rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 103–116
DOI: 10.15439/2024R117
ISSN 2300-5963 ACSIS, Vol. 42

Machine Learning-Based Prediction Models for Sentiment Analysis on Online Customer Reviews: A Case Study on Airbnb

Cu Kim Long
Information Technology Center
Ministry of Science and Technology,
Lab AI 4.0, AIRC, VNU-ITI
Hanoi, Vietnam
longck.2006@gmail.com

Le Bao Ngoc
Norwich Business School
University of East Anglia
Norwich, United Kingdom
witty.gem2211@gmail.com

Vijender Kumar Solanki
Stanley College Of Engineering & Technology for Women
Hyderabad, TG, India
vijendersolanki@ieee.org

Nguyen Viet Anh
Dental School, Hanoi University of Business and Technology
Hanoi, Vietnam
vietanh.bsrhm@gmail.com

Luu Hoang Bach
Vinmec Research Institute
Hanoi, Vietnam
luuhoangbach711@gmail.com

Cu Ngoc Son
Faculty of Information Technology
Hanoi National University of Education
Hanoi, Vietnam
stu745105077@hnue.edu.vn

***Abstract*—In the last decade, with the rise of sharing economy, in particular Airbnb, customers are not merely buyers but also actively share their thoughts and experiences toward goods and services. Sentiment analysis, a sophisticated technological approach, has emerged as a pivotal tool to extract people's opinions as well as sentiments from written language. On the other hand, assessing the price of a listing has always been a daunting task for hosts and guests. While numerous pricing models for Airbnb have been proposed, achieving precise accuracy remains a challenge. As a result, this paper aims to investigate whether incorporating the sentiment scores derived from online customer reviews could improve the accuracy of Airbnb price prediction or not. First, online customer reviews on Airbnb are examined using natural language processing techniques to seek the guest sentiment and its association with listings prices. Once sentiment scores are calculated, they are used as an additional attribute to forecast Airbnb listings price. Several machine learning models are employed, including Linear Regression, Ridge Regression, Support Vector Machine, XGBoost and Random Forest. The experimental results show that the inclusion of sentiment scores slightly decreases model performance in the case of three Asian economies (Hong Kong, Japan and Taiwan). Overall, Random Forest without sentiment variable is the best-performing model among five models for Airbnb price prediction.**

***Index Terms*—machine learning, sentiment analysis, OCRs, price prediction, Airbnb.**

I. Introduction

AIRBNB, established in 2007, is the top pioneer in peer-to-peer (P2P) sharing platforms in the hospitality industry [1]. Airbnb has created an online solution to directly link guests looking for short-term accommodations with hosts who are in demand to lease out their homes [2]. In other words, the company works as a broker facilitating the connection between property owners and hospitality seekers. Since its foundation, Airbnb has consistently experienced a year-over-year supply growth rate of over 100% for the last decade [3], serving over 220 countries and regions worldwide [4].

Despite its unprecedented yet exponential growth, pricing has always been a major concern of Airbnb's stakeholders [5]. Determining an appropriate price for a rental property on Airbnb platform has been a challenging task for not only the tenants but also the owners. While guests need to evaluate the reasonable price of the listings to avoid being deceived, hosts also need a competitive price for their rental house to attract customers [6]. Therefore, predicting price is one of the most critical components in accommodation sharing systems such as Airbnb so both tenants and house-owners gain maximum benefits from the platform.

On the other hand, in a P2P platform, online customer reviews (OCRs) are considered to have a significant impact on not only sales but also the price of the property [7]. In the era of Web 2.0, a significant volume of data in which OCRs presents a significant challenge for any business and institution to deal with effectively. Even though peer-reviewed research has been undertaken to investigate customer reviews using text mining, there has been a lack of empirical research to enrich the forecasting model by applying textual methods. This study aims to focus on the topic of sentiment analysis on customer reviews to develop Airbnb price prediction model.

Following the discussion above, the main research questions of this study are how the association between sentiment scores derived from OCRs and Airbnb rental prices is, and how the performance of Airbnb rental price prediction models change when incorporating sentiment scores derived

©PTI 2024

from OCRs. These questions are formulated not only to fill in the gaps represented in the literature in the next chapter but also to contribute to Airbnb and the hospitality industry. **To answer research questions**, the objectives of the study are threefold: *(1) To construct and explore the guest sentiment in OCRs in three Asian economies namely Hong Kong, Japan and Taiwan; (2) To identify the association between Airbnb rental price and sentiment scores expressed in OCRs; (3) To identify the performance of Airbnb rental price prediction model with the inclusion of sentiment index from online customer reviews*.

Based on the aforementioned research questions and objectives, the following **two hypotheses** are formulated as below:

H_1: Positive sentiment expressed in OCRs is associated with an increase in Airbnb rental prices.

H_2: The inclusion of sentiment scores derived from OCRs as an explanatory variable statistically improves the predictive accuracy of Airbnb rental price prediction models.

Although Airbnb has introduced its price suggestion tools since 2012, which have been developed to "smart pricing" until now, the price prediction model still needs further improvement. The significance of this paper is underscored by the incorporation of sentiment scores derived from OCRs into forecasting Airbnb rental prices. By exploring the sentiment analysis on OCRs, this study seeks to enhance the knowledge surrounding the use of textual data for predictive modelling. The outcomes are not only for Airbnb but also for the broader context of the hospitality industry. In addition, the growth of Airbnb in Asia emphasises the need for research in Asian regions. Most of the existing papers have primarily centred on Western countries, and yet little attention is being paid to Asian regions, which is a growing and potential market for Airbnb in recent years [8]. Therefore, the focus on Asian economies of this study would enhance the scope and relevance of research in the field of Airbnb as well as the hospitality industry.

The rest of this paper is organized as follows. The relevant academic research within the field of sentiment analysis and price prediction models in Airbnb is presented in Section II. Section III discusses machine learning models used for predicting price, as well as the evaluation metrics employed to assess their performance. The building of five models including Linear regression, Ridge regression, SVM, XGBoost and Random Forest is discussed and evaluated to test the stated hypotheses in Section IV. Conclusions and future works are given in Section V.

II. Literature Review

In this section, relevant academic research within the field of sentiment analysis and price prediction models in Airbnb is presented. Firstly, it delves into the key concept of online customer reviews. Then, the literature on sentiment analysis and existing models for Airbnb price prediction is discussed. This chapter not only highlights the existing knowledge but also identifies the gaps that this study attempts to address.

A. Online customer reviews (OCRs)

Prior to the advent of online opinion-sharing platforms, the primary mode of communication among consumers was word-of-mouth. Consumer word-of-mouth has been frequently cited as one of the most crucial elements to determine the long-term success of goods and services. However, since the rise of online communities as well as communication facilitated by the Internet, there has been a new product information channel with growing popularity, where consumers share their experiences toward products and services, also known as online customer reviews (OCRs).

Online customer reviews (OCRs) refer to the evaluation of a product or service shared by customers on company or third-party websites. The importance of OCRs on consumer purchase decisions in the hospitality sector has been widely studied in the economic literature [9-10]. Besides, online hotel reviews is a reliable information source for customers as they unveil guests' feelings, attitudes and evaluations, thereby, reflecting guest satisfaction or dissatisfaction [11].

