Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach

Dinda Thalia Andariesta and Meditya Wasesa

Abstract

Purpose – *This research presents machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic using multisource Internet data.*

Design/methodology/approach – *To develop the prediction models, this research utilizes multisource Internet data from TripAdvisor travel forum and Google Trends. Temporal factors, posts and comments, search queries index and previous tourist arrivals records are set as predictors. Four sets of predictors and three distinct data compositions were utilized for training the machine learning models, namely artificial neural networks (ANNs), support vector regression (SVR) and random forest (RF). To evaluate the models, this research uses three accuracy metrics, namely root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE).*

Findings – *Prediction models trained using multisource Internet data predictors have better accuracy than those trained using single-source Internet data or other predictors. In addition, using more training sets that cover the phenomenon of interest, such as COVID-19, will enhance the prediction model's learning process and accuracy. The experiments show that the RF models have better prediction accuracy than the ANN and SVR models.*

Originality/value – *First, this study pioneers the practice of a multisource Internet data approach in predicting tourist arrivals amid the unprecedented COVID-19 pandemic. Second, the use of multisource Internet data to improve prediction performance is validated with real empirical data. Finally, this is one of the few papers to provide perspectives on the current dynamics of Indonesia's tourism demand.*

Keywords *Predictive analytics, Forecasting, Tourist arrivals, COVID-19, Internet data, Big data, Machine learning, Artificial neural network, Random forest, Support vector regression, Online forum, Search engine*

Paper type *Research paper*

Dinda Thalia Andariesta and Meditya Wasesa are both based at the School of Business and Management, Institut Teknologi Bandung, Bandung, Indonesia.

Received 16 October 2021
Revised 12 December 2021
Accepted 16 December 2021

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interest: Authors declare that there is no conflict of interest due to the publication of this paper.

1. Introduction

The increasing use of web-based platforms stimulates the growing availability of structured and unstructured data (Li *et al.*, 2021). Search engines (Bangwayo-Skeete and Skeete, 2015), online forums (Fronzetti Colladon *et al.*, 2019) and photo sharing apps (Miah *et al.*, 2017) are just a handful of applications that contribute to the increasing availability of online data. The availability of online data has attracted academics and practitioners to extract business values from it. The tourism and hospitality industries are not an exception. Tourists have used various online platforms, such as social networks, microblogs, online booking, online reviews and online forums (Li *et al.*, 2021), for their traveling purposes. The data emission from this online platform provides valuable customer behavior information (Bangwayo-Skeete and Skeete, 2015; Li *et al.*, 2017). Forecasting models have been one of the most popular use cases that can be improved by utilizing this big Internet data (Song *et al.*, 2019).

Literature on tourism demand forecasting is extensive (Li *et al.*, 2021). Most studies have been focusing on predicting international tourist flow using various quantitative methods (Song *et al.*, 2019), including time series (Ma *et al.*, 2016; Park *et al.*, 2017), econometric (Padhi and Pati, 2017), and artificial intelligence (AI) (Lv *et al.*, 2018; Sun *et al.*, 2019). In this big data era, AI-based approaches have increased popularity (Song *et al.*, 2019) and have been widely used for tourism demand forecasting due to their ability to deal with nonlinear data (Law *et al.*, 2019; Sun *et al.*, 2019; Huang and Hao, 2020). The artificial neural network (ANN), support vector regression (SVR), and random forest (RF) are among the most frequently used AI-based models (Sun *et al.*, 2019; Song *et al.*, 2019; Abellana *et al.*, 2020; Huang and Hao, 2020; Li *et al.*, 2020).

While the use of historical statistics records for forecasting purposes has already matured, forecasting models using Internet data have received increasing attention (Li and Law, 2020; Li *et al.*, 2021). Previous studies have utilized Internet data from different sources, such as search engines (Dergiades *et al.*, 2018; Li *et al.*, 2020), web traffic (Yang *et al.*, 2014; Gunter and Önder, 2016) and social media (Miah *et al.*, 2017; Starosta *et al.*, 2019), for forecasting purposes. Search engine and web traffic data provide structured time-series data, while social media generate unstructured data. Most previous studies focused on utilizing single-source Internet data with notable forecasting accuracy improvements (Bangwayo-Skeete and Skeete, 2015; Park *et al.*, 2017).

Although many studies have explored the use of Internet data to develop more accurate forecasting models, the ones that attempt to utilize combinations of several types of Internet data remain limited. Since single-source Internet data cannot comprehensively reflect tourists' attention, interests and interactions (Fronzetti Colladon *et al.*, 2019; Li *et al.*, 2021), multisource Internet data can offer a solution to address this drawback. Moreover, numerous issues and challenges are present in integrating different data sources and verifying empirical applications of multisource Internet data (Li *et al.*, 2021). Correspondingly, this research study aims to fill the gap by developing tourist arrivals forecasts using multisource and multi-categories of Internet data based on well-investigated machine learning models, namely ANN, SVR and RF. As a case study, this study opts to predict international tourist arrivals in Indonesia. Furthermore, this study corresponds to the current global tourism trend that has been affected by the travel restrictions amid the COVID-19 pandemic. In the face of an unprecedented pandemic, the applicability of Internet data and the developed machine learning solution must be reexamined. Thus, the main research question of this study is how to develop machine learning models using multisource Internet data that leads to more accurate tourist arrivals prediction during the COVID-19 pandemic.

The structure of this paper is written as follows. Section 1 provides brief background, research gap and research question. Section 2 presents a literature review on extant tourism forecasting methods and tourism demand forecasting using Internet data. The research method is explained in Section 3. Section 4 presents the case study context. Section 5 provides results and discussion. The last section provides the conclusion, implications, current limitations and future research.

2. Literature review

Existing quantitative methods for tourism forecasting can be classified into three categories: time series, econometric and AI (Song *et al.*, 2019; Li *et al.*, 2021). Time series models provide simplicity by employing a lag of Internet data as explanatory variables (Li *et al.*, 2021). This model can provide accurate predictions, notably for short-term forecasting horizons (Gunter and Önder, 2016; Park *et al.*, 2017). The most commonly used time series models include autoregressive, autoregressive integrated moving average and seasonal autoregressive integrated moving average (Song *et al.*, 2019; Li *et al.*, 2021). The econometric models are concerned with the causality of various explanatory variables (Zhou-grundy and Turner, 2015; Dergiades *et al.*, 2018). The previous studies demonstrated that econometric models can improve accuracy in more extended time horizons (Bangwayo-Skeete and Skeete, 2015; Gunter and Önder, 2016). However, all variables included in these models should be stationary to avoid spurious results (Huang *et al.*, 2017;

Dergiades *et al.*, 2018; Song *et al.*, 2019). Autoregressive distributed lag model, time-varying parameter and vector autoregression are among the most popular econometric models (Song *et al.*, 2019; Li *et al.*, 2021).

Unlike econometric models, AI-based models can describe nonlinear data without a prior understanding of the correlations between input and output variables (Song *et al.*, 2019). These models rely on built-in feature engineering, which becomes the distinct advantage when dealing with large datasets (Law *et al.*, 2019). This black box nature is often chastised for its lack of theoretical underpinning, poor interpretations of analytical outcomes and questionable explanatory value of input variables (Song *et al.*, 2019; Li *et al.*, 2021). However, AI-based approaches have been widely used because their nonlinear features can enhance forecasting performance (Law *et al.*, 2019; Sun *et al.*, 2019; Huang and Hao, 2020). The ANN is the most frequently used AI-based model, which can deal with almost any nonlinearity (Sun *et al.*, 2019; Song *et al.*, 2019). SVR is also frequently used in tourism demand forecasting due to its ability to model nonlinear data (Abellana *et al.*, 2020; Huang and Hao, 2020; Li *et al.*, 2020). Besides these two models, the RF also has grown in popularity due to its reliability and practical application in various fields (Khaidem *et al.*, 2016; Tyralis and Papacharalampous, 2017; Li *et al.*, 2020).

