Scientific Journal of Informatics
Vol. 11, No. 2, May 2024
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Nowcasting Hotel Room Occupancy Rate using Google Trends Index and Online Traveler Reviews Given Lag Effect with Machine Learning (Case Research: East Kalimantan Province)

Adelina Rahmawati[1*], Erna Nurmawati[2], Teguh Sugiyarto[3]
[1,2]Department of Statistical Computing, STIS-Polytechnic of Statistics, Indonesia
[3]Directorate of Finance, Information Technology, and Tourism Statistics, BPS-Statistics Indonesia, Indonesia

Abstract.
Purpose: The presence of a two-month lag in Hotel Room Occupancy Rate (TPK) data necessitates an alternative method to accommodate adjustments in the economic circumstances of the tourism industry. In this context, TPK is connected to the influx of tourists, making the data a valuable resource for assessing the tourism potential of a particular area. The information can be used to make informed decisions when considering investments in the local tourism industry. Therefore, this research aimed to formulate predictions for future trends using now-forecasting. The variables of Google Trends Index (IGT) and online traveler reviews considered were obtained from big data.
Methods: This research used machine learning methods with Random Forest, LSTM, and CNN-BiLSTM-Attention models in determining the best model. Meanwhile, the datasets were acquired from diverse secondary data sources. Hotel Occupancy Rooms Rate was derived from BPS-Statistics Indonesia, while additional data were collected through web scraping from online travel agency websites such as Tripadvisor.com, IGT with keywords "IKN", "hotel", and "banjir". For the sentiment variable from online reviews, lag effects of one, two, and three months were analyzed to determine the correlation with TPK. The highest correlation was selected for inclusion in the prediction model across all machine learning methods.
Result: The results showed that the use of IGT and online traveler reviews increased the precision of forecasting models. The best model of hotel TPK nowcasting was Random Forest Regression with the lowest MAPE value and accuracy of 5.37% and 94.63%, respectively.
Novelty: The proposed method showed great potential in improving the prediction of hotel TPK by leveraging new technology and extensive data sources. The correlation with TPK decreases with an increasing time lag of sentiment. Therefore, the sentiment of reviews in the current month has the highest correlation with TPK, compared to the previous one, two, or three months.

Keywords: Hotel room occupancy rate, Nowcasting, Traveler reviews, Lag effect, Machine learning
Received May 2024 / **Revised** June 2024 / **Accepted** June 2024

INTRODUCTION

The government is planning to implement Sustainable Development Goals (SDGs) through the establishment of a green economy. The objective is to enhance sustainable economic expansion, reduce poverty rates, advance social integration, and uphold environmental sustainability and resource efficiency [1].

The improvement of the tourism sector is an important effort to achieve a green economy [2], [3], which is a sustainable tourism industry. According to United Nations World Tourism Organization (NWTO), sustainable tourism considers the current and future conditions of economic, social, and environmental aspects. In this context, the variable balances the needs of visitors, local communities, industry, and environmental sustainability [4]. The tourism industry is a driver of economic growth through the multiplier effect. Therefore, tourism is expected to promote the business of providing accommodation, food and drink, as well as other services.

*Corresponding author.
Email addresses: 222011462@stis.ac.id (Rahmawati)*, erna.nurmawati@stis.ac.id (Nurmawati), teguh@bps.go.id (Sugiyarto)
DOI: 10.15294/sji.v11i2.5553

The calculation of Gross Domestic Product (GDP) category I and the provision of accommodation as well as food and beverages, requires an indicator of hotel room occupancy rate (TPK). TPK is estimated from VHTS survey and measures the ratio of occupied rooms compared to the total number available [5]. Since the concept is closely related to tourist arrivals [6], [7], [8], the data can be used to evaluate tourist potential [8]. In addition, investors used the data as a consideration of investing in the tourism sector [8].

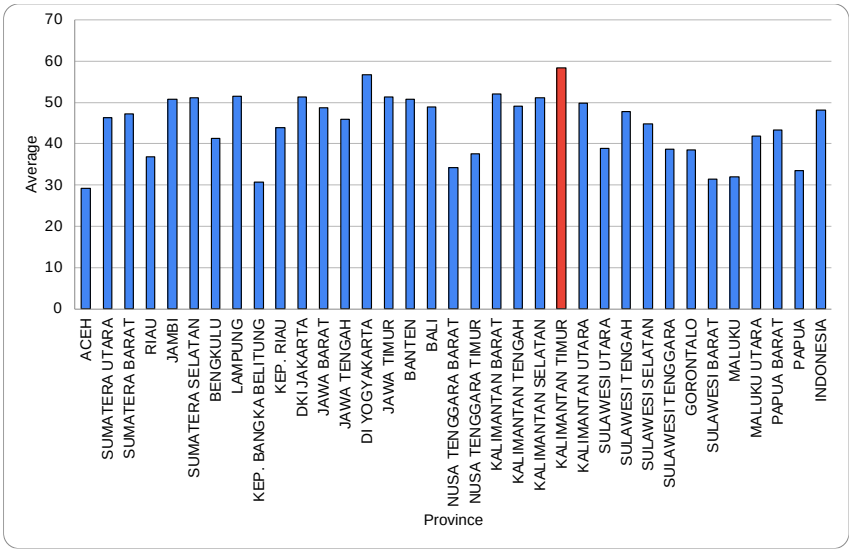


Figure 1. Average TPK value per province in Indonesia

TPK is categorized into star and non-star hotels [9] and Figure 1 shows the average from January 2023 to July 2023. The province of East Kalimantan has the highest average TPK during the period [10].

The VHT-L questionnaire is initiated at the start of the year [9] to obtain hospitality statistics on the quantity of hotels, rooms, and beds across all accommodations. Meanwhile, the VHT-S questionnaire is used to collect data on indicators, including average duration of stay, ratio of international and domestic guests, and room occupancy rate. This questionnaire is administered monthly to all-star and non-star hotel samples [9]. BPS necessitates a substantial financial investment and significant work in survey planning to produce TPK. The acquisition of the value necessitates a significant time delay [6], enabling adjustments in response to shifts in economic circumstances [5].

The development of the Internet is increasing over time, specifically in Indonesia. Therefore, the use of mobile-based applications and website services is high, including travel service applications and online hotel reservation providers. The most visited travel and tourism website in Indonesia and globally is Tripadvisor.com [11], [12].