According to [12], OCRs influence the decision-making of guests in all ages, thus, contributing to the sales revenue. This is well-illustrated empirical studies which indicated that online reviews impact early sales, as a result, can be a significant predictor of box office revenue. Similarly, OCRs and the number of reviews can be used to determine future digital camera sales by fitting a multiple linear regression. This is primarily because individuals tend to readily embrace and place trust in information shared by other peers similar to themselves. OCRs help to alleviate the perceived risk and confusion of consumers [13]. In the tourism and hospitality industry, prior research has empirically proved that OCRs have a significant influence on purchasing decisions, especially booking intentions [14]. By conducting an Analysis of Variance (ANOVA) test, online reviews affect the decision making of consumers within the hospitality industry. Employing data from a Chinese online travel agency found that an increase of 10% in traveller review ratings leads to a considerable increase of over 5% in online bookings. Positive reviews are the motivation that inspires people to travel, meanwhile, negative reviews act as an effective tool to help people avoid bad travel products [15].

In the context of Airbnb, peer-to-peer feedback or so-called OCRs is even more significant than that of traditional hotels. This is because Airbnb hosts are usually micro-entrepreneurs who are financially unable to advertise their accommodations on media such as television like hotels, thus, the online platforms serve as the exclusive mean for them to connect with their guests. Furthermore, by using textual data, hosts can delve into a more comprehensive insight on customers' experiences rather than solely depending on non-textual data such as rating scores given by guests [16]. As a result, it becomes even more compelling to investigate the sentiment of guests based on OCRs within the Airbnb ecosystem.

B. Sentiment Analysis

In the case of OCRs, sentiment analysis can serve as a methodological approach for classifying, measuring and monitoring users' emotional responses towards a product or service [17]. Realising the importance and explosive growth of OCRs on the Internet, there has been an emerging stream of research undertaken to identify the sentiment index in online textual reviews, especially in the field of hospitality. They aimed to explore the relationship between the sentiment of reviews and the listing prices, thereby understanding the role of OCRs in consumer valuation and pricing decisions. They have reinforced the sentiment analysis to aspect-based sentiment analysis, which extracts the sentiment polarities towards specific aspects of an entity within hotel reviews. As a result, this research has significantly improved the comprehensiveness and accuracy of sentiment analysis in the hospitality industry [18].

On the other hand, supervised machine learning is introduced in [19-20] for sentiment analysis as a different approach. Specifically, Naïve Bayes classification is applied to measure not only the polarity but also the subjectivity scores of user-generated contents on TripAdvisor [21]. While polarity evaluates the emotion of text, the subjectivity scores measure the subjective or objective score of text. Alternatively, the Long Short-Term Memory model is introduced in [20], which is one of the latest deep learning technologies. This has significantly improved text classification performance. The sentiment indexes are separated into past and future housing price changes. According to the paper, this model could capture the word order and dependence, which unsupervised machine learning is unable to do.

Sentiment analysis, in which sentiment polarity classification is broadly used in forecasting, including product sales forecasting [22], stock market forecasting [23], household expenditure forecasting [24]. Nevertheless, the application of sentiment analysis in hospitality forecasting literature remains uncommon [25]. For this reason, this study examines the association between OCRs and Airbnb listings' prices as well as use guest sentiment extracted from OCRs to predict Airbnb listings' prices.

C. Price Prediction on Airbnb

Pricing a listing is considered one of the most crucial business practices for any Airbnb host to master [26]. Conventionally, hosts are allowed to set their own nightly, weekly and monthly prices for their rental houses. However, Airbnb still provides suggestions to assist their hosts to set more optimal prices for the entire selling period, which is a dynamic pricing strategy called "Smart Pricing" [27]. The Smart Pricing algorithm takes into consideration various points of information, including the date of the night to price, market demand, seasonality, listings' characteristics. Once receiving a pricing tip, a host can either choose to increase, decrease or do nothing.

However, several scholars found that, in contrast to professional hosts, nonprofessional hosts appear to adopt different and less dynamic pricing strategies [28]. Therefore, identifying the determinants of price on Airbnb has received a significant concentration in recent years. Moreover, factors related to the property are also found to significantly influence listing prices such as the site and location of the property [29], cleanliness of the rooms [7], type of accommodation [30]. Besides the factors on the supply side, studies also showed that the price of Airbnb accommodation tends to decrease as the number of reviews it receives increases [26].

In parallel to factors determining price, choosing an efficient prediction model is also essential so as to give the best accuracy. In 2017, Wang and Nicolau [29] identified 25 price determinants using ordinary least squares and quantile regression. Afterwards, price prediction of Airbnb has witnessed advancements beyond traditional linear regression model. To be specific, Liu [31] conducted a study on various models by leveraging machine learning techniques to capture non-linear relationships between price and other factors. Both of their papers discovered that XGBoost yields the highest accuracy, with R^2 equals 61.8% and 63% respectively. However, Mahyoub [5] concluded that Random Forest Regressor is the most effective model with R^2 equal to 86.95%, which outperforms XGBoost regression.

The difficulties in determining the prices for Airbnb accommodations can be attributed to the inherent complexity of these properties. This is because they encompass not only a variety of functional attributes but also the social interactions between hosts and customers [27]. While studies have indicated that prices are influenced by a set of factors including host attributes and property characteristics as aforementioned, online customer reviews are largely underexplored.

Meanwhile, Ganu [32] posits that customer reviews have the capability to demonstrate reviewers' attitudes more precisely than numerical star ratings, which may be biased. In other words, unidimensional customer ratings can be significantly biased by price effects. Therefore, Lawani [7] suggested that rating scores can result in biased implications on the relationship between the scores and prices since rating scores might not accurately capture guests' opinions and sentiments regarding a good or service. All in all, this research will take into consideration the studies discussed in the literature presented to develop a price prediction model with sentiment analysis.

III. Research Methodology

This section outlines the research approaches step by-step, including data sources, data pre-processing, sentiment analysis and modelling. The technique used for sentiment analysis, which is a lexicon-based method, is introduced. After extracting guest sentiment, the chapter discusses machine learning models used for predicting price, as well as the evaluation metrics employed to assess their performance.

A. Data description

The data is retrieved from Inside Airbnb (insideairbnb.-com), which is an independent and non-commercial website that provides data collected from Airbnb for public use. Inside Airbnb encompasses not only structured data related to Airbnb listings but also unstructured data in the form of customer reviews for each listing. The data from Inside Airbnb has been largely utilised in academic research and significantly contributed to the debates around the existence as well as growth of Airbnb [33].

The dataset consists of 569,523 Airbnb listings which are managed by 14,584 hosts in three Eastern Asia economies, namely Japan, Hong Kong and Taiwan on 31 March 2023. Asian economies are chosen in this study for two reasons. Firstly, scholarly research has been geographically focused on the United States, Canada and Europe [34]. Meanwhile, only 13.4% of the studies collected their data in Asia regions. Secondly, a growing popularity of Airbnb is shifting from Europe to America, and mostly to Asia [8]. As a result, investigating the dynamics of Airbnb in Asian countries can contribute to filling the gap in the literature.

The entire workflow including data handling, pre-processing, modelling, analysis, and visualisation in this study will be executed via the R programming language.

B. Data Pre-Processing

Data collected in their raw format can be an issue for sentiment analysis and modelling as they might be formatted inconveniently such as stop words, missing data and so on. Therefore, in order to facilitate further analysis, it is essential to undertake several data pre-processing steps.

Firstly, non-English texts are detected using Google's Compact Language Detector 2 (cld2) package in R [35], which can detect 80 languages in UTF-8 text and even mixed language input. The purpose of this action is to separate English with Chinese and Japanese reviews. Secondly, reviews in Chinese and Japanese language are translated into English language using translate package in R, which translates between different languages with Google API [36]. The reason to translate non-English reviews rather than maintaining their original versions is to achieve unification and consistency in the sentiment index calculation. Since this report uses lexicon-based approaches for the sentiment score, adopting multiple dictionaries for each language would introduce variations in the scale of the sentiment score.

