AIRBNB USERS SENTIMENT THROUGH TOPIC MODELING AND SENTIMENT ANALYSIS

CHAN KUOK HONG

A thesis submitted in fulfilment of the
requirements for the award of the degree of
MSC IN DATA SCIENCE AND BUSINESS ANALYTICS

ASIA PACIFIC UNIVERSITY OF TECHNOLOGY & INNOVATION (APU)

MAY 2020

ABSTRACT

The advent of technology and rapid digitalization of businesses have opened up many possibilities for companies. One of the products of the digitalization is sharing accommodation, which is a concept that was made popular by Airbnb. As the market leader in the sharing accommodation industry, it has established itself as a notable opponent to the traditional hotel chains as an increasing number of travellers have converted to choose the accommodation provided by Airbnb over hotels due to the more competitive pricing. However, the status of the company as the pioneer and market is not unchallenged, given the fact there are few newer companies who have joined the market. Aside from the increased competition, the company is also facing a long-term issue, which is the safety and credibility of their listings. Particularly, the actual quality of the accommodation and services remains one of the biggest hindrances for the company to overcome. Although the company has implemented a review system, the system's reliability is heavily questioned due to the severe positive bias found in the reviews. And because of the positive bias, the company may not be able to recognize the actual customer experiences and area for improvement, which will eventually lead to the company stagnating. Therefore, the current study has analysed the reviews about Airbnb from two different platforms using topic modelling and fine-grained sentiment analysis to provide a more objective view of the Airbnb customers' experience. The findings revealed that the positive bias is still present in the reviews of Airbnb. Furthermore, customers from the two platforms are mentioning very different issues and have expressed very different sentiments towards the various aspects of Airbnb. Based on the findings, several recommendations were proposed for the company to further improve their understanding of the customers' experience.

CHAPTER 1

INTRODUCTION

The types of accommodation and ways people travel have been revolutionised by sharing accommodation, an idea that was made popular by Airbnb. As suggested by the concept's name, it involves the host sharing an unused space to travellers for a short-term stay in exchange for money. Due to the serious safety and credibility issues it raises, it was expected that the idea was heavily criticised in the beginning. However, through careful implementation of rating system, Airbnb has been well celebrated by the younger travellers over the past few years (Chin, 2018; Guttentag, 2015). In a recent interview, representative from the company expressed that the revenue of Airbnb has reached $1 billion at the second quarter of 2019 (Lee, 2019). Furthermore, the hotels have started to recognize the negative impact on their sales due to the growing popularity of Airbnb.

Nevertheless, it is important to bear in mind that the maintenance of Airbnb's success rely heavily on the company's reputation and its' ability to maintain a favourable brand image among consumers, even more so than traditional hotels. Other than contacting the host, Airbnb does not have concierge or direct customer services to immediately rectify any problems that the guests may face, which is also the company's strategy in keeping the operating costs lean. Furthermore, some of the guests are living with an individual whom they have just met. As a matter of fact, these hosts' behaviour is largely not monitored by any regulating bodies. Therefore, the guests are in a situation with high level of uncertainty with no means of obtaining immediate assistance should anything happen. Considering the risk, it may deter some guests from using the company's service and to opt for a hotel. However, maintaining a good reputation and ability to respond to customers' complaints quickly and effectively can serve to reduce the feeling of uncertainty. Besides, the guests may feel that their interest is safeguarded, and the hosts are held accountable for their behaviour. In other words, the company should strive to maintain a superb track record in terms of service quality and ability to respond to customers' negative experience in order to maintain the company's lead in the industry.

However, the actual reality for Airbnb is far from ideal. After a few news reports that have revealed cases of Airbnb hosts involvement physical and sexual harassment of the guests (Hohman, 2018; Levin, 2017). In response to the incidents, the company expressed that such negative experiences are rare, citing evidence from the largely positive reviews posted on their

platforms. Based on the company's response, it can be seen that the company is unaware of the actual experience that the guests had while using their services. As Airbnb regards the rating system as the evidence of the company's positive reputation, such negative news raises the concern among consumers that Airbnb's rating system may not be completely truthful and consequently be deemed unreliable by the consumers. In fact, numerous studies have questioned the reliability of the company's rating system as the reviews were found to be overwhelmingly positive (Fradkin, Grewal, and Holtz, 2018; Bridges and Vasquez, 2017). In addition to that, the company's lack of understanding of the customers' experience may cause the company to lose its' advantageous position as the market leader in the sharing accommodation industry.

Therefore, an accurate representation of the guests' experience is important and vital for Airbnb to maintain its' success and competitive edge as the pioneer in sharing accommodation. In order to obtain a more objective representation, venturing beyond the company platforms will yield a more accurate data. Hence, the current study aims to conduct a fine-grained sentiment analysis on Airbnb guests' comments extracted from both Airbnb platform and from another independent site. The inclusion of Airbnb's own platform may serve as a baseline in terms of the customers' sentiments towards Airbnb. Furthermore, the current project will also help to verify the impact of recent changes made by the company to overcome the positive bias in the review section. In addition to that, the fine-grained sentiment analysis proposed in the current project is deemed to be able to produce a more nuanced understanding of Airbnb customers' sentiments. By using the findings yield, the company will gain in-depth understanding about the customers' sentiments towards the company's service.

1.1 PROBLEM STATEMENT

The identification of the company's standpoint and areas for improvement is severely limited as the company lacks an accurate understanding of its' customers' experience. The issue is largely contributed to the two-way rating system implemented by the company in order to allow both the hosts and guests to review each after the stay. Despite its' positive intention to democratize the reviews, this system has led to the fear of retaliation from the hosts among the guests. In response to the criticisms, the company has attempted to improve the review system whereby the guests and hosts' reviews will only be revealed once both parties have submitted their reviews. However, a study conducted few years after the improvement have revealed that the issue of positive bias in the rating system remains (Porges, 2017). As of now, the

effectiveness of the changes made to overcome the positive bias in the rating system remains unknown. In addition to the positive bias, the company is still lacking in terms of understanding its' customers' experience. Furthermore, many of the current sentiment analysis techniques do not capture the customers' sentiments accurately. Specifically, the brute method of forcing human sentiments into dichotomous or trichotomous sentiments cannot provide sufficient information about the customers' experience with Airbnb.

1.2 RESEARCH OBJECTIVES

Considering the issues raised in previous section, the current project aims to perform an in-depth analysis on the customer reviews of Airbnb customers posted in two platforms. 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.

1.3 RESEARCH SIGNIFICANCE

To reiterate, Airbnb is currently facing the following issues:

- Unreliable rating system due to positive bias.
- Lack of accurate understanding on customer's experience.

Although Airbnb is still the market leader in the sharing accommodation industry, the issues mentioned above may cause the company to lose its' competitiveness over time. As mentioned, the lack of face-to-face contacts (concierge, customer service hotlines) of sharing accommodation companies amplifies the importance of having a positive reputation and reliable review system. Considering the fact that the reliability of the company's review system has been continuously questioned, it is important to investigate the effectiveness of the improvement made. In other words, if the changes were found to be ineffective, the company must formulate counteractive measures to rectify such issue as soon as possible.

In addition to the credibility issue, the company may also lose market share over time as an increasing number of customers' negative experiences left unattended due to their lack of understanding the customers' actual experiences with their service. In fact, the recent response from Airbnb in regard to the scandals of Airbnb hosts' involvement in criminal activities have indirectly confirmed that the company is unaware of the actual customers experience (Levin, 2017; Hohman, 2018). In particular, the news reports and complaints were dismissed by company's representative as individual cases and do not represent the company's quality of

service by citing the positive reviews posted on the company's site as supporting evidence for their statements. In short, the company appears to be trapped in its' self-serving bias and is blinded from what the customers are feeling towards their services and company reputation.

In addition to the business issues highlighted above, majority of the research techniques used by the hospitality research are only able to provide limited amount of information in regard to capturing customers' sentiments and deriving useful business insights. Partly due to the positive bias on the company's platform, majority of the studies conducted on Airbnb customer reviews have almost always revealed positive sentiments, which does not provide significant implications for the company to grow (Hu, Zhang, Pavlou, 2009). In the pioneering study conducted by Sthapit and Bjork (2019), the researchers used qualitative methods to extract major themes from the negative comments of Airbnb collected on another website. As a result, the study has highlighted several sources of Airbnb customers' dissatisfaction. However, the study did not indicate the extent of the customers dissatisfaction and the highly unstructured nature of qualitative findings did not translate well into actionable business insights.

Therefore, by performing an in-depth analysis on the comments of the Airbnb guests, the current project aims to overcome the issues raised in both business settings and research community. More specifically, the findings of the study analyse and quantify the findings so that an analytical model can be produced, and the customer sentiments can be accurately recognized. With an analytical model, the processes required to find out customers sentiments can be made easier and meaningful business actions can be generated. Additionally, the fine-grained sentiment analysis technique also provided more in-depth understanding in terms of the influencing factors and extent of customers' emotion polarity. Different from previous studies, the aspects and sentiments extracted in the current project are also unique to Airbnb. A set of recommendations is also generated by using the findings so that Airbnb can rectify aspects that the company is lacking.

1.4 RESEARCH SCOPE

In order to capture the customer experience of Airbnb guests, a total number of 2000 of textual comments written by previous Airbnb guests are collected from Trustpilot.com and Airbnb. The inclusion of Trustpilot.com as one of the data sources is because of the potential positive bias on Airbnb's platform, as mentioned in previous sections. Despite the efforts made to overcome the positive bias in its' rating system, the actual results remain unknown. Therefore, the comments collected from Trustpilot serve as a comparison as the site is not affiliated with

the company and hence users are more willing to share their negative experiences. Furthermore, the comparison of the customer sentiments of both platforms help to evaluate the effectiveness of the company's changes made to the rating system. Additionally, the current project also focused exclusively on the analysis on English comments only. Therefore, it is important to note that the findings are limited to the defined scope above and it may not be suitable to be implemented as the model of the customer experience of Airbnb.

1.5 RESEARCH QUESTIONS

By taking the problem statement and research objectives into consideration, there are several questions that needs to be answered. The few research questions are:

1. What are the differences between the reviews from Airbnb and Trustpilot?
2. What are the important aspects raised by the customers of Airbnb in different platforms?
3. What are the sentiments held by the customers of Airbnb towards the company's service?

1.6 STRUCTURE OF THE CAPSTONE

The report is structured into multiple chapters. The first chapter illustrates the project background and the project objectives. Then, the second chapter provides an introduction to reviews system in online environment and is followed by an extensive review on the research conducted on reviews system. Subsequently, the review system of Airbnb is introduced and relevant studies are discussed. Following the identification of the issue in Airbnb review system, the methodology used to study customer reviews in past research is discussed and reviewed.

Moving on to the methodology section, the dataset, tools, and algorithms used in the current project are introduced and explained in detailed. Procedures performed on the dataset are also outlined in this section. After that, the implementation chapter shows the coding and process of model building and optimization. Results are presented in the subsequent section and the findings are explained and interpreted in the discussion section.

CHAPTER 2

LITERATURE REVIEW

In the sections that will follow, the development of online reviews and its' implications on sharing accommodation is introduced and discussed. Following the discussion on online review system, several works performed on this regard are critically reviewed. The studies discussed have led to the research gap on the lack of research conducted on review system on sharing accommodation platform and the limitations of current methods used to analyse customer reviews. Subsequently, a number of related works on the analysis of textual reviews are reviewed so that potential improvement can be identified.

2.1 CUSTOMER REVIEWS IN ONLINE SETTINGS

The rapid increase of access to internet has given rise to the proliferation of companies where their business models operate primarily in an online environment. Perhaps, the most obvious example as well as the pioneering industry is the e-commerce industry. Following the rise of e-commerce, the rapid digitalization of various businesses led to the emergence of sharing economy where the business model operates by connecting the providers and consumers via an online platform. One of the forerunners in this industry is Airbnb. Such platform-based businesses are advantageous in terms of keeping the operating cost lean. However, the intangibility of such business models also causes high level of uncertainty. Although the hospitality industry has long relied on customer reviews to build their reputation, the lack of human touch for sharing accommodation like Airbnb can further increase the feeling of uncertainty among customers. Therefore, the intangibility of sharing accommodation companies like Airbnb makes customer reviews even more important as it is one of the main sources of reference for potential customers regarding to the quality of their products or service. Expectedly, Airbnb has implemented a review system for customers to provide their reviews after their stay and it was regarded as the main tool to reduce the feeling of uncertainty among potential customers.

Following the rise in importance for online customer reviews, an active research community is established on the study of customer reviews where most data were collected from the field of tourism and hospitality (Schukert el al., 2015). In general, online customer reviews contain 2 components, which are numerical ratings (usually on a scale) and textual

reviews. Furthermore, the numerical rating is usually presented as a numerical assessment of the product or service while the textual review complements the numerical rating with additional information (Gutt et al., 2019). Additionally, the research of online reviews can be divided into three main types of studies. The first line of research focuses on the study on textual reviews, which is also the majority in this field. These studies usually entail the categorization of the reviews into either positive or negative sentiment. In studies of textual reviews, topic modelling techniques and sentiment analysis are popular choice among the researchers. Taking the study by Zhou et al. (2014) as an example, over 4000 textual reviews were extracted from agoda.com and analysed. The sentiments were aggregated and divided into four main categories, which are positive, negative, mixture of both positive and negative, and neutral sentiment. A more recent research similar to Zhou et al (2014) was conducted by Gao and colleagues (2018) where online reviews of various restaurants were analysed to derive customer sentiments. Using similar methods, Gao et al (2018) have aggregated the overall sentiments and divided the sentiments into discrete positive and negative sentiments. In addition to that, the researchers have gone one step further by identifying potential competitors within the industry through the customer sentiments. Similar types of studies were also conducted by Dickinger et al. (2017), and Berezina et al. (2015).

Moving on to the second type of research, this line of studies focus primarily on the numerical ratings. For instance, Radojevic et al. (2017) analysed the numerical ratings of online reviews collected from Tripadvisor on several hotels. In addition to an overall rating, the website also allows the users to provide numerical rating on several aspects such as value, price, and location of their accommodation. The researchers analysed the relationship between the overall rating and the rating of each individual aspects rating using multilevel analysis. Meanwhile, Schukert et al. (2015) compared the difference in rating behaviour among English speaking and non- English-speaking customers. Additionally, the researchers also compared the difference in terms of overall rating and aspects rating behaviour of the two groups. In short, this line of study primarily focuses on the exploration of the relationship between the overall numerical ratings and aspects ratings with various variables that may cause a difference in the numerical ratings.

Finally, the third type of research includes the numerical ratings and textual reviews in its' analyses. For example, the impact of various properties of the textual reviews on the overall rating score is analysed in a study by Zhao et al. (2019). More specifically, the researchers represented the textual reviews as a vector of features such as readability, subjectivity, and

diversity so that their influence on the numerical ratings can be analysed. Meanwhile, Zhang et al. (2016) examined the potential influence of experts' reviews on the rating behaviour of customers. In simpler terms, the researchers are interested in finding out if users' rating behaviour will be influenced when exposed to reviews made by "experts", which are typically users who have made large number of reviews. The findings suggested that the exposure to others' comments indeed have an influence on the users' rating behaviour. In short, it can be seen that the numerical ratings and textual reviews are often studied separately. Nevertheless, attempts to include both numerical ratings and textual reviews in the main analyses have gradually increased in recent works. However, the review by Gutt et al (2019) suggested that textual reviews are able to provide much more rich information in most situations.

