Advancing tourism demand forecasting in Sri Lanka: evaluating the performance of machine learning models and the impact of social media data integration

Isuru Udayangani Hewapathirana

Abstract

Purpose – *This study explores the pioneering approach of utilising machine learning (ML) models and integrating social media data for predicting tourist arrivals in Sri Lanka.*

Design/methodology/approach – *Two sets of experiments are performed in this research. First, the predictive accuracy of three ML models, support vector regression (SVR), random forest (RF) and artificial neural network (ANN), is compared against the seasonal autoregressive integrated moving average (SARIMA) model using historical tourist arrivals as features. Subsequently, the impact of incorporating social media data from TripAdvisor and Google Trends as additional features is investigated.*

Findings – *The findings reveal that the ML models generally outperform the SARIMA model, particularly from 2019 to 2021, when several unexpected events occurred in Sri Lanka. When integrating social media data, the RF model performs significantly better during most years, whereas the SVR model does not exhibit significant improvement. Although adding social media data to the ANN model does not yield superior forecasts, it exhibits proficiency in capturing data trends.*

Practical implications – *The findings offer substantial implications for the industry's growth and resilience, allowing stakeholders to make accurate data-driven decisions to navigate the unpredictable dynamics of Sri Lanka's tourism sector.*

Originality/value – *This study presents the first exploration of ML models and the integration of social media data for forecasting Sri Lankan tourist arrivals, contributing to the advancement of research in this domain.*

Keywords *Tourism demand forecasting, Social media analytics, Machine learning, Support vector regression, Random forest, Artificial neural network, Sri Lanka*

Paper type *Research paper*

Isuru Udayangani Hewapathirana is based at the Faculty of Science, University of Kelaniya, Kelaniya, Sri Lanka.

Received 26 June 2023
Revised 24 August 2023
13 October 2023
Accepted 11 November 2023

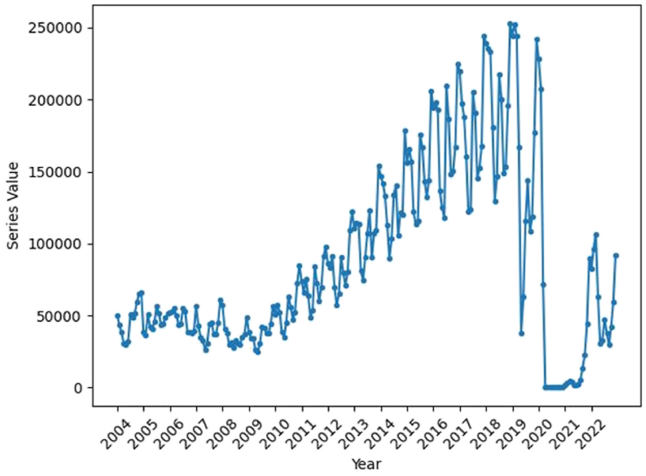
The author extends her sincere gratitude to the reviewers for their valuable feedback and acknowledges the significant contributions made in improving the manuscript. The author also acknowledges the financial support received for this study by the University Research Grant RP/03/02/09/01/2022 from the University of Kelaniya. The support provided by this grant has been instrumental in conducting and completing the research.

1. Introduction

Accurate forecasting of tourism demand is pivotal in maintaining a sustainable tourism industry. Inaccurate predictions can lead to issues such as over- or under-supply of essential services like food, accommodations and infrastructure within the destination, disrupting the delicate balance of the tourism ecosystem (Zhang *et al*., 2021). Tourism demand in Sri Lanka is inherently volatile due to external factors such as geopolitical events and economic fluctuations in key source markets. As illustrated in Figure 1, unexpected events like the 2019 Easter Sunday explosions and the economic impacts of the COVID-19 pandemic have exerted adverse effects on tourist arrivals to the country. These sudden disruptions emphasise the need for sophisticated forecasting models robust to unexpected changes.

DOI 10.1108/JTF-06-2023-0149

Figure 1 Tourist arrivals from January 2004 to December 2022



Source(s): Figure by authors

Recent advancements in forecasting tourism demand have seen a surge in the adoption of machine learning (ML) models (Höpken *et al.*, 2021; Li *et al.*, 2021; Li *et al.*, 2022). These models can potentially enhance the prediction accuracy of tourist arrivals by leveraging their ability to capture non-linear relationships and unveil hidden insights. Previous research has demonstrated the capacity of ML algorithms, such as artificial neural networks (ANNs) and support vector machines, to surpass conventional forecasting methods in prediction accuracy (Mishra *et al.*, 2021). Moreover, the accessibility of rich social media data sources provides an opportunity to enhance predictions further by incorporating additional variables derived from online reviews and tourist discussions (Fronzetti Colladon *et al.*, 2019; Li *et al.*, 2020). While ML methods and social media data have been successfully incorporated for tourism demand predictions in countries such as Thailand (Taecharungroj and Mathayomchan, 2019), China (Li *et al.*, 2020) and Indonesia (Andariesta and Wasesa, 2022), these methodologies have not yet been employed to forecast tourism demand for Sri Lanka. Thus, applying ML models with the integration of social media data for tourism demand forecasting in Sri Lanka holds particular significance. It helps to achieve more accurate predictions and make informed business decisions.

Considering tourist arrivals as the primary measure to represent tourism demand, this study focusses on two primary research objectives. The first objective aims to assess the comparative performance of ML models against conventional time series forecasting models in predicting tourist arrivals to Sri Lanka. Time series models have long been a popular tool for authors to forecast tourist arrivals to Sri Lanka (Konarasinghe, 2016). Since the tourism demand in Sri Lanka is characterised by its seasonality, influenced by the country’s tropical climate and cultural festivals (Kodippili and Senaratne, 2017), the vast majority of those papers forecast future tourist arrivals using historical tourist arrival data, often relying on the well-established seasonal autoregressive integrated moving average (SARIMA) model (Basnayake and Chandrasekara, 2022; Kodippili and Senaratne, 2017; Thushara *et al.*, 2016). Thus, the SARIMA model was chosen as the baseline traditional forecasting model to evaluate the first research objective, and three ML models were selected to assess their performance. These models include support vector regression (SVR), random forest (RF) regression and ANN models.

Social media platforms are rich real-time data sources that can provide valuable insights into tourist behaviour and trends. The second part of the study seeks to determine whether the ML models initially trained on historical tourist arrivals as their sole input can attain improved predictive accuracy when integrated with social media data. Thus, the second objective investigates the potential enhancement in predictive accuracy that can be achieved by incorporating social media data into ML models. This investigation aims to identify the potential benefits of utilising big data sources for improving the predictive accuracy of ML models.

Given the complex dynamics of Sri Lanka's tourism landscape, integrating ML methods and social media data into demand forecasting will offer profound benefits for multiple stakeholders. Sri Lanka's tourism authorities can make more informed decisions regarding resource allocation, price regulation and infrastructure development. Local businesses can better anticipate tourist flows, adjusting their services accordingly. Travellers can benefit from more accurate forecasts for planning their trips. With tourism contributing approximately 10% to Sri Lanka's gross domestic product (Samarathunga, 2020), this approach has the potential to ultimately contribute to the nation's economic growth by bolstering the tourism industry. According to our knowledge, this research is the first attempt to combine ML and social media data for predicting tourist arrivals to Sri Lanka.

2. Literature review

2.1 Traditional models for tourism demand forecasting

Several studies demonstrate the utility of conventional time series models in forecasting tourist arrivals in Sri Lanka, highlighting the significance of accounting for seasonality in prediction models. With a data set spanning the years January 2000 to February 2016 (Thushara *et al.*, 2019), forecasted the number of foreign tourists arriving in Sri Lanka using ARIMA and MDA techniques. According to their recommendations, the decomposition approach was effective in terms of the fitted models' accuracy, while the post-war dataset produced more accurate forecasts for both models. In contrast, Basnayake and Chandrasekara (2022) undertook an expansive evaluation of diverse time series models, and found that the SARIMA (1,1,11)(2,1,3)[3] model was the best for forecasting tourist arrivals. Konarasinghe (2016) used descriptive statistics, time series plots and auto-correlation functions to identify patterns in tourist arrivals from Asia, Western Europe, Eastern Europe and the Middle East regions from January 2008 to December 2014. They recommended testing various models such as moving average methods, exponential smoothing techniques, decomposition techniques, linear and non-linear trend models and circular models for forecasting. Kurukulasooriya and Lelwala (2014) employed classical time series decomposition techniques on a post-war international tourist arrival dataset and found that seasonality is a prominent component in international tourist arrivals, with June, July and October being the most popular months. Priyangika (2016) employ data from 2000 January to 2014 December to fit ARIMA and GARCH models to forecast future tourist arrivals in Sri Lanka. ARCH-1 model with optimal lag (2, 7 and 12) is identified as the most accurate model for prediction. In order to forecast international visitor arrivals to Sri Lanka from the top 10 nations, Diunugala and Mombeuil (2020) compared three different methods and found that Winter's exponential smoothing and ARIMA were the best methods to forecast tourist arrivals to Sri Lanka.

The majority of the aforementioned publications rely on tourist arrival data preceding 2019 to construct their time series models and perform forecasts. As a result, these models do not capture the impact of external events such as the COVID-19 pandemic and the subsequent economic downturn experienced after 2019, which had a profound effect on the country's tourism industry. Consequently, the forecasts generated by these models, incorporating data beyond 2019, may not accurately represent the actual future conditions. To tackle this concern, this research employs a wider temporal scope, encompassing data from 2004 until the end of 2022. This extended timeframe is selected precisely to address the temporal disparity observed in earlier research, allowing for a more comprehensive and nuanced understanding of predicting tourism demand within the context of evolving external dynamics.