Previous studies have investigated three categories of Internet data to predict tourism demand: search engine, web traffic and social media. Google Trends (Bangwayo-Skeete and Skeete, 2015) and Baidu (Huang *et al.*, 2017) are examples of search query data generated from search engines. Baidu performed better for tourism forecasting in China due to its market share advantage than Google in the region. However, Google performed better for international tourism forecasting contexts (Yang *et al.*, 2015). Google Analytics account provides web traffic data from a particular website (Yang *et al.*, 2014). Social media data can be obtained from photo-sharing applications (Miah *et al.*, 2017), online forums (Fronzetti Colladon *et al.*, 2019) and news articles (Starosta *et al.*, 2019).

In the context of forecasting using search engine data, Google Trends have been used to predict tourist demand both at the country level (Park *et al.*, 2017) and at the tourist destination level, such as tourist arrivals to five London museums (Volchek *et al.*, 2019) and US National Parks (Clark *et al.*, 2019). Besides Google Trends, several studies with forecasting context in China have utilized the Baidu index (Huang *et al.*, 2017). Highly correlated query data are a challenge in utilizing search engine data. Therefore, Li *et al.* (2017) construct a composite search index to overcome highly correlated search query data (Li *et al.*, 2017). Moreover, the corrected aggregate search volume index or adjusted index for different search languages and search platforms is preferable to the nonadjusted index (Dergiades *et al.*, 2018). Prior studies demonstrated that incorporating search engine data from Google Trends and Baidu can improve forecasting accuracy.

Other researchers have explored the use of web traffic data of destination marketing organizations to predict hotel demand (Yang *et al.*, 2014) and tourist arrivals to Vienna (Gunter and Önder, 2016). Both studies obtained web traffic data by using a Google Analytics account. Google Analytics provides two significant types of web traffic data: visitors and visits. The findings showed that web traffic data can improve the error reduction (Yang *et al.*, 2014) and improve vector autoregression models' performance in a more extended time horizon (Gunter and Önder, 2016).

In terms of social media data, Miah *et al.* (2017) used geotagged photos uploaded by tourists to Flickr, a social media for photo-sharing, to predict tourism demand in Melbourne (Miah *et al.*, 2017). Another study classified the user reviews in social media into positive and negative sentiments (Starosta *et al.*, 2019). In contrast to search engines and web traffic data, these user-generated social media data are commonly found in unstructured data. Processing textual and image data from social media require advanced data preprocessing techniques. In general, using single-source Internet data to forecast tourist demand has been explored extensively.

While using a single category of Internet data has been well studied, only a few studies explored the use of different categories of Internet data (see Table 1). In this stream, some studies combined

Table 1 Previous research of tourism demand forecasting using Internet data

Study	*Category of Internet data*	*Predictor variables*	*Predicted variable*	*Forecasting methods*	*COVID-19 context*
Yang *et al.* (2015)	Search engine	Baidu index and Google Trends	Tourist arrivals to Hainan, China	ARMA, ARMAX	No
Lv *et al.* (2018)	Search engine	Baidu index and Google Trends	Tourism demand to America, Hainan, Beijing and Jiuzhaigou China	SARIMA, MLR, SVR, SLFN, ESN, LSTM, SAEN	No
Fronzetti Colladon *et al.* (2019)	Social media and search engine	Online forum (TripAdvisor) and Google Trends	International airport arrivals to seven major European capital cities	AR, FAAR, FABM, BM	No
Gunter *et al.* (2019)	Social media and search engine	Facebook and Google Trends	Tourist arrivals to four Austrian cities	Naïve, ETS, ARMA, ADLM, MIDAS	No
Sun *et al.* (2019)	Search engine	Baidu index and Google Trends	Tourist arrivals to Beijing	KELM, ARIMAX, ANN, LSSVR	No
Li *et al.* (2020)	Social media and search engine	Online reviews (Ctrip, Qunar) and Baidu index	Tourist arrivals to Mount Siguniang, China	ARIMAX, SVR, RF	No
Huang and Hao (2020)	Search engine	Baidu index and Google Trends	Tourist arrivals to Hong Kong	DBEDBN, RW, ARIMAX, SVR, ANN, DBN, EANN	No

Note(s): ADLM = Autoregressive distributed lag model, ANN = Artificial neural network, AR = Autoregressive, ARIMAX = Autoregressive integrated moving average with exogenous, ARMA = Autoregressive moving average, ARMAX = Autoregressive moving average with exogenous, BM = Bridge model, DBEDBN = Double boosting ensemble deep belief network, DBN = Deep belief network, EANN = Ensemble artificial neural network, ESN = Echo state network, ETS = Exponential smoothing, FAAR = Factor augment autoregressive model, FABM = Factor augmented bridge model, KELM = Kernel extreme learning machines, LSSVR = Least squares support vector regression, LSTM = Long short-term memory, MIDAS = Mixed-data sampling, MLR = Multiple linear regression, RF = Random forest, RW = Random walk, SAEN = Stacked autoencoder with echo-state regression, SARIMA = Seasonal autoregressive integrated moving average, SLFN = Single-hidden Layer Feed-forward Neural Network, SVR = Support vector regression

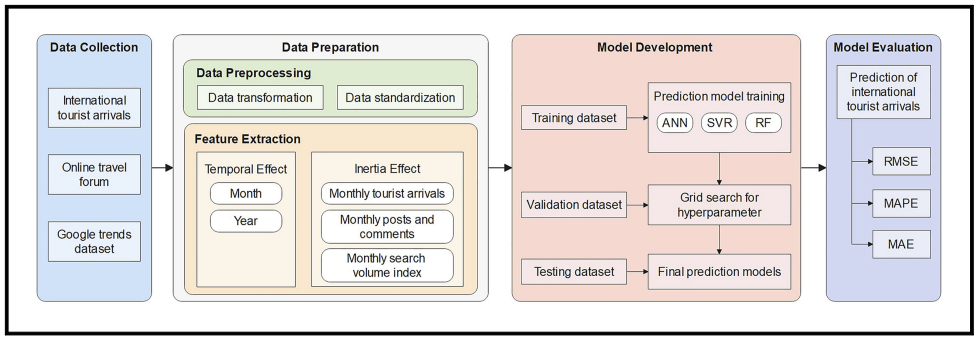
Google Trends and the Baidu index to predict tourist arrivals at the city level, such as Hong Kong (Huang and Hao, 2020), Hainan (Yang *et al.*, 2015) and Beijing (Lv *et al.*, 2018; Sun *et al.*, 2019). The results indicated that the forecasting performance of the models using combined search engine data outperformed the ones using individual search engine data. A study combined online reviews from TripAdvisor and Google Trends to predict international airport arrivals to major European capital cities (Fronzetti Colladon *et al.*, 2019). Other researchers utilized Facebook likes data and Google Trends to predict tourist arrivals to Austrian cities (Gunter *et al.*, 2019). At the destination level, online reviews from two platforms, namely Ctrip and Qunar, are combined with the Baidu index to predict tourist arrivals to Mount Siguniang China (Li *et al.*, 2020). The findings showed that better accuracy can be obtained by combining user-generated reviews from several online platforms.