The existence of online booking platforms has changed the pattern of consumer behaviour in selecting hotels [13], [14]. Consumers can filter information that suits the needs provided by the website and consider previous reviews [13]. Therefore, many research used online review analysis as a subject of hotel management [13], [14], [15]. In addition, guest reviews can also be analysed for sentiment and adopted as an additional evaluation of the quality of hotels. The use of search engines such as Google Search is also widely recognized. Since 2009, search intensity was introduced through Google Trends which provided daily, weekly, and monthly reports with explored keywords [5].

In previous research, regression models such as ARIMA and ARIMAX [5], [6] were used to predict TPK values. However, machine [16] and deep learning models were used to predict the values [13]. The models provided better accuracy compared to ARIMA/ARIMAX [17], [18], [19]. Therefore, this research adopted machine and deep learning methods. Sequential deep learning such as Long Short-Term Memory (LSTM) [20] is used to predict time series data [13], [21], [22]. LSTM algorithm improvises the vanishing gradient problem in Recurrent Neural Network (RNN) when processing long sequences [13]. In the TPK prediction [13], LSTM showed better accuracy compared to other models such as BPNN, GRNN, LSSVR, RF, and GPR. Meanwhile, in stock forecasting [23], pollution prediction [24], water level prediction [25], and

electricity load prediction [26], other models such as CNN-BiLSTM-Attention were proposed to produce better accuracy than LSTM and BiLSTM.

In previous research using CNN-BiLSTM-Attention method, PM2.5 concentration forecasting was carried out to measure the level of air pollution [24]. This method consisted of Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM) layer, and Attention layers. CNN was first proposed in 1998 [27] for character recognition in printed documents. According to [24], the algorithm was used in extracting effective features from related factors. BiLSTM was adopted in solving the exploding gradient problem in time series and identifying temporal features in two hidden layer directions [24]. Meanwhile, the attention mechanism analyses the importance of features and assigns appropriate weights [24].

Room occupancy rate of star hotel in East Kalimantan was predicted using several methods, namely Random Forest Regression (RFR) [28], [29], [30], [31], LSTM [13], [21], and CNN-BiLSTM-Attention [23], [24], [25], [26]. This research used variables derived from scraping reviews on tripadvisor.com which were considered on the sentiment scale [13], rating [13], and number of reviews [32]. The prediction also included additional variables affecting TPK such as the number of tourist visits [7], [33], average length of stay [33] and scores from Google Trends with keywords "IKN" [34] and "Hotel" [5], [6]. Tourist arrivals that influence TPK [7], [30] are also behind the addition of dummy variables and Google Trend indexes (IGT) related to pandemics [35], [36] and natural disasters [37] with the keyword "flood". The results are expected to take the best model which can be applied in predicting the value of TPK. By identifying the best predictive model, this research contributes to the accurate forecasting of TPK value.

METHODS

Problems regarding the release of TPK with time lag can be solved by prediction or nowcasting. The prediction was achieved by adding variables from Tripadvisor to increase the accuracy. The added variables included the number of reviews, rating, and sentiment scale. In addition, Tripadvisor review variable was analyzed to determine guest sentiment at star hotel in East Kalimantan.

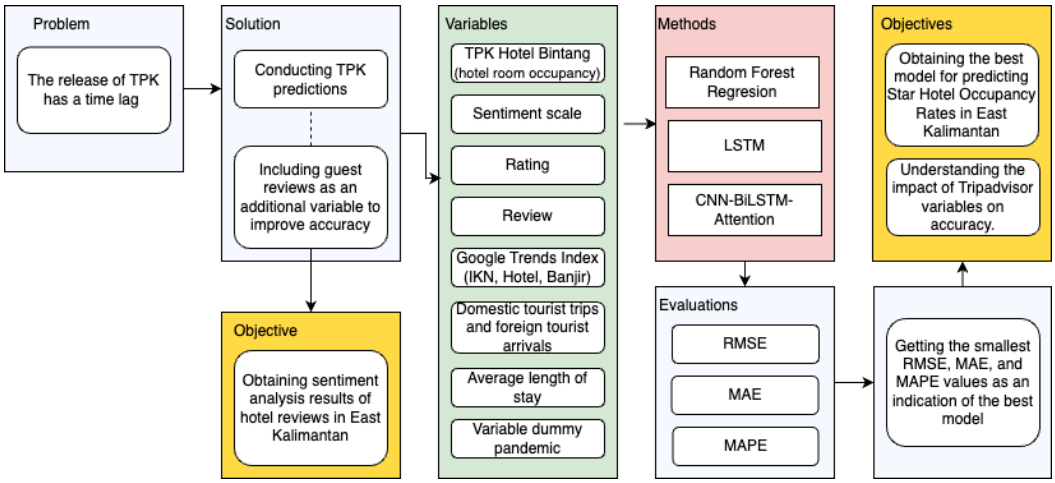


Figure 2. Research framework

The variables used include data on TPK of star hotel (TPK), the number of domestic tourist trips based on the original region (wisnus_asal), the domestic tourist trips based on the destination region (wisnus_tujuan), the number of foreign tourist arrivals (wisman), the average length of stay of domestic guests (rlm_tanus), average length of stay of foreign guests (rlm_taman), pandemic dummy variable (pandemic), IGT score with keywords hotel (IGT_Hotel), IKN (IGT_IKN), and flood (IGT_Flood) and variables derived from tripadvisor, namely the number of reviews (review), rating (rating), vader sentiment scale (vader_scale), and textblob sentiment scale (textblob_scale). For the sentiment variable from online reviews, lag effects of one, two, and three months were analyzed to determine the correlation with TPK. The highest correlation was selected for inclusion in the prediction model across all machine learning methods. The variables were included in the Random Forest Regression (RFR), LSTM, and CNN-BiLSTM-Attention models. Meanwhile, a model without the variables was tested to de determine the influence of tripadvisor.

In this research, the stages were carried out with the following steps.

Data collection
The data used are obtained from room occupancy rates of star hotels, number of foreign tourist arrivals, domestic tourist trips, and average length of stay. In addition, additional data are collected by scraping the tripadvisor.com website and IGT, as well as manually creating dummy variables.

