Subroto and Christianis *J Big Data* (2021) 8:9
https://doi.org/10.1186/s40537-020-00395-6

Journal of Big Data

RESEARCH Open Access

Rating prediction of peer-to-peer accommodation through attributes and topics from customer review

Athor Subroto[1,2*] and Marcel Christianis[1]

*Correspondence:
athor.subroto@ui.ac.id
[1] Department of Management, Faculty of Economics and Business, Universitas Indonesia, Jakarta, Indonesia
Full list of author information is available at the end of the article

Abstract

This study aims to predict customers' behavior in classifying their reviews as high rated or low rated using associated attributes and topics found in the review. Knowing customer reviewing action better can lead to a successful strategy implementation of the relevant parties related to this study such as policy to manage customer reviews by keeping their satisfaction high. We applied a big data approach on a dataset of 55,377 reviews from Airbnb listings in the top 10 most visited cities in Indonesia (based on foreign arrivals data). We used The Classification and Regression Tree Model, Random Forest Model, Least Absolute Shrinkage and Selection Operation and Logistic Regression Model, Artificial Neural Network as well as Multi-Layer Perceptron to make prediction's classification. Those models are used to identify a set of attributes and topics that will increase the chance of the review to render a high rate and a different set of attributes and topics that will lead the review to be low rated. This study found; first, attributes and topics that influence customers' odds to classify their review as high rated or low rated adhere to the understanding of Peer to Peer accommodation attributes. Second, successfully proved that customer reviews' attributes and topics could be used to predict the classification of ratings in Peer to Peer accommodation. Where for Topics, we can predict the rating using Random Forest yields 60.09% accuracy, slightly better than Artificial Neural Network (58.33%) and Multi-Layer Perceptron (58.8%). However, it seems better to use Attributes to predict the rating, where the accuracy is yielded better by applying Artificial Neural Network with 84.79% accuracy compared to Multi-Layer Perceptron with only 72.35% of accuracy.

Keywords: Rating prediction, Peer to peer accommodation, Customer review, Big data, Sharing economy, Tourism industry

Introduction

Peer-to-peer (P2P) is an emerging model in the tourism industry that has changed and disrupted the way consumers choose accommodation. P2P is a component of a bigger movement known as the "Sharing Economic," where consumers share and offer underutilized assets to other consumers [1]. P2P accommodation has established itself as a viable alternative to the traditional accommodation model and has driven away significant market share in the accommodation industry [2]. This study will focus on Airbnb, a

dominant P2P accommodation platform used across the globe. Albeit recent, the peer-to-peer accommodation model has grown at an unprecedented speed, which causes an understanding of the model to lag behind the phenomenon that is taking action in the industry. As such, many cast the understanding and perspective of traditional accommodation in order to make sense of how P2P accommodation works and the driving forces behind the decision making process of P2P accommodations' customers [3]. Yet, there is no denying that P2P accommodation must be viewed as a separate model and a fresh perspective.

Several prior studies have attempted to identify a separate set of attributes that exclusively describe P2P accommodation without being anchored to traditional accommodation perspectives. Previous research identified a new set of attributes that are commonly found in P2P accommodation and able to describe the nuanced aspects of the Airbnb experiences that are not found in traditional accommodation [4, 5]. Bridges and Vásquez have looked into how languages are used and contrasted in positive and negative customer reviews by Airbnb customers [6]. However, existing studies mainly focus on identifying by breaking down customer reviews but have yet to look further into the role of these attributes and how they can predict customer behaviors in term of giving final rating.

Given the importance of P2P accommodation attributes in understanding customers' decisions, this study sees the opportunity to fill the gap of how customers' behavior can be predicted based on the attributes they use in writing reviews. The study employs a Big Data approach by creating a predictive model on attributes extracted from a large set of Airbnb customer reviews. The Big Data's significance to this study is that the view coming from the millions of customers is considerably frank by their own willingness to share that reflects the user experiences. Thus, Big Data can offer the real like situation and robustness of the data. The studies build on top of the research by [6] by predicting customer behavior in classifying their review as high rated or low rated based on the usage of attributes as previously identified by [4, 5]. The studies will use an array of classification predictive models such as Classification and Regression Tree (CART) Model, Random Forest (RF), Least Absolute Shrinkage and Selection Operation (LASSO) Logistic Regression, Artificial Neural Network (ANN), and Multi-Layer Perceptron (MLP) to perform classification prediction. The study contributes to the P2P accommodation literature by providing a case of understanding how customers behave through reviews that they wrote. The study also contributes methodologically by showing the feasibility of creating a predictive model using attributes and topics identified from the texts' collection.

Literature review

Sharing economy and peer-to-peer accommodation

Reviews on the peer to peer accommodation have been conducted by some researchers [7–9], mostly using Airbnb related data [10–13] and through the sharing economy [14, 15]. The sharing economy refers to a global phenomenon with rapid growth potential [16]. That is where consumers are sharing and granting each other temporary access to their privately owned goods. In other words, asset owners utilize digital clearinghouses to make the most out of the unused capacity of things that people possess [17]. This

behavior is supercharged and enabled by online marketplaces that act as intermediaries between consumers. Usually, the property consumers offer are underutilized physical assets and can be provided to other consumers for monetary benefits [18]. Among the existing sharing economy model, peer-to-peer accommodation has emerged as a significant segment of the sharing economy that has impacted the tourism industry [19, 20]. One research argues that P2P accommodation may be a form of evolution in the accommodation offerings market due to hotels' inability to meet the needs and habits of their guests. Thus, most of its users consider it as a hotel substitute [21], although during weekends and holidays [22]. Those needs are predominantly in the dimensions of uncertainty, localness, communities, and personalization [23]. As well as because of the needs around modern internet technologies, distinct appeal, which centers on cost-savings, household amenities, and the potential for more authentic local experiences [24].

P2P accommodation has several distinct differences in comparison to traditional hotels. In the sharing economy model of P2P accommodation, neither supplier nor consumer is affiliated with the platform that enables the transactions. The platform simply facilitates the discovery process and mediates transactions. That contrasts with how traditional hotels work where customers will transact with an established entity that manages the accommodation service [18]. The most significant and most noticeable difference with the sharing economy model is the aspect of risk. Unlike hotels, P2P accommodation has a lack of standards and operating procedures dependent on each host's capabilities. In the traditional hotel model, there is a clear definition of the standards that are reflected in the hotel's brand, operating procedures, price, and experience.

Meanwhile, in P2P accommodation, these attributes vary from one accommodation to another. Since customers cannot immediately judge the accommodation's quality, there is a degree of caution behind choosing accommodation in the peer-to-peer model. This caution is mitigated by the review system that platforms enforced with the interest of establishing trust between hosts and potential future guests [18]. However, the guest's reviews could give challenges to the Airbnb host, such as risk, lack of privacy, and emotional stress [25].

Peer-to-peer accommodation attributes and topics

The human element is a crucial characteristic of P2P accommodation that fundamentally change how accommodation is perceived. The human element rooted in the behavior of people renting out their property to guests creates an authentic experience with the property host that conventional hotels can't replicate [5]. In understanding how P2P accommodation is perceived, previous research has attempted to find what attributes and aspects of P2P accommodation that guests care about. A study finds that service quality associated with the website, host, and facility can produce distinctive customer satisfaction effects [26]. According to [27], Airbnb and hotels' critical differences are reflected mainly through a wide variety of distinctive and similar attributes. The key differences include bringing pets and the opportunity to encounter hosts' pets, atmosphere, flexibility, value for money, and quality assurance. Most of them are strongly attracted by practical attributes and somewhat less so by its experiential attributes [28]. P2P accommodation has unique motivators that are linked to the characteristics of human elements: environmental responsibility, community, and economic benefits [4]. Financial

benefits are an aspect as some people look at the sharing economy as a cheaper alternative to traditional accommodation options. This perspective has been countered by [4], who identified that motivation for P2P accommodation also comes from reasons beyond monetary factors, such as community. Community is a value that leans towards the idea of social relationships and the practice of sharing, openness, and collaboration. It reflects guests' desire for social interactions and reflects how guests interact with locals, experience local cultures, and indulge in local culinary. The community's value can also evoke the feeling of homeliness and create a home-like atmosphere [4]. Sustainability is a factor of sharing economy where it is believed that reductions of environmental harm can be achieved through the use of underutilized assets. By using P2P accommodation, we are not spending more resources but using assets that already exist [4].

Research about the Topics in P2P accommodation identified four critical topics; they are location, amenities, host, and recommendation [5]. The location's theme covers concepts such as geographical location to the point of interest, distance to nearby landmarks, and easiness of access to the accommodation. The location's convenience is essential, which describes the reach and ease of access to major tourist attractions, transportation hubs, and points of interest from the accommodation. The theme of amenity is a broad theme that also discusses the theme of facility and room. Amenity describes the availability of basic facilities such as towels, soap, and breakfast, which guests desire but may or may not be essential to the accommodation. Facilities deal with a broader accommodation category, such as the availability of a garden, pool, or balcony at the accommodation.

Meanwhile, the room's theme describes the environment inside the room, such as space, bed, room design, cleanliness, and other decoration. An essential attribute for a room is privacy and quality of sleep. The host's theme encompasses concepts that describe host's role in facilitating an Airbnb experience to guests [5]. The host is an important theme as the host has a central role in setting the guests' experience upon their visit. The last strong theme that emerged within reviews that guests made is a recommendation. While the recommendation itself is not an attribute that directly describes P2P accommodation, it is an outcome discussed by guests of P2P accommodation as a result of the other identified themes [5].