Aside from that, most of the studies were also found to use the aspects provided by the platform where data is collected. In other words, regardless of the issues mentioned in the reviews, the reviews are analysed based on aspects (value, location, price, facilities, etc) provided on the platform. Although such method greatly simplifies the research procedures, it can lead to several limitations. Firstly, the aspects on each platform can differ and this makes the comparison between studies difficult. For instance, the aspects listed at Trivago's platform are location, value, sleep quality, rooms, and cleanliness. Meanwhile, service, location, cleanliness, and comfort are listed by hotel.com. In addition to that, most of these platforms and websites go through routine upgrades and the aspects may be changed from time to time. Consequently, the comparison of studies performed using the same platform may be incomparable in terms of the aspects evaluated.

Other than causing the difficulties in comparing, researchers have highlighted that some of the aspects listed can be vague and is subject to individuals' interpretation (McAuley and Leskovec, 2013). Taking lovehomeswap.com as an example, the site listed one aspect as "accuracy". However, no explanation is provided on the actual definition of "accuracy" in that context. Furthermore, it could be about the accuracy of the descriptions or the accuracy of the service provided. Additionally, perhaps the greatest limitation by directly using the aspects provided by the platform is that many of the listed aspects are inconsistent with what the customers are mentioning in their textual reviews. Generally, most platforms like Tripadvisor provides 5 to 8 aspects to be rated by the customers, excluding the overall rating. In regard to this, Dolnicar and Otter (2003) conducted an extensive meta-analysis on studies published between 1984 to 2000 and concluded that there are approximately 173 relevant aspects present in the customers reviews on the condition of the accommodation. Besides the identification of

the aspects, the findings also suggested that relevance of the aspects depends on the type of accommodation being reviewed. Therefore, the researchers concluded that customers usually have different sets of expectations for different types of accommodation. So, it is safe to say that customers of sharing accommodation will evaluate sharing accommodation differently from other forms of accommodation such as hotels. However, most if not all of the platforms reviewing accommodation are using a "one-size-fits-all" approach in regard to the aspects of the accommodation. Stemming from the findings above, the aspects listed by many of platforms do not reflect the actual content in the customers' reviews or what capture aspects deemed most important to the customers when it comes to sharing accommodation.

2.2 AIRBNB REVIEW SYSTEM

Since the direct application of aspects listed by these platforms are not advisable, it is important to seek for relevant works conducted on Airbnb customer reviews. As of now, Airbnb has not provided any aspects for the customers to provide their rating yet. So, after their stay, Airbnb customers provide an overall numerical rating alongside a textual review. One may question the actual importance of listing down aspects along with the textual reviews since textual reviews provide larger amount of information. In regard to this, the abundance of information available in textual reviews can be a double-edged sword. Specifically, the large amount of information in textual reviews can be rich but it also makes the understanding of customer experience more challenging and do not translate well into measurable business goals or actionable business insights.

As the platform does not provide any aspects, past research conducted on Airbnb customer reviews are the next best option to derive potential aspects. However, there is still limited amount of research that have suggested any aspects that can represent the customers experience despite a handful number studies have been conducted on Airbnb reviews. In the study by Cheng and Jin (2019), the researchers pointed out that the most important aspects to the customers of Airbnb differ across studies. Indeed, some studies have found out that interactions between the guest and host as most important (Tussyadiah and Pesonen, 2016; Lampinen and Cheshire, 2016). Meanwhile, some studies have argued otherwise, stating that Airbnb is merely another form of hotel experience with a much lower price tag, suggesting that price as the most important aspect among Airbnb customers (Guttentag, 2016). Additionally, studies have also suggested that Airbnb customers place higher importance on the practical attributes of the accommodation instead of experiential attributes (Tussyadiah, 2016). However,

the findings remain inconclusive as “location”, one of the practical attributes was found to be not statistically significant in predicting customer satisfaction. Besides, experiential attributes like “enjoyment” was found to be significant. Such contradictory findings can be attributed to the large variety of accommodations listed on the platform and individual preferences of the Airbnb customers. Unlike hotels that usually have clear target market and offerings, Airbnb has no clear standards on the type of market they serve hence what was offered can vary significantly.

Additionally, the lack of standardization is not the only issue discovered by past research. Many researchers have pointed out that the reviews posted on Airbnb platform are overwhelmingly positive (Fradkin, Grewal, and Holtz, 2018; Bridges and Vasquez, 2017). Furthermore, the positive comments make up 95% of the reviews posted. Given the large variety of listings, customers and hosts from various backgrounds and hospitality experiences, it is questionable that 95% of guests are completely satisfied with Airbnb’s services. Given the credibility issue that may be raised from the positive bias, the company changed the review policy by revealing the reviews only after both parties (guest and host) have made their reviews to overcome the fear of retaliation from the host. Although changes have been made, this may also suggest that past research conducted using reviews collected from Airbnb may lack reliability. In short, the contradictory findings and positive bias of the review section jointly contributed to the lack of understanding about the customers’ experience and severely limits the company’s ability to improve and grow. Therefore, it is very important for Airbnb to understand the actual experiences of its’ guests by using reliable methods and sources.

2.3 SENTIMENT ANALYSIS OF ONLINE REVIEWS

Sentiment analysis is one of the most common technique used to understand customer experience (Yu et al., 2013). Generally, sentiment analysis work by analysing the sentiments in a body of text and return a score that indicates the sentiment. Furthermore, it can either be performed on an entire document, a complete sentence, or an entity aspect (Vinodhini and Chandrasekaran, 2012). In the study by Nakayama and Wan (2018), sentiment analysis was conducted on reviews of customers from different cultural backgrounds to find out the different customers’ preference for restaurant entrée items. Meanwhile, the study by Guo et al. (2017) performed sentiment analysis on customer reviews on hotels and identified factors that contribute to customers’ satisfaction with their accommodation. In other words, the researchers used several identified factors in the prediction of the sentiment polarity of customers. In

addition to the accommodation-related factors, the researchers also used the customers' demographics as predictors for customers' sentiment. The findings show that there is significant difference in terms of preferences and satisfaction for their accommodation for customers from different backgrounds.

Although some interesting insights have been produced, majority of the studies that have made use of sentiment analysis have one major limitation. Most of these studies 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). This is a limitation as it assumes that human emotions can be forced into a one-dimensional scale. Furthermore, these studies have also forced the human sentiments into a dichotomous or trichotomous categories of emotions. However, it is important to note that that human emotions do not always fall neatly into one discrete category (Lichtenstein and Slovic, 2006; Tversky and Thaler, 1990). In real life, it is more often for humans to have experience mix feelings towards any products or services given the fact that all services and products can have their own pros and cons.

Aside from the fact that it does not translate well in real life, the crude division of human emotions into discrete sentiments also fall short when it comes to 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 given, the sentiments expressed do not allow a clear category of positive nor negative sentiments. Besides, it also lacks the ability to discern the positive and negative aspects of the review. Although emotions can be better represented as a multi-dimensional continuum, the categorization of emotions is performed in order to streamline the computation process given the current limitations in computation capabilities. However, the dimensionality of emotions in sentiment analysis should increase in terms of in order to provide a more accurate representation of human emotions (Inkpen et al., 2010; Esmin et al., 2012). By increasing the dimensionality, one can still represent emotions as discrete categories but with much greater nuances in comparison to the strictly positive or negative sentiments.

2.4 PLUTCHIK'S EMOTIONAL WHEEL

A handful of theories has been proposed by psychologists in terms of the representation and categorization of the wide range of human emotions. One of the most widely studied and well-supported theories are the 6 fundamental emotions proposed by Ekman (1992), and the emotional wheel by Plutchik (1994). The original theory was proposed by Ekman (1992) where there are 6 emotional states, which includes fear, sadness, anger, joy, disgust, and surprise. Building upon Ekman's theory, Plutchik (1994) extended the theory by proposing the emotional wheel and added two more emotional states, which are anticipation and trust. Besides that, Plutchik also grouped the emotional states into 4 opposing pairs of emotional states, which are fear-anger, disgust-trust, surprise-anticipation, and sadness-joy. Furthermore, the 4 pairs of emotions can be further grouped to form positive and negative emotions. The figure below is an adaptation of the emotional wheel proposed by Plutchik.

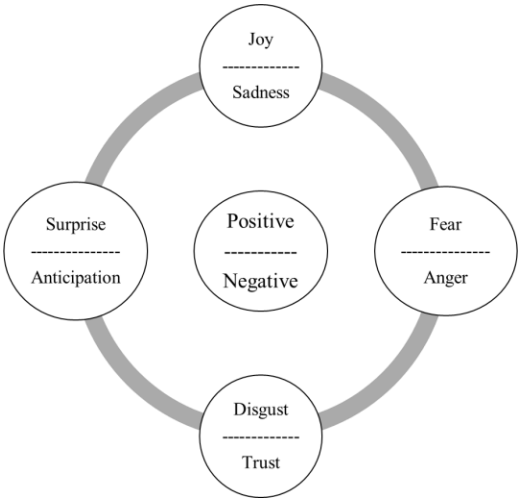


Figure 2.1 Adaptation of Plutchik's wheel of emotions

As mentioned above, the discrete representation of emotions can be improved by increasing the dimensions. Therefore, the Plutchik's emotional wheel is deemed suitable as an improved alternative to the dichotomous categories of emotions. Numerous psychological studies performed have found supporting evidence and it has gradually been incorporated into a few studies using sentiment analysis (Borth, Chen, Ji, and Chang, 2013; Chafale and Pimpalkar, 2014; Terada, Yamauchi, and Ito, 2012). The Plutchik's emotional wheel is also more suitable as the model has a more balanced amount of positive and negative emotions.

Therefore, the emotional states proposed in Plutchik's emotional wheel is used by the current study.

2.5 SUMMARY

As a summary of the findings above, the importance of online reviews on sharing accommodation is established but the current state of the research on the experience of this group of customers is still limited. Although there have been some conversations being made in this field of research, many of these studies are still very much reliant on the readily available information from websites and online platforms. As explained above, most of these studies have taken the aspects listed by the platforms where data are collected. Subsequently, sentiments analysis is also performed based on the aspects provided. However, such methods cannot provide an accurate understanding on 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). Furthermore, the sentiment analysis performed by these studies must be interpreted with caution as the positive bias in Airbnb's reviews posted on the site has not been shown to be improved (Fradkin, Grewal, and Holtz, 2018; Bridges and Vasquez, 2017).

Therefore, in order to have an accurate understanding of the customer experience of Airbnb, it has been decided that the study should refrain from using the readily available aspects. As pointed out by Dolnicar and Otter (2003), customers tend to evaluate different types of accommodations with different sets of standards. In other words, there should be set of aspects that are deemed important by the customers of Airbnb when they are using the company's services. To the best of the researcher's knowledge, no study has been conducted yet. Furthermore, there should also be a range of sentiments experienced by the customers on the different aspects of Airbnb.

Hence, the data collection should go beyond the company's own website but to include other independent sites while the Airbnb platform can act as a baseline. Following the extraction of data, the extraction of aspects unique to Airbnb textual reviews also provides results that are unique to the Airbnb customers. Specifically, the findings can reveal the set of expectations that customers have when using the services of Airbnb. With the unique set of aspects, sentiment analysis based on the emotional wheel provides a more nuanced understanding of the emotions experienced by the customers when interacting with the company's various forms services.

CHAPTER 3

METHODOLOGY

The flowchart below illustrates the steps taken from data collection to data validation. The following section introduces the dataset as well as the data extraction methods, and tools used. Subsequently, the data preparation steps are explained. Following the explanation for dataset and the preparatory steps, the data modelling algorithms are introduced and explained in detail. Lastly, the evaluation metrics used to evaluate the model performance are also introduced.

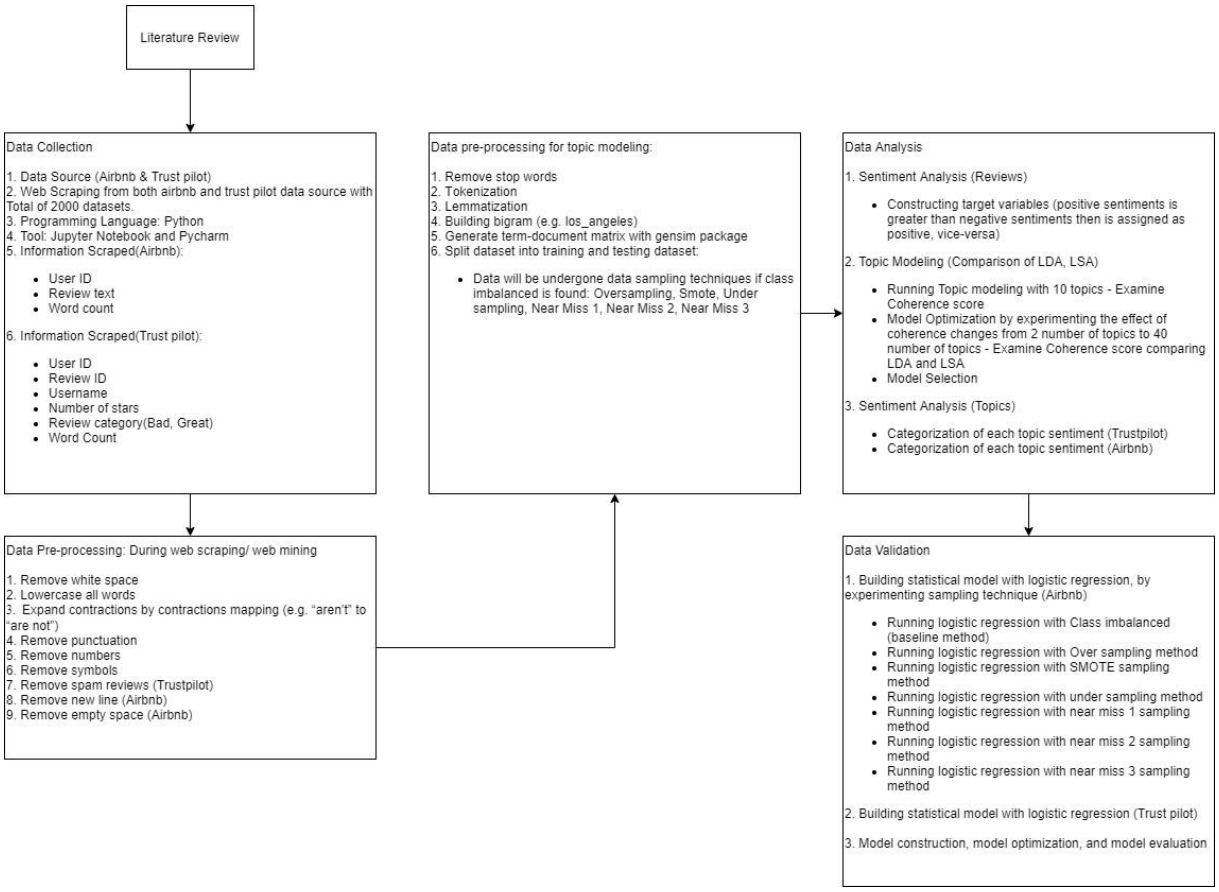


Figure 3.1 Process diagram of project methodology

3.1 DATASET AND DATA COLLECTION METHODS

As mentioned in the previous paragraphs, the potential presence of the positive bias in the Airbnb platform makes the reliability of the reviews at Airbnb platform questionable. In order to achieve greater objectivity, the reviews are also collected from another independent review site, Trustpilot. In doing so, the results from the reviews of Airbnb can become the baseline as well as to verify if the changes made by Airbnb has indeed improved the review system. As the data are available online, the extraction of the data is performed through web mining with scripts written in Python programming language.