Finally, pre-processing steps are performed, following the processes recommended in prior research [17]:

(1) Lowercase: Every character in each review is converted into lowercase. R is a case sensitive program language, and 'Visit' is different from 'visit' due to character coding; therefore, it is essential to convert textual data into lowercase. For example, taking one review in the dataset, we have:

The original sentence: *"The apartment is in a very convenient location. Host was extremely helpful and the apartment was great. Located in a very calm and nice area, extremely convenient. Would come again for sure!"*.

All cases are transformed into lowercase: *"the apartment is in a very convenient location. host was extremely helpful and the apartment was great. located in a very calm and nice area, extremely convenient. would come again for sure!"*.

(2) Tokenisation: Each review is split into tokens in the form of single words or terms. At the same time, white spaces and punctuation are removed. This is an important step to remove stop words or unnecessary characters: *"the apartment is in a very convenient location host was extremely helpful and the apartment was great located in a very calm and nice area extremely convenient would come again for sure"*.

(3) Stop words removal: This phrase removes stop words, which may be articles, pronouns, prepositions, etc. These words frequently occur but do not add meaning to a sentence, meaning that they do not express any sentiment when applied to lexicon resources. Thus, removing stop words would reduce the noise before text processing. In English, stop words could be 'an', 'the' or 'is'. To remove, we use a stop-words list that is already available on R: *"apartment convenient location host extremely helpful apartment great located calm nice area extremely convenient come again sure"*.

(4) Stemming: Stemming is the technique that involves removing word suffixes to extract the root form of words. This is commonly used in text mining because it simplifies the textual data without causing significant loss of information. For example, "extremely" in the sentence is converted to "extreme": *"apartment convenient location host extreme helpful apartment great locate calm nice area extreme convenient come again sure"*.

To understand the effects of text pre-processing on the comments, key statistics about the length of words in comments are demonstrated in table below.

TABLE I. DESCRIPTIVE STATISTICS ABOUT THE OCRS AFTER AND BEFORE TEXT PRE-PROCESSING

	Before text pre-processing	**After text pre-processing**
Average number of words	40	22
Median number of words	26	15
Shortest comment	1 word	1 word
Number of words 1st quantile	12	7
Number of words 3rd quantile	52	29
Longest comment	2905 words	651 words

As we can see from the Table I, unnecessary characters and stop words are removed to reduce the noise of the dataset since they do not contain information and express sentiment. Quantitatively, the average number of words in

comments has been reduced by nearly a half while the longest comment dropped by about 77.6%. The use of clean data facilitates faster training and thus, enables the implementation of multiple experiments even though limited computational resources are available.

C. Sentiment Extraction

After pre-processing procedures, the sentiment score is constructed from online customer reviews. As discussed in the previous section, while sentiment scores can be extracted in different ways, it is essential to acknowledge that each method has its strengths and limitations. Based on the objectives of this paper, lexicon-based approach is chosen to extract sentiment from texts. There is a wide range of dictionaries that were developed for lexicon-based method, in which each of them offering unique features and attributes. To compare these dictionaries, Al-Shabi [37] has evaluated the performance of the five most well-known lexicons used in sentiment analysis. The results show that VADER (Valence Aware Lexicon and Sentiment Reasoner) demonstrates the highest accuracy in both positive and negative classification.

For this study, VADER is applied to calculate the sentiment score of reviews. VADER is a rule-based sentiment analysis tool to detect sentiment in social media texts. When comparing the classification accuracy, it was found that VADER outperforms individual human raters with Classification Accuracy scores equal 0.96 and 0.84 respectively as it considers both polarity and intensity of emotion. With VADER, the sentiment score, or compound score, is calculated by adding up the valence scores of individual words in the lexicon, which is subsequently normalised to range from –1 (complete negative) to +1 (complete positive). Sentiment classification is then based on this compound score as follows:

- Positive sentiment: compound score $\geq +0.05$
- Negative sentiment: compound score ≤ -0.05
- Neutral sentiment: $-0.05 <$ compound score < 0.05

After determining sentiment score for each review, the overall sentiment score for each listing is computed by taking the mean of sentiment scores associated with that listing. These final sentiment scores are then set as a new feature which is used in price prediction models.

D. Modelling and Analysis

As a reminder, there are two hypotheses that need addressing in this research. The first one is to test the positive association between sentiment scores and prices of Airbnb listings. The second one is to test whether the performance of price prediction models with sentiment score variable are

TABLE II. VARIABLE DEFINITIONS AND LABELS

Variable name	Description
Property characteristics	
Log price	Log-transformed daily price of listing (in USD Dollar)
Country	Country from which the listing is located
Room type	Type of room: (1) Entire home/apt; (2) Hotel room; (3) Shared room; (4) Private room
Accommodate	The maximum capacity of the property
Beds	The number of bed(s)
Minimum nights	The minimum number of nights stay in the listings
Quality characteristics	
Sentiment score	Sentiment score extracted from sentiment analysis on OCRs
Number of reviews	The number of reviews a listing has
Review score rating	The overall review score rating a listing has
Host characteristics	
Host response rate	The rate at which a host response to the guest
Host response time	Time a host response to the guest: (1) Within an hour; (2) Within a few hours; (3) Within a day; (4) A few days or more; (5) Unknown
Host identity verified	The identity of host is verified or not (True/False)
Host acceptance rate	The rate at which a host approves booking requests
Host total listings count	The number of listings one host own on Airbnb

TABLE III. SUMMARY STATISTICS OF NUMERICAL VARIABLES

Variable	Obs	Mean	Median	Min	Max
Log price	13,823	4.653	4.662	-2.322	11.068
Country	13,823	0.672	0.690	-0.960	1
Room type	13,823	3.936	3	1	16
Beds	13,823	2.422	2	1	42
Minimum nights	13,823	6.513	2	1	1125
Sentiment score	13,823	0.672	0.690	-0.960	1
Number of reviews	13,823	40.86	18	1	1548
Review score rating	13,823	4.624	4.750	0	5
Host response rate	13,823	0.971	1	0	1
Host acceptance rate	13,823	0.918	0.990	0	1
Host total listings count	13,823	26	12	1	748

TABLE IV. SUMMARY STATISTICS OF CATEGORICAL VARIABLES

Variable	Categories	Freq.	%
Host response time	0 (within an hour)	1,034	7.48
	1 (within a few hours)	10,678	77.24
	2 (within a day)	1,402	10.14
	3 (a few days or more)	521	3.77
	4 (unknown)	188	1.36
Host identity verified	1 (True)	12,511	90.51
	0 (False)	1,312	9.49
Room type	0 (Entire home/apt)	9,246	66.89
	1 (Hotel room)	444	3.21
	2 (Private room)	3,638	26.32
	3 (Shared room)	495	3.58
Country	0 (Hong Kong)	2,149	15.55
	1 (Japan)	8,895	64.35
	2 (Taiwan)	2,779	20.10

improved or not. In order to examine the second hypothesis, two versions of the dataset are created - one with the sentiment feature and one without the sentiment feature.

The explanatory variables chosen in this study are categorised into four main attributes:

(1) Host characteristics including host response time, host identity verified, host acceptance rate, host response rate, host total listings count;

(2) Property characteristics including country, room type, accommodates, beds, minimum nights stay;

(3) Rating-related features including number of reviews, review score ratings;

(4) Sentiment scores derived from reviews.

The descriptions and summary statistics of both predictor and explanatory variables are shown in Table II, III, IV and V.

Linear Regression [38] is employed first as a baseline model to evaluate the performance of other models. After the baseline is established, several machine learning models namely Ridge Regression [39-40], Support Vector Machine [41], XGBoost [42] and Random Forest [43] are performed.