Security issues in accommodation, according to [3] refer to guests' safety during their stay. They defined two sides of security; the first is the guests' active contribution in creating a safe environment, concerning issues such as illegal drug use, identity theft, and other unlawful activities that guests may do during their stay. Second is the hosts' ability to create a safe environment at the accommodation, which includes maintaining the accommodation well, preparing safety equipment, and having safety precautions [3].

Positive and negative customer reviews

Guests often seek advice or review from other people to justify their purchase decision, where the review score and negative sentiment are tested significant [13]. As such, positive and negative word of mouth is proven to have a strong influence on consumer purchase decisions [29]. However, negative reviews are more authentic and credible than positive reviews on Airbnb. Social words' occurrence is positively related to positive emotion in reviews but negatively related to negative emotion in reviews [30].

P2P platforms such as Airbnb adopted customer reviews as an official feature that facilitates word of mouth within the platform. Since guests who have prior experience are encouraged to share their opinions, customer reviews also become a way for guests to express their satisfaction and dissatisfaction towards the accommodation they chose. These expressions are written in words but also quantified through the usage of the rating system. Hence, ratings and reviews in P2P accommodation also act as a proxy behind guests' satisfaction and dissatisfaction [29].

The reasoning behind how customers rate in online platforms can be explained by the research of [31]. They made an observation using the J-shaped distribution in ratings. They found that the majority of consumers who write customer reviews tend to write positive reviews to express their satisfaction in the form of a five or 4-star rating on a 5 point rating scale. Meanwhile, a noticeably smaller number of negative reviews express dissatisfaction in the form of a two or one-star rating on a 5 point rating scale. The reason behind this is because people with moderate and undesirable views are less passionate to exert the time and effort to report their ratings in comparison to people who have desirable and positive aspects on their experience. This phenomenon is not unique to Airbnb as the same pattern was found on a similar study conducted on customer reviews found on Amazon, Yelp, and TripAdvisor, where customer reviews are highly skewed towards the positive sides.

Prior research using customer reviews in the accommodation industry to draw meaningful insights on P2P accommodation are definite. The study found that the overall review star rating correlates pretty well with the sentiment scores for both the title and the full content of the online customer review [32]. [33] tried to present a case of text mining on Airbnb user reviews to analyze and understand various aspects that drive customer satisfaction. In the different industries, such as airlines, using a text mining approach on the Online Customer Reviews (OCRs) can predict airline recommendations by customers, resulting in an accuracy of 79.95% [34].

Several studies of online reviews have seen a consistent positivity bias in the writing of reviews [6]. They looked into factors that contribute to the positivity bias behind customer reviews. The first reason is guests' expectations being lower for individuals' accommodations in comparison to accommodations provided by hotels. Since Airbnb properties are provided by individuals who act as hosts, in addition to the lack of standards that Airbnb properties have, guests tend to be more realistic in their expectations over what they will receive from their Airbnb stays [35]. Second, there also tends to be a more personal and personalized interaction between the host and the guest. Often, guests will be communicating with the person who owns the property directly instead of talking to customer service staff as they do with hotels [36]. Furthermore, reviews in Airbnb are not anonymous, and each review is linked to the reviewers' profile. This personal experience and lack of anonymity lead to the behavior where posting negative feedback may be difficult and awkward for guests and may lead to guests not writing negative reviews when possible [35].

Lastly, a bias towards positive reviews can be driven by Airbnb's review guidelines. Although Airbnb doesn't directly edit and censor reviews, Airbnb reserves the right to remove personal insults, contain profanity, discrimination, and generally inappropriate and against their review guidelines. Airbnb review guidelines also encourage a typically

positive and constructive review, which possibly becomes a community-developed norm that Airbnb users abide and follow [35].

There are three main findings that the study by [6] discovered regarding Airbnb customer reviews. Firstly, customer reviews written are highly positive and frequently comment on the ease of communication and the accommodation's cleanliness. Second, a negative review is rare, and when it is found, it is written in a manner that is suppressed, and complaints are sandwiched in between positive comments. Finally, when guests do not feel like writing a positive review but avoid writing negative reviews, they chose to write lukewarm reviews where the content is positive but lacks the enthusiasm that usually comes with a genuine positive review.

Method

Data collection

Data for this research consists of guest reviews of Airbnb listings from the top 10 cities in Indonesia based on cities with the highest foreign arrivals according to statistical data issued by the Indonesian Central Agency of Statistics [37]. This study focuses on the top 10 cities as a proxy that mirrors Indonesia's popular destinations to foreigners. That also reflects the dominant amount of English reviews found in the dataset. The Foreign arrival and English use for the review reflect the geographical diversity of the data drawn. The research applied a web scraping method using direct HTTP request with Javascript to get user review data from Airbnb's website. We drawn a total of 66,630 reviews from 7356 properties listed on the Airbnb website from the ten cities. 55,377 out of the 66,630 reviews are written in English (83%) and have an average length of 42 words. In general, the steps of our study mentioned earlier and will be explained subsequently after this section can be visualized in the form of a flowchart as shown in Fig. 1.

Data analysis

A very limited previous research can be found in prediction rating using customer reviews. [38] used sentiment analysis from customer reviews to predict hotel ratings where they discovered that classified reviews as positive or negative are correlated positively with numerical ratings. Some other studies used customer reviews to predict the beaches' rating [39] and [40] urged that neuro-fuzzy can be utilized for sentiment analysis and review rating prediction tasks.

Another recent study in prediction related to peer to peer accommodation was done by [41], where they found that house popularity can be predicted more effectively using a Dual-Gated Recurrent Unit (DGRU). The predictive model will attempt to classify reviews as high rated or low rated based on its associated P2P accommodation attributes as the independent variable. In creating this model, several steps are taken. The first step is to classify the data as high rated or low rated. In accordance with our positive and negative review literature, we will classify reviews into two categories; high rating and low rating. High rating reviews are reviews that have four or 5-star ratings, and low rating reviews are reviews that have one or 2-star ratings [31]. In this model, reviews with a star rating 3 will be excluded. A new column called '*highrating*' is created and given the value of 1 if the review is categorized as a high rating and the value of 0 if the review is categorized as a low rating. This prediction aims to be able to classify reviews as either a high

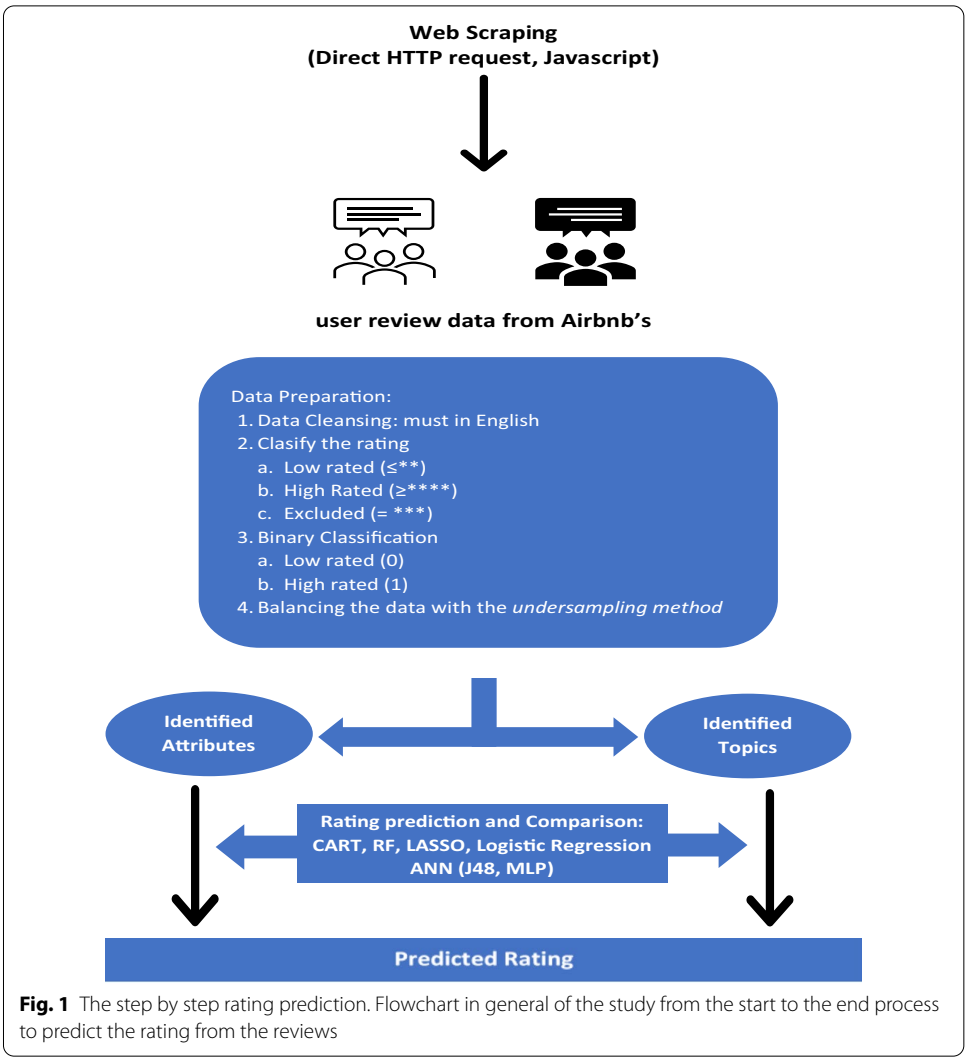

Fig. 1 The step by step rating prediction. Flowchart in general of the study from the start to the end process to predict the rating from the reviews

rating or low rating based on the binary value of 1 or 0 as registered in the binary column. Another pre-processing data judgment is to limit reviews to those written before the year 2020 and reviews with at least 5 characters or more. This step provides us with 544 negative reviews and 48,435 positive reviews to analyze.