To the best of our knowledge, developing tourism demand forecasting models using multisource Internet data, particularly with different categories of Internet data, is hard to find. Moreover, the applicability of using Internet data and the performance of existing machine learning forecasting models must be reexamined under an unprecedented COVID-19 pandemic context. This study fills the gap by utilizing two categories of Internet data, namely search engine (Google Trends) and social media (TripAdvisor travel forum), to develop prediction models that can accurately predict international tourist arrivals in the pandemic context. In addition, this study evaluates the prediction models under different combinations of Internet data and training dataset compositions.

3. Methodology

Figure 1 portrays the research framework of this study consisting of four main steps, namely (1) data collection, (2) data preparation, (3) model development and (4) model evaluation. First, we collected the data from the Indonesian Statistical Bureau (locally known as *Badan Pusat Statistik* or BPS),

Figure 1 The research framework



TripAdvisor travel forum and Google's search engine. In the second step, we conduct data preprocessing followed by feature extraction to obtain valuable and representative information from the dataset. The third step is the forecasting models development phase, followed by model evaluation at the fourth step.

Table 2 shows the specification of the prediction models, namely the predictors and predicted variables. We use four variables: temporal factors, TripAdvisor, Google Trends and international tourist arrivals. In total, we use four different sets of predictors and predicted variables that will be adopted in developing the prediction models using ANN, SVR and RF. We vary the predictors to verify that the proposed multisource Internet data can improve the prediction accuracy. Model evaluation based on root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) was used to examine out-of-sample prediction accuracy. In order to ensure the robustness of the prediction models using multisource Internet data, we constructed the models using three distinct data compositions with different lengths of training, validation and testing dataset. While different settings of data splits can affect the model's forecasting performance (Yang *et al.*, 2014), it is important to determine which data split setting will lead to the highest prediction accuracy.

3.1 Artificial neural networks

A feed-forward neural network consists of one or more input layers, one or more hidden layers and one output layer where each neuron in one layer conveys information to all neurons in the subsequent layer (Höpken *et al.*, 2020). In this study, the ANN model consists of an input layer with three neurons that represent the predictor variables, namely the previous tourist arrivals (x_1), the number of posts and comments and search volume index (x_3), and an output layer representing

Table 2 The specification of prediction models

Construct	*Attribute*	*Function*	*Model* 1	2	3	4
Temporal	Month	Predictor	v	v	v	v
	Year	Predictor	v	v	v	v
TripAdvisor	Number of posts	Predictor		v		v
	Number of comments	Predictor		v		v
Google Trends	Main entry point	Predictor			v	v
	International travel requirement	Predictor			v	v
	Tourism planning	Predictor			v	v
Tourist arrivals	International tourist arrivals in the previous month	Predictor	v			v
	Monthly international tourist arrivals	Predicted	v	v	v	v

the predicted variable, namely international tourist arrivals or (Y). The output of hidden neurons (V_L) and the international tourist arrivals (Y) can be written in Eq. (1) and (2):

$$V_1 = \sum_{i=1}^{3} h(w_{1i}x_i + b_1)$$
$$V_L = \sum_{i=1}^{3} h(w_{Li}x_i + b_L) \quad (1)$$

$$Y = \sum_{j=1}^{L} h(w_j V_j + \beta) \quad (2)$$

where w_{Li} is the input weight, x_i is the input neurons, b_L is the hidden layer threshold, w_L is the output weight, V_L is the output of hidden neurons, β is the output layer threshold, $h(x)$ is the activation function and Y is the output neuron (international tourist arrivals). Figure 2 shows the structure of the feed-forward neural network.

3.2 Support vector regression

Support vector machine (SVM) is a machine learning algorithm that maps data in high-dimensional feature space through a nonlinear mapping function (Li *et al.*, 2020). SVM classifies training data vectors ($\overrightarrow{x_i}$) into two segments (y_i) that are represented in Eq. (3).

$$G = \left(\overrightarrow{x_i}, y_i\right); \overrightarrow{x_i} \in R^n; y_i = -1 \ \ or \ \ 1; i = 1, 2, .., N \quad (3)$$

where $\overrightarrow{x_i}$ is the training data vectors ($\overrightarrow{x_i} = (x_1, x_2, x_3)$ with x_1 = the previous tourist arrivals, x_2 = the number of posts and comments, x_3 = search volume index), N is the number of training data and n is the input space dimension represented by the number of predictor variables. The training data vectors $\overrightarrow{x_i}$ classified by a hyperplane $\overrightarrow{w}.\overrightarrow{x} + b = 0$, which satisfy the following equations:

$$y_i - \left(\overrightarrow{w}.\overrightarrow{x} + b\right) \leq \varepsilon; if \ \ y_i = 1; i = 1, 2, \ldots, N$$
$$y_i - \left(\overrightarrow{w}.\overrightarrow{x} + b\right) \geq -\varepsilon; if \ \ y_i = -1; i = 1, 2, \ldots, N \quad (4)$$

where $\overrightarrow{w}$ is the weight vector, φ: $R^n \rightarrow R^m$ is the mapping of input space (R^n) to high-dimensional space (R^m), b is a constant and ϵ is the incentive loss function.
In Figure 3, we draw two parallel lines $\overrightarrow{w}.\overrightarrow{x} + b = 1$ for one segment and $\overrightarrow{w}.\overrightarrow{x} + b = -1$ for the other segment. In SVR, the model seeks a hyperplane to fit the given training data points with the

Figure 2 The structure of the feed-forward neural network

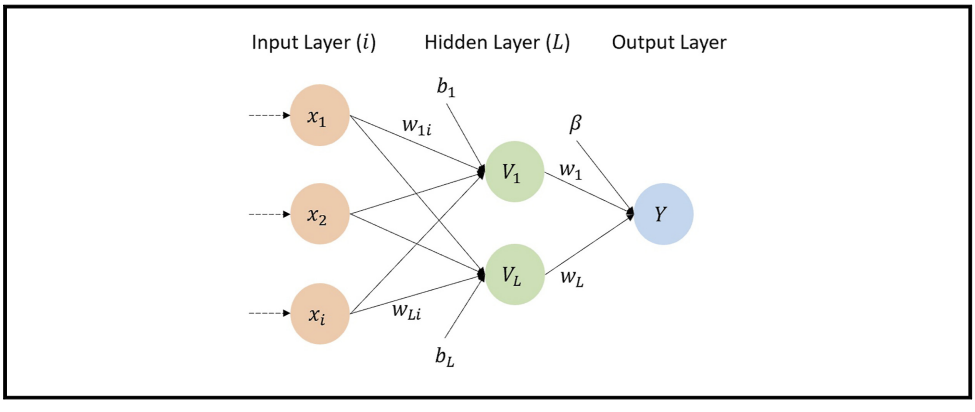
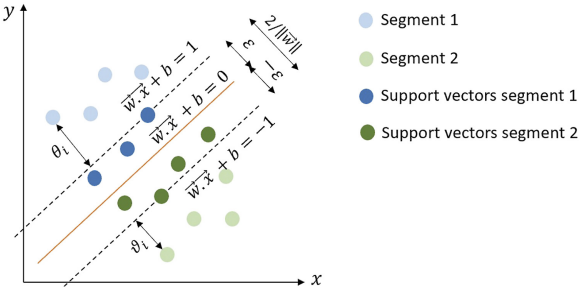