3.2 DATA PRE-PROCESSING PROCEDURES

After the data is obtained through web mining, a series of pre-processing procedures are carried out on the raw dataset. The procedures are:

- Remove white space
- Lowercase all words
- Expand contractions (e.g. “aren’t” to “are not”)
- Remove punctuations
- Remove numbers
- Remove symbols
- Remove spam reviews (Trustpilot)
- Remove new line (Airbnb)
- Remove empty space (Airbnb)
- Perform word count

These steps are carried out so that the next analysis can be carried out in a more efficient manner and the unnecessary symbols will not cause any issue to the subsequent analyses. A word cloud is also generated to inspect for differences between the reviews extracted from the different platforms. After that, the dataset goes through another series of steps to generate the input for the modelling algorithm. The steps are:

- Remove stop words
- Tokenization
- Lemmatization
- Generate term-document matrix
- Split dataset into training and testing dataset

3.1 ALGORITHMS USED

In the new few sections, several algorithms used for sentiment analysis, topic modelling, and statistical modelling are introduced. The fundamentals and their application on the current project are explained so that the output can be better understood.

3.2 LATENT DIRICHLET ALLOCATION (LDA)

According to the original author of the algorithm, Blei and colleagaues (2003), stated that Latent Dirichlet Allocation (LDA) assumes that several hidden topics are present in any body of but with varying level of weights. From the name of the algorithm, "Dirichlet" is actually a type of probability distribution while the word "latent" connotates the meaning of something hidden. In other words, LDA extracts the hidden topics by using the probability distribution of the words within each document it analyses (Blei et al., 2003). The figure below is an illustration of all the parameters of LDA.

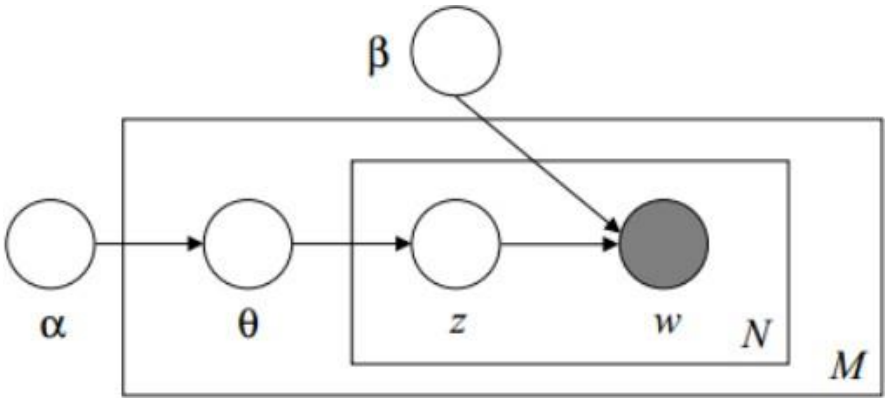


Figure 3.2 Parameter of LDA

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.

Referring to the diagram above, the position of the various parameters indicates their level of influence on the algorithm. For example, parameters α and β are placed at the outermost rectangle, which means that these two parameters' level of influence limited at the document level only. As specified above, α indicates the number of topics in a document while β is the number of words in each topic. Therefore, a high α indicates high number of topics in a document while a high β indicates that there is a high number of words associated with a particular topic. In most situations, a term-document matrix is required as in an input for the implementation of LDA. A term-document matrix can be visualised as a table where each row represents a document while each column represents the words appeared in all the documents combined. In the simplest form of LDA, each element will contain raw counts of each word's appearance in each document. However, such method is unreliable, so most common practices replace raw count with Term Frequency–Inverse Document Frequency (TF-IDF) normalized where the weight of each word is considered. A word with high weightage will have high occurrence in one document while low occurrence across the entire corpus.

Based on the list of parameters above, the most important parameter is the number of topics, k. With the term-document matrix, LDA assigns words into a user-defined number, *k* number of topics. At the same time, two matrices that contains two types of probabilities will be generated by LDA. The first type of probability 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:

$$\text{topic}, t_i = p\ (\text{topic}, t_i \mid \text{document}, d) \tag{3.1}$$

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

$$\text{word}, w_i = p\ (\text{word}, w_i \mid \text{topic}, t_i) \tag{3.2}$$

Subsequently, the two probabilities generated will be multiplied to determine the probability of each topic producing a specific word. Then, the assignment of the words will be based on the product of the two probabilities where words with the highest corresponding probability will be assigned. This iterative process assigns each word into one topic and stops until the assignment of words has produced a meaningful output. By applying the algorithm onto the dataset of the current study, the textual reviews collected is analysed to produce a number of topics where each topic contain several keywords deemed to be most associated with

the particular topic. Based on the output, each topic can be profiled into various aspects of Airbnb.

3.3 LATENT SEMANTIC ANALYSIS (LSA)

Another popular choice of topic modelling is the Latent Semantic Analysis (LSA). While it is similar to LDA in terms of input and output, the underpinnings of LSA is different from LDA. As a variant of Principal Component Analysis (PCA), LSA is actually a dimension reduction algorithm that works based on linear algebra while LDA uses Dirichlet probability. Taking the term-document matrix as an input, LSA performs topic modelling by decomposing each document where each document (in rows) is represented by the weight of the words (in columns) in the matrix.

3.4 LEXICON BASED SENTIMENT ANALYSIS

Following the analyses using topic modelling, a lexicon-based sentiment analysis is performed, and each aspect generated are categorised using the Plutchik's emotional wheel. Lexicon based sentiment analysis works by taking a bag of words as input and calculate the sentiment scores of each word to return a sentiment score. In order to do so, it requires a collection of words that are labelled with negative or positive sentiment scores (Jurek, Mulvenna, and Bi, 2015). To be specific, a dictionary that contains sentiment score labels is compared against the bag of words obtained from the raw corpus. As shown in the diagram below, each individual word from the corpus will be matched against the lexicon. Subsequently, any overlap (words that are present in the lexicon) will be selected. Then, the sentiment scores of the selected words will be summed to return an overall sentiment score.

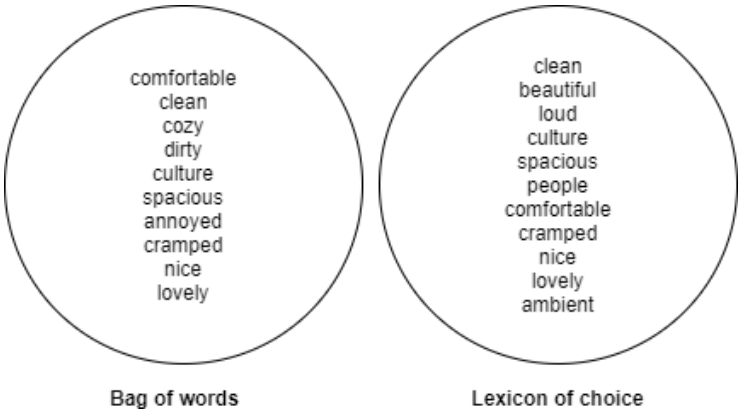


Figure 3.3 Pairing of the bag of words and lexicon of choice

Figure 3.4 Overlap of word between bag of words and lexicon

Following the overall sentiment score, the sentiment score will be used to categorize each aspect into one of the 8 sentiments in Plutchick's emotional wheel. For the purpose of the current project, the emotional lexicon developed by National Research Council (Mohammad and Turney, 2003) is used. This lexicon is adopted because it is deemed as more nuanced in terms of capturing the variety of human sentiments by having 8 emotional states. In doing so, the type of sentiments experienced by Airbnb customers on the various aspects of the company's services.

3.5 LOGISTIC REGRESSION

In order to validate the model produced through topic modelling, a statistical test is performed. For the purpose of this project, logistic regression is chosen as the predictive modelling approach. As a variant of the linear regression, logistic regression has the same theoretical underpinnings and model parameters with linear regression. It is similar to linear regression where the model uses the combined effects of the variable impact, b_n, and feature values, X_n to predict the likelihood of the outcome variable. However, it is different from linear regression whereby the main purpose of logistic regression is to produce the probability of the outcome variable, Y_i variable and not predicting any continuous value (Sperandei, 2014). The equation of logistic regression and the parameters mentioned above are denoted as below:

$$(Probability\ of\ outcome)\ Y_i = b_1X_1 + b_2X_2 + \ldots b_nX_n + b_0 \quad (3.3)$$

In addition to that, logistic regression is chosen as the statistical model due to the constraints of the target variable as the customers' sentiments are categorised into discrete

emotional states. Furthermore, the flexibility to allow both continuous and categorical variables in its' prediction makes it especially useful for the current project. Besides, the potential issue of non-linearity due to the presence of categorical variables can also be overcome by logistic regression through the use of logarithmic values in its prediction.

3.6 EVALUATION METRICS

In order to evaluate the models formed, several model evaluation metrics have been chosen for each algorithm. For the purpose of evaluating topic modelling models, topic coherence is chosen as another metric to evaluate the output of topic modelling. To reiterate, topic modelling assigns words into topics based on the likelihood of certain topics producing such keywords. Hence, in every topic extracted by the algorithm, few words will have higher scores. Topic coherence works by measuring the semantic similarity between the few high scoring words in any topic. Furthermore, the higher the topic coherence score, the more meaningful or "human interpretable" are the topics. In short, the perplexity and topic coherence are used to select the best model based on mathematical criteria that is complemented with meaningful information that translates well into business requirements.

In addition to the evaluation of topic models, the logistic regression models produced are also evaluated with few metrics. The metrics are:

- Accuracy
- Receiver operating characteristic (ROC) / Area under the curve (AUC)
- Precision and recall
- f1 score

The accuracy of the model measures the model's prediction accuracy by calculating the ratio between the total number of correct predictions over the total number of predictions made. The formula is as below:

$$Accuracy = Total\ TP + Total\ TN / Total\ number\ of\ predictions \quad (3.4)$$

As a visualization on the model accuracy, the AUC (area under the curve) plots the model sensitivity and specificity. The model will also be compared to the baseline model to see the amount of improvement made by the algorithm. In general, the higher the percentage of AUC, the better the model. In addition to the model accuracy, the model recall is also evaluated, which is the model's ability to detect all the relevant cases. The formula to calculate recall is as

below, which is the number of true positives over the total of true positives (TP) and false negatives (FN).

$$Recall = TP/ (TP + FN) \quad (3.5)$$

However, being able to detect is not sufficient. Therefore, the precision of the model is also taken into consideration. Model precision is the model's ability to detect ONLY the relevant cases without making too many false alarms. The formula to obtain the model precision is the number of true positives divided by the total of true positives and false positives.

$$Precision = TP / (TP + FP) \quad (3.6)$$

With the usage of precision of recall, it is important to have a model that is sensitive enough to detect all relevant cases and yet stringent enough to not label all cases as positive data. In other words, a balance between precision and recall is important for an optimal model. In this case, f1 score measures the balance between precision and recall. Therefore, f1 score is also included in the evaluation of the statistical model produced.

3.7 SUMMARY

In a nutshell, the current project chose topic modelling and fine-grained sentiment analysis as the main analytical approach based on the consideration that the data contains primarily unstructured textual data that will require it to be represented in a numerical manner in order to be analysed. The regression algorithm is chosen in order to validate the topic models into analytical model as well as to predict the customer sentiments based on the topic's models extracted.

CHAPTER 4

EXPERIMENTATION

4.1 DATA COLLECTION

1000 customer reviews are collected from Airbnb and Trustpilot which results in a total number of 2000 customer reviews from both platforms. The sample size is kept at 2000 to ensure that the model findings remain interpretable and does not consume too much computing resource during implementation and modelling. As explained in the introduction and literature review, there is a positive bias in the reviews section of Airbnb. The collection of data is performed using web scraping technique. Furthermore, the program for web scraping is written in Python programming language. As explained before, the inclusion of two different platforms is to form a comparison in terms of aspects and customer sentiments between the company's own site and another independent site.

4.2 DATA PRE-PROCESSING FOR TOPIC MODELING

Although there are two different data sources, the data pre-processing steps performed on both datasets are identical. The pre-processing begins with importing the dataset, as shown in the screenshot below.

```
# setting up testing and training sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=27)
```

```
print("training set: ", X_train.shape, y_train.shape)
print("testing set: ", X_test.shape, y_test.shape)
```

```
training set:  (799, 7) (799,)
testing set:  (200, 7) (200,)
```

Figure 4.1 Import dataset

Then, the stopwords are removed from the original dataset so that such words that may cause some issues in the term-document matrix are removed to ensure only important words are retained.

```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['etc','give','go','to','the','this','not','of'])
```

Figure 4.2 Stop words removal

Following the removal of stopwords, all the reviews are tokenized so that each passage of reviews will be converted to individual tokens.

```
def tokenize_word(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))  # deacc=True removes punctuations

def tokenize_corpus(data):
    data_words = list(tokenize_word(data))
    return data_words

data_words = tokenize_corpus(X_train.review_text.values.tolist())
```

Figure 4.3 Tokenize documents

The next step is to build Bigrams so that the entity within the reviews can be recognized during the lemmatization process. For example, directly lemmatize the names of location (e.g. Los Angeles) may result in inaccurate results.

```
# Build the biagram and trigram model with gensim Phrases()
bigram = gensim.models.Phrases(data_words, min_count=5, threshold=100) # higher threshold fewer phrases.
trigram = gensim.models.Phrases(bigram[data_words], threshold=100)

# Faster way to get a sentence clubbed as a trigram/bigram
bigram_mod = gensim.models.phrases.Phraser(bigram)
trigram_mod = gensim.models.phrases.Phraser(trigram)

# See trigram example
print(trigram_mod[bigram_mod[data_words[0]]])
```

```
['place', 'as', 'described', 'and', 'having', 'an', 'almost', 'all', 'day', 'check', 'in', 'counter', 'really', 'make
s', 'the', 'process', 'lot', 'easier', 'fell', 'ill', 'during', 'one', 'of', 'the', 'days', 'in', 'malacca', 'and', '
the', 'property', 'manager', 'was', 'very', 'kind', 'to', 'extend', 'my', 'checkout', 'time', 'closer', 'towards', 'm
y', 'coach', 'departure', 'time', 'nice', 'condominium', 'with', 'great', 'pool', 'facilities', 'and', 'view', 'love
', 'the', 'breeze', 'in', 'the', 'common', 'space', 'and', 'overall', 'experience', 'will', 'be', 'lot', 'nicer', 'on
ce', 'all', 'the', 'renovations', 'and', 'construction', 'in', 'the', 'surrounding', 'is', 'complete']
```

Figure 4.4 Building bigrams

After the bigrams are generated, the bigrams are lemmatized to return the words into their root from. Furthermore, only nouns, adjectives, verbs, and adverbs are retained in the dataset.

```
def preprocess_data(data_words):
    # Remove Stop Words
    data_words_nostops = remove_stopwords(data_words)

    # Form Bigrams
    data_words_bigrams = make_bigrams(data_words_nostops)

    # Initialize spacy 'en' model, keeping only tagger component (for efficiency)
    # python3 -m spacy download en
    nlp = spacy.load('en', disable=['parser', 'ner'])

    # Do lemmatization keeping only noun, adj, vb, adv
    data_lemmatized = lemmatization(data_words_bigrams, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV'])

    return data_lemmatized

data_lemmatized = preprocess_data(data_words)
print(preprocess_data(data_words)[:1])
```

```
[['place', 'describe', 'almost', 'day', 'check', 'counter', 'really', 'make', 'process', 'lot', 'easy', 'fall', 'ill
', 'day', 'malacca', 'property', 'manager', 'kind', 'extend', 'checkout', 'time', 'closer', 'coach', 'departure', 'ti
me', 'nice', 'condominium', 'great', 'pool', 'facility', 'view', 'love', 'breeze', 'common', 'space', 'overall', 'exp
erience', 'lot', 'nice', 'renovation', 'construction', 'surround', 'complete']]
```

Figure 4.5 Lemmatization of bigram

Finally, the last step is to generate the term-document matrix that will become the input for the topic modelling algorithm later.

```
def prepare_corpus(doc_clean):
    """
    Input  : clean document
    Purpose: create term dictionary of our courpus and Converting list of documents (corpus) into Document Term Matrix
    Output : term dictionary and Document Term Matrix
    """
    # Creating the term dictionary of our courpus, where every unique term is assigned an index. dictionary = corpora.D
    dictionary = corpora.Dictionary(doc_clean)
    # Converting list of documents (corpus) into Document Term Matrix using dictionary prepared above.
    doc_term_matrix = [dictionary.doc2bow(doc) for doc in doc_clean]
    # generate LDA model
    return dictionary,doc_term_matrix
```

Figure 4.6 Generate term-document matrix

4.3 DATA ANALYSIS (TOPIC MODELING)

The current section details all the steps taken in order to build topic models. However, the screenshots used are the topic construction steps taken for data extracted from Airbnb. Although there may be some differences in the figures, the steps taken are nonetheless identical for both platforms. Differences in findings are also highlighted and discussed throughout the entire section. The first step of the data analysis begins by loading of all the required packages such as re, spacy, genism, pandas, and NumPy for text processing and data manipulation.