The following pseudocode initiates the scripting environment, cleans the reviews dataset by removing rows with a rating of 0, rating of 3, the year 2020, and comments with length less or equal to 5, see Table 1.

Table 2 shows a pseudocode to creates a binary classification stored in high rating column. If row has rating of 5 or rating of 4, high rating is classified as 0. If row has rating of 2 or rating of 1, high rating is classified as 1.

To find the attributes, text-mining techniques are utilized to tokenize reviews and isolate relevant and impactful terms using pseudocode in Table 3 below.

Further processing is also conducted to prepare the tokenized datasets with attributes to match the required structure to perform predictive analysis. That process is executed using pseudocode available in Table 4.

Table 1 Pseudocode to initiates the scripting environment, cleans the reviews dataset

```
Start
READ datasets.csv
INIT reviews as datasets.csv
SET year in data with format "%Y"
FOR every row in reviews
        IF rating > 0 THEN
                DELETE row
        IF rating != 3 THEN
                DELETE row
        IF year = 2020 THEN
                DELETE row
        FUNCTION nchar with argument comment
                Pass In: comment
                Count the length of comment in row
                Pass Out: length of comment
        END FUNCTION
        If length of comment <= 5 THEN
                DELETE row
ENDIF
```

Table 2 Pseudocode to creates a binary classification

```
INIT high rating in reviews
FOR every row in reviews DO
        IF rating = 5 or rating = 4 THEN
                SET high rating as 0
        IF rating = 2 or rating = 1 THEN
                SET high rating as 1
ENDIF
```

Table 3 Pseudocode to tokenized and pre-process tokens from reviews dataset

```
INIT Quanteda package
INIT tokenized_reviews to store tokenized comments from reviews dataset
FUNCTION tokens from Quanteda package
        Pass In: comment reviews
        Tokenize each comment in the dataset reviews
        Pass Out: tokenized reviews
SET tokenized_reviews as tokenized reviews from tokens function
FUNCTION tokens_tolower from Quanteda package
        Pass In: tokenized_reviews
        Lowercase all tokens in tokenized_reviews
        Pass Out: Lower cased reviews
SET tokenized_reviews as lower-cased tokens
FUNCTION tokens_select from Quanteda package
        Pass In: tokenized_reviews and stop words
        Remove stop words from tokens
        Pass Out: reviews with stop words removed
SET tokenized_reviews as tokens with stop words removed
FUNCTION tokens_ngrams from Quanteda package
        Pass In: tokenized_reviews and integer 2 representing n of ngrams
        Create ngrams of 2 words in reviews
        Pass Out: reviews with ngrams of 2
SET tokenized_reviews as tokens with ngams
```

Due to the overwhelming amount of positive reviews in comparison to negative reviews, using the current dataset will inevitably lead to a biased prediction. This is a common problem in datasets known as imbalances in the dataset [42]. The method

Table 4 Pseudocode to create Document Frequency Matrix (DFM) and data frame from the tokens

```
INIT tm package
FUNCTION dfm from Quanteda package
        Pass In: tokenized_reviews
        Create document frequency matrix from tokens
        Pass Out: document frequency matrix
FUNCTION removeSparseTerms from tm package
        Pass In: document frequency matrix
        Remove sparse terms within document frequency matrix
        Pass Out: document frequency matrix with sparse terms removed
FUNCTION convert to data frame
        Pass In: document frequency matrix
        Convert document frequency matrix as data frame
        Pass Out: data frame
FUNCTION cbind to bind two datasets
        Pass In: high rating and data frame
        Concatenate high rating to each row in data frame
        Pass Out: data frame with high rating
```

used to balance the data in this analysis is the undersampling method. In the undersampling method, members of the majority dataset will be eliminated until the data set gets balanced based on a pre-specified selection criterion, see Table 5.

The next step, we employed (CART Model), Random Forest Model, (LASSO), and Logistic regression to predict the rating classification based on the identified attributes using pseudocode in Tables 6, 7, and 8 as well as in Table 9 below accordingly.

To test the predictive model, see Table 10, the study creates training model with random forest algorithm to perform prediction on the prepared testing dataset.

Finally, for comparison, we will also do another prediction using Artificial Neural Network (ANN) since it can produce a better result in prediction purposes with big data as suggested by [43] via the RWeka package, see Table 11.

Table 5 Pseudocode to balance the dataset between majority binary class and minority binary class using the undersampling method

```
INIT unbalanced package
INIT reviews_balance to store results from balancing reviews
FUNCTION ubUnder from unbalanced package
        Pass In: data frame
        Use undersampling method to balance majority high rating class with minority high rating class
in reviews data frame
        Pass Out: data frame balanced with undersampling method
SET reviews_balance as balanced data frame
```

Table 6 Pseudocode to execute the CART Model and plot decision tree

```
INIT rpart package
INIT tree to store CART decision tree
FUNCTION rpart
        Pass In: reviews_balance
        Pass Out: decision tree
SET tree as decision tree
SET complexity parameter as best complexity parameter from tree
FUNCTION plot
        Pass In: tree and complexity parameter
        Plot decision tree with data from tree using tree size of complexity parameter
        Pass Out: Plotted decision tree
```

Table 7 Pseudocode to execute the Random Forest Model and plot essential variables based on Random Forest algorithm

```
INIT randomForest package
INIT random_forest to store Random Forest results
FUNCTION randomForest
        Pass In: reviews_balance
        Pass Out: random forest model from reviews dataset
SET random_forest as result from random forest model
FUNCTION varImpPlot
        Pass In: random_forest
        Plot random forest model to show important variables
        Pass Out: nothing
```

Table 8 Pseudocode to execute LASSO Regression to find Minimum Lambda to be used in Logistic Regression

```
INIT glmnet package
INIT matrix to store converted training data as matrix format
INIT class to store converted classification as numerical variable
INIT lasso_regression to store results from grid search to find optimal value of lambda
FUNCTION model.matrix
        Pass In: reviews_balance
        Pass Out: model matrix
FUNCTION as.numeric
        Pass In: high rating in reviews_balance
        Pass Out: numerical class based on high rating
FUNCTION cv.glmnet
        Pass In: matrix and class
        Execute grid search for optimal value of lambda
        Pass Out: lasso_regression
FUNCTION plot
        Pass In: lasso_regression
        Plot grid to show all possible values of lambda and position of optimal value of lambda
        Pass Out: nothing
SET lambda_min as optimal value of lambda from lasso_regression
```

Table 9 Pseudocode to execute Logistic Regression to find Odds Ratio and Logistic Regression Coefficient for each variable

```
INIT coefficient
INIT odds ratio
FUNCTION glmnet
        Pass In: matrix, class, and lambda_min
        Create logistic regression model from matrix, numeric, and optimal value of lambda
        Pass Out: logistic_regression
FUNCTION coef
        Pass In: logistic_regression
        Create coefficient for each variable based on the logistic regression model
        Pass Out: Data frame for logistic regression coefficient
FUNCTION exp
        Pass In: logistic_regression
        Create odds ratio for each variable based on the logistic regression model
        Pass Out: data frame for logistic regression odds ratio
PRINT coefficient
PRINT odds ratio
```

Results

Using attributes to predict rating classification in customer review

This study wants to see what customer behavior can be predicted with the identified

Table 10 Pseudocode to train, test and execute prediction based on model trained

```
INIT caret package
INIT dataset from reviews
INIT training as 80% of dataset
INIT testing as 20% of dataset
FOR every row in testing
        DELETE high rating
INIT caret
INIT fit control
FUNCTION trainControl
        Pass In: "repeatedcv" as train control method with 10 folds repeat
        Pass Out: Training Control
SET fit control as training control
INIT model
FUNCTION train
        Pass In: training with method "random forest" and training control with settings in fit control
        Create model with training dataset
        Pass Out: trained model
SET model as trained model
FUNCTION predict
        Pass In: testing dataset and model
        Execute prediction on testing dataset using model
        Pass Out: predictions
PRINT predictions
```

topics and attributes from Airbnb customer reviews. As we found in the analysis of positive and negative reviews, customer ratings act as a proxy that reflects customers' satisfaction with the service (Bridges & Vasquez, 2018). Attempting to predict customer ratings may provide a glimpse into predicting customer satisfaction. Hence, we will look into predicting guests' behavior in the rating classification of reviews. A binomial classification predictive model is created to see how the existence of certain attributes or topics can increase the odds of a review in Airbnb to be categorized as a high rating or low rating. We will try to predict whether or not an Airbnb user will give a high or low rating in their review. The methods used are the classification and regression tree (CART), random forest, and least absolute shrinkage and selection operator (LASSO) logistic regression based on the bag of words text-mining approach [44].

