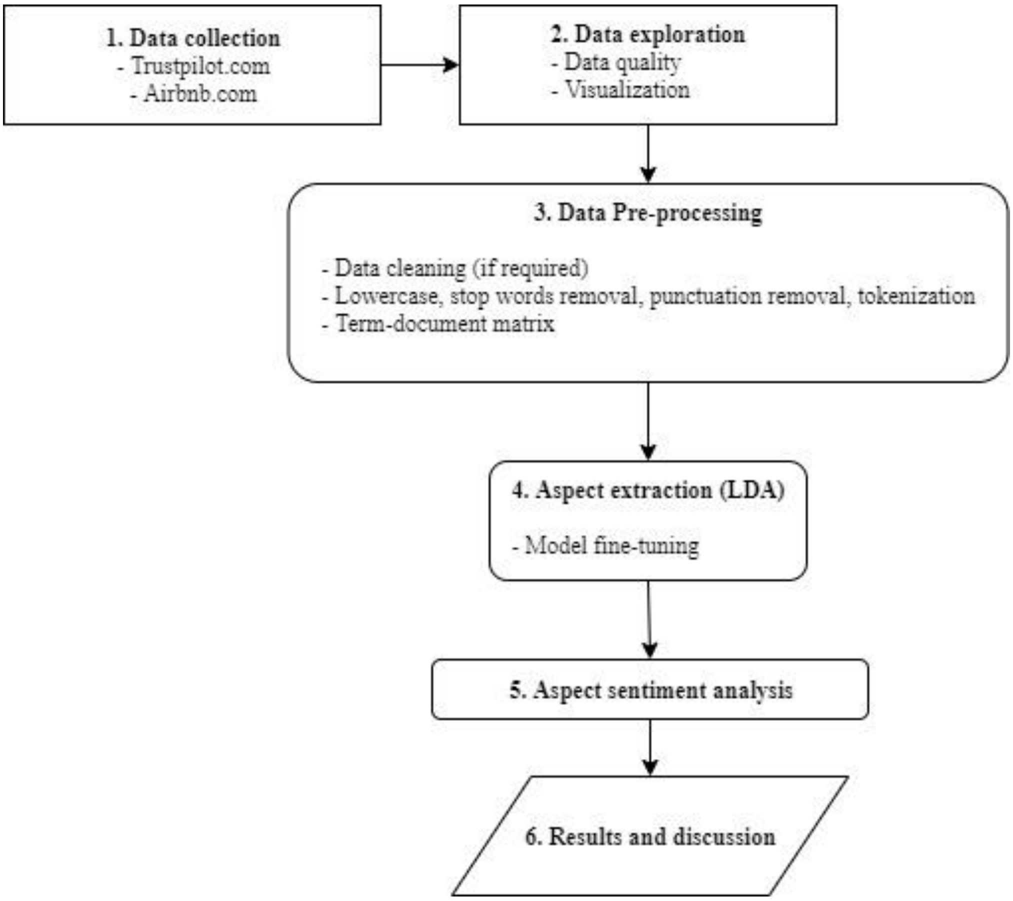


Figure 3.1.1 Flowchart of overall methodology

3.2 PROPOSED ALGORITHM AND ANALYSIS

The algorithms and analysis techniques are selected based on the data type and research objectives. In the following sections, the topic modelling algorithm and its' application to the dataset will be explained. Furthermore, the sentiment analysis technique for the current study will also be introduced.

3.3 LATENT DIRICHLET ALLOCATION (LDA)

LDA is a widely used topic modelling algorithm that can help process large body of text and extract a user-defined number of topics from the processed body of text (Guo et al., 2007). Based on Blei et al. (2003), LDA assumes that any body of text contains a mixture of several latent topics with varying level of weights. Furthermore, the detection of such latent topics can be performed through the analysis of the word's distribution within the text body (Blei et al., 2003). Figure 3.3.1 is an illustration of the parameters of LDA.

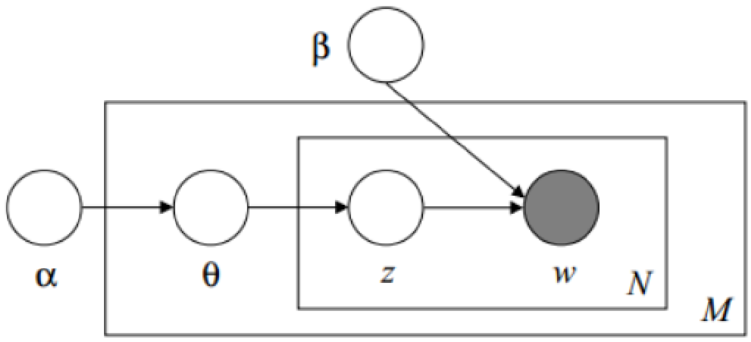


Figure 3.3.1 Graphical illustration of LDA parameters (Blei et al., 2003)

Based on the diagram above, LDA consists of the following important parameters:

- M the largest and outermost rectangle represents the total number of documents.
- N the second largest rectangle is the total number of words in the document.
- α is the mixture of topics within a document.
- β is the words that are categorised into each topic.
- θ represents the distribution of words in the document.
- w represents any specific word that is in a document.
- z represents the topic assigned to a word.