Figure 3 The margin and decision boundary of the support vector machine



fitting function $f(\vec{x}) = \vec{w}.\vec{x} + b$ by minimizing the regularized risk function $(1/2)\|\vec{w}\|^2 + C\sum_{i=1}^{N}(\theta_i + \vartheta_i)$, where C is the regularization parameter, θ_i and ϑ_i are distances from actual value y_i to the boundary values of ε. Thus, the nonlinear mapping function $f(\vec{x})$ can be generated by applying the Lagrange multiplier (Yao *et al.*, 2021),

$$f\left(\vec{x}\right) = \sum_{i=1}^{N}(\alpha_i - \beta_i)K\left(\vec{x_i}, \vec{x}\right) + b \tag{5}$$

where $f(\vec{x})$ is the prediction of tourist arrivals, $\vec{x_i}$ is the training data vectors ($\vec{x_i} = (x_1, x_2, x_3)$ with $x_1 =$ the previous tourist arrivals, $x_2 =$ the number of posts and comments, $x_3 =$ search volume index), α_i and β_i are Lagrange coefficients, $K(\vec{x_i}, \vec{x})$ is the Kernel function and b is the constant.

3.3 Random forest

A RF has grown in popularity due to its high reliability and practical application in various fields (Khaidem *et al.*, 2016; Tyralis and Papacharalampous, 2017; Li *et al.*, 2020). This model combines the classification and regression tree and bagging method to improve the accuracy (Breiman, 2001). Figure 4 portrays the process of RF.

First, training subsets are randomly selected from the training dataset. Second, trees are randomly generated and trained by using the training subsets. The parent node splits into two daughter nodes, and the information impurity due to this split can be written by

$$\Delta g(N) = g(N) - P_L g(N_L) - P_R g(N_R) \tag{6}$$

where $g(N)$ is the Gini impurity measure in node N, P_L is the population proportion of the left daughter node N_L and P_R is the population proportion of the right daughter node N_R.
Third, each tree predicts the testing dataset, and the prediction results generated by all trees are averaged to obtain the final output of tourist arrivals prediction. The final output of RF is as follows:

$$\widehat{y} = \frac{1}{N_{trees}}\sum_{i=1}^{N_{trees}} y_i \tag{7}$$

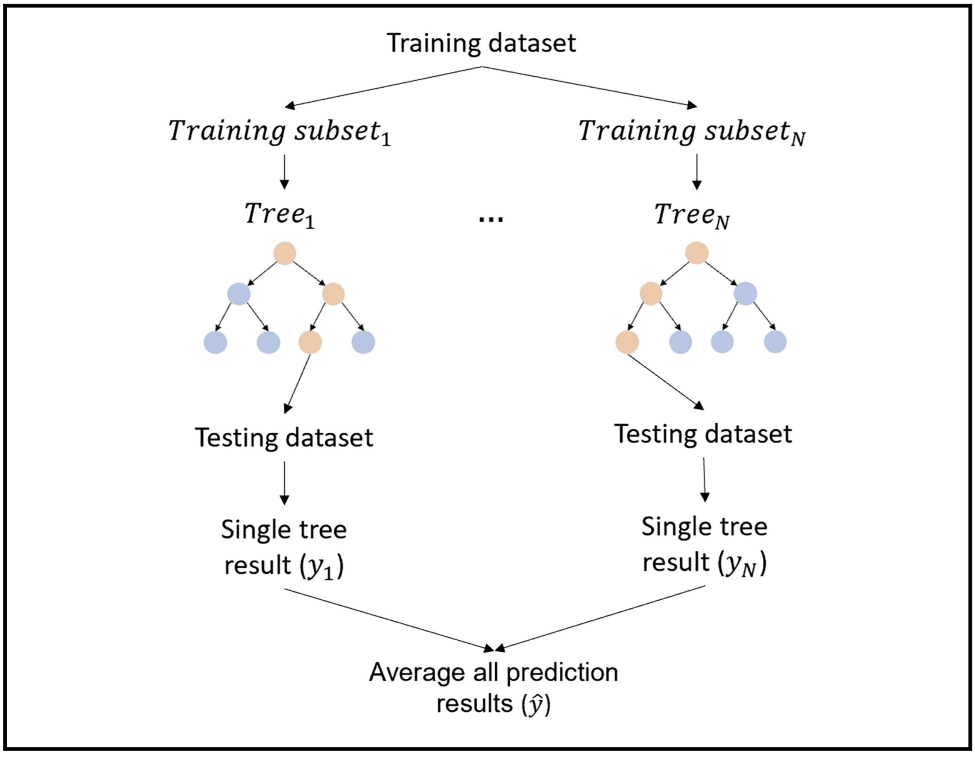
where $\widehat{y}$ is the final output, N_{trees} is the number of trees and $\widehat{y_i}$ is the result of a single tree.

4. Case study

4.1 Data collection

As a case study, we analyze international tourist arrivals to Indonesia during the COVID-19 pandemic. First, we collected tourism data from the Indonesian Statistical Bureau Indonesia or

Figure 4 The rationale of random forest



BPS from January 2017 until June 2021. Next, we collect the data from a global online tourism platform, TripAdvisor. Table 3 shows the data sample of the Indonesia travel forum in TripAdvisor. The dynamic interactions within the online forums can be seen from the number of posts and comments that vary every day and covers diverse topics (Fronzetti Colladon *et al.*, 2019). More than 43,000 posts were obtained, with 243,000 comments from users.

Table 4 shows the selected Google Trends keywords used in this study. The keywords are categorized into three topics: main entry point, international travel requirement and tourism planning. The search volume index represents search interest with values ranging from 0 to 100, a value of 100 as the search keyword's peak popularity.

Table 3 Data sample of Indonesia travel forum in TripAdvisor

Variable	*Data type*	*Data example*
Forum	String	Bali
Topic	String	"is bali safe for vacation?"
Link of post	String	https://www.tripadvisor.com/ShowTopic-g294226-i7220-k13419945-Is_bali_safe_for_vacation-Bali.html
Author of post	String	Olivia
Link of the author's profile	String	https://www.tripadvisor.com/Profile/viva99slot? tab = forum
Posting date	Date	Dec 10, 2020
Number of comments	Integer	25
Last comment by user	String	SW0590
Link of the commenter's profile	String	https://www.tripadvisor.com/Profile/SW0590? tab = forum

Table 4 Google Trends keywords

Topic	*Keyword*
Main entry point	Ngurah Rai International Airport
	Soekarno-Hatta International Airport
	Batam ferry terminal
	Bali
	Jakarta
	Batam
International travel requirement	Passport Indonesia
	Visa Indonesia
Tourism planning	Indonesia hotel
	Indonesia resort
	Indonesia restaurant
	Indonesia travel

Table 5 summarizes the descriptive statistics of the datasets. The statistics consist of monthly international tourist arrivals, daily posts and comments in the Indonesia travel forum, and the monthly search volume index of the selected keywords.

Figure 5 portrays all variables utilized for developing the prediction models. The international tourist arrivals have been experiencing significant declines since February 2020 due to the government's travel restrictions amid COVID-19. During the outbreak, the interaction in travel forums and the popularity of selected search keywords also decreased.