The first analysis is using the CART algorithm to create a classification. In this analysis, a complete decision tree is built, and the optimum complexity parameter (*CP*) minimum error is identified [44]. Each *CP* relative to the size of the tree and its relative error are plotted (see Fig. 1). In creating this tree, the column 'highrating', which has the classification of the high and low rating of a review, will be the dependent variable. Meanwhile, the remaining attributes will become the independent variable. Note that in this model, 'highrating' is binary, while the independent variable is continuous and takes the frequency of the attribute as its value.

We can visualize the decision tree, which is pruned based on the identified *CP* value. From Fig. 2 above, we obtained 0.00096 as the most optimum *CP*, and using this *CP* value; we will be able to identify at most a decision tree with the size of 12 branches. The result of the decision tree can be seen in Fig. 3.

Another classification that we will use to extend the analysis is by using the random forest algorithm. By inputting the dataset into the randomForest package on R, we are able to obtain the plotted Random Forest Model (see Fig. 4).

Table 11 Pseudocode to create training and testing dataset using J48 Classifier and MLP Classifier

```
INIT RWeka package
INIT dataset from reviews
INIT training as 80% of dataset
INIT testing as 20% of dataset

FUNCTION J48
        Pass In: training dataset
        Create J48 Classifier using training dataset
        Pass Out: J48 Classifier Model
SET J48_Model as classifier model output from J48 function
FUNCTION predicts
        Pass In: testing dataset and J48_Model
        Execute prediction using J48 Classifier Model
        Pass Out: predictions
FUNCTION confusionMatrix
        Pass In: J48 predictions
        Evaluate fitness of J48 Classifier and prediction accuracy
        Pass Out: J48 prediction results
PRINT J48 prediction results
SET MLP to store MLP Classifier
FUNCTION make_Weka_classifier
        Pass In: Path "weka/classifiers/functions/MultilayerPerceptron"
        Loads MutilayerPerceptron to load Weka's MLP model
        Pass Out: MLP Classifier
SET MLP as MLP Classifier
SET MLP_Model
FUNCTION MLP
        Pass In: training dataset
        Create MLP Model using training dataset
        Pass Out: MLP Model
SET MLP_Model as MLP Model
FUNCTION predict
        Pass In: training dataset and MLP_Model
        Run prediction on training dataset based on the provided MLP Model
        Pass Out: predictions
FUNCTION confusionMatrix
        Pass In: MLP predictions
        Evaluate fitness of MLP model and prediction accuracy
        Pass Out: MLP prediction results
PRINT MLP prediction results
```

To complete our prediction on attributes, the LASSO logistic regression algorithm for final classification using the *glmnet* package. From the LASSO logistic regression, we obtained the odds ratio of the logistic regression model for each attribute that we use as an independent variable [44]. In doing this, the first step is to find the optimal value of lambda, plotted as below (see Fig. 5).

We first conduct a LASSO regression analysis, where we obtained two values from this plot, which is the optimal value of lambda that we will use for the logistic regression. The output is two distinct datasets that give us further insights into this classification. First is obtaining an odds ratio for each variable, which indicates how the existence of a variable can increase the odds of review to be classified as a high rating. The other output is the logreg coefficient, which provides a negative coefficient and will indicate if the presence of the term can lead to the probability of a review to be classified as high rated to become low [44]. The result of this prediction is as shown in Table 12.

Using RWeka package and performing ANN techniques, we can also use the identified attributes to predict whether reviews will be classified as low ratings or high ratings.

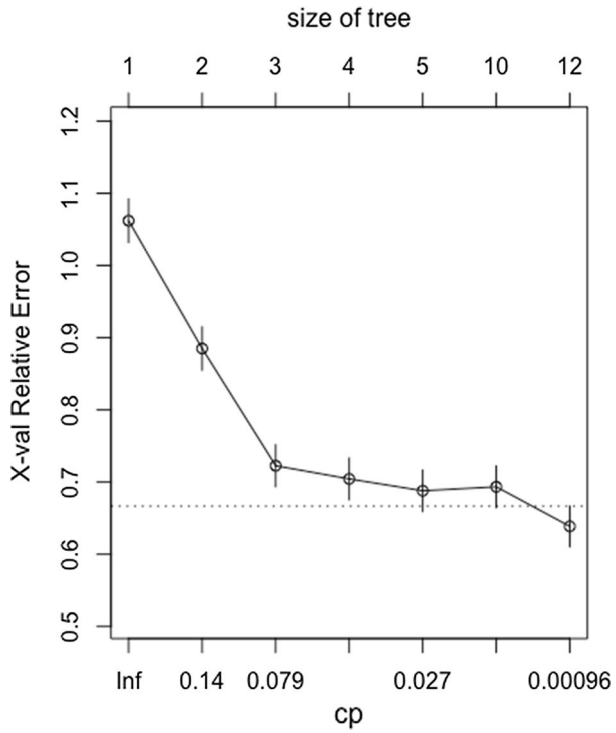

Fig. 2 Complexity Parameters for Attributes Prediction. The optimum number of branches can be determined by the minimum error and the optimum *CP*

First, using the same dataset, we partition the data into the training dataset and testing dataset in an 80% to 20% ratio using the createDataPartition function in R for fairness. We load the J48 Classifier from RWeka and create a prediction model using the Classifier and our training dataset. After we obtain the prediction model, we use it to predict against our training dataset. The result is as follows (Fig. 6).

Next, maintaining the same method, we made another prediction, but this time using the Multilayer Perceptron model or MLP for short. We use the same training and testing dataset for fairness in comparison. The result is as follows (Fig. 7).

As seen in the results above, the J48 Classifier prediction has 84.79% accuracy, while the MLP prediction has 72.35% accuracy. The results can be summarized in Table 13.

Both predictions boast a high and satisfactory accuracy level, which validates the idea that attributes can be used to predict customer behavior and satisfaction as expressed through their intention in classifying ratings. This encourages the notion of investigating attributes further as a means to understand the behavior and satisfaction of consumers in P2P accommodation.

Using topics to predict rating classification in customer review

Another analysis that we will conduct is to make the same rating classifications but with topics as the variable. This analysis used the same dataset as before, but with an extra step in the pre-processing data stage. The goal is to combine related features from the document features matrix and group its frequency based on the topic that the feature is

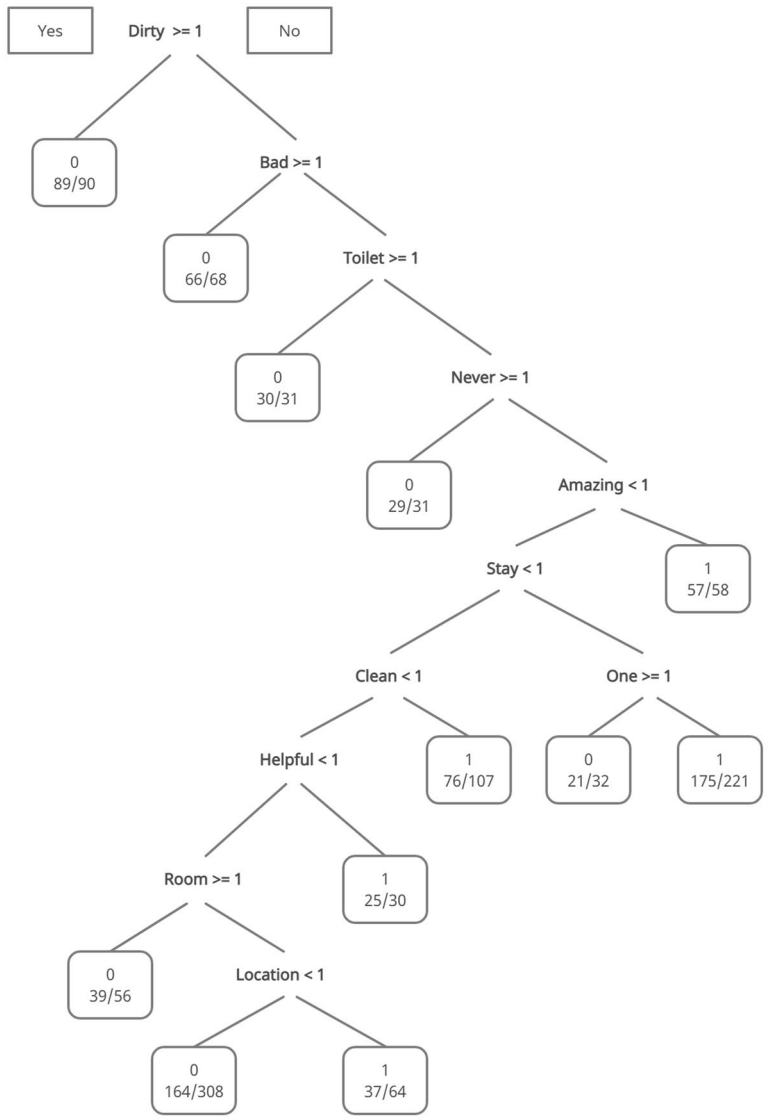

Fig. 3 Decision Tree for Attributes Prediction. This figure shows words such as 'dirty,''bad,''toilet,''never' are plotted, which are used by unhappy customers

associated with, and the result can be seen in Table 14. The list of topics and attributes associated with each topic is based on research done by (3–5).