By referring to the diagram above, the parameters that are placed outside and within the rectangles indicates their level of influence on the algorithm. For instance, parameters such as α and β are placed outside of the outermost rectangle, which means these parameters' level of influence are at the document level. In other words, the α determines the number of topics in a document while β determines the number of words contained in each topic. Hence, by specifying a high α, there will be a higher number of topics within a document.

In order to implement LDA on any body of text, a term-document matrix must be generated. Taking the term-document matrix as the input, LDA will assign the words into a user-defined number, *k* number of topics. Furthermore, for each document, LDA will generate two matrices that contains two types of probabilities. The first probability produced is the proportion of words from a document that is assigned to a specific topic, t_i. The formula of the probability is denoted as below:

$$\text{topic}, t_i = p\ (\text{topic}, t_i \mid \text{document}, d)$$

The second probability that will be produced is the proportion of topic assignments for a specific word in all the documents. The formula is denoted as below:

$$\text{word}, w_i = p\ (word, w_i \mid topic, t_i)$$

Following the generation of the two probabilities, the product of the two probabilities will be calculated. Based on the product, the probability of a certain topic producing a specific word can be determined. Then, the words will be assigned based on the product of the two probabilities. This iterative process will stop once the algorithm determines that the assignment of words has produced a meaningful output.

By applying the algorithm onto the dataset of the current study, the textual reviews collected will be required to undergo a series of pre-processing and be converted into a term-document matrix. Following the specification of the number of topics, *k*, the extraction of topics can be performed based on the word's distribution of the submitted term-document matrix. The output of LDA will produce the specified number of topics along with group of words that are assigned to the generated topics. By using the generated aspects, subsequent analysis can be conducted.

3.4 LEXICON BASED SENTIMENT ANALYSIS

As the sentiment analysis performed in the current study is based on the Plutchik's emotional wheel, a lexicon-based sentiment analysis is chosen. Lexicon based sentiment analysis is one of the most common approach in conducting sentiment analysis whereby the sentiment of words or phrases in a text body are calculated to return the overall sentiment. Furthermore, this technique also requires a dictionary of words that are labelled with negative or positive sentiment (Jurek, Mulvenna, and Bi, 2015).

Specifically, a lexicon-based technique involves the overlapping of a body of text that has been processed into a bag of words with another dictionary of sentiment-labelled texts. The diagram below illustrates the pairing of the bag of words and the chosen lexicon (Figure 3.3.2). Each word in the bag of words will be matched with the lexicon and any overlap will result in a sentiment score added to the set of words. The figure in 3.3.3 shows the overlap of the bag of words and the chosen lexicon. Based on the overlapped words, the sum of the sentiment for all overlapped words will form the overall sentiment of the message.