The retrieval of reviews on Tripadvisor is based on directory data of star hotel in East Kalimantan. Out of the 73 1-star to 5-star hotels available on tripadvisor.com, 68 were taken with 3 substituted hotels with the same criteria. Hotels were taken for guest reviews and only 10,107 were successfully received. From the results, the retrieved data was cleaned and 8,664 reviews had a date period from January 1, 2014, to August 31, 2023, in sentiment analysis. For modelling TPK predictions, only the range from January 2019 to August 2023 is used. The results of scraping produce sentiment scale variables, ratings, and the number of reviews. Other variables were obtained from IGT with several keywords, namely "IKN", "hotel", and "flood". In addition, a dummy variable is also added which marks the month of COVID-19 pandemic phase.

Preprocessing
The collected data was preprocessed to obtain the values per month and the period taken for the variables to be included in the model was from January 2019 to August 2023. This process also comprised the creation of variables including:

1. Sentiment Scale
   Sentiment scale was obtained from the results of tripadvisor scraping. This process was carried out using chrome extension [38], namely TripAdvisor® Review Scraper. Hotels in the directory are replaced with the same criteria in the East Kalimantan region.
   Scraping in the form of reviews is measured for polarity using lexicon-based. The measurement results are minus, plus and zero for negative, positive, and neutral reviews. In addition, reviews are preprocessed by being translated into English, and cleaned with the cleantext library [39]. The results were analyzed using vaderSentiment [40] and textblob [41] to obtain a polarity score. In addition, reviews were mapped and the monthly results were taken into vader and textblob.
2. Guest Rating Variables
   Guest rating variables are obtained from tripadvisor.com scraping results with a value of 1 to 5. All the ratings given by guests in a particular month are calculated with hotel by taking the average value.
3. Number of Reviews
   The number of reviews is obtained from the calculation of a particular month from the scraping results.
4. Hotel IGT
   Google Trends hotel index is obtained by using a scale of 0 to 100 for the keyword "Hotel" with coverage of East Kalimantan region from January 2019 to August 2023.
5. IKN IGT
   IKN IGT is obtained by taking a scale of 0 to 100 for the keyword "IKN" with coverage of the Indonesian region from January 2019 to August 2023.
6. Flood IGT Variable
   Flood IGT is obtained by taking a scale of 0 to 100 for the keyword "flood" with coverage of the East Kalimantan region. The weekly data obtained was grouped with the average of each month.
7. Pandemic dummy variable
   Pandemic dummy variable is coded 1 and 0 for months with and without the pandemic phase. The value of 1 in this variable starts in March 2020 [42] to June 2023 [43] while is 0 in other months.

After the collection of variables, the subsequent step in preprocessing is lag setting. Vader and textblob sentiment scale were made into new variables having lag values of t, t+1, t+2, and t+3 with t being months to determine effect of guest reviews at a time lag of t months. This determination was made by taking the best correlation between TPK and sentiment scale variables with time t, t+1, t+2, or t+3.

Data segregation
This stage commences with a split where the available data is divided into training and testing data with a ratio of 80:20. Training data is used to build the best model which is tested on testing data.

Modeling and testing

The training and testing data are entered into the model to obtain the best result from Random Forest, LSTM, and CNN-BiLSTM-Attention, as shown in Figure 3.

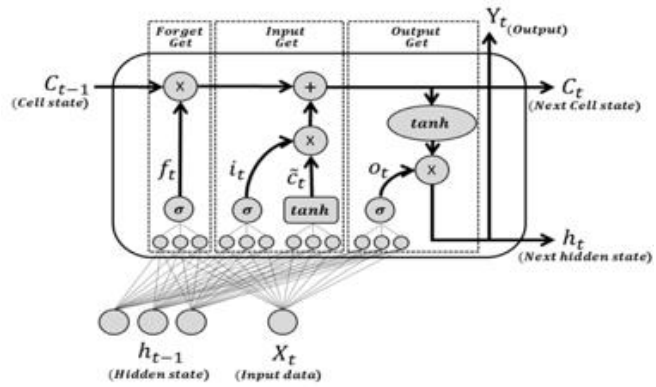


Figure 3 Structure of LSTM model[14]

LSTM modeling includes two consecutive layers to process time series data. The input layer receives data with a single time step and features including variables from TripAdvisor. The first and second layers return sequence outputs for each time step and receive more abstract representations, respectively. A dropout layer is used to prevent overfitting, and the output layer generates regression predictions with a single continuous value. The model is compiled using Adam optimizer learning rate of 0.016 and a mean squared error (MSE) loss function, with evaluation metrics such as mean absolute error (MAE) and *R-Squared*.

Hyperparameters are tuned using a 5-fold cross-validation grid search when building the LSTM model [29], [31], [44]. The values adjusted include the number of layers, batch size, and dropout rate. Additionally, the loss function is set as Mean Squared Error (MSE) to systematically explore the hyperparameter space and select the combination with the best performance. The steps of the CNN-BiLSTM-Attention modeling process are shown in Figure 4.

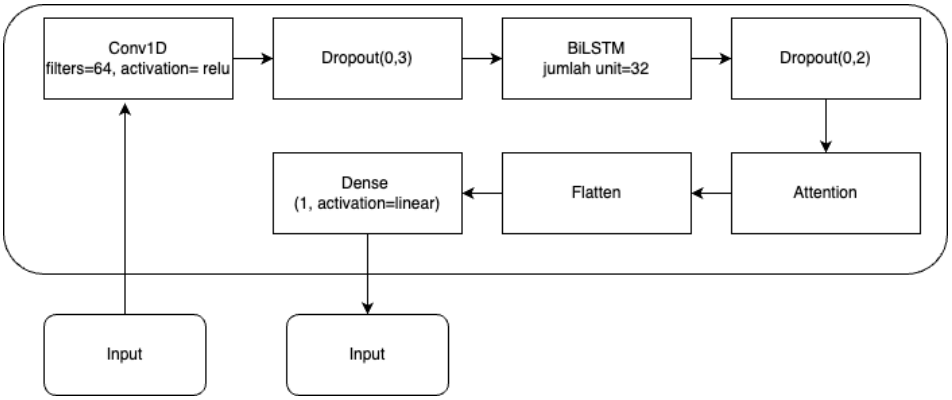


Figure 4 Structure of CNN-BiLSTM-Attention model

Testing is carried out without including variables derived from Tripadvisor, namely the number of reviews, ratings, and sentiment scale to determine the comparison of model performance.