Based on the topic distribution list, iteration can be done through the document term features matrix (DTM) for the dataset and used to find the cumulative frequency for each topic based on its associated attributes. The result is an aggregated document features matrix, which can be seen in Fig. 8.

Provided with that document features matrix (see Fig. 8), we apply the same steps and algorithm from the classification of high or low rating using the CART, Random Forest Model, and LASSO Logistic Regression. Similarly, we first built a complete decision tree and calculated the optimal *CP* value relative to the size of the tree and its relative error, see Fig. 9.

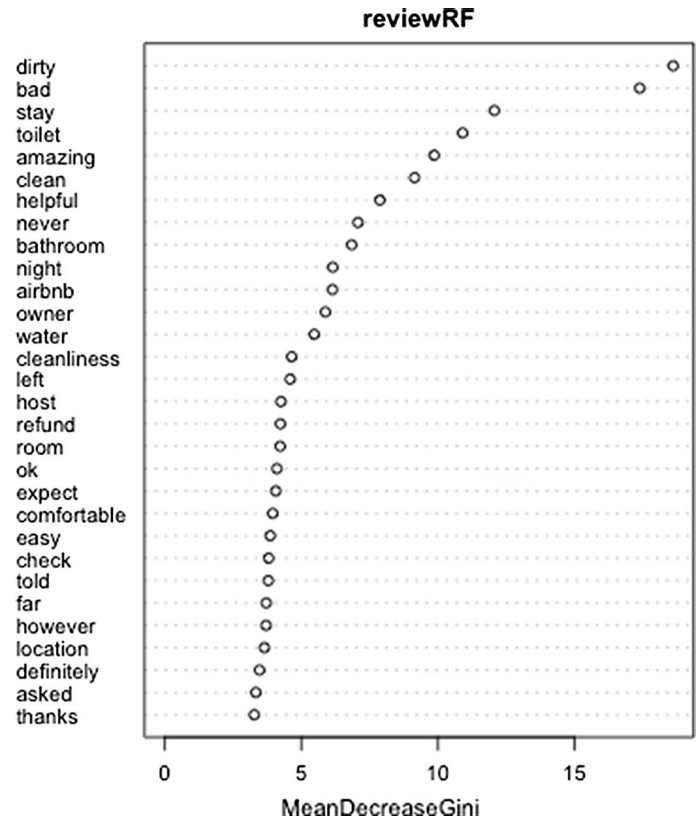

Fig. 4 Random Forest Variable Importance for Attributes Prediction. This Figure identifies 'dirty,' 'bad,' 'toilet,' and 'never' as top four most important variables

Based on the *CP* value, we visualize the decision tree for our topics, as can be seen in Fig. 10. The visualization of the decision tree shows importance being placed on the topic of the facility, room, location, and price in deciding whether a review can be classified as a high or low rating.

After the CART Model analysis, we conduct the Random Forest Model analysis as well. The output of the Random Forest analysis is as can be seen in Fig. 11.

Finally, to create our prediction, we conduct the LASSO logistic regression algorithm. Similarly, we plot our MSE values to find the optimal lambda_min that we need for our logistic regression [44], see Fig. 12.

The output of our logistic regression provides us with two data outputs, as can be seen in Table 15.

Finally, to confirm our prediction, a training model is created with the caret package. Using our aggregated topic document features matrix, we create a training model using the random forest as its method [45, 46]. Here we have the 'highrating' category as its dependent variable, which has a binary value, and the topic of location, room, amenity, host, recommendation, facility, price, security, and community as the independent variable, which has a frequency as its value. We set aside 80% of our dataset as a training set and 20% of our dataset as a testing set. This results in 870

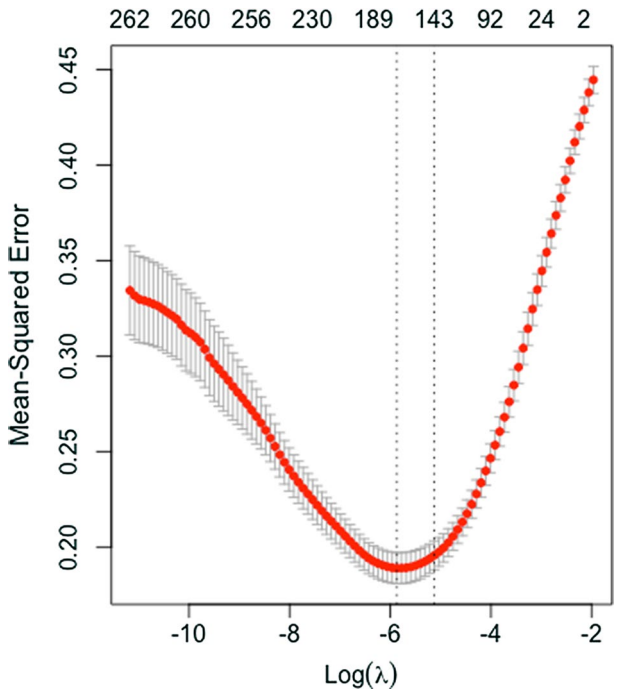

Fig. 5 Mean Squared Error for Attributes Prediction. This figure obtained two optimal values of lamda that leads to two distinct datasets

Table 12 Odds ratio and logreg coefficient for attributes prediction

Odds ratio		Logreg coefficient	
Variable	**Odds ratio**	**Variable**	**Coef**
Exactly	6.64	Dirty	−2.38
Comfy	5.08	Broken	−1.82
Amazing	4.95	Left	−1.78
Quick	4.42	Never	−1.76
Definitely	4.18	Bad	−1.75
Love	3.66	Smell	−1.72
Friendly	3.61	Owner	−1.68
Perfect	3.56	Book	−1.61
Hospitality	3.53	Ask	−1.61
Space	3.51	Disappointed	−1.59

training data and 218 test data. The result of the training model and its prediction is in Table 16.

From our testing data set, we are able to correctly predict 131 datasets from 218 data, which results in 60.09% prediction accuracy. Several examples of the prediction results are shown in Table 17.

For comparison, we also conduct prediction with the same dataset on two other prediction models: J48 Classifier and MLP, which are artificial neural network (ANN)

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  70  12
         1  21 114

               Accuracy : 0.8479
                 95% CI : (0.7931, 0.8929)
    No Information Rate : 0.5806
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6834

 Mcnemar's Test P-Value : 0.1637

            Sensitivity : 0.7692
            Specificity : 0.9048
         Pos Pred Value : 0.8537
         Neg Pred Value : 0.8444
             Prevalence : 0.4194
         Detection Rate : 0.3226
   Detection Prevalence : 0.3779
      Balanced Accuracy : 0.8370

       'Positive' Class : 0
```

Fig. 6 J48 Classifier Prediction Results for Attributes. This figure shows the 84,8% accuracy of the prediction using J48 Classifier Prediction

prediction techniques available through the RWeka package. From the J48 Classifier technique, we obtain the following result, see Fig. 13.

The J48 Classifier has 58.33% accuracy in its prediction. Meanwhile, for the MLP prediction, we obtain the following result, see Fig. 14.

The MLP prediction model has 58.8% accuracy from its prediction. The results from the three prediction methods can be summarized in the following Table 18.

As summarized, the results from the three-prediction model hover around the same range of accuracy. While the prediction result is satisfactory, the prediction result is relatively low, which indicates that the training model can be improved by including more data. The immediate area of improvement is to explore a more representative balancing method other than the undersampling method used in this prediction model. Several methods that can be considered include the oversampling method, SMOTE method, and CSL method [42]. Another area of improvement that is more challenging but will significantly enrich the dataset is to obtain more negative reviews from Airbnb, which will give a better positive to negative review ratio and allow analysis with a more significant scope of data.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 87 56
         1  4 70

               Accuracy : 0.7235
                 95% CI : (0.6589, 0.7819)
    No Information Rate : 0.5806
    P-Value [Acc > NIR] : 8.912e-06

                  Kappa : 0.474

 Mcnemar's Test P-Value : 4.577e-11

            Sensitivity : 0.9560
            Specificity : 0.5556
         Pos Pred Value : 0.6084
         Neg Pred Value : 0.9459
             Prevalence : 0.4194
         Detection Rate : 0.4009
   Detection Prevalence : 0.6590
      Balanced Accuracy : 0.7558

       'Positive' Class : 0
```

Fig. 7 MLP Prediction Results for Attributes. This figure shows the 72,4% accuracy of the prediction using MLP

Table 13 Prediction model summary for attributes prediction

Prediction model	Accuracy (%)
J48	84.79
MLP	72.35

One interesting comparison to make is comparing the accuracy results between using topics as the predictor variable and attributes as the other predictor variable. The comparison is shown in the following Table 19.

As clearly shown above, the accuracy of prediction while using attributes is significantly higher. This can be due to using attributes as the predictors to provide the prediction model with more information than using topics. This could be the case as topic information is obtained from aggregating attributes, and in retrospect, this may dilute through the process. Meanwhile, attributes information is kept at its raw form and directly used to train the prediction model. That gives us a hint towards attributes being a more substantial data point in understanding customer behavior in P2P accommodation.