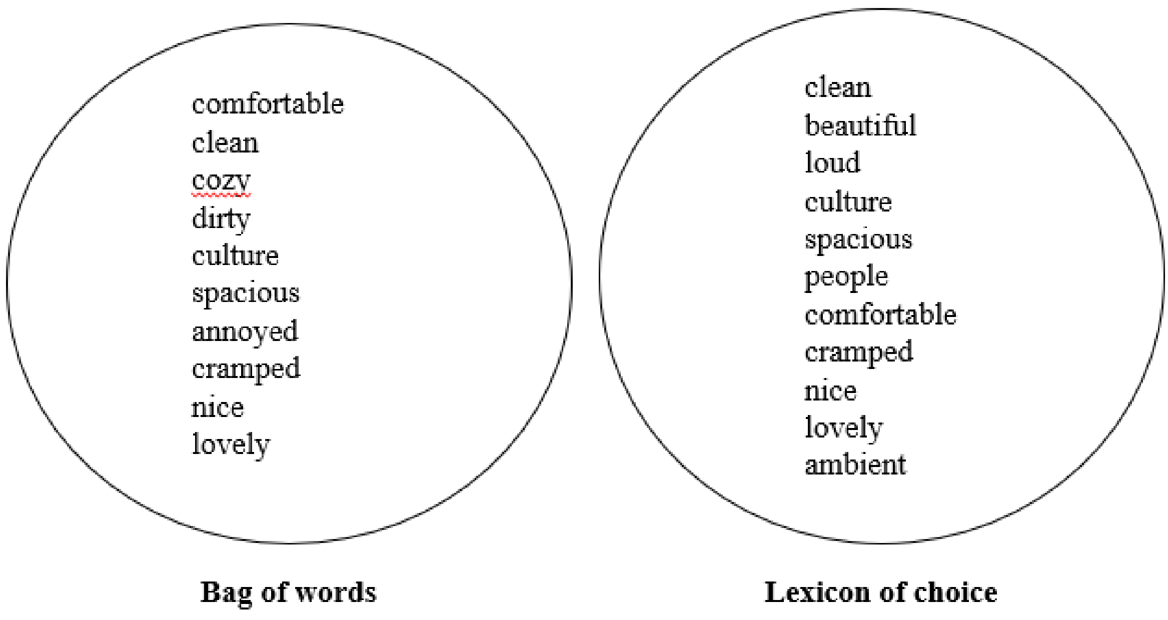


Figure 3.3.2 Pairing of the bag of words and lexicon of choice

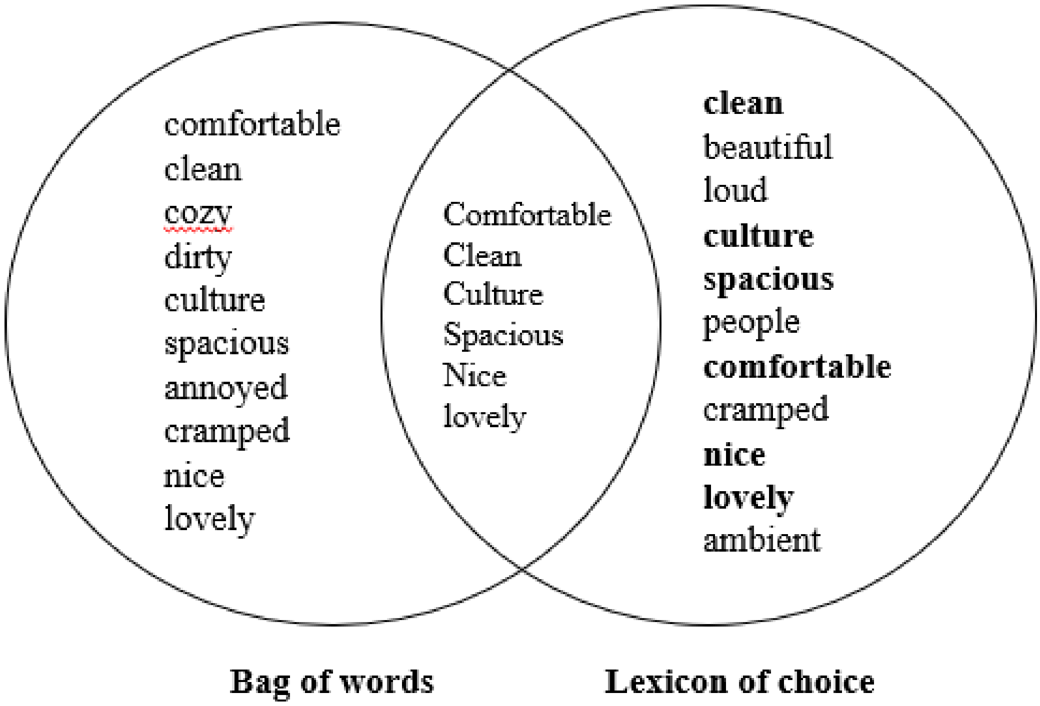


Figure 3.3.3 Overlap of words between bag of words and lexicon

As mentioned in previous paragraphs, the sentiments will be categorized based on Plutchik's emotional wheel. Therefore, the emotional lexicon developed by the National Research Council (Mohammad and Turney, 2003) will be adopted in the sentiment analysis of the current project. This lexicon is deemed suitable for the current project as the sentiment's labels are based on the 8 emotional states listed in the emotional wheel. By using this technique, the level and valence of the customers sentiments on various aspects of Airbnb's service can be understood and shed more light on the customers experience.

3.5 DATA COLLECTION

150 customer reviews on Airbnb will be collected in both Trustpilot.com and Airbnb.com and the complete dataset will consist of a total number of 300 customer reviews. As explained in the introduction and literature review, there is a positive bias in the reviews section of Airbnb. In order to obtain a more accurate representation of the customer experience, another independent site, which is Truspilot.com, is selected to obtain customer reviews about the service of Airbnb. The collection of the data will be performed through web mining technique.

3.6 DATA EXPLORATION

Collected data will be compiled and saved into two separate files, each for different sources of data. Then, several steps to inspect the quality of the data is carried out to gain an overview of the nature of the data collected so that the data preparation can be conducted more efficiently. Analytical and visualisation tools such as word clouds or bar charts will be generated to gain a better understanding of the trends within the data. Potential quality issues such as missing data or inconsistencies within the data will also be screened and noted for subsequent cleaning.