TABLE V. MULTIPLE LINEAR REGRESSION RESULTS FOR PREDICTING LOG PRICE

Coefficient	Estimate	Std. Error	T-statistic	p-value
(Intercept)	4.514	0.139	32.436	0.000 (***)
host_response_time_1	0.116	0.026	4.452	0.000 (***)
host_response_time_2	-0.058	0.318	-1.808	0.07
host_response_time_3	-0.075	0.043	-1.743	0.08
host_response_time_4	-0.436	0.123	-3.552	0.000 (***)
host_identity_verfied_1	-0.025	0.023	-1.101	0.271
room_type_1	-0.338	0.038	-8.930	0.000 (***)
room_type_2	-0.323	0.016	-20.548	0.000 (***)
room_type_3	-1.153	0.036	-31.704	0.000 (***)
country_1	0.268	0.020	13.150	0.000 (***)
country_2	-0.295	0.023	-13.093	0.000 (***)
accommodates	0.121	0.003	36.052	0.000 (***)
beds	0.001	0.005	0.149	0.882
host_acceptance_rate	0.091	0.045	2.045	0.041 (*)
review_scores_rating	0.061	0.016	3.935	0.000 (***)
host_response_rate	-0.772	0.126	-6.110	0.000 (***)
sentiment_score	0.090	0.043	2.113	0.034 (*)
number_of_reviews	-0.001	0.000	-5.751	0.000 (***)
minimum_nights	-0.000	0.000	-0.959	0.338
host_total_listings_count	-0.001	0.000	-5.170	0.000 (***)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

To test the hypothesis, five machine learning models were built to predict the log price of Airbnb properties in two cases: with and without the sentiment score variable. Linear Regression was employed first as the baseline model for model comparison in this study. Subsequently, Ridge Regression, Support Vector Machine, XGBoost, Random Forest were built. To compare the predictability performance of the models, R^2, RMSE and MAE were calculated.

E. Model Performance Evaluation

To evaluate the performance of predictive models, the dataset is partitioned into training and test sets based on the dependent variable – the price, where 75% of the dataset is for training and 25% for testing. In our case, the training set contains 10,369 records, while there are 3,454 records in the test set. First, the models are fitted into the training set to learn patterns and relationships in the data. After the models are built, they are evaluated on the test set, which contains of unseen data points.

Three metrics, R^2, RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) are used to evaluate and compare the performance of predictive models. R^2 or the coefficient of determination, which is a standard metric for evaluating regression analyses, measures how close the target variable is determined by explanatory variables, interpreted by the proportion of total variance of the regressand explained by the model. While RMSE is the standard deviation of mean prediction errors, MAE measures the average magnitude of the prediction errors.

The formula of R^2 and RMSE are as below with y_i is the actual value and $\hat{y}_i$ is the predicted value of the dependent variable.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y}_{ii})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i| \quad (3)$$

A model can be considered to be effective if its R^2 is close to 1; RMSE and MAE are close to 0.

IV. Results and Analysis

A. Customer reviews

The reviews were gathered from Inside Airbnb with a total of 569,523 reviews for the analysis. However, following the data pre-processing process, there are 507,974 reviews left. There are 13,789 listings, with an average of 37 reviews per listing. After detecting the language of each review in the dataset, it can be seen from Figure 1 that more than half of the reviews are written in English, with 254,490 reviews. This prevalence can be attributed to the global customer base of Airbnb, attracting users from worldwide, where English is commonly used for international communication. Following English are Japanese and Chinese language, with 23.6% and 24.3% of the reviews respectively.

After all the text has been translated to English, to provide more detailed information about reviews, two-word clouds are drawn based on a collection of OCRs from Airbnb listings in Hong Kong, Japan and Taiwan. These graphs help to display the keywords hidden in reviews. The larger words in the word cloud indicate higher frequency in the reviews, while the smaller words reveal less occurrence. In Figure 2 and Figure 3, 80 of the most frequent words are selected.

Using VADER sentiment analysis, the sentiment polarity and sentiment intensity of reviews are obtained, in which polarity assigns whether reviews are negative, positive or neutral, while intensity indicates the strength of the negative or positive sentiment in the text. Figure 4 shows that almost all of the reviews were positive, accounting for approximately 94% of the reviews, followed by neutral and negative reviews with only 3.8% and 2.2% respectively. Positive reviews are the most dominant sentiment category, reflecting guests' overall positive experiences with Airbnb in 3 Asia countries.

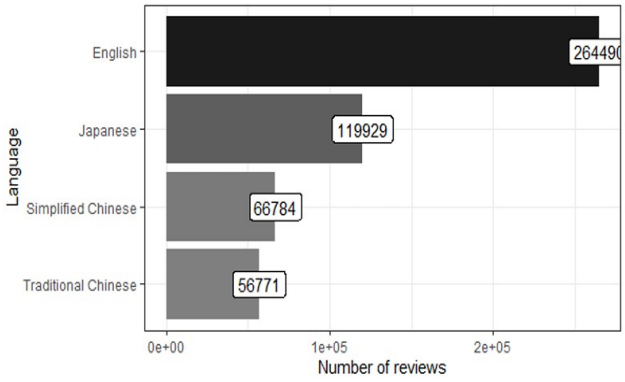

Fig. 1. Frequency of the languages in reviews written by guests on Airbnb in Hong Kong, Japan and Taiwan.

To get a sense of sentiment scores in each individual country, an ANOVA (Analysis of Variance) test was conducted to analyse the differences in sentiment scores among

Fig. 2. World cloud of OCRs on Airbnb in Hong Kong, Japan and Taiwan.

different countries. The results are illustrated in Table VI, with F-statistic equals 91.03 and the p-value is less than 0.0001. The p-value is below conventional significance level of 0.05, which indicates that there is a statistically significant difference in the means of sentiment scores among the three countries. As a result, this underscores the importance of considering country-specific factor when assessing sentiment scores.

The sentiment score for each listing was calculated by taking the mean of sentiment intensity across each review associated with that specific listing. As shown in Table VII, the average sentiment score is 0.707. However, there is some variability in the sentiment score, ranging from –0.998 (most negative) to 1 (most positive).

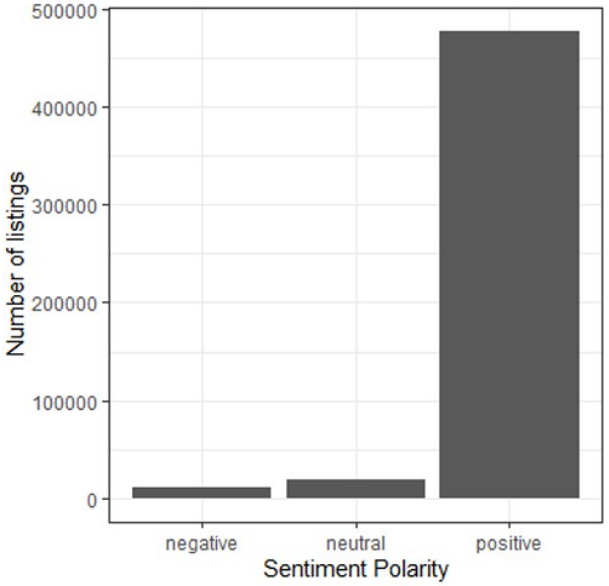

Fig. 4. Sentiment Polarity distribution of reviews per listing.

TABLE VI. ANOVA For Sentiment Scores Among Hong Kong, Japan And Taiwan

	Degrees of freedom	Sum of squares	Mean squares	F-statistic	p-value
Country	2	16	8.039	91.03	<0.0001(***)
Residuals	507971	44858	0.088		

Note: (***) denotes a 1% level of significance.