2.2 Machine learning models for tourism demand forecasting

Literature indicates that adding explanatory variables to univariate time series models may enhance forecasts (Jassim *et al.*, 2023). However, the limited availability of data types that can be used as explanatory variables in traditional forecasting models makes it challenging to implement these models. Autoregressive prediction models are used in causal econometric approaches to time series modelling (Lütkepohl and Krätzig, 2004). These models forecast future arrivals using data from past arrivals. Therefore, the target attribute, tourist arrivals, has also been employed in the model with varying time delays as input variables (Höpken *et al*., 2017). Increasingly, ML models are used in tourism research because they are more flexible and can estimate non-linear relationships without the limitations of conventional methods (Hadavandi *et al.*, 2011). This shift to ML approaches holds promise for capturing the complexities of tourist arrival dynamics, particularly in situations where multiple variables interact to influence the outcome.

The ANN, capable of handling nearly any non-linearity, is the most commonly employed ML-based model (Andariesta and Wasesa, 2022). Several times, the ANN has outperformed decomposition, exponential averaging, ARIMA and multiple regression in predicting tourism demand (Song *et al*., 2008). According to (Aslanargun *et al*., 2007), out of the ARIMA, linear ANN, multilayer perceptron (MLP) and radial basis function network (RBFN) models, those with non-linear components are superior at predicting future tourist arrivals. Contrary to the latter conclusion, Yu and Schwartz (2006) discovered that complex models employing fuzzy time series and grey theory are unlikely to produce more accurate forecasts than traditional simple models. They suggest that forecasters in the tourism industry should only employ these two methods with serious consideration. Support vector machine (SVM) is applied for tourism demand analysis by (Mishra *et al*., 2021). SVM solves classification, non-linear regression estimation and prediction problems. SVR model is a type of SVM that can be used to predict a continuous value. SVR was more effective than ARIMA models in predicting tourism demand, according to empirical evidence (Song *et al*., 2008). The RF is another AI-based model that has gained popularity due to its stability and interpretability of results concerning a variety of disciplines (Andariesta and Wasesa, 2022).

2.3 Tourism demand forecasting with social media data

In recent research, it has been found that incorporating social media data can significantly enhance the accuracy and predictive power of tourist arrival forecasting models. For instance, Höpken *et al*. (2021) demonstrated that incorporating Google Trends data improved prediction performance compared to traditional autoregressive approaches relying solely on past arrivals. Similarly, Wickramasinghe and Ratnasiri (2021) utilise regionally and temporally disaggregated tourism data along with Google search data to predict tourism demand in Sri Lanka for the post-COVID period. The forecasting models used included time series models such as SARMAX, ARIMA, ARIMAX and SARIMA models. A multisource Internet data approach is recommended for predicting international tourist arrivals to Indonesia during COVID-19 by Andariesta and Wasesa (2022), where ML models such as ANNs, SVR and RF were trained by incorporating data from TripAdvisor travel forum and Google Trends as predictors, considering temporal factors, posts and comments, search queries index and previous tourist arrivals records. Höpken *et al*. (2017) compared autoregressive and big data-based approaches for tourism demand prediction, demonstrating that the inclusion of big data information sources significantly improved prediction accuracy. The empirical findings by Li *et al*. (2020) suggest that incorporating data from multiple sources such as the Baidu search engine and online review platforms such as Ctrip and Qunar provides a more comprehensive and accurate representation of tourist behaviour and preferences. Adding big data information sources such as destination price levels and web search traffic per sending country to historical tourist arrivals and employing ML techniques, such as K-nearest-neighbours have proven beneficial to the accuracy of the prediction and the identification of non-linear relationships when predicting tourist arrivals, Höpken *et al*. (2017); Gunter and Önder (2016) examined on how adding search engine data could enhance the accuracy of forecasts as indicated by an overall reduction in errors such as root mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error for longer forecasting horizons such as 3, 6 and 12 months ahead.

The results of prior studies taken together highlight the importance of including social media data in tourism forecasting models, as it not only improves the precision of estimates of tourist arrivals but also offers an improved understanding of tourist behaviour.

2.3.1 Ethical considerations of using social media data. While integrating social media data offers substantial benefits to tourism forecasting, it also raises common concerns and requires careful practical considerations. Ethical challenges encompass issues of data privacy, consent and user anonymity, particularly as data extracted from social media often involves personal opinions and preferences. Adhering to platforms' terms of use and ensuring data anonymization are pivotal to upholding ethical research practices (Moreno *et al.*, 2013). Moreover, the dynamic nature of social media necessitates acknowledging data biases, potential misinformation and data noise. Researchers must address issues of data quality and reliability while effectively managing the diverse and unstructured nature of social media data (Reda and Zellou, 2023). It is essential to implement robust data pre-processing techniques and validation processes to mitigate such challenges and improve the data's credibility. Researchers must also grapple with the ever-evolving legal landscape governing data usage, including intellectual property rights, copyright and jurisdiction-specific regulations. Large social media datasets may be compiled (e.g. by scraping or APIs) to enable high-powered analyses that are mostly immune to the self-selection biases encountered in direct research; nevertheless, the morality and legality of such practices are debatable (Fiesler *et al.*, 2020). Striking a balance between research goals and legal compliance ensures that the insights derived from social media data contribute responsibly to the field. By demonstrating a commitment to responsible data acquisition, privacy protection and rigorous data processing, researchers can harness the power of social media data while ensuring the rights and interests of individuals are upheld. This will pave the way for more accurate, insightful and ethically sound tourism forecasting practices.

2.4 Moving window approaches

Multiple studies investigate the use of moving window validation for forecasting the arrival of tourists (Adil *et al.*, 2021; Hu *et al.*, 2021). Moving window validation is a method of cross-validation that uses a rolling window of data to predict the next period. The accuracy of the model is then evaluated by comparing the forecast to the actual data. Depending on the forecasting model and the duration of the movable window, the accuracy of the forecasts may vary. A moving window-based validation strategy is less sensitive to changes in time-series data than a one-time method because the moving window method continually updates forecasts based on the most recent information (Arlot and Celisse, 2010). Moreover, because a moving window approach does not require storing the entire dataset in memory, it is more suitable and efficient for large datasets. This study also integrates a moving window validation methodology to rigorously evaluate the predictive performances of the developed models.

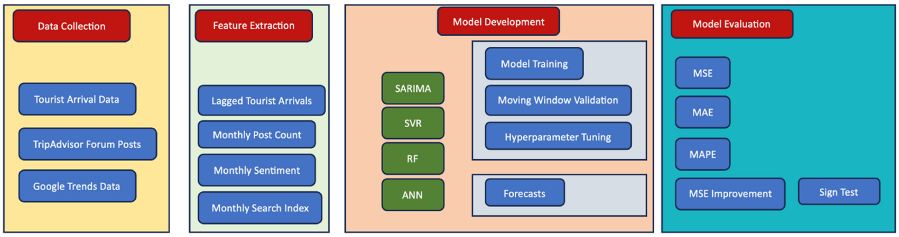
3. Methodology

The research framework for this study is illustrated in Figure 2 and consists of four main stages: (1) collecting data, (2) extracting features, (3) developing models and (4) evaluating the models.

3.1 Data collection

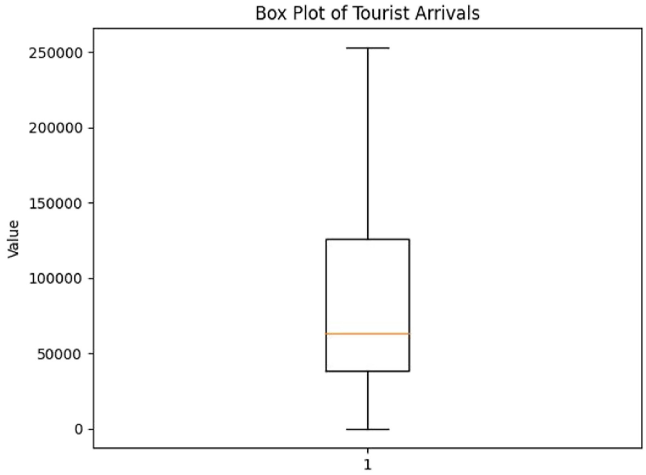
3.1.1 Tourist arrival data. The graphical representation depicted in Figure 1 displays the number of tourists who visited Sri Lanka during the period spanning from January 2004 to December 2022. The tourist arrivals observed between the year 2004 and the middle of 2011 show a non-constant pattern with a downward trend. The decrease in arrivals could potentially be attributed to the harmful impact of terrorism that was encountered during that time within the region. A gradual upward trend is observed from the end of 2011 to the beginning of 2019. Based on the boxplot of the data presented in Figure 3, it can be observed that the tourist arrivals depict a positively skewed

Figure 2 Research overview



Source(s): Figure by authors

Figure 3 Boxplot of tourist arrivals



Source(s): Figure by authors

distribution. This suggests that the majority of tourist arrivals are of a relatively small magnitude, while a minority of them are of a larger magnitude. The dataset was free of any outliers and missing values. The tourist arrivals exhibited a minimum value of 0 in April 2000 up to November 2020, and a maximum value of 253,169 in December 2018. The mean number of tourists arriving in Sri Lanka was 87,757. Between 2004 and 2022, a cumulative sum of 20,008,582 tourists arrived, with a corresponding standard deviation value of 64,461.

3.1.2 Google Trends data. Google Trends data was acquired from the website, https://trends.google.com/trends, utilising the search terms that are explained in the subsequent paragraph. Monthly Google data were collected to align with the availability of tourist arrival data, which are also reported on a monthly basis. The Google dataset can be accessed from January 2004 onwards and is presented in a normalised index format with a value range of 0–100. The computation of the index involves the division of the aggregate count of queries for a specific search term within a designated geographical region by the total count of searches conducted within the same geographic location over a specified duration. In the current study, global search query patterns were chosen to predict international tourist arrivals without limiting the predictions to a specific arrival destination. This normalisation accounts for the variability in the total search volume and helps in making the data comparable across different search terms and time periods. Given the focus of this research is on capturing the trend and identifying patterns in the search data, the normalised index, even if it may reduce absolute variability, is indeed sufficient for the research's purpose.