Model evaluation

The model built and optimized is used to predict TPK on testing data to obtain MAPE and RMSE.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2} \quad (2)$$

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

where:

n = the number of observations
y_i = the value of observations
$\hat{y}_i$ = predicted value of observations

The MAPE and RMSE are compared between Random Forest, LSTM, and CNN-BiLSTM-Attention models to determine the best model for TPK case.

RESULTS AND DISCUSSIONS

Sentiment analysis

Based on scraping reviews on Tripadvisor, the data cleaned is analyzed for sentiment with lexicon-based, namely Textblob and Vader. The sentiment results from guest reviews are classified into positive, negative, and neutral categories as shown in Figure 6.

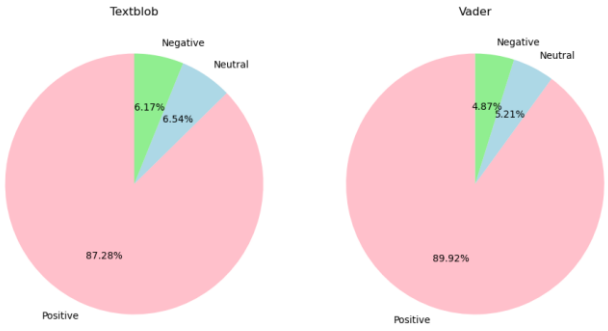


Figure 6. Sentiment comparison

Based on Figure 6, most reviews have a positive sentiment. In the analysis using textblob, the percentage of positive, negative, and neutral sentiments in guest reviews reached 87.28%, 6.17%, and 6.54%, respectively. Meanwhile, the analysis using vader is also dominated by positive, negative, and neutral sentiments with a percentage of 89.92%, 4.87%, and 5.21%, respectively.

Figure 7. Online traveler reviews Word cloud

Figure 7 is a word cloud of online traveler reviews on tripadvisor.com website that shows some of the words in guest reviews, such as room, hotel, breakfast, good, and staff. In addition, the word cloud includes all reviews in multiple languages. From the various types of word cloud, Indonesian is more dominant than other languages. The appearance of Balikpapan and Samarinda also shows that most of the star hotels are centered in the two cities.

Tpk prediction modeling

Data per month was used in modeling with a period of January 2019 to August 2023. The variables include TPK, number of domestic tourist trips based on origin (East Kalimantan), domestic tourist trips based on

destination (East Kalimantan), foreign tourist arrivals, average length of stay of foreign tourists, average length of stay of domestic tourists, average length of stay of all tourists, google trends hotel index (accommodation type), google trends hotel index (search term), google trends IKN index, google trends flood index, pandemic, rating, vader sentiment scale, textblob sentiment scale, and number of reviews. Meanwhile, certain factors are excluded while establishing the optimal latency using Pearson correlation, as shown in Figure 8.

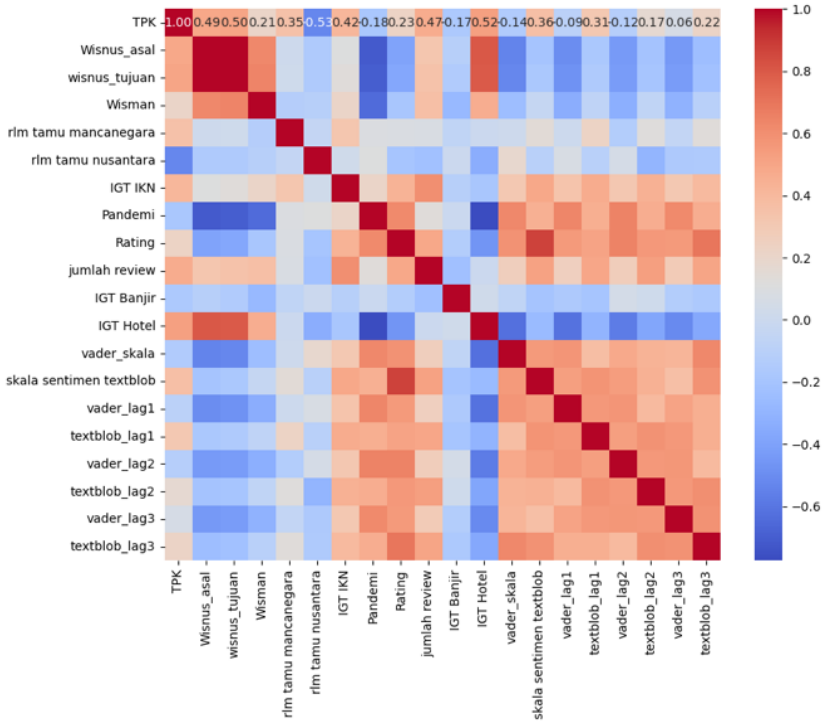


Figure 8. Correlation among variables (20 variables)

Several lags were selected with the highest absolute correlation of TPK on textblob and vader_scale. The textblob scale has a higher correlation of 0.363 compared to t+1, t+2, and t+3 with correlations of 0.302, 0.174, and 0.219 respectively. On vader scale, the highest value has an absolute correlation of 0.144 with the comparison of t+1, t+2 and t+3 at 0.093, 0.118, and 0.058, respectively. Similarly, the treatment produces clean variables to be used in modeling, as shown in Table 1.

Table 1. Descriptive statistics

variable	count	mean	std	min	max
TPK	56	54.11	9.74	26.31	67.52
Wisnus_asal	56	442024.18	268998.43	85523	1061653
wisnus_tujuan	56	453722.98	266324.54	91121	999931
Wisman	56	120.71	155	0	526
rlm tamu mancanegara	56	2.38	0.51	1.1	3.69
rlm tamu nusantara	56	1.67	0.16	1.45	2.11
IGT IKN	56	10.58	12.96	1.25	57.25
Pandemi	56	0.71	0.46	0	1
Rating	56	4.57	0.35	3.5	4.96
jumlah review	56	81.43	74.43	4	322
IGT Banjir	56	9.53	5.2	2.75	30.6
IGT Hotel	56	47.58	13.17	24.5	80.4
vader_skala	56	0.73	0.08	0.49	0.94
textblob_skala	56	0.34	0.07	0.12	0.42

Among the variables listed in Table II, TPK was predicted by the model. Meanwhile, variable x included all or a subset that is not associated with reviews.