Table 14 Topic distribution and attributes example

Topic	Attributes	Topic	Attributes	Topic	Attributes
Location	Location	Facility	Villa	Price	Price
	Walk		Room		Money
	Amaze		Apartment		Free
	Close		House		Worth
	Night		Pool		Spend
Amenity	Soap	Room	Clean	Recommendation	Recommend
	Snack		Comfortable		Perfect
	Utensil		Bed		Highly
	Pancake		Big		Lot
	Juice		Spacious		Wonderful
Host	Host	Security	Safe	Community	Love
	Time		Security		Family
	Helpful		Privacy		Lovely
	Staff		Secure		Enjoy
	Friendly		Guard		Feel

```
# A tibble: 6 x 12
  highrating rating document location amenity facility  room price recommendation  host security
       <dbl>  <dbl> <chr>       <dbl>   <dbl>    <dbl> <dbl> <dbl>          <dbl> <dbl>    <dbl>
1          0      2 text1           0       0        2     0     0              0     1        0
2          0      2 text2           2       0        1     0     0              2     5        1
3          0      2 text3           3       0        6     2     0              3     2        0
4          0      1 text4           3       0        3     1     0              0     2        2
5          0      1 text5           1       0        1     4     0              2     2        0
6          0      2 text6           5       0        7     9     0              2     6        0
# … with 1 more variable: community <dbl>
```

Fig. 8 Topic Aggregated Document Features Matrix. This figure shows the cumulative frequency for each topic based on its associated attributes

Discussion

This study has used a series of methods for predicting and understanding consumers' behavior in rating reviews based on P2P accommodation attributes and topics. The CART model for attributes in customer reviews shows words such as 'dirty,' 'bad,' 'toilet,' 'never' are plotted, which are used by unhappy customers, and those who would give a low rating for their review; see Fig. 3. Based on our understanding of these attributes from previous analysis, we can imply that these words most likely point to customers' complaints on cleanliness, especially in the bathroom and toilet area. Furthermore, these selections of words also imply that customers may not want to revisit the accommodation, implied from the word 'never'.

We also find that the classification resulted from the Random Forest Model (can be seen in Fig. 4) is inline with the CART model and also identified 'dirty,' 'bad,' 'toilet,' and 'never' as important variables. The Random forest model also identified several other words related to negative customer perception, such as 'refund'. On the other hand, the majority of the variables identified by the Random forest model lean towards a positive description of the accommodation, such as from the word 'stay,' 'amazing,' and 'helpful.'

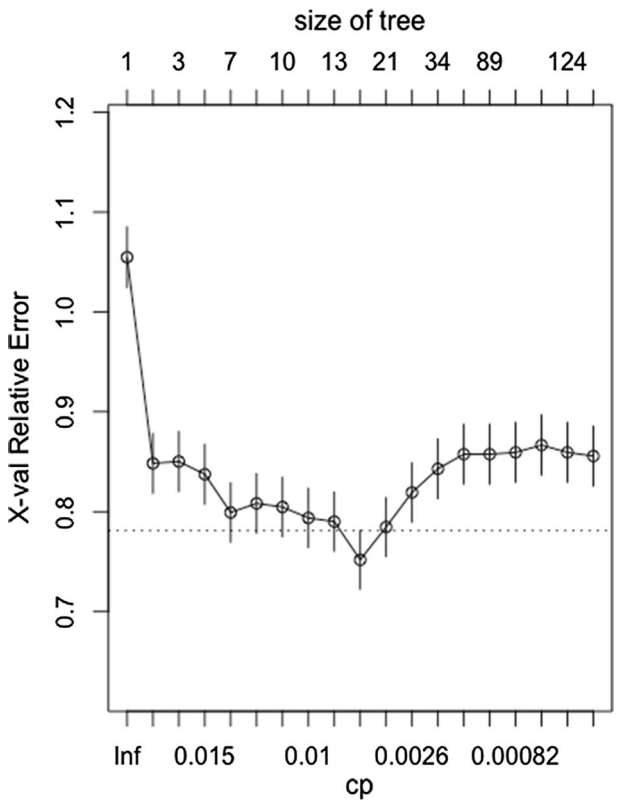

Fig. 9 Complexity Parameter for Topics Prediction. The optimum number of branches can be determined by the optimal *CP* value relative to the size of the tree and its relative error

Finally, the result of the logistic regression for attributes (see Table 12) can be interpreted as follow. Looking at the odds ratio output, we see that when the word 'exactly' is present in a review, there are 6.64 times more odds that the review will have a high rating. The same interpretation can be applied for the variable 'comfy,' 'amazing,' 'quick,' and so on where the presence of each variable can provide a review with 5.08, 4.95, and 4.42 odds, respectively, for the review to be rated high.

Meanwhile, looking at the logreg coefficient table (see Table 12), we can see which variable may reduce the probability of a customer review to have a high rating. The variable with the greatest impact for a review to be rated low is 'dirty,' and this conforms with classification from the CART and Random Forest Model. Other variables identified include 'broken,' 'left,' 'never,' 'bad,' 'smell,' 'owner,' 'book,' 'ask,' and 'disappointed' which are all words that we have identified being used in negative reviews as described on our previous analysis.

When looking at topics, we see a similar pattern between the CART model (Fig. 10) and the Random Forest Model (Fig. 11), where we identified facility, room, and location as an important variable. There is one difference as the random forest model didn't emphasize 'Price' as a variable. The Random Forest Model output, however, greatly validates and confirms the TF-IDF ranks for each topic that we obtain from our topic distribution analysis in Table 14. Similar to the topic distribution analysis, the topic of the facility, room, host, location, and recommendation are greatly more dominant compared

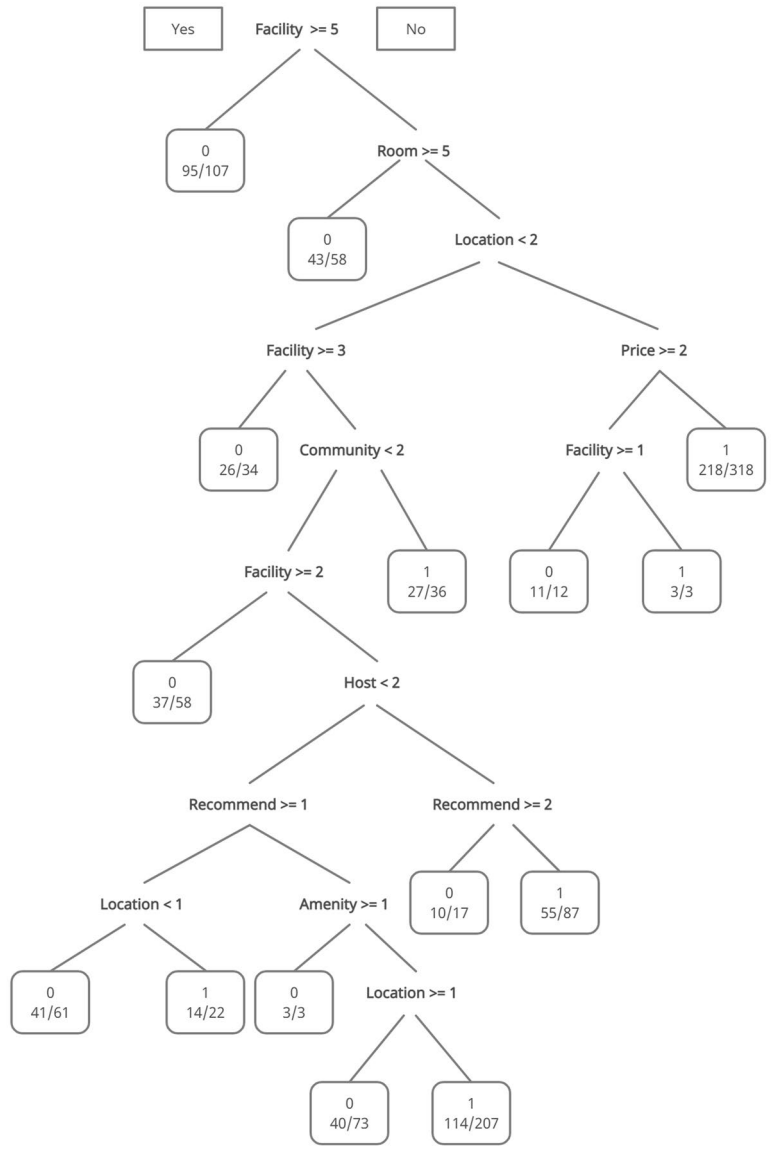

Fig. 10 Decision Tree for Topics Prediction. This Figure shows importance being placed on the topic i.e. facility, room, location, and price in deciding whether a review can be classified as a high or low rating

to the topic of community, price, security, and amenity, which implies greater attention required to be placed on those topics.

The result of the LASSO logistic regression for topics gives us exciting results, see Table 15. First, we look into the odds ratio output for each topic. The inclusion of the topic of location in a review can increase the odds of the review to be highly rated by 1.20 times. That matches our previous topic distribution analysis, where the topic of the location has the highest TF-IDF value. Surprisingly, the topic of community also has the chance to increase the odds of a review to be rated highly by 1.04 times. This contrasts with our previous analysis, where the usage of attributes related to the community is limited and not as widely discussed among reviews.