Three variables were extracted from the data collected using Google Trends. They are three monthly series, GTFlights, GTHotels and GTVisa, representing search volumes in Google Trends for three popular keywords, "Sri Lanka Flights', 'Sri Lanka Hotels' and "Sri Lanka Visa', respectively. The selection of relevant keyword phrases is a crucial step in incorporating search data into forecasting models (Park *et al*., 2017). These three keywords have been shown to correlate well with Sri Lankan tourist arrivals in a recent study (Wickramasinghe and Ratnasiri, 2021) and hence selected for the current study.

3.1.3 TripAdvisor data. Utilising a data scraping method facilitated by the Google API, specific details from TripAdvisor posts were gathered to construct a comprehensive dataset. The collected data includes the title of the post, date posted, name of forum and post content. It is important to note that during this data collection process, user details were not collected. As a result, the extracted data remains completely anonymous, ensuring that no personally identifiable information is accessible. This stringent approach aligns meticulously with the platform's terms of use and other pertinent legal requirements, upholding the ethical and privacy considerations of utilising such data sources.

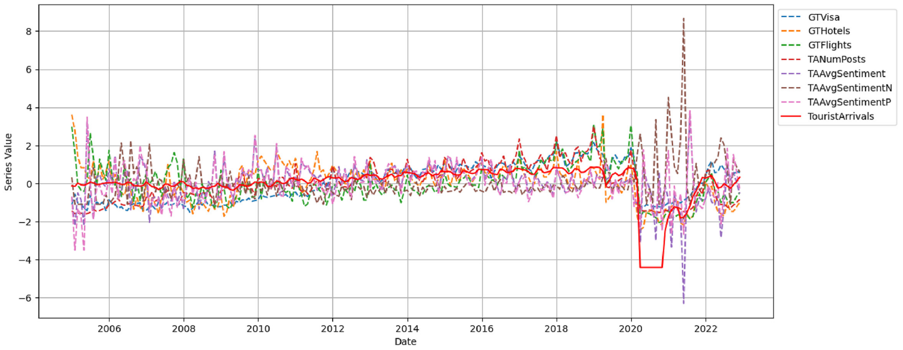
A total of 56,089 TripAdvisor post topics from the Sri Lanka forum have been gathered, and four monthly time series from January 2004 to December 2022 were extracted using this dataset. One series is the total number of forum posts per month (TANumPosts). In relation to the remaining series, a sentiment score was allocated to every post through the utilisation of VADER, a sentiment calculation approach based on lexicon and rules (Hutto and Gilbert, 2014). The adaptability and efficacy of the VADER algorithm in enhancing ML models across diverse domains have been extensively deliberated (Jiang *et al*., 2017; Marutho *et al*., 2022; Rajangam *et al*., 2022). The VADER algorithm provides three sentiment scores for each post; the positive score, the negative score and the overall sentiment score as the sum of the positive and negative sentiments. The accuracy of the sentiment scores was tested by cross-validating a subset of randomly selected posts with human annotations to ensure its reliability. The consistency and high agreement between the algorithm-generated sentiment scores and the human annotations validated the accuracy and robustness of the sentiment analysis. Based on the scores, the mean negative, positive and overall sentiment score is computed for every month. Three monthly series are derived from the sentiment information. The average monthly positive sentiment rating (TAAvgSentimentP), average monthly negative sentiment rating (TAAvgSentimentN) and the average monthly overall sentiment rating (TAAvgSentiment) of Sri Lanka-related posts on the TripAdvisor travel forum.

Figure 4 illustrates the log-transformed tourist arrival series (as detailed in Section 3.1.1.), accompanied by three series from Google trends data and four series from TripAdvisor data. These datasets were standardised for visualisation by subtracting the mean and dividing by the standard deviation. Importantly, it is evident that the depicted series follows a consistent trend with tourist arrivals, rendering them suitable for integration into the ML models to forecast tourist arrivals accurately. To substantiate these observations, Pearson correlations between the tourist arrival series and others were calculated and are presented in Table 1. As expected, tourist arrivals exhibit a negative correlation with the negative sentiment scores of TripAdvisor posts while displaying a positive correlation with the other series. This observation provides insight into the potential data preparation steps before fitting the models. Further discussions in this regard are provided in Sections 3.2. and 3.3.

3.2 Machine learning models

The experiments conducted in this research use three ML models, SVR, RF and ANN, and one traditional time series model, the SARIMA model.

Figure 4 Time series of social media features with tourist arrivals



Source(s): Figure by authors

Table 1 Pearson correlations between feature series and tourist arrivals

Series	*Correlation*
GTVisa	0.5506
GTHotels	0.5708
GTFlights	0.5698
TANumPosts	0.6119
TAAvgSentiment	0.3024
TAAvgSentimentN	−0.2534
TAAvgSentimentP	0.2283

Source(s): Table by the author

3.2.1 Seasonal autoregressive integrated moving average model. The SARIMA model (Box *et al.*, 1974) is a widely used time series forecasting model incorporating autoregressive, differencing and moving average components. It is an extension of the ARIMA model that accounts for seasonal patterns in the data. In this study, the “auto.arima” function available in Python is used to optimise the parameters of the SARIMA model. Overall, the parameters of the model to be optimised are the number of seasonal autoregressive (P) and moving-average (Q) terms, non-seasonal autoregressive (p) and moving-average (q) terms and seasonal and non-seasonal differencing terms, D and d. The “auto.arima” function evaluates various combinations of seasonal and non-seasonal ARIMA parameters on the training dataset. It selects the model with the lowest Akaike information criterion (AIC) value, indicating the best fit to the data.

3.2.2 SVR model. SVR model (Smola and Schölkopf, 2004) utilises the principles of SVMs for regression tasks. SVR aims to find a hyperplane that best fits the training data while also considering a specified error tolerance. The SVR model can capture linear and non-linear relationships between input features and target variables through kernel functions. The SVR model utilises a kernel function to transform the input features into a higher-dimensional space, enabling the modelling of non-linear relationships between the input and output variables. This allows the SVR to capture complex patterns and make accurate predictions. The choice of kernel function plays a crucial role in determining the mapping of the data points. Different kernel functions, such as the linear kernel, polynomial kernel, radial basis function (RBF) kernel and sigmoid kernel, offer various mappings with different properties. To tune the hyperparameters of the SVR model, the GridSearchCV function provided by the Python scikit-learn library is used. The hyperparameters tuned in this study are C (the regularisation parameter), epsilon (tolerance) and the kernel function.

3.2.3 Random forest model. The RF regression model (Biau and Scornet, 2016) is a powerful ensemble learning algorithm that combines multiple decision trees to make accurate predictions. In this model, each decision tree is constructed using a random subset of the training data and a random subset of the input features. The RF model benefits from two key properties: randomness and ensemble. The randomness in feature selection and data sampling helps to reduce overfitting and increase model generalisation. By combining the predictions from multiple trees, the RF model can capture complex relationships and handle noisy or irrelevant features. The RF model involves several hyperparameters that can be tuned to optimise the model's performance. These include the maximum number of features considered for each split, the maximum depth of the decision trees and the number of trees in the forest. The maximum number of features parameter determines the maximum number of features considered at each split point. This parameter plays a crucial role in controlling the diversity and randomness of the individual decision trees. Limiting the number of features helps to reduce overfitting and improves the model's generalisation ability. The maximum depth parameter sets the maximum depth or level of each decision tree in the forest. It regulates the complexity and capacity of the model. A larger maximum depth allows the model to capture more intricate patterns in the data and increases the risk of overfitting. The number of trees parameter specifies the total number of decision trees in the RF. Increasing the number of trees generally leads to better model performance, up to a point where the improvement becomes marginal. This study performs hyperparameter tuning using GridSearchCV in Python to identify the optimal values for these hyperparameters.

3.2.4 Artificial neural network. An ANN (Tealab, 2018) is a computational model inspired by the structure and functioning of biological neural networks. It consists of interconnected artificial neurons, also known as nodes or units, organised into layers. Each neuron receives inputs, processes them and produces an output. The network's architecture, which includes the number of layers and the number of neurons in each layer, can vary depending on the problem being addressed. The number of hidden layers, learning rate and alpha are important hyperparameters to tune when training ANNs. The number of hidden layers determines the depth and complexity of the network, allowing it to learn intricate patterns and representations. The learning rate controls the step size at each iteration during the training process, influencing the speed of convergence and the quality of the learnt weights. Alpha, also known as the regularisation parameter, helps prevent overfitting by adding a penalty term to the loss function. This study utilises GridSearchCV function of Python scikit-learn library to optimise these hyperparameters.

3.3 Experimental framework

This study has two research objectives, as discussed in Section 1. The first objective is to determine whether ML models are superior to conventional forecasting models for forecasting tourist arrivals. The second objective is to determine if incorporating social media data improves the foresting accuracy of ML models. The following sections, 3.3.1. and 3.3.2., discuss the two experimental setups used to achieve the desired research outcomes respectively.

3.3.1 Research objective 1 (RO1). The dataset containing only monthly tourist arrivals spanning from 2004 to 2022 is used to assess RO1. Log-transformation is a common approach to mitigate non-stationarities in time series data before model fitting, as supported by existing literature (Basnayake and Chandrasekara, 2022). Thus, $\log_{10}(\text{series}+10)$ transformation of the tourist arrivals series is utilised before model fitting. A numerical increment of 10 was implemented to rectify the '0' entries in the recorded data during the time period spanning from April 2020 to November 2020. The SARIMA model is the conventional benchmark forecasting model, whereas three ML models, SVR, RF and ANN are employed as proficient alternatives. The SARIMA methodology employs the logarithmically transformed time series of tourist arrivals to estimate the model. To fit the ML models, distinct time lags of the tourist arrival series that have been log-transformed are calculated and included as input variables (also known as features) for the model. A maximum of six lags are employed for this purpose.