The number of domestic tourist trips and foreign tourist arrivals has a positive impact on increasing TPK. The presence of tourists causes a demand for accommodation service businesses. In this research, domestic tourists with the destination of East Kalimantan are used as variable x. Generally, the trips are carried out within the province, and the demand for hotel room usage is affected. The average length of travel is directly proportional to the opportunity of using accommodation services.

Pandemic and IGT_Flood are assumed to have an impact on reducing demand for hotel use. In disaster situations, tourism activities and population mobility trips are reduced. Conversely, IGT_IKN and IGT_Hotel can have a positive influence on travel to East Kalimantan. There is a centralized IKN development affecting the movement of people. In this context, Google searches are also assumed to be interested in IKN or traveling to East Kalimantan.

Reviews and ratings are important aspects used to describe sustainable tourism. The framework for measuring sustainable tourism developed by UNWTO includes the social dimension obtained from the perspective of visitors [4]. In this research, reviews and ratings from hotel users are used as explanatory variables. Positive reviews and ratings are an effective form of promotion for the sustainability of the tourism industry. Moreover, the need to measure visitor satisfaction has been stated for the sustainability of tourism industry.

The origin tourist, destination tourist, IGT IKN, IGT hotel, and reviews are positively correlated with TPK with a value above 0.4 in the sufficient category [45].

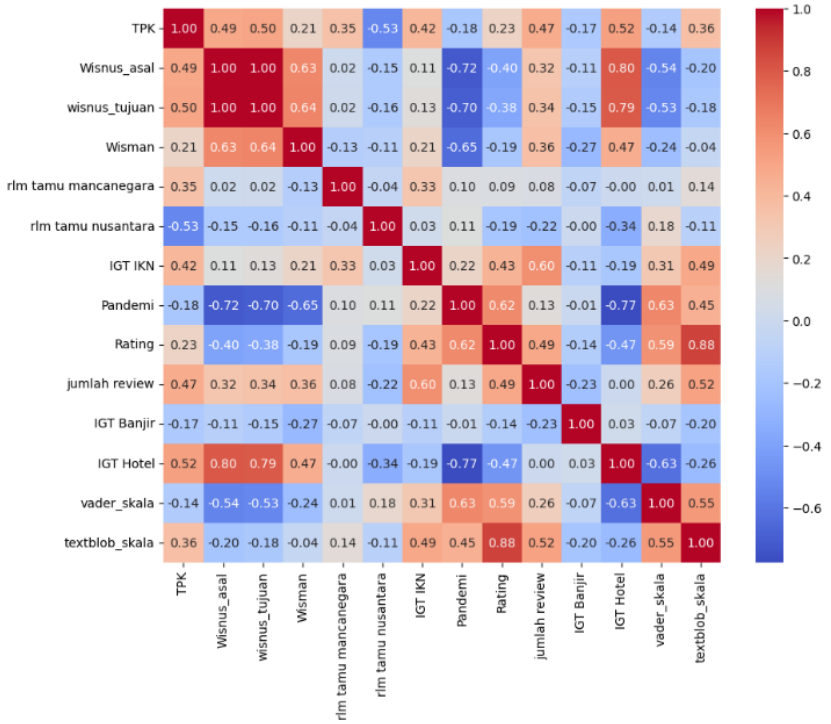


Figure 9. Correlation among variables (14 variables)

The y variable is the data on TPK of star hotel in East Kalimantan from January to August 2023 in Figure 10.

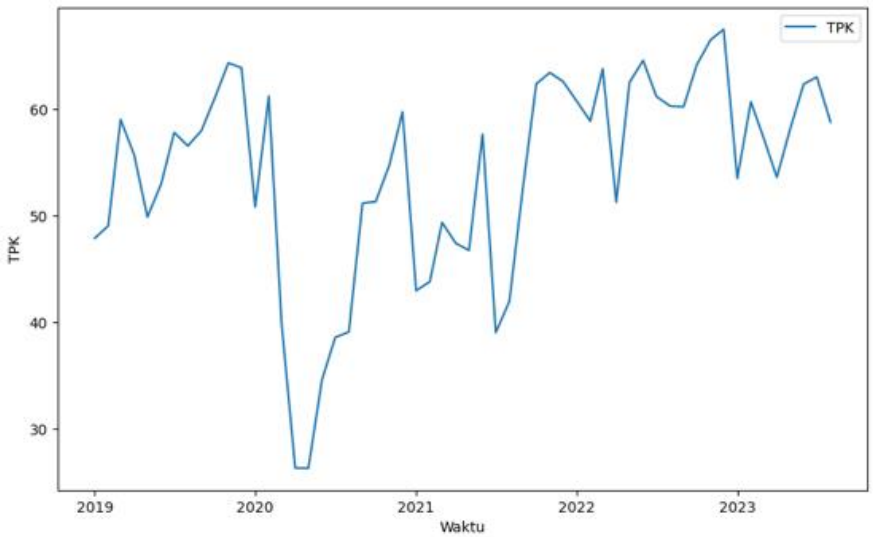


Figure 10. TPK from January 2019-August 2023

The modeling stage commences by testing machine learning using the Random Forest Regression model. This model adopts a 5-fold cross-validation grid search with the following hyperparameters.

Table 2. Hyperparameter Random Forest Regression

Hyperparameter	Interval
Number of trees	10, 50, 100
Depth	None, 5, 10, 20
Min sample split	2, 5, 10
Min sample leaf	1, 2, 4

From the various parameters used in Table III, the best model is obtained with the number of trees, depth, minimum sample split, and minimum sample leaf being 50, none, 5, and 1, respectively. The prediction model using a random forest regressor for all variables is shown in Figure 11

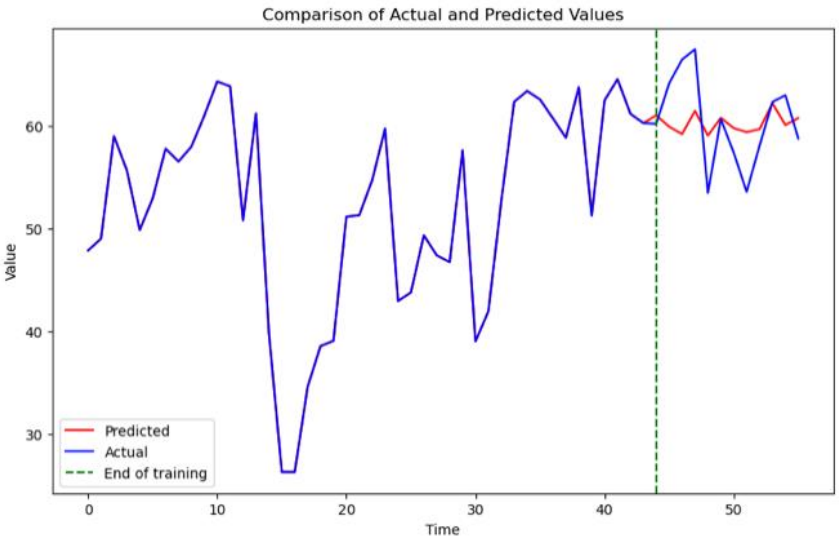


Figure 11. Random Forest Regression model

RFR model that includes all variables and features is shown in Figure 12.