3.7 DATA PRE-PROCESSING

The textual reviews collected from both sources will go through a series of steps to prepare the data for subsequent analysis. Furthermore, the procedures below will be carried out using the Natural Language Toolkit (NLTK) library in Python.

- Converting all words to lowercase.
  - All collected text reviews will be converted to lowercase.
- Stop words removal.
  - Stop words such as the, a, or, that do not add any informational value will be removed from the data.
- Punctuations removal
  - Any punctuation or symbols that do not contribute to the subsequent analyses will be removed.
- Tokenization
  - Each text reviews in the datasets will be tokenized
- Term-document matrix
  - A term-document matrix will be generated for LDA analysis

3.8 ASPECTS EXTRACTION USING LDA

Taking the term-document matrix created in previous step, the potential aspects underlying the customer reviews will be extracted. Nonetheless, the number of aspects extracted will be specified as 5 topics, given the most common number of aspects in many platforms are around 5. Following the creation of the base model using 5 topics, several other models will also be created to be compared. For the purpose of the current study, the topic coherence, C_v score, will be used as the evaluation metric for model's performance.

3.8.1 MODEL FINETUNING

As mentioned in the previous section, several models will be fine-tuned, and the performance will be compared against the base model. Based on recent findings, there are three main hyperparameters that can be adjusted to optimize the performance of the models, which are:

- The number of topics, k
- Hyperparameter alpha, α
- Hyperparameter beta, β

In order to compare the potential impact of each parameter, the parameters listed above will be adjusted one at a time while keeping the other two constants. Results will be visualised and the model that produces the highest topic coherence which is indicated with the highest C_v score, will be selected as the best model. Thus, the resulting number of aspects and words from the model will be used for the subsequent sentiment analysis. As the topics are unnamed, manual profiling will be performed for each extracted aspect by referring to the keywords. The table below shows an example of the output and manual profiling of LDA generated output.

Table 3.8.1.1 Example of LDA generated topics and words

Topic 1 (Response)	**Topic 2 (Environment)**	**Topic 3 (Value)**	**Topic 4 (Amenities)**	**Topic 5 (Experience)**
question	location	price	Wi-Fi	friendly
email	noise	value	heater	annoyed
call	city	cheap	kitchen	fun

Additionally, a bootstrapping algorithm is used to extract more relevant words for each aspect to ensure that there is sufficient amount of words to be analysed for subsequent sentiment analysis. The boot-strapping algorithm calculates the dependencies between the aspects and words through Chi-square statistics. Then, words will be assigned into the aspect that have shown the highest dependencies (Wang et al., 2010).

3.9 SENTIMENT ANALYSIS ON EACH ASPECT

Following the extraction of aspects along with the relevant words in previous step, sentiment analysis will be conducted. As explained above, the NRC emotional lexicon will be used to measure the polarities of customer sentiments within the dataset. Nonetheless, it is important to note that the NRC has one drawback as it only provides sentiment labels for adjectives. In order to overcome this shortcoming, the adjectives closely related to the keywords extracted in previous step will be selected and be used for the estimation of customer sentiment polarities. Meanwhile, if the extracted keywords are adjective, the sentiment scores will be calculated automatically.

For each topic, the level of the 8 sentiments polarities will be calculated. Take the first topic in Table 3.8.1.1 for example, the customers may show varying level for the 8 types of sentiments for different aspects, as illustrated below.