Import library

```
import re
import numpy as np
import pandas as pd
from pprint import pprint

# Gensim
import gensim
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
from gensim.models import CoherenceModel

# spacy for lemmatization
import spacy

# Plotting tools
import pyLDAvis
import pyLDAvis.gensim  # don't skip this
import matplotlib.pyplot as plt
%matplotlib inline

# Enable logging for gensim - optional
import logging
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.ERROR)

import warnings
warnings.filterwarnings("ignore",category=DeprecationWarning)
```

Figure 4.7 Load all required packages

Then, the processed dataset is loaded and displayed to verify the correct file is used. There is a total number of 1000 of reviews collected.

```
# Import dataset
airbnbDF = pd.read_csv("C:\\Users\RexPC\\Desktop\\Capstone\\TP034717\\dataset\\scraped_text_air_bnb_4.csv")
```

```
airbnbDF.head()
```

	user_id	user_name	review_text	word_count
0	147929067	Vin Li	great location really quick replies well equip...	58
1	147929067	Vin Li	the host is very responsive and helpful the ap...	102
2	147929067	Vin Li	i am quite disappointed with the host due to t...	131
3	147929067	Vin Li	the suites near kl sentral station are well eq...	75
4	147929067	Vin Li	very suitable to do a business trip just like ...	84

```
len(airbnbDF)
```

```
999
```

Figure 4.8 Load dataset

After the dataset is loaded, a target variable is constructed by first running a sentiment analysis on the dataset and each review is categorised into either positive or negative sentiment.

Getting target variable

```
import lbsa

# refine word_count
airbnbDF['word_count'] = airbnbDF['review_text'].apply(lambda x: len(x.split()))

nrc_lexicon = lbsa.get_lexicon('opinion', language='english', source='nrc')

airbnbDF['positive_sentiment'] = airbnbDF['review_text'].apply(lambda x: nrc_lexicon.process(x).get('positive'))
airbnbDF['negative_sentiment'] = airbnbDF['review_text'].apply(lambda x: nrc_lexicon.process(x).get('negative'))

airbnbDF['sentiments'] = airbnbDF['review_text'].apply(lambda x: nrc_lexicon.process(x))
airbnbDF['target'] = airbnbDF['sentiments'].apply(lambda x: 1 if x.get('positive') >= x.get('negative') else 0)
```

Figure 4.9 lexicon-based sentiment analysis to create target variable

As shown in the screenshot below, the categorisation of the review sentiment is based on the overall sentiment score. The target variable is a binary variable with 1 as positive while 0 is negative.

```
airbnbDF.head()
```

	user_id	user_name	review_text	word_count	positive_sentiment	negative_sentiment	sentiments	target
0	147929067	Vin Li	great location really quick replies well equip...	59	5	1	{'positive': 5, 'negative': 1}	1
1	147929067	Vin Li	the host is very responsive and helpful the ap...	102	10	0	{'positive': 10, 'negative': 0}	1
2	147929067	Vin Li	i am quite disappointed with the host due to t...	131	6	7	{'positive': 6, 'negative': 7}	0
3	147929067	Vin Li	the suites near kl sentral station are well eq...	75	4	1	{'positive': 4, 'negative': 1}	1
4	147929067	Vin Li	very suitable to do a business trip just like ...	85	9	0	{'positive': 9, 'negative': 0}	1

```
print(airbnbDF.shape)
```

```
(999, 8)
```

Figure 4.10 Target variable created

After the creation of the target variable, the distribution of the two sentiments are first inspected. Displayed in the output below, there are much more positive comments in the current dataset, which consists of comments extracted from Airbnb platform. This may have resonated with the studies who have highlighted the positive bias on the platform. Despite the much smaller number, there are also negative comments posted on the platform. This may be

indicative that the changes made to the system may have started to see some improvements in terms of customer's willingness to voice their dissatisfaction on the platform.

```
unique_val = pd.DataFrame(airbnbDF)
```

```
unique_val.target.value_counts()
```

```
1    943
0     56
Name: target, dtype: int64
```

```
str(float("{:.2f}".format(len(airbnbDF[airbnbDF.target == 1]) / 999 *100))) + "%"
```

```
'94.39%'
```

```
str(float("{:.2f}".format(len(airbnbDF[airbnbDF.target == 0]) / 999 *100))) + "%"
```

```
'5.61%'
```

Figure 4.11 Sentiment distribution of reviews from Airbnb

Similar steps were also performed on the reviews extracted from Trustpilot, but the distribution of customer sentiments appears to be much more balanced when compared to Airbnb. As shown below, there are approximately 30% of negative comments and 60% of positive comments. This piece of finding appears to add on the evidence that the customers are much more willing to share their negative experience in platforms other than the company's own platform.

```
print(trustpilotDF.shape)
```

```
(999, 11)
```

```
# factorize target variable
char_cabin = trustpilotDF["target"].astype(str)     # Convert target to str
trustpilotDF["target"] = pd.Categorical(char_cabin)  # factorize the target var
```

```
my_tab = pd.crosstab(index=trustpilotDF["target"],  # Make a crosstab
                     columns="count")                  # Name the count column
# my_tab.columns = ["class1","class2","class3"]
my_tab.index= ["negative","positive"]
my_tab
```

col_0	count
negative	350
positive	649

Figure 4.12 Sentiment distribution of reviews from Trustpilot

Moving on, the dataset is split into 80:20 ratio for topic modelling. The code below shows the splitting of the dataset.

Splitting the dataset into training (80%) and testing (20%)

```
import pandas as pd
from sklearn import datasets, linear_model
from sklearn.model_selection import train_test_split
from matplotlib import pyplot as plt
from sklearn.utils import resample
```

```
y = airbnbDF.target
```

```
# columns = "user_id user_name review_text word_count positive_sentiment negative_sentiment sentiments".split()
# X = pd.DataFrame(airbnbDF, columns=columns)
```

```
X = airbnbDF.drop('target', axis=1)
```

```
# setting up testing and training sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=27)
```

```
print("training set: ", X_train.shape, y_train.shape)
print("testing set: ", X_test.shape, y_test.shape)
```

```
training set:  (799, 7) (799,)
testing set:  (200, 7) (200,)
```

Figure 4.13 Splitting dataset to train topic models

As displayed below, the class imbalance issue remains for both training and testing set whereby each set contains more than 90% positive reviews and approximately 5-6% negative reviews. Although the severe class imbalance may pose issue to the predictive modelling, the dataset is left as its' original form during topic modelling. Steps to overcome the class imbalance issue were taken during the predictive modelling stage.

Figure 4.14 Sentiment distribution for dataset from Airbnb (left) and Trustpilot (right)

Both datasets are saved before the construction of topic models, as shown in the screenshot of codes below.

```python
# Saving dataset for topic modeling
import pickle
with open('X_train.pkl', 'wb') as f:
    pickle.dump(X_train, f)

with open('y_train.pkl', 'wb') as f:
    pickle.dump(y_train, f)

with open('X_test.pkl', 'wb') as f:
    pickle.dump(X_test, f)

with open('y_test.pkl', 'wb') as f:
    pickle.dump(y_test, f)
```

Figure 4.15 Saving training and testing dataset

Before the topic modelling is performed, a word cloud is also generated to inspect the general themes within the reviews collected from both sites. Based on the word clouds generated, there are some noticeable difference in regard to the themes of the reviews from both sites. On one hand, the reviews from Airbnb have primarily mentioned about the accommodation and its' amenities while the reviews from Trustpilot have focused on their experiences with the host and company. As the differences can already be seen from the word clouds, the next analyses will further verify if there is significant difference between the two sites.

Figure 4.16 Word cloud for reviews from Airbnb (left) and Trustpilot (right)

As explained in the previous sections, one of the most important parameters for topic modelling is the number of topics. In the first implementation, the number of topics is first set to 10 and the coherence score is observed. The implementation of both LDA and LSA models produced the output below where the assignment of each keywords into a particular topic is displayed.

Running LDA, LSA and PLSA model and see if word distribution per topic looks sensible

1. First approach - Run various models with 10 number of topics and examine the coherence scores

```
number_of_topics=10
words=10

#LSA Model
print("Building LSA Model:")
lsi_model = create_gensim_lsa_model(data_lemmatized,number_of_topics,words)
lsi_model.save('lsi_train.model')

# LDA Model
print("\n")
print("Building LDA Model:")
lda_model = create_gensim_lda_model(data_lemmatized,number_of_topics,words)
lda_model.save('lda_train.model')
```

```
Building LSA Model:
[(0, '0.397*"place" + 0.369*"stay" + 0.202*"clean" + 0.195*"host" + 0.186*"good" + 0.171*"apartment" + 0.168*"great"
+ 0.157*"also" + 0.156*"room" + 0.141*"would"'), (1, '-0.669*"place" + 0.271*"apartment" + 0.249*"room" + 0.178*"wate
r" + -0.166*"great" + 0.147*"bed" + 0.132*"check" + 0.107*"good" + 0.103*"day" + 0.095*"issue"'), (2, '0.358*"stay" +
-0.303*"apartment" + -0.254*"nice" + -0.208*"good" + -0.203*"host" + 0.191*"room" + -0.189*"really" + -0.177*"bed" +
0.168*"day" + 0.165*"check"'), (3, '-0.597*"stay" + 0.405*"place" + 0.313*"room" + 0.192*"good" + -0.189*"apartment"
+ -0.157*"great" + 0.145*"water" + -0.130*"host" + 0.089*"house" + 0.086*"area"'), (4, '0.456*"good" + -0.280*"apartm
ent" + 0.254*"room" + -0.226*"place" + 0.207*"nice" + 0.169*"location" + 0.158*"stay" + 0.158*"clean" + -0.146*"get"
+ -0.134*"day"'), (5, '0.432*"host" + -0.324*"apartment" + 0.299*"also" + -0.237*"stay" + -0.224*"place" + 0.159*"hug
e" + 0.136*"house" + -0.114*"issue" + -0.114*"good" + 0.105*"want"'), (6, '0.506*"host" + -0.321*"also" + -0.285*"pro
vide" + -0.265*"bed" + 0.245*"good" + 0.174*"check" + -0.156*"pool" + 0.112*"clean" + -0.109*"place" + 0.105*"apartme
nt"'), (7, '0.525*"clean" + -0.295*"stay" + -0.249*"host" + -0.218*"room" + 0.218*"apartment" + 0.216*"also" + 0.178
*"property" + 0.171*"great" + 0.164*"location" + 0.129*"unit"'), (8, '0.330*"good" + -0.288*"room" + -0.285*"great" +
-0.284*"nice" + 0.223*"stay" + 0.214*"house" + -0.209*"host" + -0.171*"check" + -0.134*"also" + 0.131*"get"'), (9, '
0.354*"great" + -0.251*"host" + -0.250*"unit" + -0.243*"clean" + 0.233*"apartment" + -0.210*"nice" + -0.206*"bed" +
0.177*"good" + 0.145*"get" + 0.142*"also"')]
```

```
Building LDA Model:
[(0, '0.021*"early" + 0.014*"new" + 0.012*"room" + 0.012*"safe" + 0.012*"pretty" + 0.011*"drive" + 0.011*"leave" + 0.
011*"take" + 0.011*"store" + 0.011*"smell"'), (1, '0.024*"get" + 0.024*"time" + 0.020*"house" + 0.019*"friendly" + 0.
019*"thing" + 0.019*"little" + 0.018*"check" + 0.017*"much" + 0.015*"bit" + 0.014*"look"'), (2, '0.032*"perfect" + 0.
031*"next" + 0.022*"trip" + 0.021*"large" + 0.019*"cook" + 0.018*"group" + 0.018*"local" + 0.018*"landlord" + 0.017*"
see" + 0.017*"nearby"'), (3, '0.042*"parking" + 0.025*"seem" + 0.023*"future" + 0.018*"worth" + 0.018*"value" + 0.017
*"modern" + 0.015*"process" + 0.015*"money" + 0.014*"thoughtful" + 0.013*"detail"'), (4, '0.042*"bring" + 0.023*"towe
r" + 0.022*"cool" + 0.018*"city" + 0.018*"window" + 0.017*"best" + 0.015*"holiday" + 0.014*"decorate" + 0.013*"channe
l" + 0.012*"traveller"'), (5, '0.055*"place" + 0.047*"stay" + 0.031*"clean" + 0.029*"host" + 0.026*"great" + 0.020*"g
ood" + 0.019*"also" + 0.019*"nice" + 0.018*"apartment" + 0.018*"location"'), (6, '0.032*"shopping" + 0.025*"return" +
0.022*"awesome" + 0.021*"attraction" + 0.021*"stain" + 0.020*"expectation" + 0.020*"reach" + 0.018*"choice" + 0.014*"
lift" + 0.014*"furniture"'), (7, '0.036*"style" + 0.021*"ceiling" + 0.019*"walkable" + 0.019*"flight" + 0.016*"consta
ntly" + 0.016*"remember" + 0.015*"feature" + 0.015*"thought" + 0.014*"staying" + 0.013*"washing"'), (8, '0.030*"room"
+ 0.020*"unit" + 0.019*"floor" + 0.018*"bed" + 0.017*"bathroom" + 0.013*"use" + 0.012*"especially" + 0.012*"water" +
0.010*"could" + 0.010*"shower"'), (9, '0.017*"buy" + 0.017*"construction" + 0.015*"building" + 0.014*"quality" + 0.01
4*"kuche" + 0.014*"cost" + 0.012*"complex" + 0.011*"never" + 0.011*"know" + 0.011*"description"')]
```