4.2 Data preparation

This phase consists of data preprocessing and feature extraction. In the data preprocessing, we transform all data into monthly data. We performed a three-month moving average for Google Trends data that smoothed out popularity trends to filter noise. In the last data preprocessing step, we perform data standardization using Eq. (8).

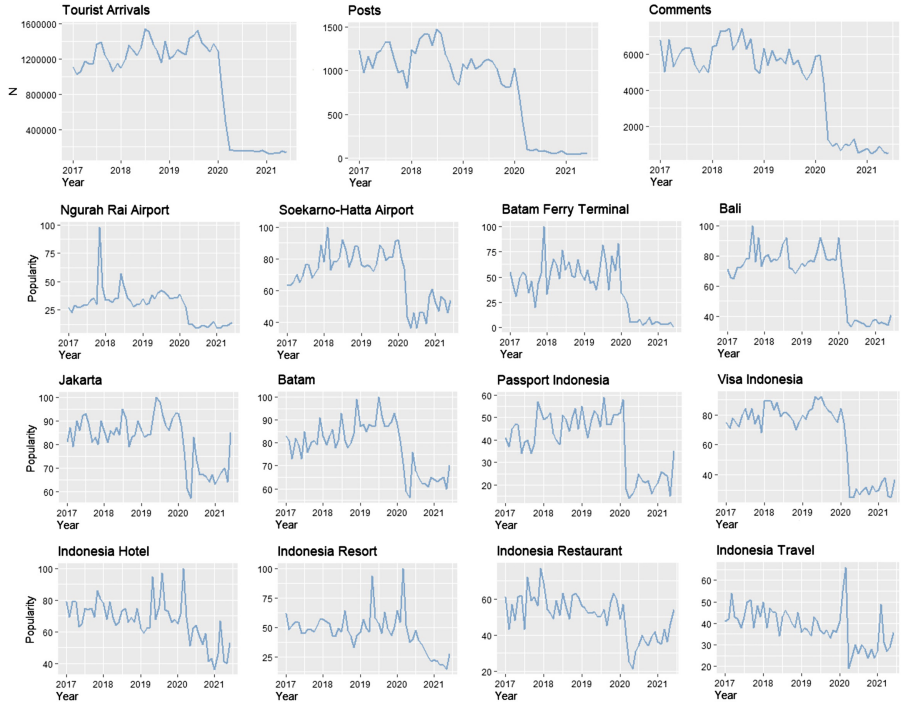
$$X_{transformed} = \frac{X - \overline{X}}{\sigma} \tag{8}$$

where X is the original value, $\overline{X}$ is the mean and σ is the standard deviation.

Table 5 Descriptive statistics of the datasets

Data source	*Variable*	*Count*	*Mean*	*Std. dev*	*Min*	*Max*
Indonesian Statistical Bureau Indonesia (https://www.bps.go.id/)	Tourist arrivals	54	940,902.98	518,460.03	115,765	1,547,231
TripAdvisor (https://www.tripadvisor.com/ShowForum-g294225-i7219-o5320-Indonesia.html)	Posts	1,642	26.54	17.70	0	73
	Comments	1,642	148.08	100.55	0	728
Google Trends (https://trends.google.com/)	Ngurah Rai International airport	54	29.11	14.94	9	98
	Soekarno-Hatta International airport	54	69.78	15.86	36	100
	Batam ferry terminal	54	38.94	25.64	0	100
	Bali	54	65.85	19.97	33	100
	Jakarta	54	81.80	10.35	57	100
	Batam	54	78.48	10.77	56	100
	Passport Indonesia	54	39.30	13.25	14	59
	Visa Indonesia	54	65.98	23.19	25	92
	Indonesia hotel	54	67.07	13.38	36	100
	Indonesia resort	54	46.33	15.78	14	100
	Indonesia restaurant	54	50.63	11.54	21	77
	Indonesia travel	54	38.63	8.96	19	66

Figure 5 All variables for developing the prediction models



For processing time series data using the machine learning method, we extract two temporal features in this study: month and year. These variables are converted into dummy variables that aim to prevent information duplication. The second feature is the inertia variable or lag feature, which describes the value of the data in the previous month. We extract the inertia variable for all data categories, including tourist arrivals, search volume index, number of posts and comments.

4.3 Model development

We split the entire dataset into three segments: training, validation and testing datasets. We decompose the training datasets into three partitions (see Figure 6), namely (1) January 2017–April 2020 (the period when COVID-19 starts to gain popularity and infect Indonesian citizens), (2) January 2017–August 2020 (the period when the government implemented international travel restrictions) and (3) January 2017–December 2020 (the period when the government extend the international travel restrictions and implement wide-scale social restrictions).

The model parameters are optimized through a hyperparameter grid search (Lijuan and Guohua, 2016; Bi *et al.*, 2020). First, we optimized the learning rate and the number of hidden layers for the

Figure 6 Composition of training, validation and testing datasets

Data composition	2017	2018	2019	2020											2021					
				Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
1	Training							Validation								Testing					
2	Training												Validation			Testing					
3	Training															Validation			Testing		

ANN model. Second, three parameters, namely the regularization parameter (*C*), Kernel and epsilon (ϵ), are optimized for the SVR model. Lastly, grid search for the RF model is performed by considering the number of variables randomly sampled at each split (Mtry), the number of trees (*N* trees) and the maximum nodes. Table 6 shows the results of the hyperparameters optimization.

4.4 Model evaluation

Evaluation of model performance is an inseparable step in developing prediction models. The difference between the predicted and actual values refers to the prediction error (Li *et al.*, 2017). We evaluate the prediction performance using two scale-dependent errors, namely RMSE and MAE, and a percentage error, namely MAPE, which can be calculated using Eq. (9)–(11).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{10}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{y_i} \tag{11}$$

where y_i is the actual, and $\hat{y}_i$ is the predicted value of tourist arrivals.

5. Results and discussion

Tables 7 and 8 summarize the accuracy of all prediction models in terms of RMSE and MAE. From a total of 36 models, the prediction models utilizing multisource Internet data perform consistently better than the other models using single or even no Internet data predictors. The superiority of the multisource Internet data is also consistent across different data compositions. This finding

Table 6 Hyperparameter optimization

Data composition	*Method*	*Hyperparameter*	*Model 1*	*Model 2*	*Model 3*	*Model 4*
1	ANN	Learning rate	0.01	0.01	0.1	0.1
		Hidden layer	8	7	7	10
	SVR	C	0.01	0.01	0.01	0.01
		Kernel	Sigmoid	Sigmoid	Sigmoid	Sigmoid
		Epsilon	0.025	0.025	0.05	0.05
	RF	Mtry	5	4	5	4
		N trees	10	30	50	10
		Maximum nodes	5	10	10	5
2	ANN	Learning rate	0.1	0.1	0.01	0.1
		Hidden layer	4	3	1	3
	SVR	C	0.01	0.01	0.01	0.01
		Kernel	Sigmoid	Sigmoid	Sigmoid	Sigmoid
		Epsilon	0.025	0.025	0.025	0.025
	RF	Mtry	5	5	5	4
		N trees	30	10	30	10
		Maximum nodes	10	10	10	5
3	ANN	Learning rate	0.01	0.1	0.1	0.01
		Hidden layer	2	7	3	2
	SVR	C	0.01	0.01	0.01	0.01
		Kernel	Sigmoid	Sigmoid	Sigmoid	Sigmoid
		Epsilon	0.05	0.025	0.025	0.025
	RF	Mtry	3	5	5	4
		N trees	10	40	10	10
		Maximum nodes	5	10	5	10