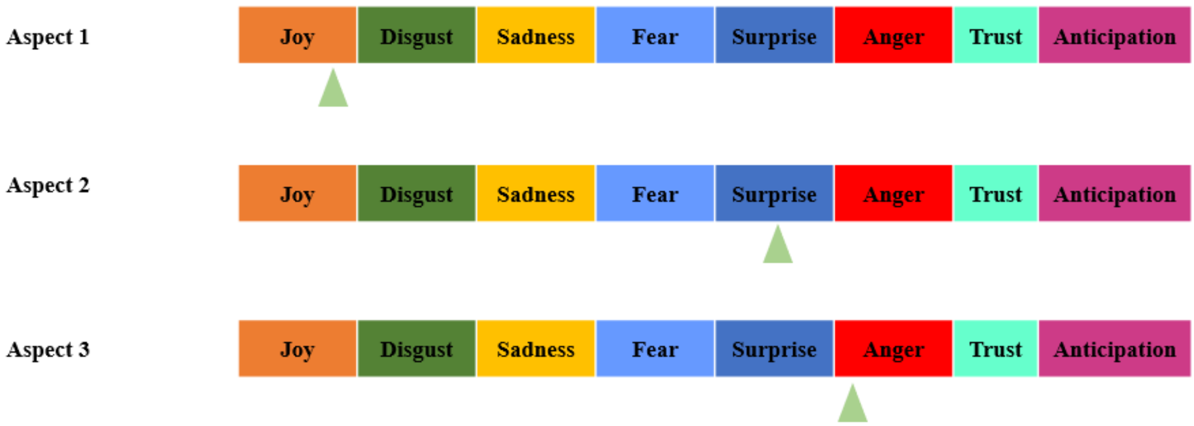


Figure 3.9.1 Illustration of aspect-based sentiment analysis

4.0 CONCLUSION

As a summary, the current project aims to fill in the knowledge gap in regard to the customer experience of Airbnb. Unlike most studies that have concentrated on only one source for the customer reviews, the current project takes the Airbnb site and another independent site into consideration while presenting the customer experience with Airbnb. Furthermore, the current project also aims to generate a set of aspects that are unique to Airbnb without using the pre-defined aspects by most platforms and sites that may not always reflect what the customers are mentioning in the reviews. In addition to the practical implications above, the current study may contribute to the research in sentiment analysis by using a more nuanced categorization of sentiments. As a result, the results derived from the current study will also be more meaningful as compared to the dichotomous or trichotomous division of sentiments.

5.0 RESEARCH PLAN

Capstone phase 1 will start from the planning and write up of the research proposal.

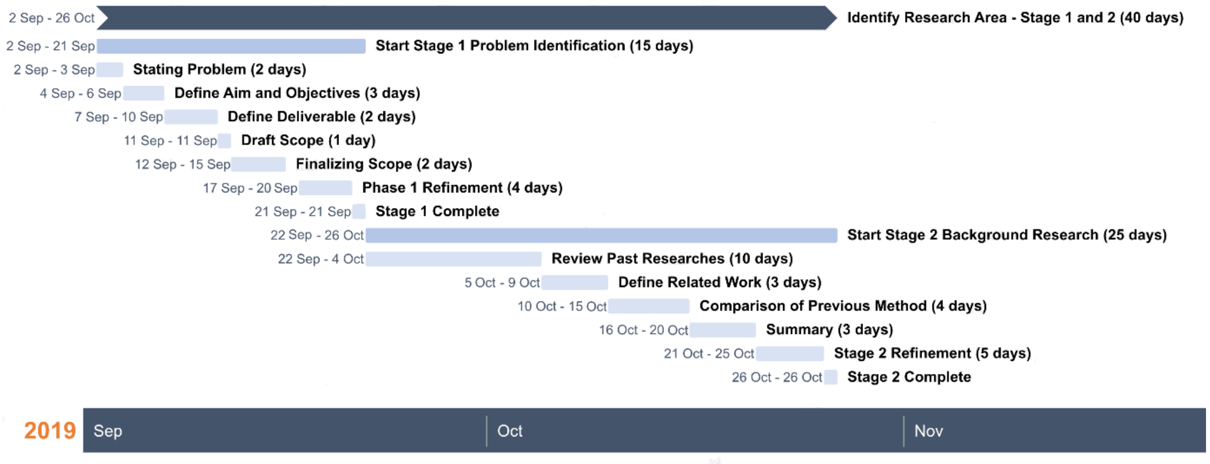


Figure 5.0.1 Gantt Chart: Phase 1 Identify Research Area

Capstone phase 2 will start from the data collection stage. Analysis of the data will consume most of the project time. At the same time, the write up will also be planned.

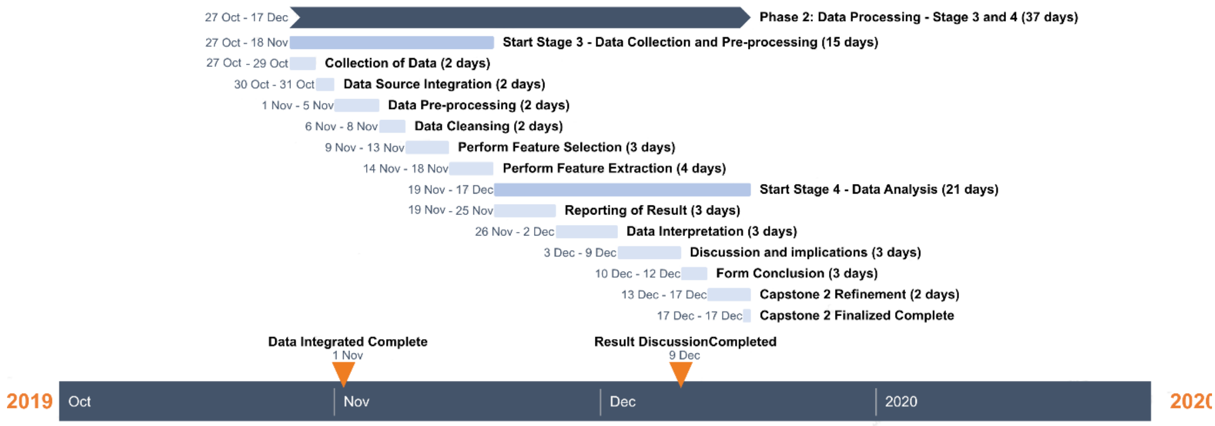


Figure 5.0.2 Gantt Chart: Phase 2 Data Processing

6.0 REFERENCES

Barron, K., Kung, E. and Proserpio, D. (2018). The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. *SSRN Electronic Journal*, 1-72. doi: 10.2139/ssrn.3006832