Figure 4.17 Implementation of first LDA and LSA models

In terms of the evaluation of the models, the model perplexity and topic coherence are used but higher priority is given to topic coherence due to its' association with interpretability. As shown in below screenshots, the models with 10 topics have achieved moderate model perplexity and topic coherence. Similar findings have also been found for reviews extracted from Trustpilot. The table below summarizes the topic coherence score for the two models.

```
coherence_model_lda = CoherenceModel(model=lda_model, texts=data_lemmatized, dictionary=dictionary, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score on LDA: ', coherence_lda)
```

```
Coherence Score on LDA:  0.4155596486530186
```

```
coherence_model_lsi = CoherenceModel(model=lsi_model, texts=data_lemmatized, dictionary=dictionary, coherence='c_v')
coherence_lsi = coherence_model_lsi.get_coherence()
print('\nCoherence Score on LSA/LSI: ', coherence_lsi)
```

```
Coherence Score on LSA/LSI:  0.38455248742708326
```

Figure 4.18 Topic coherence of LDA and LSA models (Airbnb)

```
coherence_model_lda = CoherenceModel(model=lda_model, texts=data_lemmatized, dictionary=dictionary, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print('\nCoherence Score on LDA: ', coherence_lda)
```

```
Coherence Score on LDA:  0.4259732071930532
```

```
coherence_model_lsi = CoherenceModel(model=lsi_model, texts=data_lemmatized, dictionary=dictionary, coherence='c_v')
coherence_lsi = coherence_model_lsi.get_coherence()
print('\nCoherence Score on LSA: ', coherence_lsi)
```

```
Coherence Score on LSA:  0.340973900733201
```

Figure 4.19 Topic coherence of LDA and LSA models (Trustpilot)

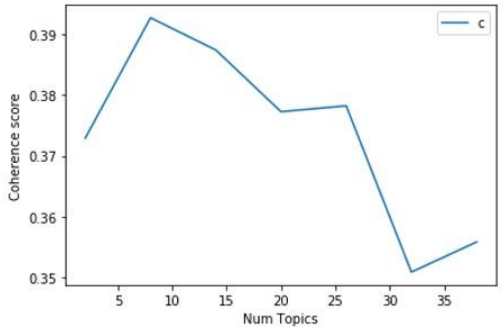
Table 4.1 Topic Coherence of initial models

Source	Model	Number of topics	Topic Coherence (C_v)
Airbnb	LDA	10	0.42
	LSA	10	0.38
Trustpilot	LDA	10	0.43
	LSA	10	0.34

In order to find other potentially better models, models with number of topics ranging from 2 to 40 are produced. Using the method suggested by Kamal (2018), a plot which shows the changes in topic coherence is used to discover the optimal number of topics.

```
model_list_lda, coherence_values_lda = compute_coherer
<
```

```
limit=40; start=2; step=6;
x = range(start, limit, step)
plt.plot(x, coherence_values_lda)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```




```
limit=40; start=2; step=6;
x = range(start, limit, step)
plt.plot(x, coherence_values_lsa)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```

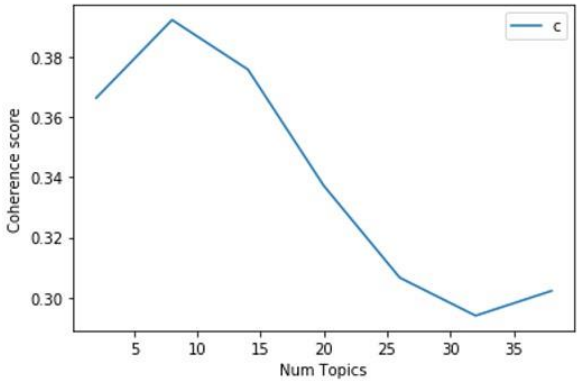


Figure 4.20 Topic coherence plot for Airbnb reviews LDA (left) and LSA (right)

Although one can choose for the number of topics that produces the highest topic coherence, it is still important to note that having too many topics can result in many redundancy in keywords and unneccessarily complex models (Kamal, 2018). Therefore, Kamal (2018) have suggested that the point in which the model ceases to show rapid growth in topic coherence is the point where optimal number of topics is located. Based on the plots above, the gain in topic coherence appears to plateau when approaching 10 topics.

On the other hand, the reviews extracted from Truspilot appears to have higher coherence score for models with higher amount of topics. Furthermore, LDA and LSA models have also provided very different findings. As shown below, LDA model has suggested around 14 topics for the highest coherence score while LSA model has suggested around 10 topics to achieve maximum coherence score. In this case, the actual coherence score is compared between LDA and LSA models to determine the best topic model to be used in the subsequent analysis.

```
limit=40; start=2; step=6;
x = range(start, limit, step)
plt.plot(x, coherence_values_lda)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```

```
limit=40; start=2; step=6;
x = range(start, limit, step)
plt.plot(x, coherence_values_lsa)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```

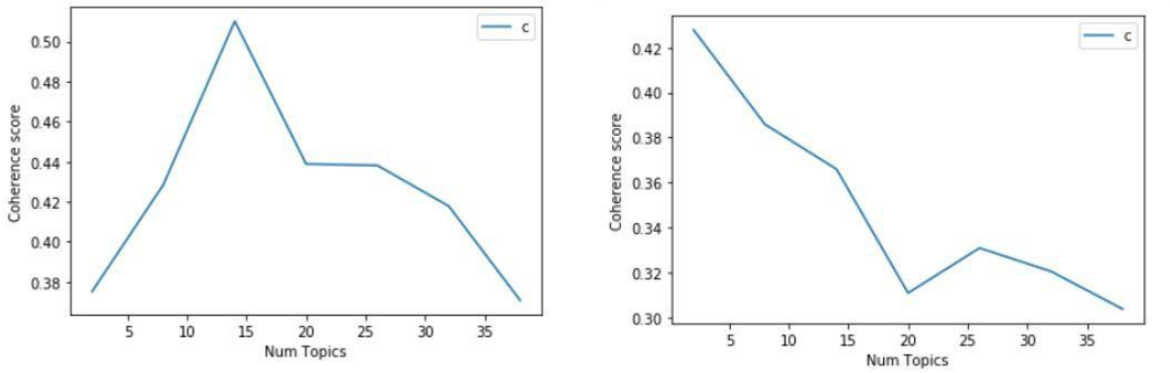


Figure 4.21 Topic coherence plot for Trustpilot reviews for LDA (left) and LSA (right)

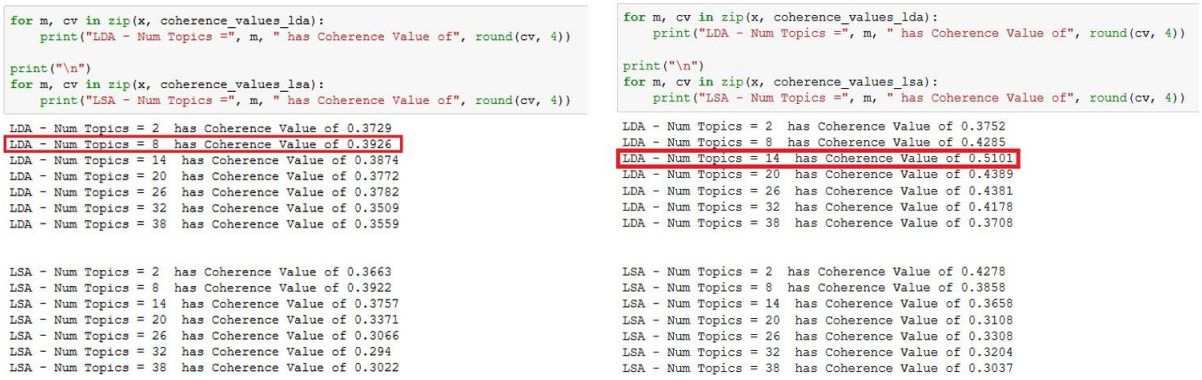


Figure 4.22 Screenshot of models coherence score of Airbnb (left) and Truspilot (Right)

By looking at all the topic coherence score, it appears that the most parsimonious yet highest topic coherence score for Airbnb reviews is the LDA model with 8 topics. Therefore, the model is selected as the best model to proceed for the subsequent analyses. Meanwhile, the model with highest topic coherence score for Trustpilot reviews is the LDA model with 14 topics. So, the model is selected for the subsequent analyses.

Table 4.2 Topic coherence scores of all models

	Airbnb		Trustpilot	
Model	**Number of Topics**	**Topic Coherence (C_v)**	**Number of Topics**	**Topic Coherence (C_v)**
LDA	2	0.3729	2	0.3752
	8	**0.3926**	8	0.4285
	14	0.3874	**14**	**0.5101**
	20	0.3772	20	0.4389
	26	0.3782	26	0.4381
	32	0.3509	32	0.4178
	38	0.3559	38	0.3708
LSA	2	0.3663	2	0.4278
	8	0.3922	8	0.3858
	14	0.3757	14	0.3658
	20	0.3371	20	0.3108
	26	0.3066	26	0.3308
	32	0.2940	32	0.3204
	38	0.3022	38	0.3037

After the selection of model, each topic should be profiled according to the keywords assigned to them. As shown below, the top scoring keywords are listed for each topic of the model.

Selecting LDA with 8 topics (it has the highest coherence and more promosing result)

```
# Select the model and print the topics
optimal_model_lda = model_list_lda[1]
model_topics = optimal_model_lda.show_topics(formatted=False)
pprint(optimal_model_lda.print_topics(num_words=8))
optimal_model_lda.save('lda_optimal.model')
```

```
[(0,
  '0.021*"beautiful" + 0.015*"see" + 0.015*"design" + 0.014*"photo" + '
  '0.014*"local" + 0.013*"station" + 0.012*"visit" + 0.012*"want"'),
 (1,
  '0.022*"room" + 0.014*"water" + 0.013*"get" + 0.013*"bed" + 0.013*"night" + '
  '0.013*"even" + 0.012*"first" + 0.012*"use"'),
 (2,
  '0.017*"problem" + 0.016*"try" + 0.014*"property" + 0.012*"toilet" + '
  '0.012*"dirty" + 0.012*"point" + 0.011*"different" + 0.011*"price"'),
 (3,
  '0.052*"place" + 0.043*"stay" + 0.032*"clean" + 0.027*"host" + 0.024*"great" '
  '+ 0.024*"good" + 0.017*"location" + 0.016*"also"'),
 (4,
  '0.028*"apartment" + 0.023*"large" + 0.020*"bring" + 0.019*"floor" + '
  '0.015*"pool" + 0.015*"quiet" + 0.013*"bed" + 0.013*"city"'),
 (5,
  '0.023*"home" + 0.019*"pool" + 0.017*"feel" + 0.016*"car" + 0.015*"morning" '
  '+ 0.013*"could" + 0.013*"help" + 0.013*"perfect"'),
 (6,
  '0.020*"return" + 0.017*"extra" + 0.017*"ever" + 0.016*"stain" + '
  '0.015*"happy" + 0.014*"trip" + 0.012*"photo" + 0.011*"beautiful"'),
 (7,
  '0.020*"shampoo" + 0.020*"communicate" + 0.013*"cockroach" + 0.012*"extra" + '
  '0.011*"slipper" + 0.011*"sufficient" + 0.010*"walkable" + 0.010*"plate"')]
```

Figure 4.23 LDA topics of Airbnb reviews

The profiling of the topics is also performed on the data extracted from Trustpilot based on the keywords. The table below summarizes the topic profiles alongside the keywords assigned to the topics.

Selecting LDA with topic 14 (it has the highest coherence and more promosing result)

```
# Select the model and print the topics
optimal_model_lda = model_list_lda[2]
model_topics = optimal_model_lda.show_topics(formatted=False)
pprint(optimal_model_lda.print_topics(num_words=14))
optimal_model_lda.save('lda_optimal.model')
```

```
[(0,
  '0.029*"together" + 0.024*"imagine" + 0.023*"assure" + 0.022*"fraudulent" + '
  '0.022*"procedure" + 0.021*"unknown" + 0.016*"violate" + 0.016*"reality" + '
  '0.016*"polite" + 0.014*"opportunity" + 0.012*"specialist" + 0.011*"unhappy" '
  '+ 0.011*"damage" + 0.010*"utensil"'),
 (1,
  '0.037*"whole" + 0.025*"resolution" + 0.023*"deposit" + 0.023*"manager" + '
  '0.022*"base" + 0.021*"evidence" + 0.020*"nowhere" + 0.019*"hide" + '
  '0.019*"shock" + 0.018*"line" + 0.017*"space" + 0.016*"dog" + 0.015*"submit" '
  '+ 0.015*"difficult"'),
 (2,
  '0.000*"nature" + 0.000*"narrative" + 0.000*"political" + 0.000*"perpetuate" '
  '+ 0.000*"partially" + 0.000*"ounce" + 0.000*"oppose" + 0.000*"noncare" + '
  '0.000*"prime" + 0.000*"implore" + 0.000*"instantly" + 0.000*"ideal" + '
  '0.000*"propose" + 0.000*"incidence"'),
 (3,
  '0.040*"reimburse" + 0.027*"phishing" + 0.022*"model" + 0.021*"cut" + '
  '0.017*"act" + 0.014*"super" + 0.014*"tour" + 0.012*"glad" + '
  '0.011*"identity" + 0.011*"slow" + 0.011*"expire" + 0.010*"afraid" + '
  '0.010*"dreadful" + 0.007*"scammed"'),
 (4,
  '0.052*"guy" + 0.040*"positive" + 0.040*"item" + 0.027*"learn" + '
  '0.027*"past" + 0.023*"human" + 0.021*"tiny" + 0.014*"video" + '
  '0.014*"corona" + 0.013*"prevent" + 0.012*"general" + 0.012*"trustpilot" + '
  '0.011*"basement" + 0.011*"summer"'),
 (5,
  '0.039*"airbnb" + 0.035*"host" + 0.019*"stay" + 0.015*"get" + 0.014*"refund" '
  '+ 0.012*"place" + 0.012*"review" + 0.012*"day" + 0.012*"money" + '
  '0.012*"guest" + 0.012*"book" + 0.011*"use" + 0.010*"take" + 0.010*"leave"'),
 (6,
  '0.022*"would" + 0.022*"book" + 0.022*"cancel" + 0.022*"call" + 0.018*"time" '
  '+ 0.016*"reservation" + 0.016*"airbnb" + 0.015*"try" + 0.015*"ask" + '
  '0.014*"account" + 0.014*"booking" + 0.014*"say" + 0.014*"tell" + '
  '0.013*"email"'),
 (7,
  '0.085*"key" + 0.056*"hard" + 0.042*"appear" + 0.022*"fire" + 0.021*"law" + '
  '0.020*"sad" + 0.017*"upset" + 0.017*"attend" + 0.016*"connect" + '
  '0.015*"sofa_be" + 0.015*"violation" + 0.014*"supply" + 0.011*"deserve" + '
  '0.010*"remote"'),
 (8,
  '0.046*"helpful" + 0.038*"kind" + 0.034*"bit" + 0.030*"quite" + '
  '0.029*"conversation" + 0.024*"stand" + 0.024*"instruction" + '
  '0.020*"honestly" + 0.019*"best" + 0.018*"strongly" + 0.018*"pandemic" + '
  '0.012*"manager" + 0.012*"concerned" + 0.012*"chain"'),
 (9,
  '0.035*"card" + 0.034*"fee" + 0.033*"word" + 0.031*"easy" + 0.030*"access" + '
  '0.025*"text" + 0.023*"traveller" + 0.019*"policy" + 0.016*"info" + '
  '0.016*"communicate" + 0.015*"virus" + 0.014*"similar" + 0.014*"useless" + '
  '0.014*"log"'),
 (10,
  '0.064*"inform" + 0.040*"decision" + 0.036*"profile" + 0.032*"city" + '
  '0.028*"free" + 0.024*"hang" + 0.023*"count" + 0.022*"bill" + 0.022*"final" '
  '+ 0.018*"video" + 0.017*"guarantee" + 0.017*"tax" + 0.017*"community" + '
  '0.016*"stuff"'),
 (11,
  '0.059*"clean" + 0.041*"bed" + 0.040*"dirty" + 0.034*"door" + '
  '0.024*"bathroom" + 0.022*"room" + 0.016*"private" + 0.015*"walk" + '
  '0.013*"window" + 0.013*"air" + 0.013*"large" + 0.012*"towel" + 0.012*"old" '
  '+ 0.012*"lock"'),
 (12,
  '0.082*"location" + 0.053*"shower" + 0.044*"unit" + 0.033*"stick" + '
  '0.031*"hot" + 0.027*"evening" + 0.023*"level" + 0.022*"cold" + 0.020*"old" '
  '+ 0.017*"proceed" + 0.017*"glitch" + 0.017*"nasty" + 0.016*"pm" + '
  '0.014*"sense"'),
 (13,
  '0.050*"disappointed" + 0.043*"rather" + 0.041*"therefore" + 0.027*"pocket" '
  '+ 0.022*"resort" + 0.020*"emergency" + 0.018*"compensate" + 0.013*"voice" + '
  '0.009*"satisfy" + 0.009*"fool" + 0.008*"declare" + 0.008*"withhold" + '
  '0.007*"paste" + 0.006*"next_morne"')]
```