The estimation of each model was performed using a moving window out-of-sample forecasting technique that discards distant observations and considers recent observations to build the model. In other words, an initial sample of data from $t = 1, 2, \ldots, w$, is used to estimate the models and generate out-of-sample forecasts h' steps ahead of the sample. The window of size w is then incrementally advanced by s time periods, and the models are re-estimated using data from $t = w + s, w + s + 1, \ldots, w + s + w$. In this experimental setup, we use $w = 180$, $s = 12$ and $h' = 12$. This experimental setup allows us to obtain model forecasts for years 2018, 2019, 2020, 2021 and 2022. The model parameters are optimised dynamically at each stage of the rolling window of size w months using a hyperparameter grid search. Table 2 displays the outcomes of optimising the parameters of the SARIMA model, and Table 3 displays the optimal hyperparameters of the ML models. The SVR, RF and ANN models employed for this set of experiments are named RF_0, SVR_0 and ANN_0, respectively.

3.3.2 Research objective 2 (RO2). To evaluate the second research objective (RO2), three ML models, SVR (Section 3.2.2), RF (Section 3.2.3) and ANN (Section 3.2.4), are employed with two different sets of features to predict the log-transformed series of tourism arrivals. The construction of the log-transformed tourist arrival series is outlined in Section 3.2.1. The first set of features comprises the monthly tourist arrivals, ranging from lag 1 to lag 6. In addition to the six historical tourism arrival series in the first feature set, the seven series derived from social media data are included in the second feature set. As explained in Sections 3.1.2 and 3.1.3, the aforementioned series comprises three monthly series that correspond to search volumes in Google Trends for three distinct keywords (GTVisa, GTHotels and GTFlights), and four series that correspond to the monthly count of the number of posts (TANumPosts), average monthly positive sentiment rating (TAAvgSentimentP), average monthly negative sentiment rating (TAAvgSentimentN) and the average monthly overall sentiment rating (TAAvgSentiment) of Sri Lanka-related posts on the TripAdvisor travel forum. Furthermore, three temporal lags for each social media series are also incorporated into the feature set of the second dataset. The selection of the optimal number of lags was based on the travel planning data obtained from the 2018–2019 survey of Sri Lankan outbound international tourists, which was conducted by the Sri Lanka Tourism Development Authority. As per the survey findings, a majority of tourists, i.e. over 80%, initiate the process of pre-planning their trips by fixing their travel schedule well in advance, typically ranging from one to three months. As a result, three temporal lag series for the social media-derived features are calculated.

Estimating each model is executed through a technique of out-of-sample forecasting using a moving window, which resembles the methodology explained in Section 3.3.1. In contrast, in this instance, the window shall progress while retaining the most recent observations from the past. Specifically, the window size will not remain constant as in the previous scenario. Table 3 displays the outcomes of optimising the hyperparameters. The SVR, RF and ANN models employed using the first set of features are named RF_1, SVR_1 and ANN_1, while the SVR, RF and ANN models employed using the second set of features are named RF_2, SVR_2 and ANN_2, respectively.

Table 2 Parameter optimisation of SARIMA model

Forecast period	p	*d*	*q*	P	*D*	Q
2018	0	1	2	3	0	1
2019	3	1	0	1	0	2
2020	1	0	0	1	1	1
2021	0	1	0	0	1	0
2022	1	1	0	1	0	1

Source(s): Table by the author

Table 3 Hyperparameter optimization of ML models

Forecast period	*Model*	*Hyperparameters*	*Model_0*	*Model_1*	*Model_2*
2018	SVR	Kernel	linear	linear	linear
		C	10	10	0.01
		Epsilon	0.01	0.01	0.01
	RF	Maximum features	4	4	10
		Number of trees	10	10	10
		Maximum depth	5	5	10
	ANN	Hidden layers	1	1	3
		Learning rate	0.01	0.01	0.1
		Alpha	0.01	0.01	0.001
2019	SVR	Kernel	linear	linear	linear
		C	1	1	0.1
		Epsilon	0.01	0.01	0.1
	RF	Maximum features	6	6	6
		Number of trees	30	30	10
		Maximum depth	5	10	5
	ANN	Hidden layers	2	1	3
		Learning rate	0.01	0.01	0.01
		Alpha	0.01	0.01	0.01
2020	SVR	Kernel	linear	linear	linear
		C	1	1	0.1
		Epsilon	0.001	0.01	0.01
	RF	Maximum features	5	4	10
		Number of trees	50	10	10
		Maximum depth	10	5	10
	ANN	Hidden layers	1	1	1
		Learning rate	0.01	0.01	0.01
		Alpha	0.01	0.01	0.01
2021	SVR	Kernel	linear	linear	linear
		C	1	1	100
		Epsilon	0.01	0.01	0.001
	RF	Maximum features	5	5	10
		Number of trees	10	10	10
		Maximum depth	10	30	10
	ANN	Hidden layers	2	1	1
		Learning rate	0.01	0.01	0.1
		Alpha	0.01	0.01	0.0001
2022	SVR	Kernel	linear	linear	linear
		C	1	1	0.01
		Epsilon	0.001	0.001	0.1
	RF	Maximum features	4	5	10
		Number of trees	30	10	30
		Maximum depth	10	None	10
	ANN	Hidden layers	1	1	1
		Learning rate	0.01	0.01	0.1
		Alpha	0.01	0.01	0.0001

Source(s): Table by the author

3.4 Feature engineering

To pre-process the features for the ML methods, scaling techniques are applied to the features within each window in order to mitigate potential discrepancies in scale amongst them. Specifically, three scaling techniques are applied during the process of hyperparameter tuning within the window, and the scaling technique that yields the best performance for the validation set is selected. The objective of this step is to automatically determine the most suitable scaling approach for that selected time window, fostering coherence between variables and contributing to the optimal performance of the model. The techniques employed include Min-Max scaling (Equation (1)), Standard Scaling (Equation (2)) and Robust Scaling (Equation (3)). For an original feature, denoted as X, the transformed feature, $X_{transformed-method}$, using each of these methods can be computed as follows:

$$X_{transformed-MinMax} = \frac{(X - X\_min)}{(X\_max - X\_min)}, \tag{1}$$

$$X_{transformed-Standard} = \frac{(\mathrm{X} - \mathrm{X\_mean})}{\mathrm{X\_std}}, \tag{2}$$

$$X_{transformed-Robust} = \frac{(\mathrm{X} - \mathrm{X\_median})}{IQR}, \tag{3}$$

In the aforementioned equations, X_{min} is the minimum value of the feature, X_{max} is the maximum value of the feature, X_{mean} is the mean of the feature, X_{std} is the standard deviation of the feature, X_{median} is the median of the feature and IQR is the interquartile range, which is the difference between the 75th percentile and the 25th percentile of the feature.

3.5 Performance evaluation

For all models trained in this study, predictions are obtained for each month in the years 2018, 2019, 2020, 2021 and 2022. Critical to the development of prediction models is the evaluation of model performance. It permits to evaluate the accuracy as well as reliability of the model's predictions. The prediction error, which represents the difference between the predicted and actual values, behaves as a performance metric for models (Li *et al.*, 2017). MSE and MAE are employed to quantify the magnitude of prediction errors in this study. The MSE is determined by averaging the squared differences between the predicted and actual values, whereas the MAE is determined by averaging the absolute differences. MSE is particularly valuable as it emphasises the accuracy of the forecasts by penalising larger errors more significantly, which is well-suited for minimising the impact of outliers and extreme values. MAE, on the other hand, provides an intuitive representation of the average magnitude of forecast errors, offering insights into the overall forecast quality (Plevris *et al.*, 2022). Mean absolute percent error (MAPE) is a percentage-based error metric that quantifies the average relative difference between predicted and actual values. MAPE offers a valuable perspective on the accuracy of the forecasts, providing a practical understanding of the predictive performance relative to the actual values (Tofallis, 2015). The following are the equations for calculating MSE (Equation 9), MAE (Equation 10) and MAPE (Equation 11):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{5}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y_i}|}{y_i} \tag{6}$$

Here y_i represents the actual values, $\widehat{y_i}$ represents the predicted values, and n represents the number of samples.

In order to compare the performance of any two models, this study also utilises the forecasting performance improvement (Li *et al.*, 2020). For example, the improvement of using the SVR model compared to the SARIMA model with respect to the MSE can be calculated as,

$$\text{Performance Improvement} = \frac{\text{MSE(SARIMA) - MSE(SVR)}}{\text{MSE(ARIMA)}} \times 100\% \tag{7}$$

Furthermore, it is also assessed if one model outperforms the other in a statistically significant manner using the sign test (Dixon and Mood, 1946). The sign test is a non-parametric statistical test that is particularly suitable when the data does not meet the assumptions required for parametric tests. It allows to statistically evaluate whether one model produces greater forecast errors compared to another model. In this study, the sign test is applied to compare the prediction errors calculated from each model.

4. Results

Table 4 presents a comprehensive analysis of the forecasting performance metrics, performance improvements and sign test results for the three ML models, as part of the experiments undertaken to evaluate RO1. For 2020; 2021, all three ML models consistently outperformed the SARIMA model across all metrics, MSE, MAE and MAPE. This shows that the ML models exhibited higher accuracy and precision in predicting tourist arrivals during these specific years. For 2019, the ML models, SVR_0 and RF_0, demonstrated better performance than the SARIMA model. This observation is particularly relevant as 2019 was marked by unforeseen events such as the Easter Sunday attacks and the global COVID-19 pandemic, significantly impacting tourist demand. The ability of those ML models to produce more reliable and robust forecasts in the presence of such unforeseen occurrences indicates their effectiveness in capturing and adapting to abrupt changes and disruptions due to geopolitical events and economic fluctuations in the tourism industry of Sri Lanka. However, the findings indicate that no significant improvements were observed in forecasting visitor demand for 2018 and 2022 when employing ML models.