Table 7 RMSE of the prediction models

Model	Data composition	Predictors: 1 (Temporal + Previous arrivals)	2 (Temporal + TripAdvisor)	3 (Temporal + Google Trends)	4 (Temporal + Previous arrivals + TripAdvisor + Google Trends)
ANN	Data composition 1	410,507.90	286,340.00	263,504.20	115,814.80
	Data composition 2	244,910.67	161,509.29	122,716.81	62,873.48
	Data composition 3	48,094.28	23,014.18	20,578.84	*11,698.45*
SVR	Data composition 1	251,740.00	172,068.00	198,134.90	164,414.70
	Data composition 2	180,286.49	85,727.30	108,142.37	78,241.15
	Data composition 3	55,374.35	31,004.33	58,953.95	*28,175.02*
RF	Data composition 1	676,674.60	349,367.50	154,126.95	55,156.28
	Data composition 2	349,704.78	20,014.93	100,762.79	19,084.13
	Data composition 3	289,465.10	38,689.82	11,798.04	*10,334.02**

Note(s): The italic figures indicate the best performing model across different data compositions and predictors sets within a similar prediction model, and * indicates the best performing model across different data compositions, predictors sets and prediction models

Table 8 MAE of the prediction models

Model	Data composition	Predictors: 1 (Temporal + Previous arrivals)	2 (Temporal + TripAdvisor)	3 (Temporal + Google Trends)	4 (Temporal + Previous arrivals + TripAdvisor + Google Trends)
ANN	Data composition 1	386,762.20	268,013.40	244,880.00	107,909.30
	Data composition 2	243,863.25	160,187.97	122,073.73	61,053.17
	Data composition 3	46,388.95	19,068.90	17,136.90	*10,686.56*
SVR	Data composition 1	241,113.60	158,033.00	188,119.90	156,366.70
	Data composition 2	179,089.45	84,890.89	106,825.07	77,057.46
	Data composition 3	44,949.64	26,717.46	57,775.94	*26,049.87*
RF	Data composition 1	649,825.46	330,855.86	144,193.22	34,691.26
	Data composition 2	349,419.29	17,407.48	98,328.00	16,871.83
	Data composition 3	289,244.74	34,160.84	10,816.20	*9,930.24**

Note(s): The italic figures indicate the best performing model across different data compositions and predictors sets within a similar prediction model, and * indicates the best performing model across different data compositions, predictors sets and prediction models

indicates the robustness of using the multisource Internet data approach. Furthermore, all prediction models trained using data composition 3 yielded the best RMSE and MAE compared to those trained using data compositions 1 and 2. The RMSE and MAE significantly improve when we incorporate more data within the outbreak.

In line with the RMSE and MAE results, Table 9 shows that the prediction models trained using data composition 3 have the lowest MAPE compared to those using other data compositions. These findings indicate that the prediction models trained using sufficient data covered unexpected events, such as the COVID-19, will positively influence the prediction accuracy of the developed models. As noted by the previous study, researchers must develop forecasting models that can account for unforeseen events (Qiu *et al.*, 2021). Overall, the RF model incorporating all predictors trained using data composition 3 has the highest prediction accuracy.

Discussing the impact of different predictors sets on prediction accuracy, prediction models trained using multisource Internet data perform better in predicting tourist arrivals than those trained using single-source Internet data and previous tourist arrivals. The ANN 4 and RF 4 models that use a complete set of predictors consistently outperformed the other three models. However, using a complete set of predictors in the SVR models leads to the best RMSE and MAE, but not for MAPE. By utilizing data composition 3, SVR 2 model has a slightly better MAPE than SVR 4 model.

Table 9 MAPE of the prediction models

Model	Data composition	Predictors: 1 (Temporal + Previous arrivals)	2 (Temporal + TripAdvisor)	3 (Temporal + Google Trends)	4 (Temporal + Previous arrivals + TripAdvisor + Google Trends)
ANN	Data composition 1	292.21%	202.89%	185.69%	82.11%
	Data composition 2	186.68%	122.97%	93.89%	47.40%
	Data composition 3	34.51%	14.63%	13.13%	*7.70%*
SVR	Data composition 1	181.62%	118.69%	141.74%	118.08%
	Data composition 2	137.28%	64.32%	81.30%	59.34%
	Data composition 3	33.52%	*19.03%*	42.49%	19.57%
RF	Data composition 1	488.66%	248.91%	109.15%	26.74%
	Data composition 2	266.34%	13.95%	76.44%	13.46%
	Data composition 3	211.23%	25.93%	8.00%	*7.09%**

Note(s): The italic figures indicate the best performing model across different data compositions and predictors sets within a similar prediction model, and * indicates the best performing model across different data compositions, predictors sets and prediction models

The SVR 2 model using data composition 3 has greater prediction error variations but a better average of percentage errors than the SVR 4 model.

Evaluating the accuracy of the models utilizing single-source Internet data, Google Trends data resulted in better forecasts than online forum data for ANN and RF models. In contrast, online forum data yielded better forecasts than Google Trends data in the SVR model. The training complexity of Google Trends data might be higher than online forum data due to the greater number of attributes. In addition, the training complexity of SVM is indeed high (Cervantes *et al.*, 2007). Despite the good theoretic foundations and accuracy, SVM does not perform well when the dataset contains more noise (Sarker, 2021). However, no single method can outperform other methods in all forecasting contexts (Li *et al.*, 2020), and not all Internet data variables will improve the accuracy (Yang *et al.*, 2015).

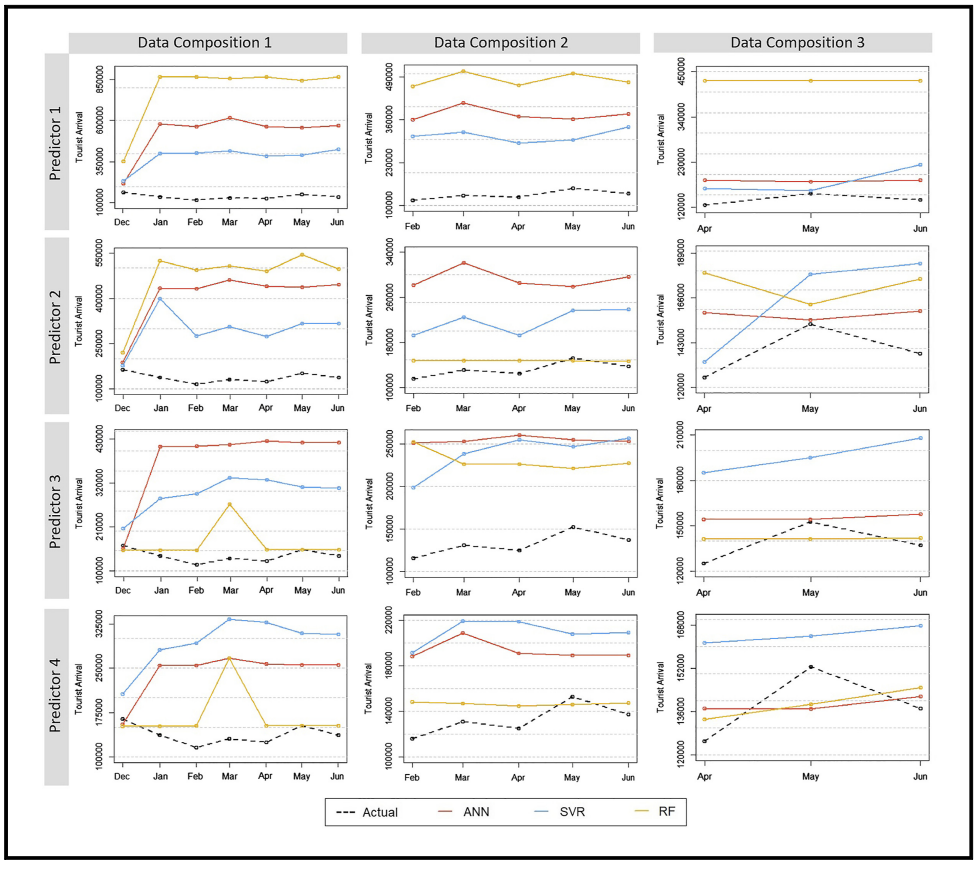
Figure 7 visually portrays the models' prediction results compared to the actual record of the international tourist arrivals in Indonesia. The training set of data composition 1 covers only two months of the pandemic (March to April 2020), resulting in a premature model's learning process leading to inaccurate forecasts with many overestimation cases. By utilizing data composition 2, the prediction results of the RF model improve when using predictors sets 2 and 4. However, the prediction results of this model have not captured the dynamics of tourist arrivals. Meanwhile, the results significantly improved when we applied data composition 3 and predictor set 4 for ANN and RF models. At the same time, the SVR model with data composition 3 cannot produce good predictions if we append Google Trends data due to increasing model complexity. In general, the prediction accuracy improves when we increase the training dataset covering the COVID-19 period and utilize a complete set of predictors.