Figure 4.24 LDA topics of Trustpilot reviews

4.4 DATA ANALYSIS (SENTIMENT ANALYSIS)

Following the extraction of topics and the selection of topic models for both platforms, a lexicon-based sentiment analysis is performed on each extracted topic. Take Airbnb for instance, the 8 extracted topics are analysed in order to reveal the sentiments held by customers towards all the mentioned aspects in their reviews. The screenshot below shows that the selected topic model is used.

Selecting LDA with 8 topics (it has the highest coherence and more promosing result)

```
# Select the model and print the topics
optimal_model_lda = model_list_lda[1]
model_topics = optimal_model_lda.show_topics(formatted=False)
pprint(optimal_model_lda.print_topics(num_words=8))
optimal_model_lda.save('lda_optimal.model')
```

```
[(0,
  '0.021*"beautiful" + 0.015*"see" + 0.015*"design" + 0.014*"photo" + '
  '0.014*"local" + 0.013*"station" + 0.012*"visit" + 0.012*"want"'),
 (1,
  '0.022*"room" + 0.014*"water" + 0.013*"get" + 0.013*"bed" + 0.013*"night" + '
  '0.013*"even" + 0.012*"first" + 0.012*"use"'),
 (2,
  '0.017*"problem" + 0.016*"try" + 0.014*"property" + 0.012*"toilet" + '
  '0.012*"dirty" + 0.012*"point" + 0.011*"different" + 0.011*"price"'),
 (3,
  '0.052*"place" + 0.043*"stay" + 0.032*"clean" + 0.027*"host" + 0.024*"great" '
  '+ 0.024*"good" + 0.017*"location" + 0.016*"also"'),
 (4,
  '0.028*"apartment" + 0.023*"large" + 0.020*"bring" + 0.019*"floor" + '
  '0.015*"pool" + 0.015*"quiet" + 0.013*"bed" + 0.013*"city"'),
 (5,
  '0.023*"home" + 0.019*"pool" + 0.017*"feel" + 0.016*"car" + 0.015*"morning" '
  '+ 0.013*"could" + 0.013*"help" + 0.013*"perfect"'),
 (6,
  '0.020*"return" + 0.017*"extra" + 0.017*"ever" + 0.016*"stain" + '
  '0.015*"happy" + 0.014*"trip" + 0.012*"photo" + 0.011*"beautiful"'),
 (7,
  '0.020*"shampoo" + 0.020*"communicate" + 0.013*"cockroach" + 0.012*"extra" + '
  '0.011*"slipper" + 0.011*"sufficient" + 0.010*"walkable" + 0.010*"plate"')]
```

Figure 4.25 Selection of optimal LDA model for Airbnb

As shown in the screenshot below, the input of the sentiment analysis contains the list of topics alongside the keywords assigned to each topic. Also explained in previous section, the lexicon-based sentiment analysis is performed by taking the list of words and compare them against a labelled lexicon to obtain the sentiment score. In other words, each topic will contain a bag of words and each bag of words is overlapped with the labelled lexicon.

Trustpilot Lexicon Sentiment Analysis

```
import numpy as np
import pandas as pd

# Reloading lexicon sentiment analysis csv built on trust pilot
airbnb_lexicon_df = pd.read_csv("airbnb_lexicon_sentiment.csv")
```

```
airbnb_lexicon_df.head()
```

	Topic	W1	W2	W3	W4	W5	W6	W7	W8
0	1	beautiful	see	desgin	photo	local	station	visit	want
1	2	room	water	get	bed	night	even	first	use
2	3	problem	try	property	toilet	dirty	point	different	price
3	4	place	stay	clean	host	great	good	location	also
4	5	apartment	large	bring	floor	pool	quiet	bed	city

Figure 4.26 Input of lexicon-based sentiment analysis (Airbnb)

The code below shows the inner workings of the lexicon-based sentiment analysis. Each word within every topic is compared against the lexicon and to obtain the sentiment score. Then, the sentiment score of each word is summed to finally obtain the overall sentiment that best represent the entire topic. The overall sentiment score results in each topic being categorised into one of the 8 emotional states in the Plutchik's emotional wheel.

```
j =[]
for i in range(8):
    new_dict = get_lexicon_model('topic' + str(i + 1))
    print('topic' + str(i + 1) + ': ' + str(new_dict))
    j.append(new_dict)
```

```
topic1: {'anger': 0, 'anticipation': 0, 'disgust': 0, 'fear': 0, 'joy': 1, 'sadness': 0, 'surprise': 0, 'trust': 0}
topic2: {'anger': 0, 'anticipation': 0, 'disgust': 0, 'fear': 0, 'joy': 0, 'sadness': 0, 'surprise': 0, 'trust': 0}
topic3: {'anger': 0, 'anticipation': 0, 'disgust': 2, 'fear': 1, 'joy': 0, 'sadness': 1, 'surprise': 0, 'trust': 0}
topic4: {'anger': 0, 'anticipation': 1, 'disgust': 0, 'fear': 0, 'joy': 2, 'sadness': 0, 'surprise': 1, 'trust': 2}
topic5: {'anger': 0, 'anticipation': 0, 'disgust': 0, 'fear': 0, 'joy': 0, 'sadness': 1, 'surprise': 0, 'trust': 0}
topic6: {'anger': 0, 'anticipation': 1, 'disgust': 0, 'fear': 0, 'joy': 1, 'sadness': 0, 'surprise': 0, 'trust': 1}
topic7: {'anger': 0, 'anticipation': 1, 'disgust': 1, 'fear': 0, 'joy': 2, 'sadness': 0, 'surprise': 1, 'trust': 1}
topic8: {'anger': 0, 'anticipation': 0, 'disgust': 0, 'fear': 0, 'joy': 0, 'sadness': 0, 'surprise': 0, 'trust': 1}
```

Figure 4.27 Categorisation of each topic sentiment (Airbnb)

Trustpilot Lexicon Sentiment Analysis

```
import numpy as np
import pandas as pd

# Reloading lexicon sentiment analysis csv built on trust pilot
trustpilot_lexicon_df = pd.read_csv("trust_pilot_lexicon_sentiment.csv")
```

```
trustpilot_lexicon_df.head()
```

	Topic	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14
0	1	together	imagine	assure	fraudulent	procedure	unknown	violate	reality	polite	opportunity	specialist	unhappy	damage	utensil
1	2	whole	resolution	deposit	manager	base	evidence	nowhere	hide	shock	line	space	dog	submit	difficult
2	3	nature	narrative	political	perpetuate	partially	ounce	oppose	noncare	prime	implore	instantly	ideal	propose	incidence
3	4	reimburse	phishing	model	cut	act	super	tour	glad	identity	slow	expire	afraid	dreadful	scammed
4	5	guy	positive	item	learn	past	human	tiny	video	corona	prevent	general	trustpilot	basement	summer

Figure 4.28 Input of lexicon-based sentiment analysis (Trustpilot)

By looking at the output below, the customer sentiment for Truspilot appears to be much more diverse for each aspect extracted. The table below summarizes the type of sentiment derived from the sentiment analysis for each topic mentioned by the reviewers from Trustpilot.

```
j =[]
for i in range(14):
    new_dict = get_lexicon_model('topic' + str(i + 1))
    print('topic' + str(i + 1) + ': ' + str(new_dict))
    j.append(new_dict)
```

```
topic1: {'anger': 3, 'anticipation': 2, 'disgust': 3, 'fear': 2, 'joy': 0, 'sadness': 2, 'surprise': 0, 'trust': 2}
topic2: {'anger': 1, 'anticipation': 1, 'disgust': 0, 'fear': 3, 'joy': 0, 'sadness': 0, 'surprise': 1, 'trust': 1}
topic3: {'anger': 0, 'anticipation': 1, 'disgust': 0, 'fear': 0, 'joy': 0, 'sadness': 0, 'surprise': 0, 'trust': 0}
topic4: {'anger': 1, 'anticipation': 2, 'disgust': 1, 'fear': 2, 'joy': 1, 'sadness': 2, 'surprise': 0, 'trust': 0}
topic5: {'anger': 0, 'anticipation': 0, 'disgust': 0, 'fear': 1, 'joy': 0, 'sadness': 0, 'surprise': 0, 'trust': 1}
topic6: {'anger': 1, 'anticipation': 1, 'disgust': 0, 'fear': 0, 'joy': 1, 'sadness': 1, 'surprise': 2, 'trust': 1}
topic7: {'anger': 0, 'anticipation': 1, 'disgust': 0, 'fear': 0, 'joy': 0, 'sadness': 1, 'surprise': 0, 'trust': 1}
topic8: {'anger': 3, 'anticipation': 1, 'disgust': 0, 'fear': 2, 'joy': 0, 'sadness': 2, 'surprise': 1, 'trust': 2}
topic9: {'anger': 0, 'anticipation': 0, 'disgust': 0, 'fear': 2, 'joy': 2, 'sadness': 2, 'surprise': 0, 'trust': 3}
topic10: {'anger': 1, 'anticipation': 0, 'disgust': 0, 'fear': 0, 'joy': 0, 'sadness': 0, 'surprise': 0, 'trust': 3}
topic11: {'anger': 0, 'anticipation': 0, 'disgust': 0, 'fear': 0, 'joy': 0, 'sadness': 1, 'surprise': 0, 'trust': 3}
topic12: {'anger': 0, 'anticipation': 0, 'disgust': 1, 'fear': 0, 'joy': 1, 'sadness': 0, 'surprise': 0, 'trust': 1}
topic13: {'anger': 2, 'anticipation': 0, 'disgust': 1, 'fear': 1, 'joy': 0, 'sadness': 1, 'surprise': 0, 'trust': 1}
topic14: {'anger': 1, 'anticipation': 1, 'disgust': 2, 'fear': 1, 'joy': 1, 'sadness': 2, 'surprise': 2, 'trust': 1}
```

Figure 4.29 Categorisation of each topic sentiment (Trustpilot)

4.5 DATA VALIDATION (STATISTICAL MODELLING)

Stemming from the findings above, the topic range for Airbnb appears to be much narrower than the reviews from Trustpilot. It is somewhat expected as the positive bias has been highlighted by previous studies. However, the current findings also appears to have confirmed that the recent changes made by Airbnb on its' review system to overcome the positive bias issue does not seem to be producing significant changes. As a step to validate the models above, a logistic regression model is used. Additionally, the statistical model is built with the negative sentiment as the reference class. In other words, the regression model predicts the likelihood of the sentiment of the customer review as negative while combining the effects of the variables.

Building classifier to training data (seen data) to perform predictive modelling

```
train_vecs = []
for i in range(len(X_train)):
    top_topics = optimal_model_lda.get_document_topics(doc_term_matrix[i], minimum_probability=0.0)
    topic_vec = [top_topics[i][1] for i in range(14)]
    topic_vec.extend([len(X_train.iloc[i].review_text)]) # review word count
    train_vecs.append(topic_vec)
```

```
X_train.iloc[2].review_text
```

```
'airbnb is not a safe place to host they promise they verify renters and cover them if they damage your property and
they do not beware of opening up your home and renting on this platform they do not offer insurance coverage for dama
ges as they promise they allow fake reviews to be left up as a form of extortion from fake renters they are impossibl
e to get a hold of they transfer you from one person to another and never get back to you once you encounter a renter
who damages your home violates your rules and steals from your property airbnb block you as a host never have i seen
such practices as a host and user of airbnb for over seven years i now understand why the fellow hosts i know no long
er open up their homes to renters on this platform airbnb is notorious for this behaviour and insurance companies hav
e caught on to this and are no longer covering homes that are on this platform beware if you think your insurance wil
l cover you if you use airbnb they probably will not and neither will airbnb so if you make a few thousand dollars bu
t then a renter burns your home down because of careless behaviour and not following your house rules you are out of
luck for millions renters beware'
```

Figure 4.30 Model training with training dataset

The implementation of the logistic regression model on Trustpilot data was more straightforward as class imbalance is less of an issue for this dataset. Furthermore, using sklearn, a machine learning package in Python, the regularization of regression models are handled automatically. Therefore, the resulting model is the best possible model using the current dataset.