Bair, D., Wright, P. (2017). What do millennial travelers want in a vacation? [Online] Available from: https://www.bostonglobe.com/lifestyle/travel/2017/05/03/what-millennial-travelers-want-new-study-finds-out/mRCbCdZzhtOyFVU7jN7sVM/story.html [Accessed: 12th December 2019]

Berezina, K., Bilgihan, A., Cobanoglu, C. and Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. Journal of Hospitality Marketing & Management, 25(1), 1-24. doi: 10.1080/19368623.2015.983631

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3, 993-1022. [Online] Available from: http://www.jmlr.org/papers/v3/blei03a.html [Retrieved: 11th December 2019]

Bridges, J. and Vásquez, C., 2017. If nearly all Airbnb reviews are positive, does that make them meaningless? Current Issues in Tourism, 21(18), pp.2057-2075. doi: https://doi.org/10.1080/13683500.2016.1267113

Borth, D., Chen, T., Ji, R. and Chang, S.F. (2013). 'Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content.' *Proceedings of the 21st ACM international conference on Multimedia*, October 2013. Barcelona: ACM, 459-460.

Chafale, D., Pimpalkar, A. (2014). Review on developing corpora for sentiment analysis using plutchik's wheel of emotions with fuzzy logic. *International Journal of Computer Sciences and Engineering (IJCSE)*, 2(10), 14-18.

Chin, C. (2018). Millennial travellers are sick of social media influencers. [Online] Available from: https://www.thestar.com.my/lifestyle/travel/2018/10/05/millennial-travellers-social-media-influencers [Accessed: 12th December 2019]

Dickinger, A., Lalicic, L. and Mazanec, J. (2017). Exploring the generalizability of discriminant word items and latent topics in online tourist reviews. *International Journal of Contemporary Hospitality Management*, 29(2), 803-816. doi: 10.1108/IJCHM-10-2015-0597

Dolnicar, S. and Otter, T. (2003). 'Which hotel attributes matter? A review of previous and a framework for future research.' *Proceedings of the 9th Annual Conference of the Asia Pacific Tourism Association*, 1-21. [Online]. Available from: http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1281&context=commpapers [Retrieved: 12th December 2019]

Esmin, A.A., De Oliveira Jr, R.L. and Matwin, S. (2012). Hierarchical classification approach to emotion recognition in twitter. In 2012 *11th International Conference on Machine Learning and Applications*, 2, 381-385. December 2012: IEEE.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4), 169-200. doi: 10.1080/02699939208411068

Fradkin, A., Grewal, E., Holtz, D., & Pearson, M. (2015). 'Bias and reciprocity in online reviews: Evidence from field experiments on Airbnb.' *Proceedings of the 18th ACM Conference on Economics and Computation*, 18th to 22nd June. New York: ACM. doi:10.1145/2764468.2764528

Fry, R. (2018). Millennials projected to overtake Baby Boomers as America's largest generation. [Online]. Available from: https://www.pewresearch.org/fact-tank/2018/03/01/millennials-overtake-baby-boomers/ [Accessed: 12th December 2019]

Gitto, S. and Mancuso, P. (2017). Improving airport services using sentiment analysis of the websites. *Tourism management perspectives*, 22, 132-136. doi: 10.1016/j.tmp.2017.03.008

Gao, S., Tang, O., Wang, H. and Yin, P. (2018). Identifying competitors through comparative relation mining of online reviews in the restaurant industry. *International Journal of Hospitality Management*, 71, 19-32. doi: 10.1016/j.ijhm.2017.09.004

Ganu, G., Elhadad, N. and Marian, A. (2009). 'Beyond the stars: improving rating predictions using review text content.' *WebDB 12th International Workshop on the Web and Databases*, 1st -6th June. Rhode Island: USA.

Geetha, M., Singha, P. and Sinha, S., 2017. Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis. *Tourism Management*, 61, 43-54. doi: 10.1016/j.tourman.2016.12.022

Guttentag, D. (2015). Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. *Current issues in Tourism*, 18(12), 1192-1217. doi: 10.1080/13683500.2013.827159

Guo, Y., Barnes, S.J. and Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, 59, 467-483. doi: 10.1016/j.tourman.2016.09.009

Hohman, M. (2018). Airbnb Apologizes for 'Insufficient' Response After Host Breaks Through Window While Guests Sleep. People. [Online]. Available at:

https://people.com/home/airbnb-host-attacks-his-renters-los-angeles/ [Accessed: 3rd June 2019]

Hu, N., Zhang, J., Pavlou, P. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, 52(10), 144–147. doi:10.1145/1562764.1562800

Inkpen, D. and Strapparava, C. (2010). 'Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.' *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. June 2010. California: USA.