Predicting tourist arrivals during the COVID-19 period is a nontrivial task. In nonroutine circumstances, we cannot rely only on standard historical statistical records to develop accurate forecasts. Nevertheless, alternative data are available. Search engine and online forum data are user-generated data that can be acquired publicly. This study has demonstrated that multisource Internet data can significantly improve the prediction accuracy of tourist arrivals under travel restrictions during the pandemic. This study confirms the usefulness of multisource Internet data for increasing the accuracy of tourist arrival predictions.

6. Conclusion and future works

This research presents machine learning models to predict international tourist arrivals in Indonesia during the COVID-19 using multisource Internet data, namely the TripAdvisor travel forum and Google Trends. The results show the positive impact of combining multisource Internet data to

Figure 7 Prediction results of international tourist arrivals in Indonesia



improve forecasting performance. Prediction models utilizing a combination of predictors from an online travel forum and a search engine have better accuracy than those using the predictor from a single source of Internet data, either the online travel forum only or search queries only. Moreover, our models have better performance than the prediction model that only uses historical tourist arrivals statistical records.

In developing the model, we decompose the training datasets into three partitions, namely (1) January 2017–April 2020 (the period when COVID-19 starts to gain popularity and infect Indonesian citizens), (2) January 2017–August 2020 (the period when the government implemented international travel restrictions) and (3) January 2017–December 2020 (the period when the government extended the international travel restrictions and implemented wide-scale social restrictions). The result indicates that the prediction model using the third training set performs best. These results are consistent across all investigated prediction models. Note that this third training set has the most extensive coverage of the pandemic situation. Thus, using more training sets covering the phenomenon of interest, such as COVID-19, will improve the prediction model's learning process and accuracy. In conclusion, the complete set of predictors and the third data composition applied to the RF model yielded the best prediction performance compared to ANN and SVR models.

Compared to the previous studies using the search query and online forum to predict tourist arrivals (Fronzetti Colladon *et al.*, 2019; Sun *et al.*, 2019; Huang and Hao, 2020), this study offers three contributions. First, this study pioneers the practice of a multisource Internet data approach in predicting tourist arrivals amid the COVID-19 pandemic. Second, this study has validated the use of multisource Internet data to improve prediction performance. Third, this is one of the few papers to provide perspectives on the current state of Indonesia's tourism demand.

In terms of managerial implications, the presented forecasting models can help tourism decision-making in many contexts, such as pricing strategies, allocating resources, planning tourism infrastructures and developing emergency plans (Li *et al.*, 2018; Sun *et al.*, 2019). The accurate forecasts reinforce the foresight capabilities of tourism decision-makers and policymakers, which can help the government to make better corresponding decisions in unexpected situations, such as the COVID-19 pandemic. Moreover, the fast-growing Internet data allows managers for in-depth analysis of visitor activities, interests and interactions, as well as their influence on tourism demand forecasting. The Internet data usage in tourism demand analysis offers several advantages, including timeliness, low cost (since it is open to the public) and good predictive power. Lastly, Internet data may help overcome survey data consumers' sample size constraints (Yang *et al.*, 2015).

Not without limitations, this study opens for further research opportunities. First, this study only focuses on international tourist arrivals in Indonesia. The selected keywords are limited and solely represent this country's public interests and attention. Thus, further studies can investigate other search queries and travel forums relevant to their specific contexts. Furthermore, future studies can explore the application of multisource Internet data for different countries or destinations. Second, this study only uses two data variables extracted from an online forum. Other variables extracted from online forums, such as the sentiment index, which provides an overview of public response, can also be incorporated. In addition, more external factors can be further examined as input for the prediction model. Other data sources, such as Facebook, Twitter and other online forums, can be explored to enrich the training data during prediction model development.

References

Abellana, D.P.M., Rivero, D.M.C., Aparente, M.E. and Rivero, A. (2020), "Hybrid SVR-SARIMA model for tourism forecasting using PROMETHEE II as a selection methodology: a Philippine scenario", *Journal of Tourism Futures*, Vol. 7 No. 1, pp. 78-97, doi: 10.1108/JTF-07-2019-0070.

Bangwayo-Skeete, P.F. and Skeete, R.W. (2015), "Can Google data improve the forecasting performance of tourist arrivals?", *Mixed-data Sampling approach'Tourism Management*, Vol. 46, pp. 454-464, doi: 10.1016/j.tourman.2014.07.014.

Bi, J., Liu, Y. and Li, H. (2020), "Annals of tourism research daily tourism volume forecasting for tourist attractions", *Annals of Tourism Research*, Vol. 83, p. 102923, doi: 10.1016/j.annals.2020.102923.

Breiman, L. (2001), "Random forests", *Machine Learning*, Vol. 45, pp. 5-32, doi: 10.1201/9780367816377-11.

Cervantes, J., Li, X. and Yu, W. (2007), "SVM classification for large data sets by considering models of classes distribution", *Proceedings - 2007 6th Mexican International Conference on Artificial Intelligence, Special Session, MICAI 2007*, pp. 51-60, doi: 10.1109/MICAI.2007.27.

Clark, M., Wilkins, E.J., Dagan, D.T., Powell, R., Sharp, R.L. and Hills, V. (2019), "Bringing forecasting into the future: using Google to predict visitation in US national parks", *Journal of Environmental Management*, Vol. 243, pp. 88-94, doi: 10.1016/j.jenvman.2019.05.006.

Dergiades, T., Mavragani, E. and Pan, B. (2018), "'Google Trends and tourists' arrivals: emerging biases and proposed corrections", *Tourism Management*, Vol. 66, pp. 108-120, doi: 10.1016/j.tourman.2017.10.014.

Fronzetti Colladon, A., Guardabascio, B. and Innarella, R. (2019), "Using social network and semantic analysis to analyze online travel forums and forecast tourism demand", *Decision Support Systems*, Vol. 123, p. 113075, doi: 10.1016/j.dss.2019.113075.