Load built classifier with unseen data

```
kf = KFold(5, shuffle=True, random_state=42)
lr_model_F1, lr_model_accuracy, lr_model_precision, lr_model_recall, lr_model_roc  = [], [], [], [], []

for train_ind, val_ind in kf.split(X2, y2):
    # Assign CV IDX
    X_train, y_train = X2[train_ind], y2[train_ind]
    X_val, y_val = X2[val_ind], y2[val_ind]

    # Scale Data
    scaler = StandardScaler()
    X_train_scale = scaler.fit_transform(X_train)
    X_val_scale = scaler.transform(X_val)

    # Logisitic Regression
    # though, sklearn handles regularization by default, i will implement newton-cg to further handle L2 or no penalty
    logistic_reg_model = log_model1.fit(X_train_scale, y_train)
    logistic_classifier = logistic_reg_model.predict(X_val_scale)

    lr_model_F1.append(f1_score(y_val, logistic_classifier, average='binary'))
    lr_model_accuracy.append(metrics.accuracy_score(y_val, logistic_classifier))
    lr_model_precision.append(metrics.precision_score(y_val, logistic_classifier, average='binary'))
    lr_model_recall.append(metrics.recall_score(y_val, logistic_classifier, average='binary'))
    lr_model_roc.append(roc_auc_score(y_val, logistic_classifier))
```

Figure 4.31 Building logistic regression model

The metrics used for model evaluation are produced after running the model, as shown in the screenshot below.

```
print("Logistic Regression Accuracy: %.3f (+/- %.3f)" % (np.mean(lr_model_accuracy), np.std(lr_model_accuracy)))
print("Logistic Regression Precision: %.3f (+/- %.3f)" % (np.mean(lr_model_precision), np.std(lr_model_precision)))
print("Logistic Regression Recall: %.3f (+/- %.3f)" % (np.mean(lr_model_recall), np.std(lr_model_recall)))
print("Logistic Regression f1 score: %.3f (+/- %.3f)" % (np.mean(lr_model_F1), np.std(lr_model_F1)))
print("Logistic Regression Roc: %.3f (+/- %.3f)" % (np.mean(lr_model_roc), np.std(lr_model_roc)))
```

```
Logistic Regression Accuracy: 0.660 (+/- 0.062)
Logistic Regression Precision: 0.703 (+/- 0.062)
Logistic Regression Recall: 0.822 (+/- 0.044)
Logistic Regression f1 score: 0.757 (+/- 0.046)
Logistic Regression Roc: 0.597 (+/- 0.062)
```

Figure 4.32 Model evaluation metrics

Based on the results below, the model resulted from the reviews of Airbnb appears to have produced average results across all metrics where all metrics have scores ranging from 0.6 to 0.7. However, the model appears to perform fairly well in detecting the relevant cases as the recall for training and testing set are 0.94 and 0.82 respectively. In other words, the model was able to detect at least 80% of the relevant cases. Interestingly, the model accuracy, precision,

and model ROC are higher in the testing set. This is unusual as training set generally produce higher performance given the fact that testing set is unseen data. Furthermore, the sklearn library also performs cross-validation by default. Therefore, a mixture of the testing and training dataset is performed to produce the final model. In this case, a more detailed discussion on model improvement is provided in the discussion section and move on to the results of regression models produced from Airbnb datasets.

Table 4.3 Logistic regression model performance (Trustpilot)

Dataset	Accuracy	Precision	Recall	f1 score	ROC
Training set	0.661	0.670	0.944	0.784	0.539
Testing set	0.660	0.703	0.822	0.757	0.597

Several sampling techniques are used for the reviews from Airbnb as there is serious class imbalance issue for the dataset. A model is produced with the original data to serve as a baseline for the other models. Oversampling and variants of under-sampling methods are experimented on the imbalanced dataset and the results are displayed as below. In short, oversampling overcomes the issue of class imbalance by duplicating more samples from the minority class to match the number of majority class. On the other hand, under-sampling select data from the majority class to be removed to match the number of minority class. In this project, the variants of under-sampling using the nearest neighbour rule are used.

The imbalanced data and oversampled models have resulted in 100% recall and over 90% of precision and accuracy. This appear as a fairly good performance, but the ROC has suggested otherwise, with only 0.50. In this case, the findings may be suggesting that these models have included more irrelevant cases in its' detection. As a result of labelling more cases as positive data, this resulted in more actual positive data being included. However, the ROC indicates that these predictions are not significantly better than the baseline model, which is based on pure chance.

On the other hand, the models produced using under-sampling methods are performing better in terms across all the metrics. As there are multiple metrics used to evaluate the models, more emphasis is given to certain metrics when the context of the research is taken into consideration. In this case, being able to correctly detect and recognize the type of sentiments

experienced by the customers are equally important. Therefore, striking a balance between precision and recall are equally important for the project's objectives in providing an understanding of the customer sentiments. So, the f1 score should be given higher weightage in the evaluation of the models. Furthermore, to ensure that the model produced from this project produces significant improvement is also important to the conclusion that can be drawn. Hence, the ROC should also be taken into consideration. Stemming from this notion, the model produced using Near Miss-2 appears to be the best performing model across the important metrics mentioned above.

	Accuracy	Precision	Recall	f1 score	Roc	Accuracy	Precision	Recall	f1 score	Roc
	Training set					Testing set				
	Original Imbalanced data									
Logistic Regression	0.946	0.954	0.991	0.972	0.6	0.94	0.94	1	0.969	0.5
	Oversampling									
Logistic Regression	0.846	0.864	0.824	0.843	0.847	0.94	0.94	1	0.969	0.5
	Oversampling – SMOTE synthetic technique									
Logistic Regression	0.857	0.871	0.837	0.854	0.857	0.94	0.94	1	0.969	0.5
	Under sampling									
Logistic Regression	0.818	0.801	0.792	0.79	0.803	0.818	0.801	0.792	0.79	0.803
	NearMiss-1									
Logistic Regression	0.752	0.747	0.789	0.741	0.781	0.752	0.747	0.789	0.741	0.781
	NearMiss-2									
Logistic Regression	0.807	0.797	0.84	0.806	0.817	0.807	0.797	0.84	0.806	0.817
	NearMiss-3									
Logistic Regression	0.797	0.777	0.831	0.786	0.81	0.797	0.777	0.831	0.786	0.81

Figure 4.33 Logistic regression model performance (Airbnb)

Following the selection of the best model, each variable within the models are also evaluated to form an analytical model. As shown in the screenshot below, the amount of words was found to be the most important variable in predicting the sentiment of the customers.

```
feature_importance_airbnb=pd.DataFrame(np.dstack((np.array(np.array(columns)),model_coeff[1]))[0], columns=columns2)
feature_importance_airbnb.sort_values(by='Coefficient/Beta', ascending=False)
```

	Feature Variable	Coefficient/Beta
8	Word_Count	1.3484863012954555
2	Topic_3	0.8738165417310791
3	Topic_4	0.8588108623460611
0	Topic_1	0.6212181309548562
7	Topic_8	0.5740957002359528
6	Topic_7	0.17887907535790198
4	Topic_5	0.08432128436708435
5	Topic_6	0.04838530626053989
1	Topic_2	0.012120508800618261

Figure 4.34 Variable importance for reviews from Airbnb

The importance of each variable in predicting the customer sentiments for reviews from Trustpilot is also summarized in a table below.

```
feature_importance_airbnb=pd.DataFrame(np.dstack((np.array(np.array(columns)),model_coeff[0]))[0], columns=columns2)
feature_importance_airbnb.sort_values(by='Coefficient/Beta', ascending=False)
```

	Feature Variable	Coefficient/Beta
1	Topic_2	0.466770521275747
10	Topic_11	0.34312293172150254
4	Topic_5	0.32327073334094986
13	Topic_14	0.28291185764281146
11	Topic_12	0.2794788001521407
3	Topic_4	0.26066696525258404
2	Topic_3	0.18634297614294884
8	Topic_9	0.16851479016601068
14	Word_Count	0.1328330289144526
5	Topic_6	0.1275317491158501
6	Topic_7	0.11114689026274638
9	Topic_10	0.06940677147336265
7	Topic_8	0.05733567758948624
0	Topic_1	0.033144664015943605
12	Topic_13	0.001891135721275526

Figure 4.35 Variable importance for reviews from Trustpilot

4.6 SUMMARY

Based on the results above, the most prominent finding was the different amount and types of information extracted from the two different platforms, despite the fact that all the reviews are discussing about the same company. Specifically, the type of topics extracted from reviews from Airbnb has produced a much narrower focus, which revolves around the accommodation. On the other hand, topics mentioned in the reviews from Trustpilot have revealed a much more diverse set of topics but most of them revolves around the customer service and problem resolution of the company. In terms of the sentiment analysis, there are also more variety of sentiments detected from the reviews of Trustpilot. Following the extraction of topics and sentiment analysis, the topic models are used to produce regression models. When compared, the regression model produced using the reviews from Airbnb has produced greater performance. Even after tuning, the regression model produced using Trustpilot reviews only managed to achieve an average performance. Considering the greater diversity in the Trustpilot dataset, the average performance of the model may be attributed to the large amount of variance. Alternatively, it requires a much larger dataset to sufficiently train the model. More detailed explanations of the findings are provided in the next few sections.

CHAPTER 5

RESULT AND ANALYSIS

Based on the profiles and the associated keywords, it can be seen that most Airbnb reviews are related to the accommodation such as location, amenities, level of comfort, and value. At a glance, the keywords appear to suggest that the overall sentiment is leaning towards the more positive end on most of the topics mentioned in the guests' reviews. Nonetheless, the sentiment analysis in subsequent section will confirm the customers' sentiment.

Table 5.1 Topics extracted from Airbnb reviews

Topic	Topic Name	Keyword examples
0	Experience	Visit, local, design, see, beautiful
1	Comfort	Room, bed, water, night
2	Expectations	Problem, price, dirty, toilet, property, different
3	Value	Place, location, stay, clean, great
4	Location	Apartment, city, quiet, floor, large
5	Service	Help, pool, home, feel, morning
6	Travel	Return, happy, photo, trip
7	Amenities	Shampoo, slipper, sufficient, plate, extra

The topics mentioned by reviewers from Trustpilot, on the other hand, are much more diverse. The topics also included other aspects of Airbnb such as the company culture, company's problem resolution, and customer service. Based on the keywords, it seems to suggest that the reviewers of Trustpilot have a more negative sentiment towards the company, which seems to validate the positive bias again. Stemming from the findings, it may be possible that the customers that are satisfied with the accommodation and amenities provided generally do not engage with the company's customer service. Therefore, these customers will not have any comments for the company's handling of customers inquiries or complaints. On the other hand, if the customers are not satisfied with the accommodation have already held negative sentiment towards the company. Furthermore, with the existing negative sentiment, any dissatisfaction caused by the customer service will only aggravates the matter. As a result, more

negative sentiments may be detected in regard to the company itself instead of the accommodation. This hypothesis appears to be supported by the findings from the sentiment analysis.

Table 5.2 Profile of 14 topics from Trustpilot reviews

Topic	Topic Name	Keyword examples
0	Experience	Polite, damage, unhappy, utensils, procedure, fraudulent
1	Problem resolution	Manager, resolution, hide, deposit, evidence, submit
2	Company culture	Implore, political, narrative, incidence, ideal
3	Credibility	Reimburse, identity, scammed, phishing, dreadful, tour
4	Safety	Positive, corona, human, prevent, basement
5	Company	Review, guest, book, Airbnb, place, refund, money
6	Convenience	Email, cancel, call, time, booking, reservation, account
7	Service	Appear, connect, upset, supply, key, remote
8	Communication	Kind, instruction, concerned, conversation, manager
9	Access to information	Access, easy, fee, card, info, policy, text, log
10	Listing	City, profile, video, free, stuff, guarantee, community
11	Amenities	Bed, room, towel, clean, bathroom, window, private
12	Accommodation	Location, level, old, nasty, unit, shower, hot, cold
13	Customer service	Compensate, emergency, resort, withhold, satisfy

Based on the results of the sentiment analysis, it can be seen that most of the sentiment expressed for the 8 aspects are positive sentiments such as anticipation, joy, and trust. However, there are aspects where it cannot be categorised into any of the 8 emotional states. In this case, the aspect is labelled as neutral. Similar to what was described in past research, not many negative sentiments (disgust, anger, sadness) are detected from the reviews extracted from Airbnb. By looking at the findings, the most prominent sentiment is anticipation, which has been detected in several aspects of the topic model.

Table 5.3 Sentiment of each topic (Airbnb)

Topic	Topic Name	Overall sentiment
0	Experience	Joy (1)
1	Comfort	Neutral
2	Expectations	Disgust (2), fear (1), sadness (1)
3	Value	Anticipation (1), joy (2), surprise (1), trust (2)
4	Location	Sadness (1)
5	Service	Anticipation (1), joy (1), trust (1)
6	Travel	Anticipation (1), disgust (1), joy (2), surprise (1), Trust (1)
7	Amenities	Trust (1)

Meanwhile, the most prominent emotional state for reviews from Trustpilot is anger as it appears in almost all of the extracted topics. This has confirmed the suspicion raised by the current project where there is a serious positive bias on Airbnb's review. At the same time, the confirmation of this suspicion also raises a serious concern for the company as the company is unaware of the negative experiences that the customers had with the company and subsequently leading to be company being stagnant in its' growth in the ever-competitive industry.

Table 5.4 Sentiment of each topic (Trustpilot)

Topic	Topic Name	Keyword examples
0	Experience	Anger(3), anticipation(2), disgust(3), fear(2), sadness(2), trust(2)
1	Problem resolution	Anger(1), anticipation(1), fear(3), surprise(1), trust(1)
2	Company culture	Anticipation(1)
3	Credibility	Anger(1), anticipation(2), disgust(1), fear(2), joy(1), sadness(2)
4	Safety	Fear(1), trust(1)
5	Company	Anger(1), anticipation(1), joy(1), sadness(1), surprise(2), trust (1)
6	Convenience	Anticipation(1), trust(1)
7	Service	Anger(3), anticipation(1), fear(2), sadness(2), surprise(1), trust(2)
8	Communication	Fear(2), joy(2), sadness(2), trust(3)
9	Access to information	Anger(1), trust(3)
10	Listing	Sadness(1), trust(3)
11	Amenities	Disgust(1), joy(1), trust(1)
12	Accommodation	Anger(2), disgust(1), fear(1), surprise(1), trust(1)
13	Customer service	Anger(1), anticipation(1), disgust(2), fear(1), joy(1), sadness(2), surprise(2), trust(1)

Following the extraction of the topics, the various topics are converted into variables of the regression model to predict the customer sentiments. Aside from the prediction, the importance of each variables is also analysed and listed in the table below. Intuitively, longer reviews were found to have greater importance in the prediction of customer review sentiments as longer reviews provide richer information. In addition to the importance, the word count also has a positive coefficient. Therefore, it may suggests that customer tend to provide greater amount of details and write longer reviews when their experience is negative.

Additionally, the next most important variable is the expectation of the Airbnb guests have about their accommodation. Such comments are usually about the different quality of the accommodation shown in the listing and the actual condition of the accommodation when

guests check in. Based on the sentiments detected about this aspect, all of the sentiments are negative. This may be suggesting that there are a number of guests who are disappointed with the actual quality of the accommodation after checking in. In addition to that, the aspect about location only detection negative sentiment as well. Meanwhile, the remaining aspects extracted from Airbnb has predominantly positive sentiment.

Table 5.5 Importance of each variable in predicting customer sentiments (Airbnb)

Importance	Variable
1	Word Count
2	Expectations
3	Value
4	Experience
5	Amenities
6	Travel
7	Location
8	Service
9	Comfort

On the other hand, the most important aspect from Trustpilot is problem resolution. Furthermore, the accommodation related aspects also appear to have lesser importance in the prediction of the customer sentiment, except aspects about the listing and amenities. Considering the fact that the most prominent sentiment detected from the reviews, it may be suggesting that majority of the customers have an issue with the way Airbnb handled their complaints or inquiries. Furthermore, the next few important variables are about the listing and the safety of the accommodation. Although positive sentiment is detected in both of these aspects, negative sentiments are also equally prominent. Particularly, Airbnb should look into the sentiment of fear being detected in the aspects about safety as the operation of the company depends heavily on the guests feeling safe while using their service.