Fig. 11 Random Forest Variable Importance for Topics Prediction. This figure shows the top importance words i.e. facility, room, host, and location

As we look at the logreg coefficient (see Table 15), we can see the topic where most customer complaints lay. The topic of amenity has the highest coefficient and shows that complaints regarding amenity are the greatest factor that can lead to dissatisfied customers and hence, low rating review. This confirms the understanding of negative reviews for an amenity that shows the importance customers placed in having a clean and well-maintained supply of amenity. This also conforms with the previous attributes prediction where the variable 'dirty' is the most impactful variable in classifying a review as low rated.

The prediction also shows that attributes contain richer information in creating prediction in comparison to topics. Prediction summary results (Table 19) indicated that attributes have an accuracy of 72–85% while prediction results with topics have accuracy at the range of 60%. This posed the argument of P2P accommodation attributes as an excellent avenue in studying consumer behavior and satisfaction in P2P accommodation.

This prediction gives us more understanding of how each attribute influences users' intention to give their reviews a particular rating. As mentioned before, ratings can act as a proxy that reflects customers' satisfaction and dissatisfaction [6]. From this prediction

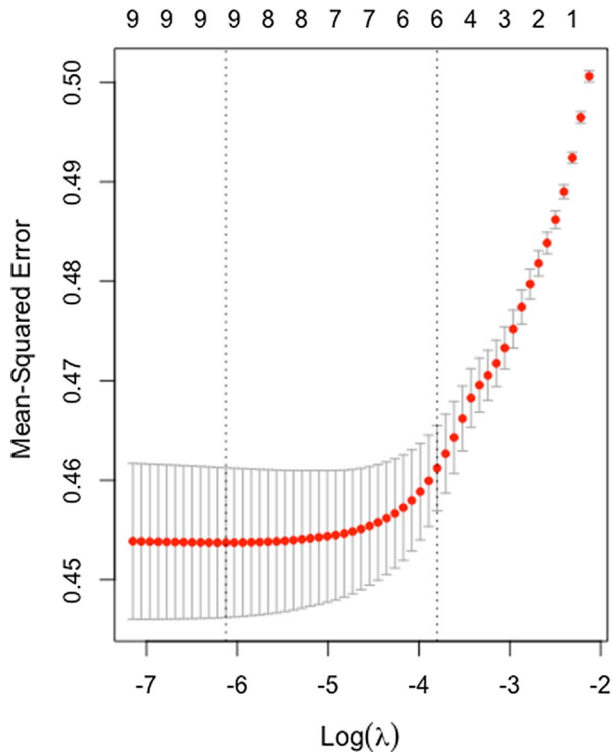

Fig. 12 Mean Squared Error for Topics Prediction. This Figure plots the MSE values to find the optimal lambda that is needed for the logistic regression

Table 15 Odds ratio and logreg coefficient for topics prediction

Odds ratio		Logreg coefficient	
Variable	**Odds ratio**	**Variable**	**Coef**
Location	1.20	Amenity	−0.82
Community	1.04	Price	−0.53
Host	1.00	Security	−0.26
Recommendation	0.91	Facility	−0.25
Room	0.89	Room	−0.11
Facility	0.78	Recommendation	−0.10
Security	0.77	Host	0.00
Price	0.59	Community	0.04
Amenity	0.44	Location	0.18

Table 16 Topics testing set prediction result

Dataset	1088
Training set	870 (80%)
Testing set	218 (20%)
Correct prediction	131
Prediction accuracy	60.09%

Table 17 Testing set prediction document example

Document	Original	Prediction	Result
Text1	0	0	Correct
Text2	1	0	Incorrect
Text3	1	0	Incorrect
Text4	0	0	Correct
Text5	0	1	Incorrect

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 73 55
         1 35 53

               Accuracy : 0.5833
                 95% CI : (0.5145, 0.6498)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.008522

                  Kappa : 0.1667

 Mcnemar's Test P-Value : 0.045201

            Sensitivity : 0.6759
            Specificity : 0.4907
         Pos Pred Value : 0.5703
         Neg Pred Value : 0.6023
             Prevalence : 0.5000
         Detection Rate : 0.3380
   Detection Prevalence : 0.5926
      Balanced Accuracy : 0.5833

       'Positive' Class : 0
```

Fig. 13 J48 Classifier Prediction Results for Topics. This figure shows the 58.3% accuracy of the Topics prediction using J48 Classifier Prediction

study, we can look into which attributes and topics can increase the odds of high rating reviews, which implies increases in customer satisfaction.

Conclusion and implication for further research

In this research, we have created and attempted to make a prediction using our dataset, thus acknowledged the potential and opportunity for a more extensive prediction model with the dataset available and data points identified from the analysis of topics and attributes in this research. This research will propose several predictive model areas that can be pursued in future research.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 65 46
         1 43 62

               Accuracy : 0.588
                 95% CI : (0.5192, 0.6543)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.005822

                  Kappa : 0.1759

 Mcnemar's Test P-Value : 0.832107

            Sensitivity : 0.6019
            Specificity : 0.5741
         Pos Pred Value : 0.5856
         Neg Pred Value : 0.5905
             Prevalence : 0.5000
         Detection Rate : 0.3009
   Detection Prevalence : 0.5139
      Balanced Accuracy : 0.5880

       'Positive' Class : 0
```

Fig. 14 MLP Prediction Results for Topics. This figure shows the 58.8% accuracy of the prediction using MLP

Table 18 Prediction model summary for topics prediction

Prediction model	Accuracy (%)
RandomForest	60.09
J48	58.33
MLP	58.8

Table 19 Prediction model summary for topics and attributes prediction

Predictor	Prediction model	Accuracy (%)
Topics	RandomForest	60.09
Topics	J48	58.33
Topics	MLP	58.8
Attributes	J48	84.79
Attributes	MLP	72.35

Firstly, the prediction model that will be of great interest for future research, is in accordance with behavioral intention. In the topics and attributes analysis, this research identified common occurrences where users show behavior such as repurchase intention

from their statement to revisit the same accommodation, and word of mouth from recommendation to potential guests. The opposite behavior is also shown when customers write warnings and discouragement future guests to choose a particular accommodation. Future studies can quantify these behaviors and use them as Independent Variables to predict Airbnb's customer behavior in accordance with the behavioral intention theory.

Secondly, the prediction model is to create predictions based on relationships between guests and hosts, creating bias in review reporting [47]. From analyzing the available data in the dataset, we notice that relationships between hosts and guests can be mapped since each review consists of data on who the reviewer is, who the reviewee is, and on which accommodation the review is given. Equipped with knowledge and understanding of topics and attributes of P2P accommodation from this research, the relationship between hosts and guests can be mapped, and their behavior and perception towards certain topics or attributes can be predicted. For example, a relationship can be mapped to see the number of times guests choose accommodation for a specific set of features such as the distance and proximity towards a point of interest or the completeness of amenities. Based on this relationship, a prediction model can be created to predict which attributes will encourage guests to choose accommodation.

Thirdly, the prediction model is to create a training set that can predict ratings of accommodation. Using topics and attributes identified in this research, along with other data available from the dataset, it is possible to construct a predictive model to predict ratings precisely.

Other than the above proposal for the prediction model, there are several areas of improvement for the predictive analysis's technical conduct that can be adopted for future research. The first area of improvement is the issue of an imbalanced dataset and the approach to balancing the data. The phenomenon of a positivity bias in customer review is well known, and imbalanced data is an inevitable characteristic that datasets such as Airbnb's will have. The solution to this issue is to balance the dataset through sampling methods. There are several popular methods of sampling to balance datasets such as undersampling, oversampling, synthetic minority oversampling technique (SMOTE), and cost-sensitive learning (CSL) [42]. This research currently uses the undersampling method; however, it is recommended for a better representation of the data to consider both undersampling and oversampling methods since each has pros and cons. Furthermore, SMOTE and CSL can be an advanced sampling method that can represent the dataset better.

The second area of improvement is exploring and using other predictive models for analysis other than CART, random forest, and LASSO logistic regression method. The three methods are chosen for the current research due to the independent variable being binary and the prediction to be a form of classification. However, to conduct predictive models for scenarios such as predicting rating, other methods will be required. Other methods to consider are Multinomial Logistic Regression, Ordinal Logistic Regression, and Multinomial Naïve Bayes as potential prediction methods.

The third area of improvement is to use the semantic parsing approach. The current prediction analysis in this research uses the bag of words approach where all words are analyzed as a single token, and the order of the token does not matter. A more extensive

study can utilize semantic parsing where the model will also consider word sequence and word usages such as nouns, verbs, and hierarchical word structure.

Regardless of future research possibilities, this research has proved that P2P accommodation attributes can be used to predict customer behavior in classifying ratings for their review. This act of classification serves as a proxy in predicting customer satisfaction based on how they describe their stay experience. Furthermore, this research also identifies which set of attributes and topics leads to increased high rated reviews (satisfaction). Conversely, another set of attributes to avoid can lead to an increase in low rated reviews (dissatisfaction). These insights contribute to the industry of P2P accommodation as a model to understand which attributes of P2P accommodation must be optimized to drive higher satisfaction from customers. Nevertheless, P2P accommodation attributes and prediction models are an extensive and in-depth research area where possibilities of approaches are limitless.