TABLE VII. Descriptive Statistics On Sentiment Scores And Prices Of Listings

Variable	Mean	Median	Maximum	Minimum	Standard Error
Sentiment Score	0.707	0.807	1	-0.998	0.297
Price	172.07	105.86	64,109.08	0.1	838.65
Log_price	4.65	4.66	11.07	-2.32	0.86

In terms of prices of listings, the highest price is 64,109.08 USD, reflecting the presence of luxury listings. On the other hand, the lowest price recorded is 0.1 USD, possibly a promotional offer by the host or the platform. Besides that, the average price (172.07) is higher than its median (105.86), indicating that the distribution of price is skewed to the right due to some extreme values. According to Osborne and Overbay (2004), extreme values or also known as outliers can affect even simple analyses and model performance, therefore, a logarithmic transformation is applied to the prices to deal with positive skewness, as illustrated in Figure 5 and Table VII, where the mean and median of log price are fairly similar. Log_price is then used as the target variable in building models.

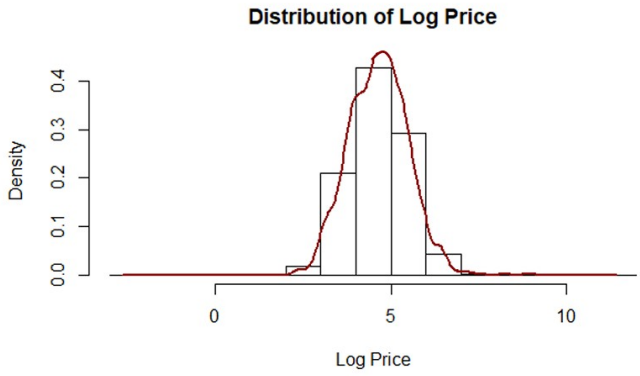

Fig. 5. Distribution of Price and Log Price.

To examine the association between sentiment scores and rental prices, the scatter plot between these two numeric variables is drawn as in Figure 6 and Figure 7. From the chart, it appears that the data shows an uphill pattern in Hong Kong, Japan and Taiwan, which suggests higher sentiment scores is associated with higher prices of properties. Notably, there is no high-priced property exhibiting overall negative reviews in all three countries.

In a statistical context, we use Pearson's correlation to measure the linear association between two numerical variables. In Hong Kong and Taiwan, the correlation coefficients are 0.06 and 0.07 respectively, indicating a weak positive linear association between price and sentiment score. On average, higher sentiment scores on Airbnb are slightly associated with higher listings prices in both countries, however, the relationship is not strong.

On the other hand, the correlation coefficient between price and sentiment is –0.01 in Japan, which opposes to the case of Hong Kong and Taiwan. To be specific, this value is no statistically significant difference from zero, suggesting that there is almost no linear relationship between price and sentiment in this country. In other words, the sentiment expressed in Airbnb reviews does not have any significant impact on the listing prices in Japan. Overall, there is a statistically weak relationship between sentiment score and price variables in the three countries.

In Figure 7, the correlation coefficients between sentiment scores and log-transformed prices are higher compared to those between sentiment scores and actual prices in both Hong Kong and Taiwan, with 0.13 and 0.18 respectively. This is because log transformation mitigates the issues of outliers which are the cases with extremely high prices in these two countries. Nevertheless, the correlation coefficient remains unchanged in Japan. This suggests that there is little or no meaningful linear correlation between the scores and the price variable in Japan, regardless of whether the price is log-transformed or not.

B. Performances of the models

As already mentioned in the early section, the sentiment score per listing is then integrated into the listing dataset as one of the features for price prediction. To summary, this paper employed various machine learning models to not only predict price but also examine the association between target feature price and a number of explanatory features. As the distribution of price is positively skewed, it is decided to apply logarithmic transformation to the price before modelling. The log transformation not only makes data closer to normal distribution but also mitigates the impact of outliers.

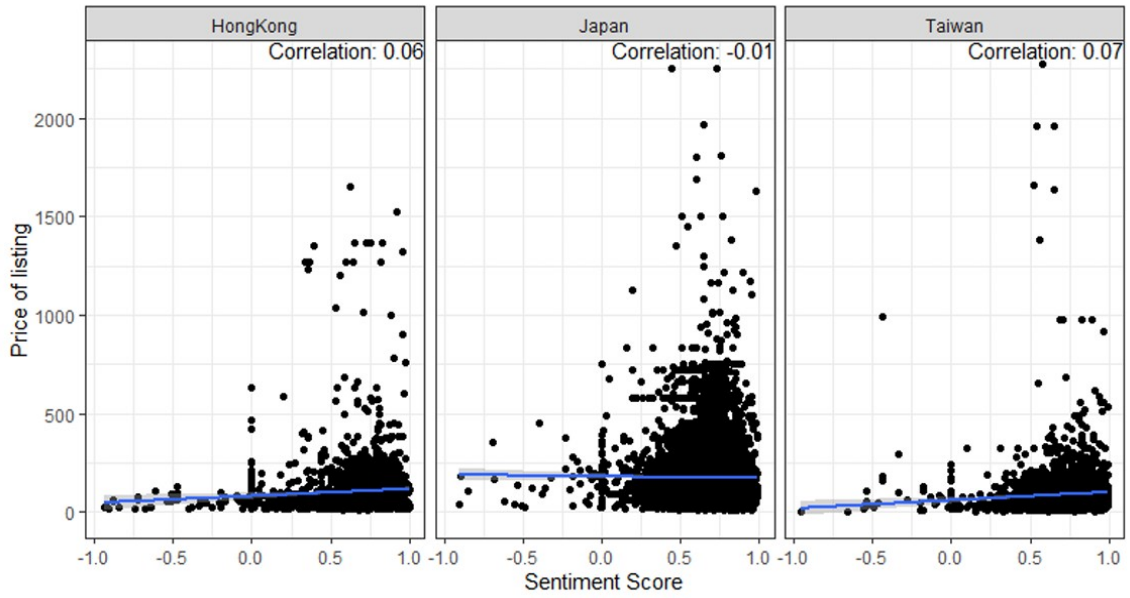

Fig. 6. Scatter Plot of Sentiment Score and Price of listing on Airbnb in Hong Kong, Japan and Taiwan.

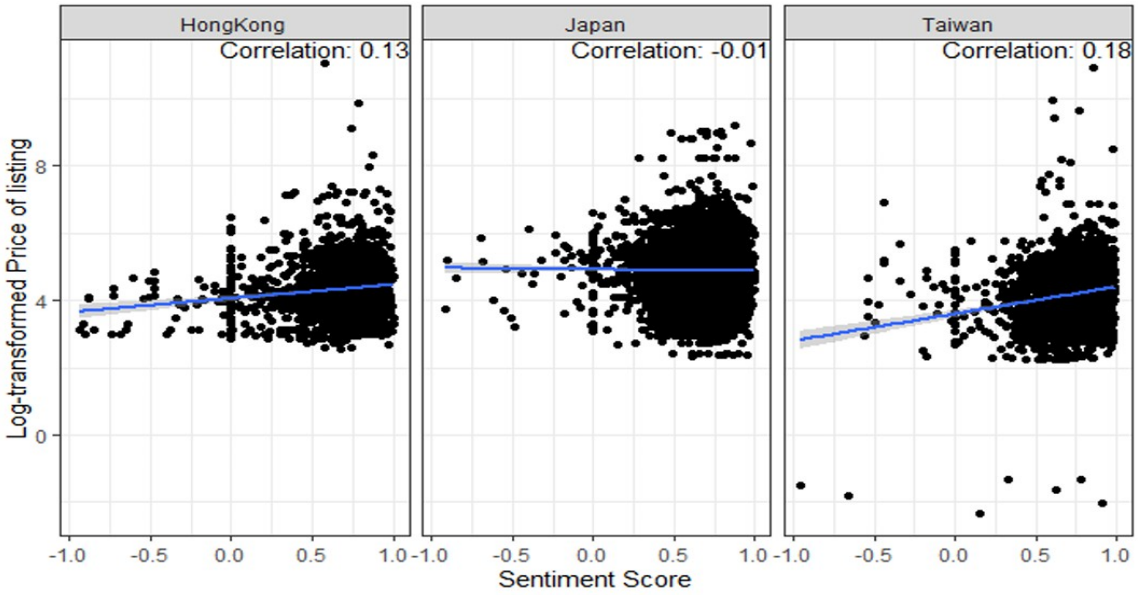

Fig.7. Scatter Plot of Sentiment Score and Log-transformed Price of listing on Airbnb in Hong Kong, Japan and Taiwan.