Table 5.6 Importance of each variable in predicting customer sentiments (Trustpilot)

Importance	Variables
1	Problem resolution
2	Listing
3	Safety
4	Customer service
5	Amenities
6	Credibility
7	Company culture
8	Communication
9	Word count
10	Company
11	Convenience
12	Access to information
13	Service
14	Experience
15	Accommodation

5.1 SUMMARY

Although the results have provided some useful insights about the various issues mentioned in the reviews of the customers, it is important to note that the sentiments detected from both platforms are still vastly different in terms of the distribution of negative and positive sentiment. Particularly, there is still a positive bias in the reviews of Airbnb while Trustpilot has a more balanced amount of both negative and positive reviews. Furthermore, the aspects mentioned by the customers from both platforms are also very different. Based on the findings, the difference may be attributed to the fact that the guests are still unwilling to disclose their negative experience on the platform and have turned to other independent sites to post about their experience with Airbnb. Considering the fact that negative comments are often given more weight, it is important for Airbnb to actively seek out the negative comments in order to rectify the issues and make improvements as soon as possible.

CHAPTER 6

DISCUSSION AND CONCLUSION

6.1 INTRODUCTION

In this section, the results reported are critically discussed and the potential directions that the company should look into are also pointed out. Furthermore, the limitations of the study are also mentioned alongside the potential extension for future research to fill in the research gap of the current project.

6.2 DISCUSSION

As a summary of the findings above, the positive bias in Airbnb review system highlighted by several researchers have yet to be overcome as the reviews extracted from the platform still contains a vast majority of positive reviews. Meanwhile, the reviews extracted from Trustpilot shows are more balanced distribution in terms of the customer sentiments as the emotional states detected from the reviews are much more diverse. Nonetheless, it is still important to note that the valence of negative sentiments is still stronger than the positive sentiments. In other words, the reviews from Trustpilot revealed that there are happy customers of Airbnb, but the unhappy customers feel much more strongly about their negative experiences.

Additionally, the comments from Trustpilot were also found to be much longer and detailed than the reviews posted on Airbnb platform. Combining with the presence of positive bias, the review system of Airbnb will appear much less credible and reliable. Moreover, as the negative comments were also longer, customers are also more likely to believe in such comments as research has shown that negative comments are given more weightage (Camilleri, 2017). All in all, the findings suggest that the situation is unfavourable for Airbnb should the customers seek out reviews from independent sites like Trustpilot. It is likely that they will be deterred from trying the company's service. Other than the findings above, the topic models have also revealed the type of sentiments held by the customers on various aspects of the company. There are few aspects that are recurring among the reviews from both platforms, which are the experience, amenities, and service. Based on the keywords, the experience mentioned by both platforms are usually referring to the experience the customers had when interacting with the company.

On one hand, the reviews posted on Airbnb revealed that the customers are content with Airbnb as it allows the customers to travel and experience the local culture. Furthermore, only the sentiment of joy can be detected on this aspect. On the other hand, much more sentiments are detected among the reviews from Trustpilot, but the most prominent emotions are anger and disgust. Some of the keywords are mostly negative in nature, such as “fraudulent” or “damage”. Furthermore, the valence of such negative sentiments is also higher than the positive sentiment detected from the Airbnb platform. The aspect “service” is also present in both platforms where the customers are discussing the service provided by the Airbnb hosts. Similar patterns in sentiments is also detected as Airbnb customers are generally satisfied with the hosts’ service while negative sentiments are more prominent for reviews from Trustpilot. Lastly, amenities are also an aspect that were brought up by reviewers from both platforms. Expectedly, reviews from Trustpilot lean towards a more negative sentiment in comparison to the reviews from Airbnb.

Other than that, the results section has also briefly mentioned that the reviews from Airbnb has focused primarily on the accommodation related aspects while reviews from Trustpilot have mentioned more aspects that are not related to the accommodation. Moreover, the main theme of the reviews from Trustpilot mainly revolves around Airbnb’s handling and response towards customers’ complaints or negative feedbacks. For instance, aspects such as problem resolution, company culture, communication, and customer service were not mentioned in the reviews of Airbnb. In addition to that, almost all of these aspects absent in Airbnb reviews have more prominent negative sentiments where the most prominent sentiment is anger.

The aforementioned findings may be suggesting few things regarding the company. Assuming all the reviews on Airbnb were truthful, this may suggest that the customers are generally satisfied with the accommodation offered by most of the Airbnb listings in terms of the price, value for money, amenities, and services offered by the host. This was supported by the positive sentiment detected among reviews from both Airbnb and Trustpilot in regard to the accommodation. Furthermore, since the reviews from Airbnb mainly mentioned about the accommodation related aspects, the sentiments are almost always positive. However, the findings from Trustpilot suggests that majority of the customers are not satisfied when it comes to the interaction with the company. These interactions include interaction with the customer service to lodge a complaint or to obtain a refund. Unlike the traditional hotels with a concierge, Airbnb guests can only rely on the host or the company’s customer service for assistance should

any issue arise during their travel. Based on the findings, the company is not performing up to the customers' standards on this regard.

In regard to this pattern in the reviews, two factors may be able to offer an explanation. As mentioned before, the two-way review system implemented by Airbnb is known to cause customers to feel reluctant to share their negative feedbacks due to the fear of retaliation (Bolton, Greiner, and Ockenfels, 2013). Although the company has made some changes to the review system, some of the customers may still be unaware of such changes. Therefore, the customers are still reluctant to share their negative experience on the company's platform. Alternatively, a worse possibility may be due to the customers' perception that the company do not care about their negative experiences. Specifically, the customers may have felt extremely frustrated at the customer service and have arrived at a conclusion that the company does not and will not take customers' negative experiences seriously. Regardless of the reason, the company's priority is to devise ways that can improve the ways it handles customers' complaints as well as the ways the company can support the customers throughout their entire purchase journey.

Furthermore, about the customer sentiments, there is another interesting finding where the sentiment of anticipation was found to be present in almost all of the aspects derived from both Trustpilot and Airbnb. Given the fact that negative sentiment was the most prominent customer sentiment detected from Trustpilot, it is interesting to see that most of the customers are still looking forward for the company to perform well on many aspects other than providing quality accommodation. Building from this piece of finding, it is likely that the customers still have certain expectations for the company. Put it another way, the brand name of Airbnb still holds some value to the consumers, and they have the perception that the company should have the capability to perform up to a certain standard.

Another important aspect mentioned in the reviews of Trustpilot is the issue of safety. In this aspect, two polarising sentiments were detected, which are fear and trust. They are polarising as fear occurs due to the presence of uncertainty. This is understandable as nature of Airbnb's business model increases the amount of uncertainty due to the lack of concierge and the possibility to be living with strangers from vastly different cultural and language backgrounds. Meanwhile, the feeling of trust can only occur when the level of uncertainty is low, which is contradicting to the sentiment of fear. This aspect is especially important as the business model of Airbnb has always been criticised for the potential safety and credibility

issues. The presence of trust among customers is an encouraging sign for the company. Nonetheless, the company must also look into the source of fear among the customers.

Moving on to findings of the regression models, the model produced using reviews from Airbnb has produced better results in terms of the detection of customer sentiment despite the severe class imbalance issue. It appears that under-sampling method is able to overcome the issue fairly well. Nevertheless, the model produced using reviews of Trustpilot may be further improved by including more data or increase the number of features in the model. In the article by Van der Alst et al (2010), such pattern in classification algorithm may be attributed to underfitting, which can occur when the model is unable to capture the variability of the dataset.

In other words, the model is not complex enough to represent the amount of variance in the dataset. Some methods have been proposed to overcome such issue, which are the penalty method or the early-stopping method (Murakoshi, 2005; Ruppert and Carroll, 2000; Loughrey and Cunningham, 2005) The penalty method is akin to regularization where the bias-variance balance is achieved by adding a cost term to the error function. However, the selection of cost term is still largely subjective. In regard to this, the article by Jabbar and Khan (2015) proposed the usage of *k*-fold cross validation in order to select the best cost term. Meanwhile, the early-stopping approach involves the specification of a point or the number of iterations for the model to stop the training (Prechelt, 1998). Typically, a part of the training set is used to build the model and form the parameters. Then, the rest of the training set is used to monitor the amount of error produced by the newly trained model at various points. In other words, the early-stopping method involves taking a subset of the training dataset just to find out the best number of iterations that can minimise overfitting and underfitting (i.e. produce the least amount of error). After the best point is determined, the model is trained using that number of iterations. However, this method may require a much larger dataset as part of the training dataset will be used to determine the best number of iterations. Furthermore, the division of the training and testing dataset must also be performed with caution. If the training dataset is vastly different from the testing data, the best number of iterations determined using the training dataset may not be suitable to be used on the testing dataset.

Nevertheless, there are also areas that the current project is limited. One of the limitations of the current project is the fact that the analytical model produced is limited to the dataset used in the current project. In other words, both platforms have much larger amount of reviews that have yet to be included in the current project's analyses. Therefore, the actual

applicability of the model onto the customers of Airbnb is limited. However, only including a limited number of reviews is necessary as the computing resources and interpretability of the model is also important. Aside from this, the study is also limited where only English reviews are included in the reviews. As a company that receives customers from almost the entire world, it is also important for Airbnb to also understand the non-English speaking consumers. In fact, the cultural difference between the English and non-English speaking consumers may also influence the ways they evaluate the services of Airbnb as well as the ways they provide their reviews. In regard to these limitations, future research may extend the analyses of the current project by including more features such as the reviewers' demographics or compare the review patterns among different language users.

6.3 LIMITATIONS AND FUTURE DIRECTIONS

Nevertheless, there are also areas that the current project is limited. One of the limitations of the current project is the fact that the analytical model produced is limited to the dataset used in the current project. In other words, both platforms have much larger amount of reviews that have yet to be included in the current project's analyses. Therefore, the actual applicability of the model onto the customers of Airbnb is limited. However, only including a limited number of reviews is necessary as the computing resources and interpretability of the model is also important. Aside from this, the study is also limited where only English reviews are included in the reviews. As a company that receives customers from almost the entire world, it is also important for Airbnb to also understand the non-English speaking consumers. In fact, the cultural difference between the English and non-English speaking consumers may also influence the ways they evaluate the services of Airbnb as well as the ways they provide their reviews. In regard to these limitations, future research may extend the analyses of the current project by including more features such as the reviewers' demographics or compare the review patterns among different language users.

6.4 FUTURE RECOMMENDATIONS

In terms of the recommendations to the company, it is vital for Airbnb to consider steps that will encourage the customers to provide their feedback on their platform, despite the review may be negative. In fact, studies have shown that the presence of negative reviews can actually increase the credibility of the review system (Gutt et al., 2019). As a step to encourage the customers to provide their reviews on the platform, some token of appreciation or a badge may

be rewarded to customers who have accumulated a certain amount of reviews. By doing so, it can allow the review provided by the customer to be given a higher weightage as they are the experienced users of Airbnb. Additionally, as the company has made changes to prevent the host from retaliating against the customers for posting a negative feedback, perhaps the company should ensure that all of the customers are informed that their negative feedback will not result in any form of retaliation from the host. Such information may be displayed on the page of where the customers leave their reviews.

Aside from improving the platform, Airbnb should also look into potential ways that the company can enhance their customer service or increase the ways they can assists the customers. By thinking from the perspectives of the customers, the lack of human touch of Airbnb can cause a lot of uncertainty and customers travelling to a foreign city or country can further amplify the sense of uncertainty. Should any issues happen, the company should have channels where the travellers can reach out to seek assistance. As highlighted in the findings about Trustpilot, majority of the reviews raised revolves around the customers' experience when interacting with the company or ways the company handled customer complaints. Although Airbnb has expressed that the company's primary role is to provide a platform that connects the host and the guests, the negative experience that the customers cannot be overlooked. Perhaps, in addition to providing a platform, the company should also look into other value-added services where the customers can feel that their interest is safeguarded by the company. By doing so, it also differentiates itself from the other sharing accommodation platforms that also focus solely in connecting the host and the guests.

6.5 CONCLUSION

To reiterate, the main objectives of the current project is to uncover the main aspects that are mentioned by the customers of Airbnb in their reviews and to perform a fine-grained sentiment analysis on identified aspects. The current project has fulfilled the objectives as the major topics raised by the customers from different platforms have been identified. Furthermore, the customers' sentiments towards those topics have also been revealed. In addition to the identification of the major topics and customer sentiments, the statistical modelling has also enabled the findings to be converted to a computational model that allows the customers' sentiments to be predicted based on the important features.

REFERENCES

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

Bolton, G., Greiner, B., Ockenfels, A. (2013). Engineering trust: reciprocity in the production of reputation information. *Manage. Sci*. 59 (2), 265–285

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.

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

Camilleri, A.R. (2017). The presentation format of review score information influences consumer preferences through the attribution of outlier reviews. *Journal of Interactive Market*, 39, 1–14.

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.

Cheng, M. and Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, 76, 58-70.

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]

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

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.

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

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

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

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

Guttentag, D. (2016). Why tourists choose Airbnb: A motivation-based segmentation study underpinned by innovation concepts, *Journal of Travel Research*, 57(1), 1-19.

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

Gutt, D., Neumann, J., Zimmermann, S., Kundisch, D. and Chen, J. (2019). Design of review systems–A strategic instrument to shape online reviewing behavior and economic outcomes. *The Journal of Strategic Information Systems*, 104-117.

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

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

Jabbar, H. and Khan, R.Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study), *Computer Science, Communication and Instrumentation Devices*, 163-172.

Jurek, A., Mulvenna, M.D. and Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4(1), 1-13.

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.

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

Lampinen, A. and Cheshire, C. (2016). Hosting via Airbnb: Motivations and financial assurances in monetized network hospitality. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, San Jose: CA, May 7th – 12th.

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] June 2019]

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

Loughrey, J. and Cunningham, P. (2005). Using early-stopping to avoid overfitting in wrapper-based feature selection employing stochastic search. Trinity College Dublin, Department of Computer Science.

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.

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

Murakoshi, K., 2005. Avoiding overfitting in multilayer perceptrons with feeling-of-knowing using self-organizing maps. *BioSystems*, 80(1), 37-40.

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

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

Prechelt, L. (1998). Early stopping-but when?. In Neural Networks: Tricks of the trade Springer, Berlin, Heidelberg.

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

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]

Ruppert, D. and Carroll, R.J. (2000). Theory & Methods: Spatially-adaptive Penalties for Spline Fitting, *Australian & New Zealand Journal of Statistics*, 42(2), 205-223.

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

Sperandei, S., 2014. Understanding logistic regression analysis. *Biochemia medica: Biochemia medica*, 24(1), pp.12-18.

Sthapit, E., Bjork, P. (2019). Sources of distrust: Airbnb guests' perspectives. *Tourism Management Perspectives*, 245-253. doi: 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.

Tussyadiah, I.P. and Pesonen, J. (2016). Impacts of peer-to-peer accommodation use on travel patterns. *Journal of Travel Research*, 55(8), 1022-1040.

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

Van der Aalst, W.M., Rubin, V., Verbeek, H.M.W., van Dongen, B.F., Kindler, E., and Günther, C.W. (2010). Process mining: a two-step approach to balance between underfitting and overfitting. *Software & Systems Modeling*, 9(1), 87.

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]

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

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

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

Zhou, L., Ye, S., Pearce, P.L. and Wu, M.Y. (2014). Refreshing hotel satisfaction studies by reconfiguring customer review data. *International Journal of Hospitality Management*, 38, 1-1