Jurek, A., Mulvenna, M.D. and Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics,* 4(1). [Online]. Available from: https://security-informatics.springeropen.com/articles/10.1186/s13388-015-0024-x [Accessed: 12th December 2019]

Kirilenko, A.P., Stepchenkova, S.O., Kim, H. and Li, X. (2018). Automated sentiment analysis in tourism: Comparison of approaches. *Journal of Travel Research*, 57(8), 1012-1025. doi: 10.1177/0047287517729757

Lee, J, L. (2019). Airbnb says second-quarter revenue topped $1 billion. [Online]. Available from: https://www.reuters.com/article/us-airbnb-results/airbnb-says-second-quarter-revenue-topped-1-billion-idUSKBN1W4001 [Accessed: 12th December 2019]

Levin, S, T (2017). Airbnb sued by woman who says she was sexually assaulted by 'superhost'. The Guardian. [Online]. Available from: https://www.theguardian.com/technology/2017/jul/27/airbnb-guest-sexual-assault-allegation [Accessed: 3rd June 2019]

Lekach, S. (2018). Airbnb wants to assure you its vision is good for society. [Online]. Available from: https://mashable.com/2018/01/25/airbnb-open-letter-vision-facebook/ [Accessed: 12th December 2019]

Long, C., Zhang, J., Huang, M., Zhu, X., Li, M. and Ma, B. (2014). Estimating feature ratings through an effective review selection approach. *Knowledge and information systems*, 38(2), 419-446. doi: 10.1007/s10115-012-0495-8

Lichtenstein, S. and Slovic, P. eds. (2006). The construction of preference. Cambridge University Press.

McAuley, J. and Leskovec, J. (2013). 'Hidden factors and hidden topics: understanding rating dimensions with review text.' *Proceedings of the 7th ACM conference on Recommender systems*, 165-172, October 2013. New York: ACM.

Mohn, T. (2014). The rising wave of millennial travellers. [Online]. Available form: https://www.forbes.com/sites/tanyamohn/2014/11/08/the-rising-wave-of-millennial-travelers/#6d9b2c572bc8 [Accessed: 12th December 2019]

Mohammad, S.M. and Turney, P.D. (2013). Nrc emotion lexicon. National Research Council, Canada.

Nakayama, M. and Wan, Y. (2018). Is culture of origin associated with more expressions? An analysis of Yelp reviews on Japanese restaurants. *Tourism Management*, 66, 329-338. doi: 10.1016/j.tourman.2017.10.019

Pan, B., MacLaurin, T. and Crotts, J.C. (2007). Travel blogs and the implications for destination marketing. *Journal of Travel Research*, 46(1), 35-45. doi: 10.1177/0047287507302378

Porges, S. (2017). All Airbnb Hosts Should Use This Chrome Extension for Screening Guests. [Online]. Available: https://www.forbes.com/sites/sethporges/2017/07/27/all-airbnb-hosts-should-use-this-chrome-extension-for-screening-guests/#20edc3f24081 [Accessed: 12th December 2019]

Plutchik, R. (1994). The Psychology and Biology of Emotion. Harper-Collins, New York, NY.

Radojevic, T., Stanisic, N. and Stanic, N. (2017). Inside the rating scores: a multilevel analysis of the factors influencing customer satisfaction in the hotel industry. *Cornell Hospitality Quarterly*, 58(2), 134-164. doi: 10.1177/1938965516686114

Schuckert, M., Liu, X. and Law, R. (2015). A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently. *International Journal of Hospitality Management*, 48, 143-149. doi: 10.1016/j.ijhm.2014.12.007

Sigala, M. (2017). Collaborative commerce in tourism: implications for research and industry. *Current Issues in Tourism*, 20(4), 346-355. doi: 10.1080/13683500.2014.982522

Sthapit, E., Bjork, P. (2019). Sources of distrust: Airbnb guests' perspectives. *Tourism Management Perspectives*, 245-253. doi: https://doi.org/10.1016/j.tmp.2019.05.009

Terada, K., Yamauchi, A. and Ito, A. (2012). Artificial emotion expression for a robot by dynamic color change. *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 314-321, September 2012. Paris: IEEE.