Gunter, U. and Önder, I. (2016), "Forecasting city arrivals with Google Analytics", *Annals of Tourism Research*, Vol. 61, pp. 199-212, doi: 10.1016/j.annals.2016.10.007.

Gunter, U., Önder, I. and Gindl, S. (2019), "Exploring the predictive ability of LIKES of posts on the Facebook pages of four major city DMOs in Austria", *Tourism Economics*, Vol. 25 No. 3, pp. 375-401, doi: 10.1177/1354816618793765.

Höpken, W., Eberle, T., Fuchs, M. and Lexhagen, M. (2020), "Improving tourist arrival prediction: a big data and artificial neural network approach", *Journal of Travel Research*, Vol. 60 No. 5, pp. 998-1017, doi: 10.1177/0047287520921244.

Huang, B. and Hao, H. (2020), "A novel two-step procedure for tourism demand forecasting current issues in method and practice", *Current Issues in Tourism*, Vol. 24 No. 9, pp. 1199-1210, doi: 10.1080/13683500.2020.1770705.

Huang, X., Zhang, L. and Ding, Y. (2017), "The Baidu Index: uses in predicting tourism flows – a case study of the Forbidden City", *Tourism Management*, Vol. 58, pp. 301-306, doi: 10.1016/j.tourman.2016.03.015.

Khaidem, L., Saha, S. and Dey, S.R. (2016), "Predicting the direction of stock market prices using random forest", (May). available at: http://arxiv.org/abs/1605.00003.

Li, H., Hu, M. and Li, G. (2020), "Annals of Tourism Research Forecasting tourism demand with multisource big data", *Annals of Tourism Research*, Vol. 83, p. 102912, doi: 10.1016/j.annals.2020.102912.

Li, X. and Law, R. (2020), "Network analysis of big data research in tourism", *Tourism Management Perspectives*, Vol. 33, p. 100608, doi: 10.1016/j.tmp.2019.100608.

Law, R., Li, G., Fong, D.K.C. and Han, X. (2019), "Tourism demand forecasting: a deep learning approach", *Annals of Tourism Research*, Vol. 75, pp. 410-423, doi: 10.1016/j.annals.2019.01.014.

Li, S., Chen, T., Wang, L. and Ming, C. (2018), "Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index", *Tourism Management*, Vol. 68, pp. 116-126, doi: 10.1016/j.tourman.2018.03.006.

Li, X., Law, R., Xie, G. and Wang, S. (2021), "Review of tourism forecasting research with internet data", *Tourism Management*, Vol. 83, p. 104245, doi: 10.1016/j.tourman.2020.104245.

Li, X., Pan, B., Law, R. and Huang, X. (2017), "Forecasting tourism demand with composite search index", *Tourism Management*, Vol. 59, pp. 57-66, doi: 10.1016/j.tourman.2016.07.005.

Lijuan, W. and Guohua, C. (2016), "Knowledge-Based Systems Seasonal SVR with FOA algorithm for single-step and multi-step ahead forecasting in monthly inbound tourist flow", *Knowledge-Based Systems*, Vol. 110, pp. 157-166, doi: 10.1016/j.knosys.2016.07.023.

Lv, S.X., Peng, L. and Wang, L. (2018), "Stacked autoencoder with echo-state regression for tourism demand forecasting using search query data", *Applied Soft Computing Journal*, Vol. 73, pp. 119-133, doi: 10.1016/j.asoc.2018.08.024.

Ma, E., Liu, Y., Li, J. and Chen, S. (2016), "Anticipating Chinese tourists arrivals in Australia: a time series analysis", *Tourism Management Perspectives*, Vol. 17, pp. 50-58, doi: 10.1016/j.tmp.2015.12.004.

Miah, S.J., Vu, H.Q., Gammack, J. and McGrath, M. (2017), "A big data Analytics method for tourist behaviour analysis", *Information and Management*, Vol. 54 No. 6, pp. 771-785, doi: 10.1016/j.im.2016.11.011.

Padhi, S.S. and Pati, R.K. (2017), "Quantifying potential tourist behavior in choice of destination using Google Trends", *Tourism Management Perspectives*, Vol. 24, pp. 34-47, doi: 10.1016/j.tmp.2017.07.001.

Park, S., Lee, J. and Song, W. (2017), "Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data", *Journal of Travel and Tourism Marketing*, Vol. 34 No. 3, pp. 357-368, doi: 10.1080/10548408.2016.1170651.

Qiu, R.T.R., Wu, D.C., Dropsy, V., Petit, S., Pratt, S. and Ohe, Y. (2021), "Visitor arrivals forecasts amid COVID-19: a perspective from the Asia and Pacific team", *Annals of Tourism Research*, Vol. 88, p. 103155, doi: 10.1016/j.annals.2021.103155.

Sarker, I.H. (2021), "Machine learning: algorithms, real-world applications and research directions", *SN Computer Science*, Vol. 2 No. 3, doi: 10.1007/s42979-021-00592-x.

Song, H., Qiu, R.T.R. and Park, J. (2019), "A review of research on tourism demand forecasting", *Annals of Tourism Research*, Vol. 75, pp. 338-362, doi: 10.1016/j.annals.2018.12.001.

Starosta, K., Budz, S. and Krutwig, M. (2019), "The impact of German-speaking online media on tourist arrivals in popular tourist destinations for Europeans", *Applied Economics*, Vol. 51 No. 14, pp. 1558-1573, doi: 10.1080/00036846.2018.1527463.

Sun, S., Wei, Y., Tsui, K. and Wang, S. (2019), "Forecasting tourist arrivals with machine learning and internet search index", *Tourism Management*, Vol. 70, pp. 1-10, doi: 10.1016/j.tourman.2018.07.010.

Tyralis, H. and Papacharalampous, G. (2017), "Variable selection in time series forecasting using random forests", *Algorithms*, Vol. 10 No. 4, doi: 10.3390/a10040114.

Yang, X., Pan, B., Evans, J.A. and Lv, B. (2015), "Forecasting Chinese tourist volume with search engine data", *Tourism Management*, Vol. 46, pp. 386-397, doi: 10.1016/j.tourman.2014.07.019.

Yang, Y., Pan, B. and Song, H. (2014), "Predicting hotel demand using destination marketing organization's web traffic data", *Journal of Travel Research*, Vol. 53 No. 4, pp. 433-447, doi: 10.1177/0047287513500391.

Volchek, K., Liu, A., Haiyan, S. and Buhalis, D. (2019), "Forecasting tourist arrivals at attractions: search engine empowered methodologies", *Tourism Economics*, Vol. 25 No. 3, pp. 425-447, doi: 10.1177/1354816618811558.

Yao, L., Ma, R. and Wang, H. (2021), "Baidu index-based forecast of daily tourist arrivals through rescaled range analysis, support vector regression, and autoregressive integrated moving average", *Alexandria Engineering Journal*, Vol. 60 No. 1, pp. 365-372, doi: 10.1016/j.aej.2020.08.037.

Zhou-grundy, Y. and Turner, L.W. (2015), "The challenge of regional tourism demand forecasting: the case of China", *Journal of Travel Research*, Vol. 53 No. 6, pp. 747-759, doi: 10.1177/0047287513516197.

Corresponding author

Meditya Wasesa can be contacted at: meditya.wasesa@sbm-itb.ac.id