Linear Regression was employed first as a baseline model to establish a cornerstone for understanding the association between the predictor variable – log price, and 19 response variables including sentiment score variable. Table VIII shows the empirical results of sentiment score variable (other variables are presented in Table II). From Table VIII, the coefficient estimate is around 0.090, which suggests that on average, one unit increase i.e., from 0 (neutral) to 1 (completely positive) in sentiment score, is associated with approximately 9% increase in the price of property while other variables remain constant. The p-value is below the significance level of 0.05, indicating the association between sentiment score and log price is statistically significant in the multiple regression model.

Following that, Ridge Regression, SVM, Random Forest and XGBoost models were employed. Table IX shows the results with evaluation metrics on the test set of all the five chosen models. The metrics are compared based on the baseline model. Two cases are divided, one with *sentiment_score* and one without *sentiment_score feature* in order to test the second hypothesis.

In general, when looking at the performance metrics from Table IX, it can be concluded that other models perform better than the baseline model. The baseline model – Linear regression elicits much lower accuracy, with the value of R^2

TABLE VIII. Regression Coefficient Of Sentiment Score Variable For Predicting Log Price

Coefficient	Estimate	Standard Error	T-statistic	p-value
sentiment_score	0.090	0.043	2.113	0.035 (*)

Note: (*) denotes a 5% level of significance.

TABLE IX. Performance Comparison Of Five Models With And Without Sentiment Score Feature

Model name	With sentiment_score feature			Without sentiment_score feature		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Linear Regression	0.409	0.664	0.485	0.409	0.664	0.485
Ridge Regression	0.409	0.664	0.485	0.409	0.664	0.485
Support Vector Machine	0.516	0.602	0.417	0.532	0.593	0.406
XGBoost	**0.654**	**0.509**	0.348	0.657	0.507	0.340
Random Forest	0.653	0.510	**0.339**	**0.659**	**0.505**	**0.333**

being 40.9%. Both linear regression and Ridge Regression yield the same R^2 (40.9%), RMSE (0.664) and MAE (0.485). This suggests that there might not be substantial multicollinearity in the dataset, therefore, the additional regularisation which discourages large coefficients in regression may not be necessary.

With the sentiment score feature, it is clear the performance metrics vary across models, with SVM, XGBoost and Random Forest generally performing better than Linear regression, Ridge regression. The R^2 value of SVM, Random Forest and XGBoost are significantly improved compared to the baseline model, with increases of 10.7%, 24.4% and 24.5% respectively. This indicates that there might be weak linear patterns in the dataset. While Linear regression and Ridge regression assume a linear association between independent and dependent variables, the other three models excel at handling non-linearities and higher-dimensional dataset, therefore, provide a better fit to the dataset. To conclude, XGBoost and Random Forest perform similarly and outperform other models in terms of R^2, RMSE and MAE. While XGBoost has the highest accuracy score (65.4%) and lowest RMSE (0.509), Random Forest has the lowest MAE (0.339). Though both are high-performing models, XGBoost obtains stronger overall predictive performance, whilst Random Forest allows better accuracy in terms of minimizing the magnitude of prediction errors.

To analyse the influence of sentiment score feature on the prediction accuracy of the models, the same experiment is repeated by removing the sentiment feature. From Table IX, the performance metrics of Linear regression and Ridge regression remain the same in both scenarios. However, there are slight changes in the model performance for SVM, XGBoost and Random Forest. Though the changes are not significant, when removing the sentiment score, R^2, RMSE and MAE are slightly better across three models. Overall, with an accuracy of 65.9%, Random Forest proves to be the best-performing model. Not only the highest accuracy score, but Random Forest also shows the lowest RMSE (0.505) and MAE (0.333).

According to the findings presented in Table IX, the incorporation of sentiment score did not appear to improve the models' performance for price prediction of Airbnb in three Asian countries. After including sentiment score, the performance results dropped slightly across all models. The best result is Random Forest, without the inclusion of sentiment score feature.

C. Influencing factors of price

To illustrate the significance of each feature for price prediction, especially the sentiment score feature, we extracted the importance scores generated by XGBoost and Random Forest since they are two best-performing models with the inclusion of sentiment score. For Random Forest, the importance is calculated using the mean decrease in impurity, also known as Gini impurity, which measures the quality of a split in a decision tree. For XGBoost, it calculates based on the mean squared error when creating splits in tree.

Feature importance measures help to measure the importance of each feature by which the accuracy is improved when the high-ranking feature is included and vice versa. To be specific, the higher value of a feature, the more important this feature is for the model. Figure 8 and Figure 9 shows the ranking of each feature in XGBoost and Random Forest models with the inclusion of sentiment score feature, from the highest to lowest. For both models, the accommodates feature exhibits the highest rank, signifying its paramount role as the most influential variable in determining price.

For XGBoost, the sentiment score feature holds the 6th position among 19 variables in terms of importance. This

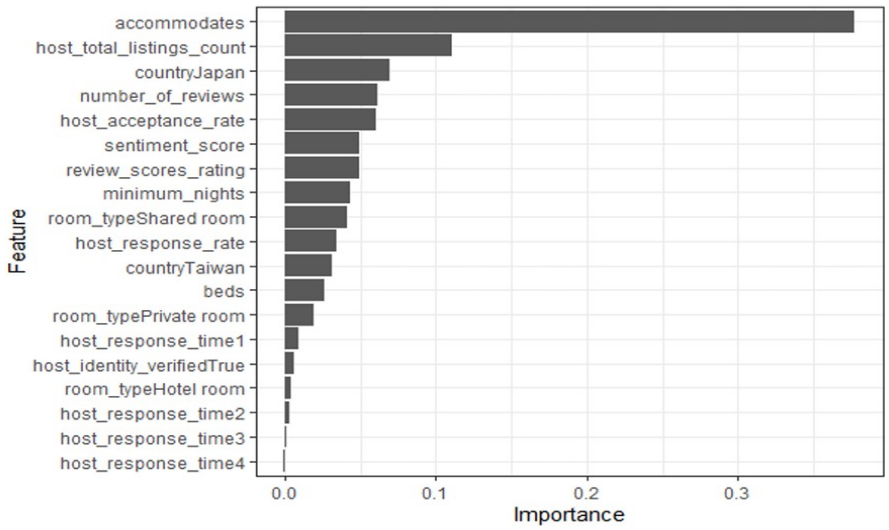

Fig. 8. The importance value for each feature in XGBoost with the inclusion of sentiment score feature.