Overall, the results presented in Table 4 highlight the enhanced forecasting capabilities of the ML models, particularly during challenging and dynamic periods characterised by unexpected events. These findings emphasise the potential of ML models to outperform traditional forecasting methods like SARIMA when dealing with volatile and uncertain situations in the tourism domain.

Table 5 provides a comprehensive comparison of the performance metrics between models trained with and without incorporating social media data for assessing RO2. Notably, the RF model incorporating social media (RF_2) data exhibits superior performance across all three performance measurements when predicting visitor arrivals in 2018, 2020 and 2022. Specifically, in terms of MSE, MAE and MAPE, the RF_2 model consistently outperforms the same model trained without social media data (RF_1). This indicates that including social media features enhances the RF model's ability to capture and incorporate additional information, leading to more accurate predictions of visitor arrivals in these specific years. Additionally, when examining the sign test *p*-values, the RF_2 model's performance in terms of MSE for the 2022 forecasts stands out significantly, further confirming the substantial improvement achieved by incorporating social media data. Conversely, for the 2018 predictions, the ANN_2 model demonstrates superior performance when trained with social media features across all three performance metrics. On the other hand, the SVR_2 models trained with all features, including social media data, do not exhibit improved performance in predicting visitor arrivals for any given year. However, the SVR_2 models for the 2018 and 2019 predictions show marginal improvements compared to those for 2020, 2021 and 2022.

In evaluating model accuracy when incorporating social media data, including such data consistently led to improved forecasting accuracy for the RF (RF_2) model compared to utilising only historical visitor arrival data (RF_1). The RF model, which leverages the ensemble learning technique, demonstrated the capability to capture complex relationships and patterns in social media data, resulting in more precise predictions. On the other hand, the SVR (SVR-2) model did not exhibit a notable performance improvement when additional features derived from social media data were incorporated. The SVR model, known for its ability to handle non-linear relationships, may have already captured the underlying patterns and dynamics from the historical visitor arrival data, rendering incorporating social media data less impactful in enhancing its predictive accuracy. The contrasting findings between the RF_2 and SVR_2 models suggest that the efficacy of incorporating social media data varies depending on the specific ML algorithm employed. The RF model's flexibility and adaptability to diverse data sources, including social media data, allowed it to exploit valuable insights and enhance its predictive capabilities. In

Table 4 SARIMA vs ML models

Forecast period	*Model*	*Hyperparameters*	*Setting*	*MSE*	*MAE*	*MAPE*	*MSE_Improvement*	*MAE_Improvement*	*MAPE_Improvement*	*p-value*
2018	SVR_0	Kernel	linear	0.0050	0.0599	0.0113	−206.95	−77.23	−0.49	0.8062
		C	10							
		Epsilon	0.01							
	RF_0	Maximum Features	4	0.0076	0.0725	0.0137	−364.59	−114.49	−0.73	0.9807
		Number of trees	10							
		Maximum depth	5							
	ANN_0	Hidden layers	1	0.0188	0.1010	0.0194	−1,044.44	−198.86	−1.30	0.9807
		Learning rate	0.01							
		Alpha	0.01							
2019	SVR_0	Kernel	linear	0.0431	0.1259	0.0256	30.38	29.39	1.08	0.6128
		C	1							
		Epsilon	0.01							
	RF_0	Maximum Features	6	0.0434	0.1482	0.0299	29.92	16.88	0.65	0.6128
		Number of trees	30							
		Maximum depth	5							
	ANN_0	Hidden layers	2	0.1414	0.3449	0.0672	−128.58	−93.46	−3.08	0.9807
		Learning rate	0.01							
		Alpha	0.01							
2020	SVR_0	Kernel	linear	1.8143	0.8266	0.7115	84.86	72.72	213.07	0.0032
		C	1							
		Epsilon	0.001							
	RF_0	Maximum Features	5	9.2318	2.6502	2.5069	22.95	12.53	33.53	0.0032
		Number of trees	50							
		Maximum depth	10							
	ANN_0	Hidden layers	1	3.9124	1.4652	1.3402	67.35	51.64	150.20	0.1938
		Learning rate	0.01							
		Alpha	0.01							

(*continued*)

Table 4 Continued

Forecast period	*Model*	*Hyperparameters*	*Setting*	*MSE*	*MAE*	*MAPE*	*MSE_Improvement*	*MAE_Improvement*	*MAPE_Improvement*	*p-value*
2021	SVR_0	Kernel	linear	0.1198	0.2996	0.0800	99.49	93.19	106.82	0.0002
		C	1							
		Epsilon	0.01							
	RF_0	Maximum Features	5	0.8544	0.7551	0.2187	96.33	82.84	92.96	0.0032
		Number of trees	10							
		Maximum depth	10							
	ANN_0	Hidden layers	2	0.7588	0.6578	0.1792	96.74	85.05	96.90	0.0193
		Learning rate	0.01							
		Alpha	0.01							
2022	SVR_0	Kernel	linear	0.0243	0.1298	0.0277	–123.57	–54.48	–0.96	0.6128
		C	1							
		Epsilon	0.001							
	RF_0	Maximum Features	4	0.0389	0.1794	0.0381	–257.16	–113.57	–2.00	0.9807
		Number of trees	30							
		Maximum depth	10							
	ANN_0	Hidden layers	1	0.0234	0.1328	0.0284	–114.86	–58.15	–1.03	0.8062
		Learning rate	0.01							
		Alpha	0.01							

Source(s): Table by the author

Table 5 Tourist arrival vs big data

Forecast period	*Model*	*Hyperparameters*	*Setting*	*MSE*	*MAE*	*MAPE*	*MSE_Improvement*	*MAE_Improvement*	*MAPE_Improvement*	p-*value*
2018	SVR_2	Kernel	linear	0.0085	0.0729	0.0140	−67.83	−21.64	−0.26	0.6128
		C	0.01							
		Epsilon	0.01							
	RF_2	Maximum Features	10	0.0052	0.0667	0.0126	32.40	7.95	0.11	0.8062
		Number of trees	10							
		Maximum depth	10							
	ANN_2	Hidden layers	3	0.0224	0.1245	0.0233	41.64	27.09	1.23	0.0730
		Learning rate	0.1							
		Alpha	0.001							
2019	SVR_2	Kernel	linear	0.0676	0.1713	0.0349	−57.42	−36.22	−0.94	0.9270
		C	0.1							
		Epsilon	0.1							
	RF_2	Maximum Features	6	0.0465	0.1487	0.0302	−13.46	−3.45	−0.12	0.9270
		Number of trees	10							
		Maximum depth	5							
	ANN_2	Hidden layers	3	0.5337	0.6574	0.1302	−418.14	−195.83	−8.49	1.0000
		Learning rate	0.01							
		Alpha	0.01							
2020	SVR_2	Kernel	linear	8.5122	2.5216	2.4048	−358.54	−197.81	−167.32	0.9270
		C	0.1							
		Epsilon	0.01							
	RF_2	Maximum Features	10	8.8823	2.5724	2.4558	4.56	3.03	5.99	0.0730
		Number of trees	10							
		Maximum depth	10							
	ANN_2	Hidden layers	1	7.5513	2.3738	2.2626	−93.47	−62.31	−92.40	0.9807
		Learning rate	0.01							
		Alpha	0.01							

(*continued*)

Table 5 Continued

Forecast period	*Model*	*Hyperparameters*	*Setting*	*MSE*	*MAE*	*MAPE*	*MSE_Improvement*	*MAE_Improvement*	*MAPE_Improvement*	*p-value*
2021	SVR_2	Kernel	linear	0.4651	0.6094	0.1739	−275.56	−101.12	−9.31	0.9807
		C	100							
		Epsilon	0.001							
	RF_2	Maximum Features	10	3.2892	1.6258	0.4566	−285.34	−113.53	−23.62	0.9998
		Number of trees	10							
		Maximum depth	10							
	ANN_2	Hidden layers	1	0.9389	0.8830	0.2239	−347.62	−146.87	−12.59	0.9998
		Learning rate	0.1							
		Alpha	0.0001							
2022	SVR_2	Kernel	linear	0.1243	0.3231	0.0679	−413.57	−149.25	−4.03	0.9998
		C	0.01							
		Epsilon	0.1							
	RF_2	Maximum Features	10	0.0100	0.0788	0.0169	71.43	52.77	1.86	0.0032
		Number of trees	30							
		Maximum depth	10							
	ANN_2	Hidden layers	1	0.3310	0.4791	0.1027	−763.14	−180.60	−6.71	0.9807
		Learning rate	0.1							
		Alpha	0.0001							

Source(s): Table by the author

contrast, the SVR model's reliance on the inherent patterns of the historical visitor arrival data limited the extent to which social media data contributed to improving its forecasting performance.

Figure 5 presents the forecasts generated by the SARIMA, SVR_0, RF_0 and ANN_0 models for the distinct forecasting horizons of 2018, 2019, 2020, 2021 and 2022, which align with the corresponding results in Table 4. In analysing the performance of the models, it is evident that the ML models showcase superior performance compared to the SARIMA model in terms of adaptability to different forecasting horizons and resilience in the face of unforeseen events. However, it is worth noting that although the SARIMA model is adept at accurately capturing the fluctuations and seasonal patterns in the tourist arrival data during 2018, it needs to be more robust when confronted with unforeseen events such as during 2019, 2020 and 2021. On the other hand, the ML models exhibit better resilience, providing more reliable predictions even in the presence of unexpected disruptions.