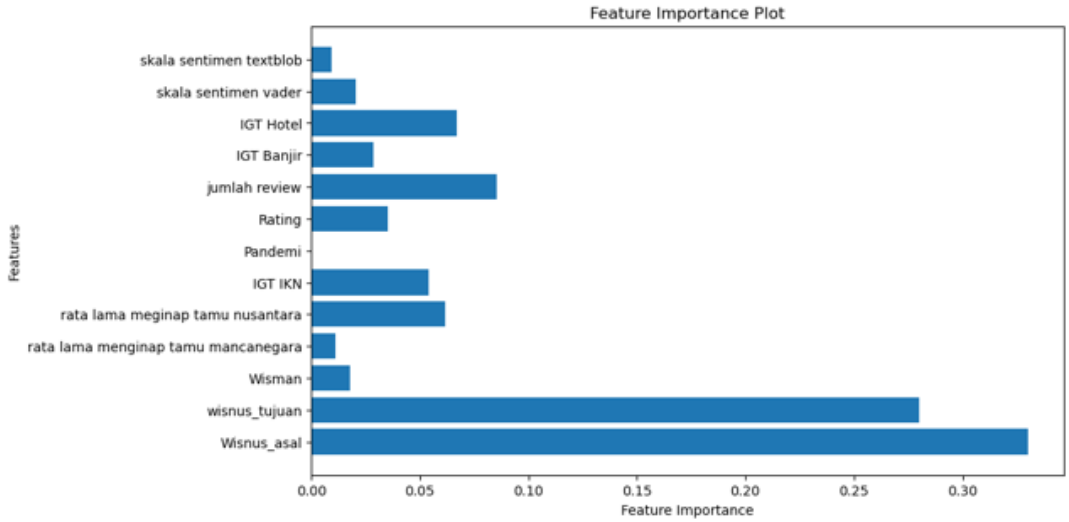


Figure 12. Feature importance of RF Regression

The origin tourist has the highest value from the feature importance output in Figure 12. The features with the highest to lowest value in order are originating tourists, destination tourists, number of reviews, hotel IGT, average length of stay of domestic guests, IGT IKN, rating, flood IGT, vader sentiment scale, foreign tourists, average length of stay of foreign guests, textblob sentiment scale, and pandemic dummy variables. The development of accommodation service businesses is affected by tourism activities. The large number of tourists provides potential for the demand for hotel facilities. Meanwhile, consumers can be influenced by ratings and reviews from previous users.

The modeling stage using RFR was tested without including variables related to Tripadvisor such as the number of reviews, ratings, textblob scale, and vader scale. The scenario produces the best parameters with the number of trees, depth, minimum sample split, and minimum sample leaf being 50, none, 5, and 1, respectively.

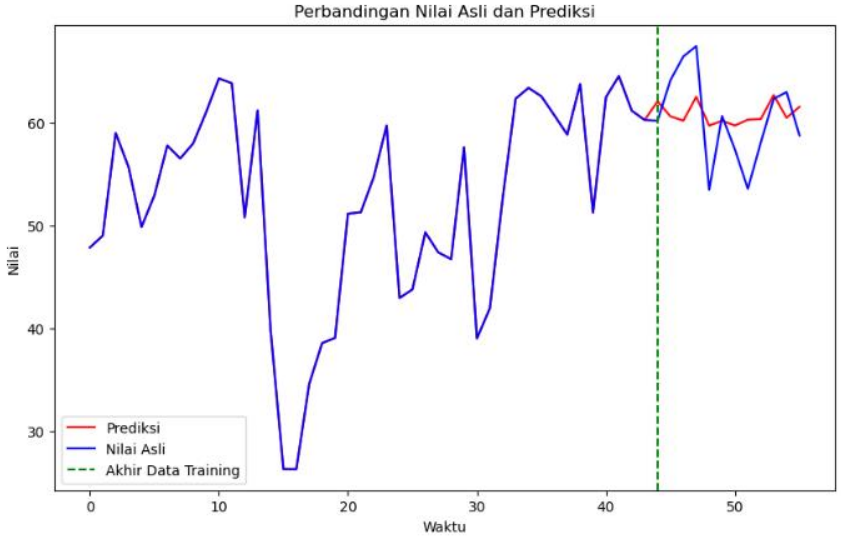


Figure 13. Random forest model without review variables

The data is also tested with other models such as LSTM and CNN-BiLSTM-Attention. Table IV shows the hyperparameter used in learning the LSTM model.

Table 3. Hyperparameter LSTM

Hyperparameter	Interval
Batch	16, 32, 64
Units	32, 50, 64
Epoch	100
Drop out rate	0.1, 0.2
Learning rate	0.001, 0.016

From the hyperparameters in Table IV, the best parameters obtained are batch size 32, dropout rate 0.2, epoch 100, and number of units 32. In addition, TPK prediction is carried out with the output in Figure 14 after learning.