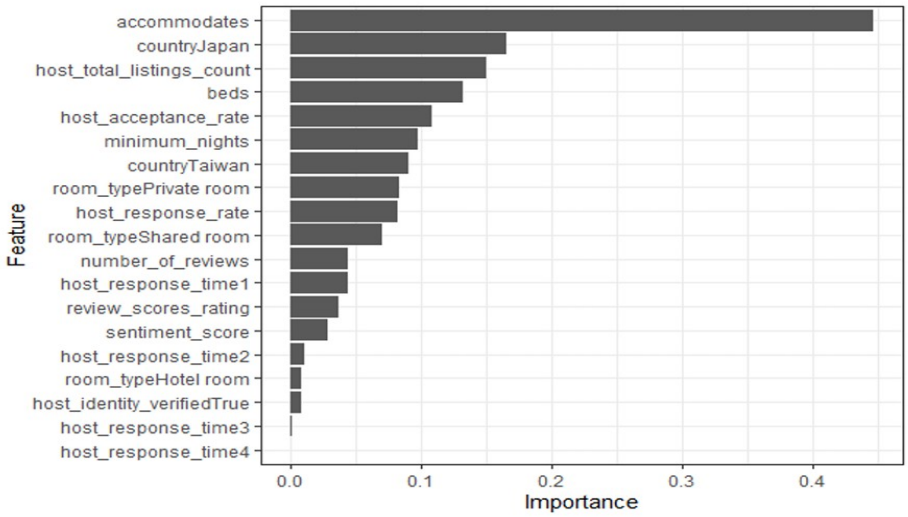

Fig. 9. The importance value for each feature in Random Forest with the inclusion of sentiment score feature.

means that this variable contributes significantly to the predictive performance of the model. On the other hand, with Random Forest, the sentiment score feature occupies a lower rank, with 14th out of 19 features. This implies that the sentiment demonstrates comparatively less importance in Random Forest performance, as it is outweighed by more than half of the total variables with greater influence.

D. The association between sentiment score and price

In order to investigate the association between sentiment scores and accommodation prices, we analyse the correlation between the two variables as well as the regression outcomes. First, the correlation between sentiment scores and Airbnb listing prices in three selected Asian countries is examined. The correlation coefficients for sentiment scores and prices are found to be slightly positive in Hong Kong and Taiwan. Subsequently, when considering log-transformed prices, the correlation coefficients remain similar patterns with slightly stronger compared to the analysis using actual prices. This indicates that higher sentiment scores are slightly associated with higher Airbnb listings prices despite a weak linear association between these two variables in Hong Kong and Taiwan. Despite a weak correlation, the observation that positive sentiment scores are associated with higher-priced listings in Hong Kong and Taiwan can be explained by the perceived quality of guests. Customers' sensitivity toward prices was found to enhance their perceived value on Airbnb. In other words, customers often assume that higher priced offerings reflect better quality. In essence, positive reviews signal previous positive guest experiences, therefore, leading potential guests to associate higher prices with superior accommodations or additional amenities. Eventually, this creates a willingness to pay more for a better experience.

On the contrary, in the context of Airbnb in Japan, the negative correlation coefficient between sentiment scores

and prices is close to 0, which is similar to that between sentiment scores and log price. This indicates that the linear correlation between these variables is likely statistically insignificant, and by that, there is no association between sentiment scores and price in Japan. In contrast to Hong Kong and Taiwan, guest sentiment in Japan, whether positive or negative, may not significantly influence Airbnb listings prices in Japan.

In general, based on the results of correlation and regression analysis, there is some support for the hypothesis that positive sentiment expressed in OCRs is associated with an increase in Airbnb rental prices. To conclude, there is no linear association between guest sentiment and prices of Airbnb listings, however, positive sentiment scores are associated with higher listings prices in Hong Kong and Taiwan despite a weak correlation. Additionally, the regression analysis across three countries shows that an increase in sentiment score is associated with a modest increase in rental prices. This result shows that prices of Airbnb listings are influenced by review scores.

From a theoretical viewpoint, this study exhibits the usefulness of leveraging text analysis on OCRs to identify the patterns of consumer sentiment as well as their behaviour. Not only enriching the insights into the behaviour preferences of consumers in Asian countries, but the study also illustrates the cultural differences when compared with existing literature on Airbnb in Western countries. This study contributes to the existing literature on the association between sentiment scores derived from OCRs and Airbnb listing prices. On practical side, this study provides an in-depth understanding of customer perceptions and behaviour toward their experiences with Airbnb hosts. On the one hand, positive contents of reviews help them to further enhance the products and services such as hygiene factors and the helpfulness of hosts.

V. Conclusions and Future Works

In conclusion, this paper has embarked on a comprehensive journey upon exploring the significance of sentiment analysis on OCRs in predicting Airbnb rental prices. As the results of content analysis, **the answers for three research questions** have been found, which are ❶ the sentiment analysis on OCRs in Airbnb in three selected Asian countries namely Hong Kong, Japan and Taiwan; ❷ the association between sentiment scores and prices of listings; ❸ the performance of Airbnb rental price prediction model with the inclusion of sentiment scores from Airbnb OCRs. In the pursuit of uncovering these research questions, integration of advanced natural language processing on sentiment analysis as well as machine learning models have been employed. The results suggest that there is a weak positive association between sentiment scores and rental prices across three countries, and the inclusion of sentiment scores into price prediction models slightly decreases their predictability. The uniqueness of this paper lies in the adoption of a large amount of data from Asian regions, which has received limited attention in existing literature. As such, it is hoped that this study enhances the understanding of not only Airbnb but also the hospitality industry in Asian countries.

This study also holds several **limitations** that open up promising avenues for future works. *Firstly,* the scope of the study is limited to only three Eastern Asia countries, namely Hong Kong, Japan and Taiwan. Therefore, future work could be extended to more countries within the region to offer a more generalised perspective on Airbnb in Asia. *Secondly,* due to the computationally expensive process, this study only built price prediction models based on the price observed on a specific date, which might neglect the dynamic fluctuations in pricing trends. This can be improved by incorporating historical pricing data over a period of time, thereby capturing any trends, seasonality and changes in demand that influence prices considering the sentiment scores. *Thirdly,* the current study has only performed the sentiment analysis on Airbnb OCRs using VADER lexicon-based approach, which may restrict the exploration of alternative techniques [44-51] (such as big data, knowledge graph, LLM, RAG and so on) that could potentially offer different insights and more accurate sentiment scores.

Acknowledgement

The authors would like to thank to Dr. Andrew Burlinson for his invaluable support to accomplish this paper to the fullest.

References

[1] Moon, H., Miao, L., Hanks, L. and Line, N.D. (2019). Peer-to-peer interactions: Perspectives of Airbnb guests and hosts. International Journal of Hospitality Management, 77, pp. 405-414.

[2] Negi, G. and Tripathi, S. (2022). Airbnb phenomenon: a review of literature and future research directions. Journal of Hospitality and Tourism Insights, doi: 10.1108/JHTI-04-2022-0133.

[3] Dogru, T., Mody, M. and Suess, C. (2019). Adding evidence to the debate: quantifying Airbnb's disruptive impact on ten key hotel markets. Tourism Management, 72, pp. 27-38.

[4] Mody, M.A., Jung, S., Dogru, T. and Suess, C. (2023). How do consumers select between hotels and Airbnb? A hierarchy of importance in accommodation choice. International Journal of Contemporary Hospitality Management, 35(4), pp. 1191-1218.

[5] Mahyoub, M., Al Ataby, A., Upadhyay, Y., and Mustafina, J. (2023). AIRBNB Price Prediction Using Machine Learning. In: 2023 15th International Conference on Developments in eSystems Engineering (DeSE), Iraq, 09-12 January, pp. 166-171.

[6] Jiang, L., Li, Y., Luo, N., Wang, J. and Ning, Q. (2022). A Multi-Source Information Learning Framework for Airbnb Price Prediction. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), USA, 28 November – 01 December, pp. 1-7.

[7] Lawani, A., Reed, M.R., Mark, T. and Zheng, Y. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. Regional Science and Urban Economics, 75, pp. 22-34.

[8] Adamiak, C. (2019). Current state and development of Airbnb accommodation offer in 167 countries. Current Issues in Tourism, 25(19), pp. 3131–3149.