Abbreviations

ANN: Artificial neural network; CART: Classification and regression tree; CP: Complexity parameter; CSL: Cost-sensitive learning; DGRU: Dual-gated recurrent unit; DTM: Document Term feature Matrix; LASSO: Least Absolute Shrinkage and Selection Operator; MLP: Multi-layer preceptron; MSE: Mean square error; OCRs: Online customer reviews; P2P: Peer to peer; SMOTE: Synthetic minority oversampling technique.

Acknowledgements

Not aplicapble

Authors' contributions

All the authors discussed and contributed to the writing of the paper. Research idea and flow, literature review, interpretation of the model result is supplied as well as the article finishing by AS. Most of the paper technicality is done by MC. Both authors read and approved the final manuscript.

Authors' information

Athor Subroto (AS) is a Senior Lecturer in the Department of Management, Faculty of Economics and Business, Universitas Indonesia. His research interests are in the area of Management Science and System Dynamics applied to policy engineering for public and private entities.

Marcel Christianis (MC). He is now working as Independent Researcher and earned a Master degree in the field of Management from Universitas Indonesia. His research interests include Computer Science, Big Data in predictive analytics.

Funding

This research is supported by Universitas Indonesia under contract: NKB-1452/UN2.RST/HKP.05.00/2020.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interest.

Author details

[1] Department of Management, Faculty of Economics and Business, Universitas Indonesia, Jakarta, Indonesia. [2] School of Strategic and Global Studies, Universitas Indonesia, Jakarta, Indonesia.

Received: 7 September 2020 Accepted: 9 December 2020
Published online: 06 January 2021

References

1. Tussyadiah IP, Zach F. Identifying salient attributes of peer-to-peer accommodation experience. J Travel Tour Mark. 2017;34(5):636–52.
2. Lee CKH, Tse YK, Zhang M, Ma J. Analysing online reviews to investigate customer behaviour in the sharing economy: the case of Airbnb. Inf Technol People. 2019;33(3):945–61.
3. Belarmino A, Koh Y. A critical review of research regarding peer-to-peer accommodations. Int J Hosp Manag. 2019;2020(84):102315.
4. Tussyadiah IP, Zach FJ. Hotels vs. peer-to-peer accommodation rentals: text analytics of consumer reviews in Portland, Oregon. SSRN Electron J. 2015;

5. Cheng M, Jin X. What do Airbnb users care about? An analysis of online review comments. Int J Hosp Manag. 2018;2019(76):58–70.
6. Bridges J, Vásquez C. If nearly all Airbnb reviews are positive, does that make them meaningless? Curr Issues Tour. 2018;21(18):2065–83.
7. Prayag G, Ozanne LK. A systematic review of peer-to-peer (P2P) accommodation sharing research from 2010 to 2016: progress and prospects from the multi-level perspective. J Hosp Mark Manag. 2018;2:12.
8. Dolnicar S. A review of research into paid online peer-to-peer accommodation: Launching the Annals of Tourism Research curated collection on peer-to-peer accommodation. Ann Tour Res. 2019;75:248.
9. Sainaghi R. The current state of academic research into peer-to-peer accommodation platforms. Int J Hosp Manag. 2020;89:102555.
10. Oskam J, Boswijk A. Airbnb: the future of networked hospitality businesses. J Tour Futur. 2016;2:22.
11. Guttentag D. Progress on Airbnb: a literature review. J Hosp Tour Technol. 2019;10:1.
12. Adamiak C. Current state and development of Airbnb accommodation offer in 167 countries. Curr Issues Tour. 2019. https://doi.org/10.1080/13683500.2019.1696758.
13. Biswas B, Sengupta P, Chatterjee D. Examining the determinants of the count of customer reviews in peer-to-peer home-sharing platforms using clustering and count regression techniques. Decis Support Syst. 2020;135:113324.
14. Wachsmuth D, Weisler A. Airbnb and the rent gap: Gentrification through the sharing economy. Environ Plan A. 2018;3:33.
15. Celata F, Hendrickson CY, Sanna VS. The sharing economy as community marketplace? Trust, reciprocity and belonging in peer-to-peer accommodation platforms. Cambridge J Reg Econ Soc. 2017;10:349.
16. Lutz C, Newlands G. Consumer segmentation within the sharing economy: The case of Airbnb. J Bus Res. 2018;88:187.
17. Geron T. Airbnb and the unstoppable rise of the share economy. Forbescom. 2013;22:9991.
18. Frenken K, Schor J. Putting the sharing economy into perspective. Environ Innov Soc Transitions. 2017;23:3–10.
19. Varma A, Jukic N, Pestek A, Shultz CJ, Nestorov S. Airbnb: Exciting innovation or passing fad? Tour Manag Perspect. 2016;20:228–37.
20. Phua VC. Perceiving Airbnb as sharing economy: the issue of trust in using Airbnb. Curr Issues Tourism. 2019;9:877.
21. Guttentag DA, Smith SLJ. Assessing Airbnb as a disruptive innovation relative to hotelsSubstitution and comparative performance expectations. Int J Hosp Manag. 2017;64:1–10.
22. Sainaghi R, Baggio R. Substitution threat between Airbnb and hotels: Myth or reality? Ann Tour Res. 2020;83:102959.
23. Mody MA, Suess C, Lehto X. The accommodation experiencescape: a comparative assessment of hotels and Airbnb. Int J Contemp Hosp Manag. 2017;29:2377.
24. Guttentag D. Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. Curr Issues Tour. 2015;18:1–26.
25. Zhang T, Bufquin D, Lu C. A qualitative investigation of microentrepreneurship in the sharing economy. Int J Hosp Manag. 2019;79:148.
26. Ju Y, Back KJ, Choi Y, Lee JS. Exploring Airbnb service quality attributes and their asymmetric effects on customer satisfaction. Int J Hosp Manag. 2019;21:892.
27. Zhang G, Cui R, Cheng M, Zhang Q, Li Z. A comparison of key attributes between peer-to-peer accommodations and hotels using online reviews. Current Issues in Tourism. 2020.
28. Guttentag D, Smith S, Potwarka L, Havitz M. Why Tourists Choose Airbnb: A Motivation-Based Segmentation Study. J Travel Res. 2018;2:90.
29. Sparks BA, Browning V. The impact of online reviews on hotel booking intentions and perception of trust. Tour Manag. 2011;32(6):1310–23.
30. Zhang J. What's yours is mine: exploring customer voice on Airbnb using text-mining approaches. J Consum Mark. 2019;1:90.
31. Hu N, Zhang J, Pavlou PA. Overcoming the J-shaped distribution of product reviews. Communications of the ACM. 2009.
32. He W, Tian X, Tao R, Zhang W, Yan G, Akula V. Application of social media analytics: A case of analyzing online hotel reviews. Online Inf Rev. 2017.
33. Joseph G, Varghese V. Analyzing Airbnb customer experience feedback using text mining. In: Big Data and Innovation in Tourism, Travel, and Hospitality: Managerial Approaches, Techniques, and Applications. 2019.
34. Lucini FR, Tonetto LM, Fogliatto FS, Anzanello MJ. Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. J Air Transp Manag. 2020;18:9.
35. Georgakopoulou A, Spilioti T. The routledge handbook of language and digital communication. In: The Routledge Handbook of Language and Digital Communication. 2015. p. 1–434.
36. Wang Z. Anonymity, social image, and the competition for volunteers: A case study of the online market for reviews. BE J Econ Anal Policy. 2010;10:1.
37. Badan Pusat Statistik (BPS). Statistik Kunjungan Wisatawan Mancanegara [International Visitor Arrival Statistics] 2019. Jakarta; 2020. https://www.bps.go.id/publication/2020/06/26/94ceb011540bd0cd73e3474c/statistik-kunjungan-wisatawan-mancanegara-2019.html.
38. Barbosa RRL, Sánchez-Alonso S, Sicilia-Urban MA. Evaluating hotels rating prediction based on sentiment analysis services. Aslib J Inf Manag. 2015;
39. Prem A, Gunasekar S, Menon DG. User Generated Big Data Analysis of Customer Ratings of Beaches in Andaman and Nicobar Islands of India. Int J Innov Technol Explor Eng 2019;9(2):4921–5. http://www.ijitee.org/wp-content/uploads/papers/v9i2/B7621129219.pdf
40. Cosma G, Acampora G. Neuro-fuzzy sentiment analysis for customer review rating prediction. In: Studies in Computational Intelligence. 2016.
41. Li Y, Wang S, Ma Y, Pan Q, Cambria E. Popularity prediction on vacation rental websites. Neurocomputing. 2020;12:6.
42. Saraswat M. Practical Guide to deal with Imbalanced Classification Problems in R. 2016. p. 1–17. https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/.

43. Subroto A, Apriyana A. Cyber risk prediction through social media big data analytics and statistical machine learning. J Big Data. 2019;6(1):50. https://doi.org/10.1186/s40537-019-0216-1.
44. Vivek S. Analyzing Customer reviews using text mining to predict their behaviour. 2018; https://medium.com/analytics-vidhya/customer-review-analytics-using-text-mining-cd1e17d6ee4e.
45. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;12:233.
46. Fonseca L. Create predictive models in R with Caret. 2019. https://towardsdatascience.com/create-predictive-models-in-r-with-caret-12baf9941236
47. Osman H, D'Acunto D, Johns N. Home and away: Why do consumers shy away from reporting negative experiences in the peer-to-peer realms? Psychol Mark. 2019;36(12):1162–75. https://doi.org/10.1002/mar.21264.

Publisher's Note