Figure 6 presents the forecasts generated by two sets of models, SVR_1, RF_1 and ANN_1 models trained using historical tourist arrivals as features, and SVR_2, RF_2 and ANN_2 models with additional features extracted from social media data. The forecasts are provided for the distinct forecasting horizons of 2018, 2019, 2020, 2021 and 2022, which correspond to the results presented in Table 5. While the SVR_1, RF_1, and ANN_1 models demonstrate a reasonable ability to produce forecasts closer to the actual series, the RF_2 model generally stands out in capturing the upward and downward trends in the data. The ANN_2 models, despite not showing significant performance improvement compared to their counterparts, demonstrate a notable ability to capture the complex dynamics of the data. They exhibit a better understanding of the upward and downward trends, suggesting their potential to capture the underlying patterns influenced by various factors, including the events occurring in 2019, 2020 and 2021.

5. General discussion and conclusion

Despite the widespread use of traditional time series models for tourism forecasting, there is a lack of research on ML-based models for forecasting tourism demand in Sri Lanka. To the best of our knowledge, we were unable to come across any research publications that estimate Sri Lankan tourist arrivals using ML-based methods. Moreover, conventional time series forecasting methods must be re-evaluated considering the unique COVID-19 pandemic environment and economic downturn using data after 2019. This work closes the knowledge gap by utilising historical time series data to develop ML-based prediction models that accurately forecast tourist arrivals. Additionally, integrating social media data, which provides valuable insights into tourist preferences, was examined to improve forecasting accuracy.

In comparison to previous research, this study expands the existing knowledge by experimenting with ML models and incorporating social media data for forecasting Sri Lankan tourist arrivals. While previous studies such as Basnayake and Chandrasekara (2022) have primarily relied on traditional time series models, this research takes a novel approach by leveraging the power of ML algorithms. By comparing the performance of three ML models (SVR, RF and ANN) against the SARIMA model, this study provides valuable insights into the effectiveness of ML models in handling unforeseen events and improving forecasting accuracy. The selection of the SARIMA model as the base model stems from its established reputation amongst traditional time series techniques for adeptly capturing seasonal trends. This choice provides a solid foundation for assessing the potential of ML models in outperforming established methods in a dynamic and often unpredictable tourism landscape.

The findings of this research indicate that ML models with lagged tourist arrivals as features, particularly the SVR and RF models, outperform the traditional SARIMA model during the years 2019, 2020 and 2021, while the ANN model outperforms in 2020 and 2021. This suggests that the seasonality captured by the SARIMA model should not be the sole consideration in forecasting Sri Lankan tourist arrivals. ML models that can provide more accurate and reliable forecasts,

Figure 5 Comparison of prediction results for SARIMA and machine learning models

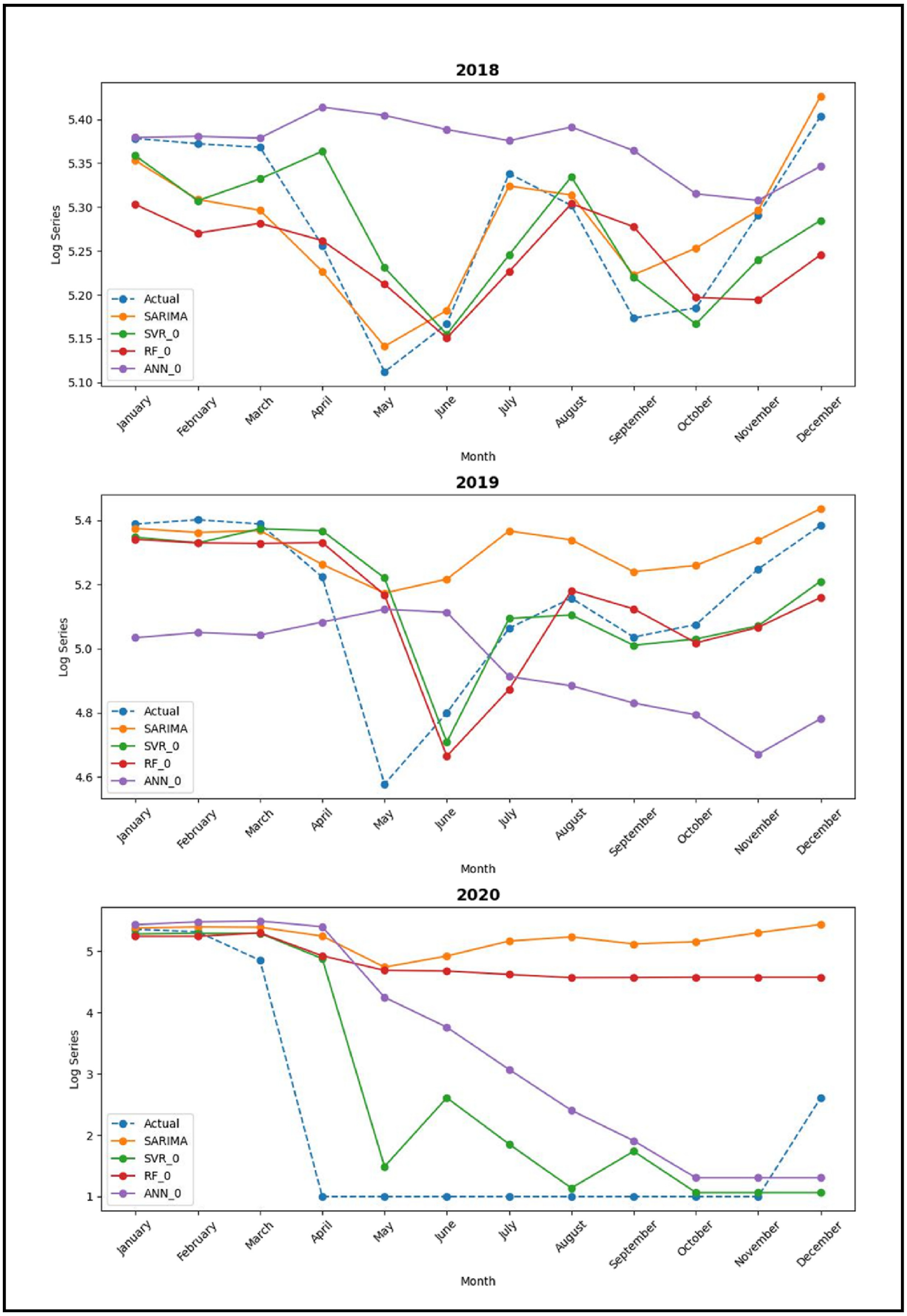
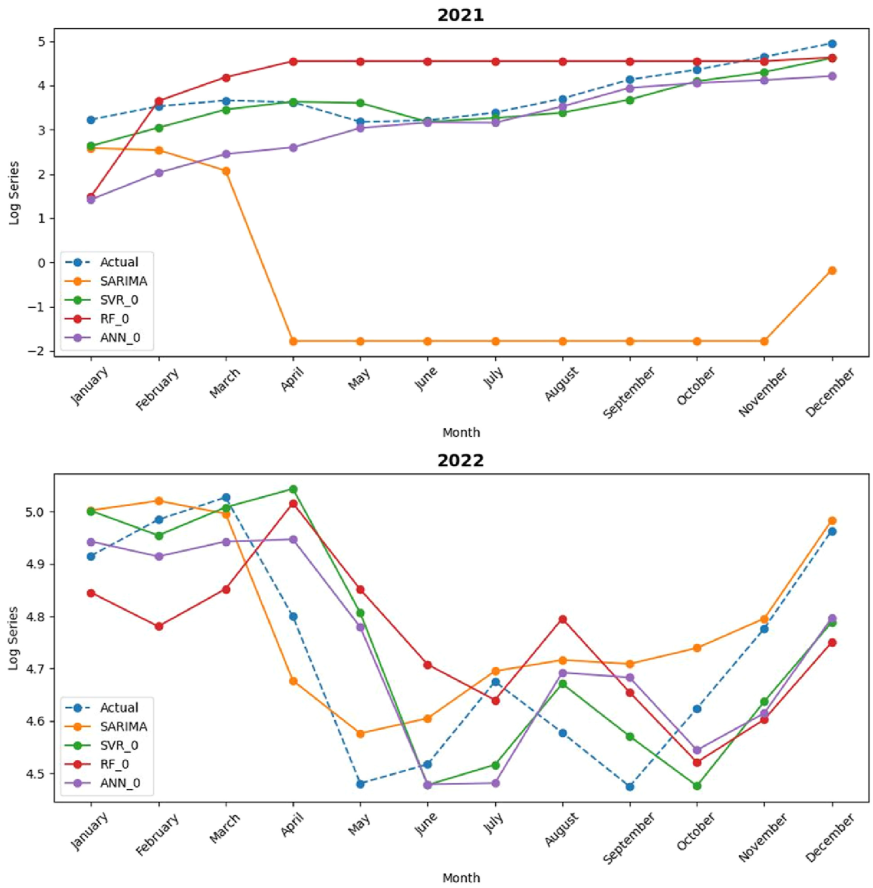


Figure 5 Continued



Source(s): Figure by authors

particularly in unforeseen events, are more suitable for this purpose in Sri Lanka's dynamic tourism environment.

Integrating social media data, specifically from platforms such as TripAdvisor and Google Trends, consistently improves the forecasting performance of the RF model. However, the SVR model does not show significant improvement with the inclusion of social media data. This aligns with the findings of Andariesta and Wasesa (2022), where SVM did not perform well when the dataset contained more noise due to the addition of social media-based features. Therefore, careful selection of the appropriate ML model is crucial as only some methods can outperform others in all forecasting contexts (Li *et al.*, 2020). Future research could explore alternative methods for incorporating social media data into ML models to enhance their predictive power. The potential advantage of the RF model's enhanced performance could be attributed to its inherent feature selection mechanism. This attribute distinguishes it from the SVM and ANN models, which lack an embedded feature selection process. Avenues for future exploration could involve training the SVM and ANN models with feature selection, which may improve their ability to incorporate social media data more effectively. This direction holds promise for refining the predictive power of these models and calls for additional development in later research projects.