[9] Chen, T., Samaranayake, P., Cen, X.Y., Qi, M. and Lan, Y.C. (2022). The Impact of Online Reviews on Consumers' Purchasing Decisions: Evidence From an Eye-Tracking Study. Frontier in Psychology, 13:865702, doi: 10.3389/fpsyg.2022.865702.

[10] Guo, J., Wang, X. and Wu, Y. (2020). Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. Journal of Retailing and Customer Services, 52, 101891, pp. 1-21.

[11] Ma, E., Cheng, M. and Hsiao, A. (2018). Sentiment analysis – a review and agenda for future research in hospitality contexts. International Journal of Contemporary Hospitality Management, 30(11), pp. 3287-3308.
[12] Chang, W.L. and Wang, J.Y. (2018). Mine is yours? Using sentiment analysis to explore the degree of risk in the sharing economy. Electronic Commerce Research and Applications, 28, pp. 141-158.
[13] Zeng, G., Cao, X., Lin, Z. and Xiao, S.H. (2020). When online reviews meet virtual reality: Effects on consumer hotel booking. Annals of Tourism Research, 81, 102860.
[14] Gavilan, D., Avello, M. and Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. Tourism Management, 66, pp. 53-61.
[15] Lee M., Jeong M. and Lee J. (2017). Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach. International Journal of Contemporary Hospitality Management, 29(2), pp. 762–783.
[16] Lee, C.K.H., Tse, Y.K., Zhang, M., and Ma, J. (2020). Analysing online reviews to investigate customer behaviour in the sharing economy: The case of Airbnb. Information Technology and People, 33(3), pp. 945-961.
[17] Amat-Lefort, N., Barravecchia, F. and Mastrogiacomo, L. (2023). Quality 4.0: big data analytics to explore service quality attributes and their relation to user sentiment in Airbnb reviews. International Journal of Quality and Reliability Management, 40(4), pp. 990-1008.
[18] Tran, T., Ba, H. and Huynh, V.N. (2019). Measuring hotel review sentiment: An aspect-based sentiment analysis approach. International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, Springer, pp. 393–405.
[19] 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, pp. 111-121.
[20] Zhu, E., Wu, J., Liu, H. and Li, K. (2022). A Sentiment Index of the Housing Market in China: Text Mining of Narratives on Social Media. The Journal of Real Estate Finance and Economics, 66, pp. 77–118.
[21] 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, pp. 111-121.
[22] Zhang, C., Tian, Y.X. and Fan, Z.P. (2022). Forecasting sales using online review and search engine data: A method based on PCA–DS-FOA–BPNN. International Journal of Forecasting, 38(3), pp. 1005-1024.
[23] Picasso, A., Merello, S., Ma, Y., Oneto, L. and Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. Expert Systems with Applications, 135, pp. 60-70.
[24] Symitsi, E., Stamolampros, P. and Karatzas, A. (2021). Augmenting household expenditure forecasts with online employee-generated company reviews. Public Opinion Quarterly, 85, pp. 463-491.
[25] Wu, D.C., Zhong, S., Qiu, R.T.R. and Wu, J. (2022). Are customer reviews just reviews? Hotel forecasting using sentiment analysis. Tourism Economics, 28(3), pp. 795-816.
[26] Gibbs, C., Guttentag, D., Gretzel, U., Morton, J. and Goodwill, A. (2018). Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. Journal of Travel and Tourism Marketing, 35(1), pp. 46-56.
[27] Ye, P., Qian, J., Chen, J., Wu, C.H., Zhou, Y., Mars, S.D., Yang, F. and Zhang, L. (2018). Customized regression model for airbnb dynamic pricing. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19 23 August, pp. 932-940.
[28] Kwok, L. and Xie, K.L. (2019). Pricing strategies on Airbnb: Are multi-unit hosts revenue pros?. International Journal of Hospitality Management, 82, pp. 252-259.
[29] Wang, D. and Nicolau, J.L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. International Journal of Hospitality Management, 62, pp. 120-131.
[30] Faye (2021). Methodological discussion of Airbnb's hedonic study: A review of the problems and some proposals tested on Bordeaux City data. Annals of Tourism Research, 86, 103079.
[31] Liu, Y. (2021). Airbnb pricing based on statistical machine learning models. In: 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML), CA, USA, 14 November 2021, pp. 175-185.
[32] Ganu, G., Elhadad, N. and Marian, A. (2009). Beyond the stars: improving rating predictions using review text content. Twelfth International Workshop on the Web and Databases, June 28, WebDB, USA.
[33] Chica-Olmo, J., González-Morales, J.G. and Zafra-Gómez, J.L. (2020). Effects of location on Airbnb apartment pricing in Málaga. Tourism Management, 77, 103981.
[34] Guttentag, D. (2019). Progress on Airbnb: a literature review. Journal of Hospitality and Tourism Technology, 10(4), pp. 814-844.
[35] Sites, D. (2013). Compact https://github.com/CLD2Owners/cld2.
[36] Mohamed, E.D. (2023). gtranslate, Github repository, https://github.-com/mohamed 180/gtranslate.
[37] Al-Shabi (2020). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. International Journal of Computer Science and Network Security, 20(1), pp. 51-57.
[38] Maulud, D., and Abdulazeez, A.M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), pp. 140-147.
[39] McDonald, G.C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), pp. 93-100.
[40] Saleh, A.M.E., Arashi, M., and Kibria, B.G. (2019). Theory of ridge regression estimation with applications. John Wiley & Sons.
[41] Wauters, M., and Vanhoucke, M. (2014) Support vector machine regression for project control forecasting, Automation in Construction, 47, pp. 92-106.
[42] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[43] Yassine Al Amrani, Mohamed Lazaar, Kamal Eddine El Kadiri (2018). Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. Procedia Computer Science. Volume 127, Pages 511-520, ISSN 1877-0509.
[44] Long, C. K., Trung, H. Q., Thang, T. N., Dong, N. T., & Van Hai, P. (2021). A knowledge graph approach for the detection of digital human profiles in big data. Journal of Science and Technology: Issue on Information and Communications Technology, 19(6.2), 6-15.
[45] Long, C. K., Van Hai, P., Tuan, T. M., Lan, L. T. H., Chuan, P. M., & Son, L. H. (2022). A novel fuzzy knowledge graph pairs approach in decision making. Multimedia Tools and Applications, 1-30.
[46] Long Cu Kim and Hai Pham Van (2018). Intelligent Collaborative Decision Model for Simulation of Disaster Data in Cities and Urbanlization. International Journal of Advanced Research (IJAR), Vol. 6, Issue 07.
[47] C. K. Long et al. (2020). A Big Data Framework for eGovernment in Industry 4.0. Open Computer Science, ISSN: 2299-1093.
[48] Hai Van Pham, Long Kim Cu, (2020). Intelligent Rule-based Support Model Using Log Files in Big Data for Optimized Service Call Center Schedule. Proceedings of International Conference on Research in Intelligent Computing in Engineering, ISBN 978-981-15-2780-7.
[49] C.K.Long et al. (2021). Disease Diagnosis in the Traditional Medicine: A Novel Approach based on FKG-Pairs. Journal of Research and Development on Information and Communication Technology, Vol. 2021(2), pp. 59-68.
[50] Pham, H. V., Long, C. K., Khanh, P. H., & Trung, H. Q. (2023). A Fuzzy Knowledge Graph Pairs-Based Application for Classification in Decision Making: Case Study of Preeclampsia Signs. Information, *14*(2), 104.
[51] Cu Kim Long, Pham Van Hai, et al. (2023). A novel Q-learning-based FKG-Pairs approach for extreme cases in decision making. Engineering Applications of Artificial Intelligence, Vol. 120, 2023, ISSN 0952-1976.