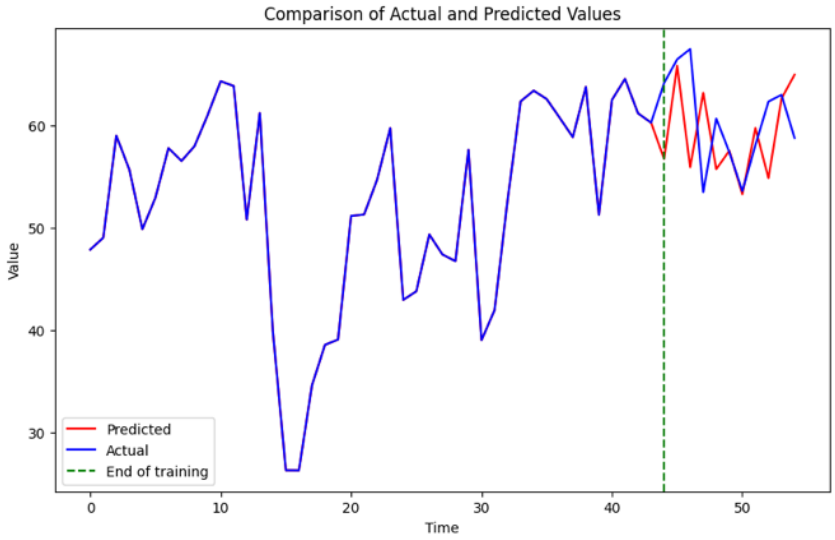


Figure 14. LSTM model with review variables

In the scenario without review, the best parameters obtained are batch size 32, dropout rate 0.1, epoch 100, and number of units 64. After learning, TPK prediction is carried out with the output in Figure 15.

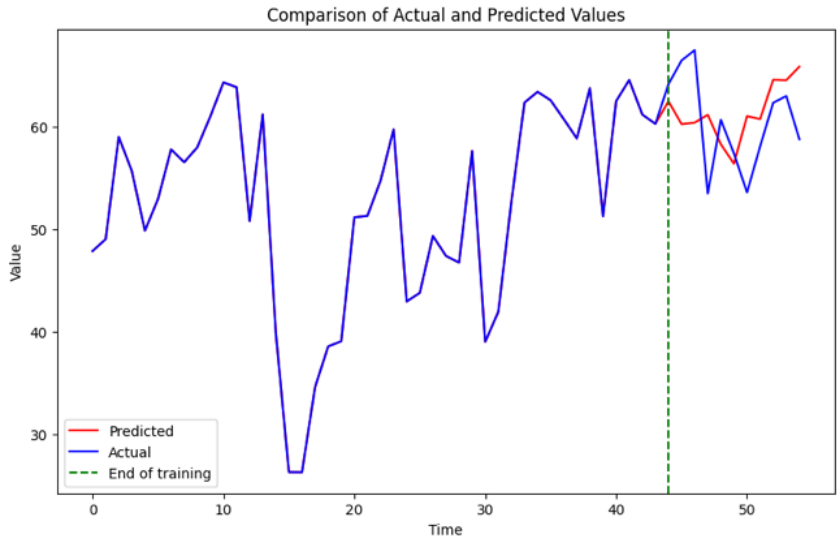


Figure 15. LSTM model without review variables

The last model to be tested is CNN-BiLSTM-Attention and the parameters used are batch size 32, dropout rate 0.2, epoch 100, number of units 32 with adam optimizer and activation relu. This modeling produces TPK predictions with the output in Figure 16.

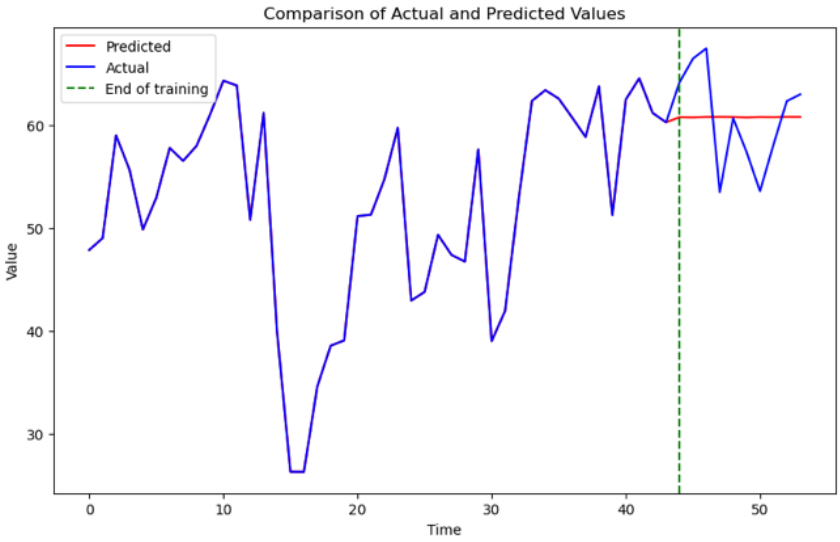
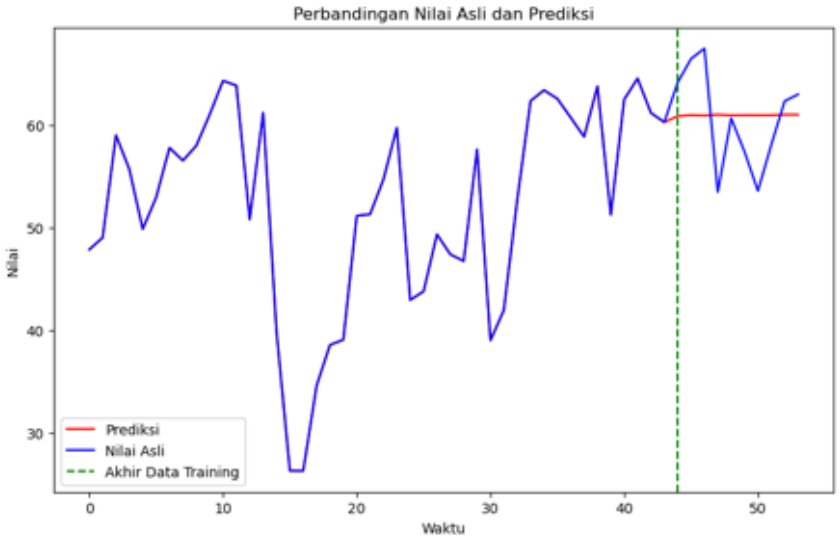


Figure 16. CNN-BiLSTM-Attention model with review variables

The scenario without reviews was also tested, resulting in the output shown in Figure 17.



Gambar 17. CNN-BiLSTM-Attention model without review variables

The evaluation of the model is based on the RMSE, MAE, and MAPE values. Table V shows the comparison of the accuracy of several machine learning models.