Figure 6 Comparative performance of machine learning models with and without social media data in predicting tourist arrivals

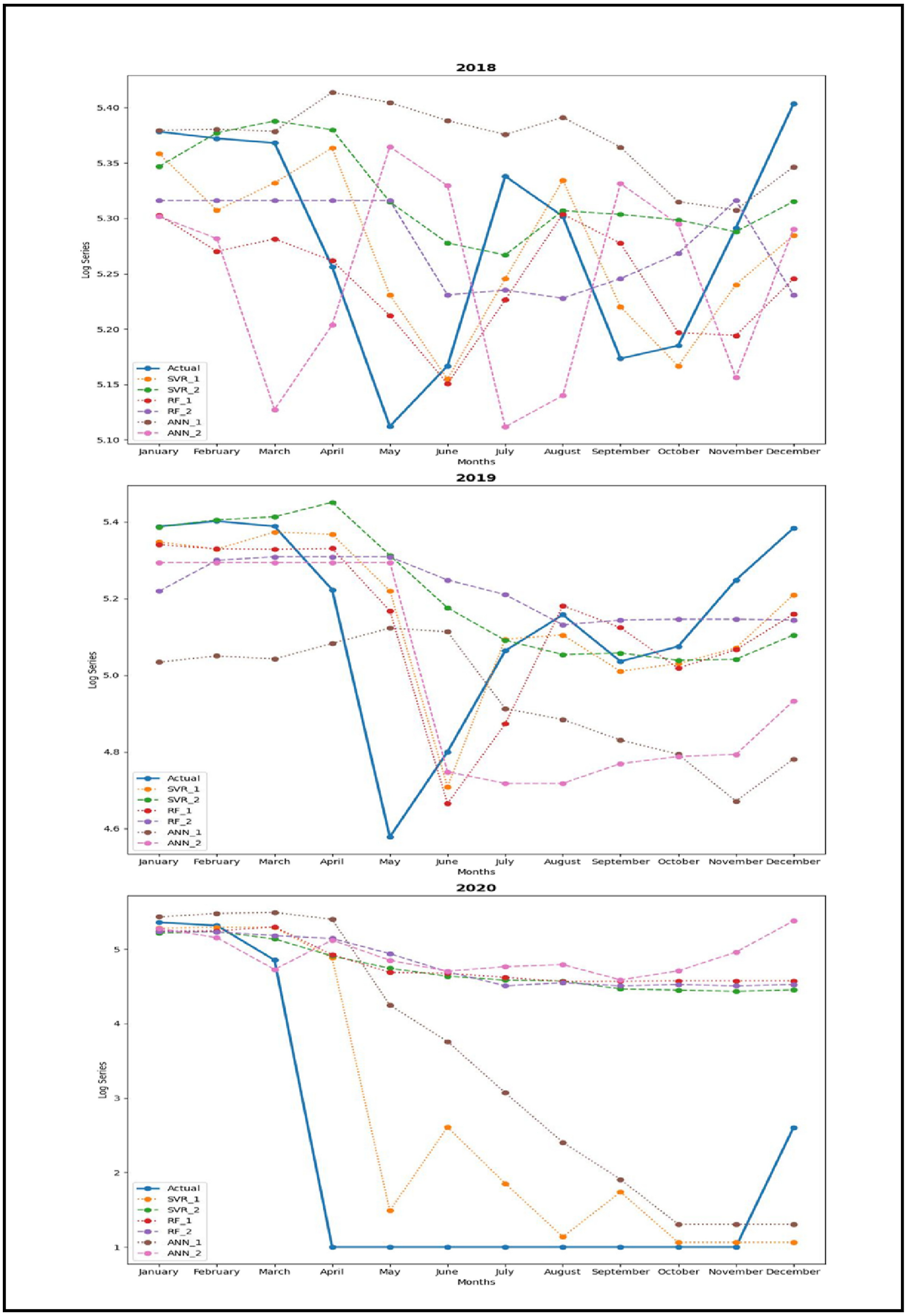
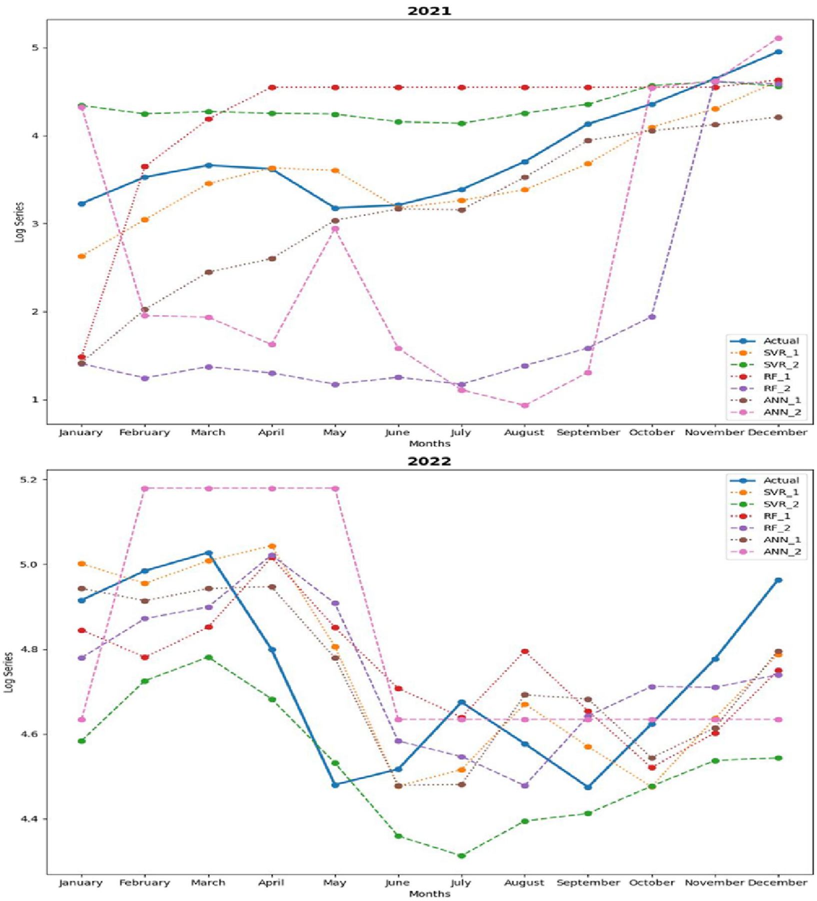


Figure 6 Continued



Source(s): Figure by authors

Despite the contributions of this study, there are a few limitations that warrant further investigation. Firstly, the constructed tourism demand forecasting models only incorporated social media data from Google Trends and TripAdvisor, along with a lagged tourism demand variable, without considering other important influencing factors. This could be a reason why the ANN model with social media data features did not yield superior forecasts. Subsequent studies could examine whether integrating Internet big data from multiple sources (Höpken *et al.*, 2017; Li *et al.*, 2017), alongside traditional influencing factors of tourism demand into one forecasting model improves forecasting accuracy. Secondly, this study only considered two types of social media data to forecast tourism demand. Future studies can extend social media data by incorporating additional types of user-generated social media data, such as information from Facebook, microblogs and Internet discussion forums, to improve the accuracy of tourism demand forecasting further. In addition, incorporating temporal features as dummy variables to incorporate seasonality in ML models could offer distinct advantages, and this avenue might be explored in future work (Balabaeva and Kovalchuk, 2019). However, this should be approached with suitable feature selection methods to ensure that the selected features contribute meaningfully to the model's performance. Lastly, as with most tourism forecasting studies, our research is based on a case study approach, fitting the models to overall international tourist arrivals. Therefore, the findings of this study should not be generalised. Our study illustrates the potential benefits of using ML models and social media data in Sri Lankan tourism forecasting. To draw more generalisable conclusions, further investigations are necessary by incorporating multiple cases focussing on tourist arrivals from different countries or arrivals to specific destinations within the country. Furthermore, other indicators of tourism demand, such as tourist expenditure (Song *et al.*, 2010) or bed-nights (Wickramasinghe and Ratnasiri, 2021), can offer insights into tourists' arrival patterns.

In conclusion, this study demonstrates the potential of ML models and the inclusion of social media data for accurately forecasting tourist arrivals in the Sri Lankan tourism industry. The findings emphasise the superiority of ML models, particularly SVR and RF, over traditional time series models in handling unexpected events and improving prediction accuracy. Integrating social media data notably enhances the RF model's performance, while the ANN effectively captures underlying data trends. This research offers practical implications for tourism attraction and destination managers, enabling them to make data-driven decisions and allocate resources more efficiently. Social media data provides valuable insights into tourist behaviour and preferences, facilitating targeted marketing strategies and personalised experiences. Future research should address data availability challenges, refine model selection, explore advanced algorithms and incorporate additional data sources to enhance forecasting capabilities in the dynamic Sri Lankan tourism industry.

References

Adil, M., Wu, J.Z., Chakrabortty, R.K., Alahmadi, A., Ansari, M.F. and Ryan, M.J. (2021), "Attention-based stl-bilstm network to forecast tourist arrival", *Processes*, Vol. 9 No. 10, p. 1759, doi: 10.3390/pr9101759.

Andariesta, D.T. and Wasesa, M. (2022 In press), "Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach", *Journal of Tourism Futures*, doi: 10.1108/JTF-10-2021-0239.

Arlot, S. and Celisse, A. (2010), "A survey of cross-validation procedures for model selection", *Statistics Surveys*, Vol. 4, pp. 40-79, doi: 10.1214/09-SS054.

Aslanargun, A., Mammadov, M., Yazici, B. and Yolacan, S. (2007), "Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting", *Journal of Statistical Computation and Simulation*, Vol. 77 No. 1, pp. 29-53, doi: 10.1080/10629360600564874.