Tversky, A. and Thaler, R.H. (1990). Anomalies: preference reversals. *Journal of Economic Perspectives*, 4(2), 201-211. doi: 10.1257/jep.4.2.201

Vinodhini, G. and Chandrasekaran, R.M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal* of *Advanced Research in Computer Science and Software Engineering*, 2(6), 282-292. [Online]. Available from: https://pdfs.semanticscholar.org/261e/26ae134b8f63270dbcacf2d07fa700fdf593.pdf [Retrieved: 12th December 2019]

Wang, H., Lu, Y., Zhai, C. (2010). Latent aspect rating analysis on review text data: a rating regression approach. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 783–792. July 2010. New York: ACM

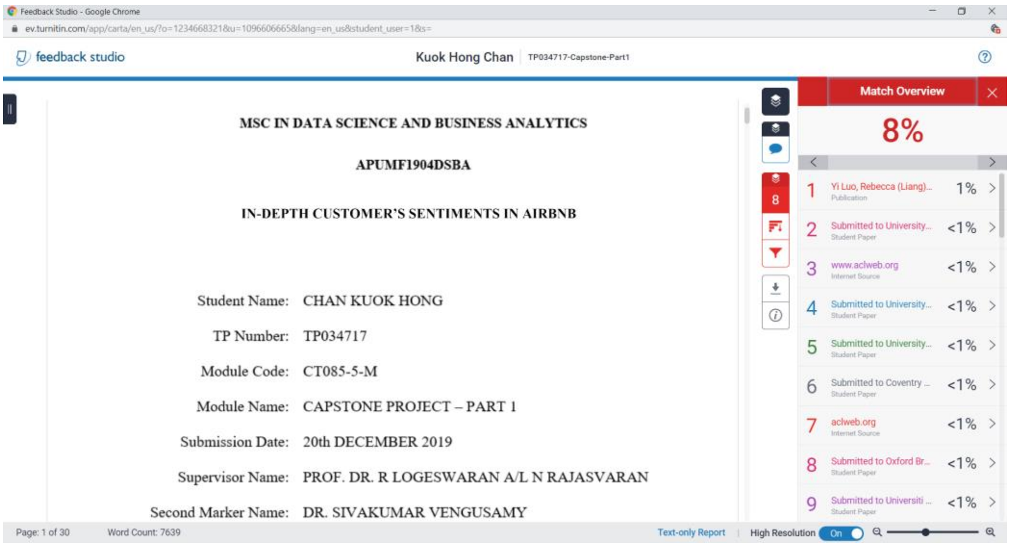
Yan, Q., Zhou, S. and Wu, S. (2018). The influences of tourists' emotions on the selection of electronic word of mouth platforms. *Tourism Management*, 66, 348-363. doi: 10.1016/j.tourman.2017.12.015

Yu, Y., Duan, W. and Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4), 919-926. doi: 10.1016/j.dss.2012.12.028

Zhao, Y., Xu, X. and Wang, M., 2019. Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management*, 76, 111-121. doi: 10.1016/j.ijhm.2018.03.017

Zhang, Z., Zhang, Z. and Yang, Y. (2016). The power of expert identity: How website-recognized expert reviews influence travellers' online rating behaviour. *Tourism Management*, 55, 15-24. doi: 10.1016/j.tourman.2016.01.004

7.0 APPENDIX – TURNITIN



Skip to Main Content

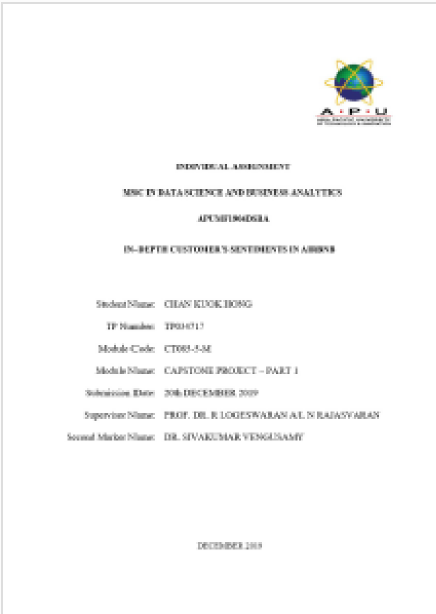
turnitin

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Kuok Hong Chan
Assignment title:	CP 1 & 2
Submission title:	TP034717-Capstone-Part1
File name:	TP034717-Capstone.docx
File size:	712.55K
Page count:	30
Word count:	7,639
Character count:	44,085
Submission date:	15-Dec-2019 11:54AM (UTC+0800)
Submission ID:	1234668321


INDIVIDUAL ASSIGNMENT

MSC IN DATA SCIENCE AND BUSINESS ANALYTICS

APUMF1904DSBA

IN-DEPTH CUSTOMER'S SENTIMENTS IN AIRBNB

Student Name: CHAN KUOK HONG
TP Number: TP034717
Module Code: CT085-3-M
Module Name: CAPSTONE PROJECT – PART 1
Submission Date: 20th DECEMBER 2019
Supervisor Name: PROF. DR. R LOGESWARAN A/L N RAJASVARAN
Second Marker Name: DR. SIVAKUMAR VENGUSAMY

DECEMBER 2019