Balabaeva, K. and Kovalchuk, S. (2019), "Comparison of temporal and non-temporal features effect on machine learning models quality and interpretability for chronic heart failure patients", *Procedia Computer Science*, Vol. 156, pp. 87-96, doi: 10.1016/j.procs.2019.08.183.

Basnayake, B.R.P.M. and Chandrasekara, N.V. (2022), "Use of change point analysis in seasonal ARIMA models for forecasting tourist arrivals in Sri Lanka", *Statistics and Applications*, Vol. 20 No. 2, pp. 103-121.

Biau, G. and Scornet, E. (2016), "A random forest guided tour", *TEST*, Vol. 25 No. 2, pp. 197-227, doi: 10.1007/s11749-016-0481-7.

Box, G.E.P., Jenkins, G.M. and MacGregor, J.F. (1974), "Some recent advances in forecasting and control", *Applied Statistics*, Vol. 23 No. 2, p. 158, doi: 10.2307/2346997.

Diunugala, H.P. and Mombeuil, C. (2020), "Modeling and predicting foreign tourist arrivals to Sri Lanka: a comparison of three different methods", *Journal of Tourism, Heritage and Services Marketing (JTHSM)*, Vol. 6 No. 3, pp. 3-13.

Dixon, W.J. and Mood, A.M. (1946), "The statistical sign test", *Journal of the American Statistical Association*, Vol. 41 No. 236, pp. 557-566, doi: 10.1080/01621459.1946.10501898.

Fiesler, C., Beard, N. and Keegan, B.C. (2020), "No robots, spiders, or scrapers: legal and ethical regulation of data collection methods in social media terms of service", *Proceedings of the 14th International AAAI Conference on Web and Social Media,* ICWSM 2020, pp. 187-196, *Icwsm*, doi: 10.1609/icwsm.v14i1.7290.

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, 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, September 2017, pp. 199-212, doi: 10.1016/j.annals.2016.10.007.

Höpken, W., Ernesti, D., Fuchs, M., Kronenberg, K. and Lexhagen, M. (2017), “Big data as input for predicting tourist arrivals”, in *Information and Communication Technologies in Tourism 2017*, Springer International Publishing, pp. 187-199, doi: 10.1007/978-3-319-51168-9_14.

Hadavandi, E., Ghanbari, A., Shahanaghi, K. and Abbasian-Naghneh, S. (2011), “Tourist arrival forecasting by evolutionary fuzzy systems”, *Tourism Management*, Vol. 32 No. 5, pp. 1196-1203, doi: 10.1016/j.tourman.2010.09.015.

Höpken, W., Eberle, T., Fuchs, M. and Lexhagen, M. (2021), “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.

Hu, M., Qiu, R.T.R., Wu, D.C. and Song, H. (2021), “Hierarchical pattern recognition for tourism demand forecasting”, *Tourism Management*, Vol. 84, 104263, doi: 10.1016/j.tourman.2020.104263.

Hutto, C.J. and Gilbert, E. (2014), “VADER: a parsimonious rule-based model for sentiment analysis of social media text”, available at: http://sentic.net/

Jassim, R.S.Al, Jetly, K., Abushakra, A. and Mansori, S.Al. (2023), “A review of the methods and techniques used in tourism demand forecasting”, *EAI Endorsed Transactions on Creative Technologies*, Vol. 9 No. 4, p. e1, doi: 10.4108/eetct.v9i31.2986.

Jiang, Z., Zheng, Y., Tan, H., Tang, B. and Zhou, H. (2017), “Variational deep embedding: an unsupervised and generative approach to clustering”, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 1965-1972, doi: 10.24963/ijcai.2017/273.

Kodippili, A. and Senaratne, D. (2017), “Forecasting tourist arrivals to Sri Lanka using seasonal ARIMA”, *An International Peer-Reviewed Journal*, Vol. 29, available at: www.iiste.org

Konarasinghe, K.M.U.B. (2016), “Time series patterns of tourist arrivals to Sri Lanka”, *Review of Integrative Business and Economics ResearchOnlineCDROM*, Vol. 5 No. 3, pp. 161-172, available at: https://ssrn.com/abstract=2860733

Kurukulasooriya, N. and Lelwala, E. (2014), *Time Series Behaviour of Burgeoning*, Vol. 1 No. 1, pp. 1-14.

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.

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

Li, X., Li, H., Pan, B. and Law, R. (2021), “Machine learning in internet search query selection for tourism forecasting”, *Journal of Travel Research*, Vol. 60 No. 6, pp. 1213-1231, doi: 10.1177/0047287520934871.

Li, Y., Lin, Z. and Xiao, S. (2022), “Using social media big data for tourist demand forecasting: a new machine learning analytical approach”, *Journal of Digital Economy*, Vol. 1 No. 1, pp. 32-43, doi: 10.1016/j.jdec.2022.08.006.

Lütkepohl, H. and Krätzig, M. (2004), *Applied Time Series Econometrics*, Cambridge University Press.

Marutho, D., Muljono Rustad, S. and Purwanto (2022), “Sentiment analysis optimization using vader lexicon on machine learning approach”, *2022 international Seminar on intelligent Technology and its applications (ISITIA)*, pp. 98-103, doi: 10.1109/ISITIA56226.2022.9855341.

Mishra, R.K., Urolagin, S., Angel, J., Jothi, A., Nawaz, N. and Ramkissoon, H. (2021), “Machine learning based forecasting systems for worldwide international tourists arrival”, *International Journal of Advanced Computer Science and Applications*, Vol. 12 No. 11, doi: 10.14569/ijacsa.2021.0121107, available at: www.ijacsa.thesai.org

Moreno, M.A., Goniu, N., Moreno, P.S. and Diekema, D. (2013), “Ethics of social media research: common concerns and practical considerations”, *Cyberpsychology, Behavior, and Social Networking*, Vol. 16 No. 9, pp. 708-713, doi: 10.1089/cyber.2012.0334.

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.

Plevris, V., Solorzano, G., Bakas, N. and Ben Seghier, M. (2022), “Investigation of performance metrics in regression analysis and machine learning-based prediction models”, *8th European Congress on Computational Methods in Applied Sciences and Engineering*, pp. 1-25, doi: 10.23967/eccomas.2022.155.

Priyangika, J. (2016), “Modelling and forecasting tourist arrivals in Sri Lanka”, *Symposium on Statistical and Computational Modelling with Applications - 2016*.

Rajangam, V., Yadav, O., Khan, F., Shukla, M. and Sangeetha, N. (2022), “VaderLogRest algorithm: an ensemble learning approach for sentiment analysis on vaccination tweets”, *2022 4th International Conference on Biomedical Engineering (IBIOMED)*, pp. 7-12, doi: 10.1109/IBIOMED56408.2022.9988439.

Reda, O. and Zellou, A. (2023), “Assessing the quality of social media data: a systematic literature review”, *Bulletin of Electrical Engineering and Informatics*, Vol. 12 No. 2, pp. 1115-1126, doi: 10.11591/eei.v12i2.4588.

Samarathunga, W. (2020), “Post-COVID19 challenges and way forward for Sri Lanka tourism”, *Available at SSRN 3581509*, doi: 10.2139/ssrn.3581509.

Smola, A.J. and Schölkopf, B. (2004), “A tutorial on support vector regression”, *Statistics and Computing*, Vol. 14 No. 3, pp. 199-222, doi: 10.1023/b:stco.0000035301.49549.88.

Song, H., Hom, H., Kowloon, H., Kong, S. and Gang, L. (2008), “Tourism demand modelling and forecasting A review of recent research”.

Song, H., Li, G., Witt, S.F. and Fei, B. (2010), “Tourism demand modelling and forecasting: how should demand be measured?”, *Tourism Economics*, Vol. 16 No. 1, pp. 63-81, doi: 10.5367/000000010790872213.

Taecharungroj, V. and Mathayomchan, B. (2019), “Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand”, *Tourism Management*, Vol. 75, pp. 550-568, doi: 10.1016/j.tourman.2019.06.020.

Thushara, S.C., Su, J.-J. and Bandaralage, J. (2016), “Forecasting international tourist arrivals to Sri Lanka using sarima approach”, Vol. 1, available at: www.gissf.com

Tealab, A. (2018), “Time series forecasting using artificial neural networks methodologies: a systematic review”, *Future Computing and Informatics Journal*, Vol. 3 No. 2, pp. 334-340, doi: 10.1016/j.fcij.2018.10.003.

Thushara, S.C., Su, J.J. and Bandara, J.S. (2019), “Forecasting international tourist arrivals in formulating tourism strategies and planning: the case of Sri Lanka”, *Cogent Economics and Finance*, Vol. 7 No. 1, p. 1699884, doi: 10.1080/23322039.2019.1699884.

Tofallis, C. (2015), “A better measure of relative prediction accuracy for model selection and model estimation”, *Journal of the Operational Research Society*, Vol. 66 No. 8, pp. 1352-1362, doi: 10.1057/jors.2014.103.

Wickramasinghe, K. and Ratnasiri, S. (2021), “The role of disaggregated search data in improving tourism forecasts: evidence from Sri Lanka”, *Current Issues in Tourism*, Vol. 24 No. 19, pp. 2740-2754, doi: 10.1080/13683500.2020.1849049.

Yu, G. and Schwartz, Z. (2006), “Forecasting short time-series tourism demand with artificial intelligence models”, *Journal of Travel Research*, Vol. 45 No. 2, pp. 194-203, doi: 10.1177/0047287506291594.

Zhang, Y., Li, G., Muskat, B. and Law, R. (2021), “Tourism demand forecasting: a decomposed deep learning approach”, *Journal of Travel Research*, Vol. 60 No. 5, pp. 981-997, doi: 10.1177/0047287520919522.

Corresponding author

Isuru Udayangani Hewapathirana can be contacted at: ihewapathirana@kln.ac.lk