Table 4. Model comparison

Model	Variables X	RMSE	MAE	MAPE (%)
RF Regression	**All variables**	**4.012**	**3.239**	**5.366**
RF Regression	Without review variables	3.969	3.362	5.638
LSTM	All variables	6.073	4.586	7.523
LSTM	Without review variables	5.025	4.277	7.188
CNN-BiLSTM-Attention	All variables	4.692	4.022	6.752
CNN-BiLSTM-Attention	Without review variables	4.713	4.031	6.788

Table V shows that the Random Forest Regression model outperforms others with the lowest RMSE, MAE, and MAPE values. Based on the analysis, the model which incorporates the variables shows a lower MAPE value of 5.366%. Therefore, this model is the most accurate in predicting TPK in East Kalimantan. In the evaluation using MAPE, the value in RFR falls within the excellent range [46]. The research suggests that Random Forest is the most effective model for predicting international visitor arrivals when compared to other machine learning methods [31].

RF is relatively simpler to train compared to LSTM and CNN-BiLSTM-Att models. The variable performs exceptionally on a relatively small amount of data [47]. Overfitting can be effectively handled through the ensemble method, providing strong predictions with limited data. RF is flexible and robust, capable of handling different types of data, including categorical and numerical features.

The inclusion of variables related to TripAdvisor such as the number of reviews, rating, and sentiment has an impact by reducing the MAPE value. In LSTM model, the addition of the variables is insufficient to enhance performance. Based on the evaluation, the addition of variables related to TripAdvisor has a significant impact on the model.

The improved performance of the Random Forest model offers valuable practical applications. The variable facilitates more precise forecasting of TPK, which can empower hotel managers to optimize resource allocation, refine pricing strategies, and enhance marketing efforts. In this context, policymakers can use the data to make smarter decisions about tourism development and infrastructure investments to ensure better planning. However, this research has potential bias and limitations in data collection since scraping was conducted on TripAdvisor. Other methods include additional platforms such as Booking.com, Agoda, and Traveloka.

CONCLUSION

In conclusion, the majority of reviews were in the form of positive sentiments based on guest reviews collected from tripadvisor.com. Among several tested models, Random Forest model had the smallest MAPE value of 5.37%, an accuracy of 94.63%, with RMSE and MAE values of 4.012 and 3.239. The MAPE of the model were in the category of excellent [46] and incorporating variables from TripAdvisor, such as the number of reviews, rating, and sentiment scale, enhanced the accuracy of the model, such as Random Forest Regression and CNN-BiLSTM-Attention. The methods and models were used as a potential solution for estimating TPK and the latest TPK data could be estimated for short-term planning purposes before the production of official statistics.

This research has limitations since the data can be augmented by adding reviews from other websites such as booking.com, agoda, and traveloka. A more precise feature selection can be performed to achieve higher accuracy. The development of a framework can facilitate the usability of predictive outcomes. In addition, the advancement of the sentiment analysis method can provide deeper insights.

REFERENCES

[1] The Government of Indonesia & GGGI, “Indonesia Green Growth Program.” Accessed: Oct. 01, 2023. [Online]. Available: http://greengrowth.bappenas.go.id/tentang-kami/

[2] D. R. Toubes and N. Araújo-Vila, “A Review Research on Tourism in the Green Economy,” *Economies*, vol. 10, no. 6, 2022, doi: 10.3390/economies10060137.

[3] Y. Shang, Y. Lian, H. Chen, and F. Qian, “The impacts of energy resource and tourism on green growth: Evidence from Asian economies,” *Resources Policy*, vol. 81, no. February, p. 103359, 2023, doi: 10.1016/j.resourpol.2023.103359.

[4] UNWTO, “Statistical Framework for Measuring the Sustainability of Tourism (SF-MST).” Accessed: May 06, 2024. [Online]. Available: www.unwto.org/tourism-statistics/statistical-framework-for-measuring-the-sustainability-of-tourism

[5] I. Ayuningtyas and I. Wirawati, “Nowcasting Tingkat Penghunian Kamar Hotel Menggunakan Google Trends,” *Seminar Nasional Official Statistics*, vol. 2020, no. 1, pp. 338–343, 2021, doi: 10.34123/semnasoffstat.v2020i1.636.

[6] F. A. Rizalde, S. Mulyani, and N. Bachtiar, "Forecasting Hotel Occupancy Rate in Riau Province Using ARIMA and ARIMAX," *Proceedings of The International Conference on Data Science and Official Statistics*, vol. 2021, no. 1, pp. 578–589, 2022, doi: 10.34123/icdsos.v2021i1.199.

[7] S. Suwanto, "Hubungan Jumlah Kunjungan Wisatawan Mancanegara dengan Rata-Rata Tingkat Penghunian Kamar Hotel Provinsi DKI Jakarta Tahun 2012—2018," *Jurnal Kepariwisataan Indonesia : Jurnal Penelitian dan Pengembangan Kepariwisataan Indonesia*, vol. 14, no. 1, pp. 9–20, 2020, doi: 10.47608/jki.v14i12020.9-20.

[8] Ramadiani, N. Wardani, A. Harsa Kridalaksana, M. Labib Jundillah, and Azainil, "Forecasting the Hotel Room Reservation Rate in East Kalimantan Using Double Exponential Smoothing," *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, 2019, doi: 10.1109/ICIC47613.2019.8985916.

[9] BPS Provinsi DKI Jakarta, *Statistik Hotel dan Tingkat Penghunian Kamar Hotel DKI Jakarta 2021*. BPS Provinsi DKI Jakarta, 2022.

[10] Badan Pusat Statistik, "Tingkat Penghunian Kamar pada Hotel Bintang 2023."

[11] SimilarWeb, "Most visited travel and tourism websites worldwide as of July 2023 (in million visits) [Graph]." Accessed: Oct. 01, 2023. [Online]. Available: https://www.statista.com/statistics/1215457/most-visited-travel-and-tourism-websites-worldwide/

[12] Similarweb, "Top Websites Ranking Most Visited Travel and Tourism Websites in Indonesia." Accessed: Feb. 19, 2024. [Online]. Available: https://www.similarweb.com/top-websites/indonesia/travel-and-tourism/

[13] Y. M. Chang, C. H. Chen, J. P. Lai, Y. L. Lin, and P. F. Pai, "Forecasting Hotel Room Occupancy Using Long Short-Term Memory Networks with Sentiment Analysis and Scores of Customer Online Reviews," *Applied Sciences (Switzerland)*, vol. 11, no. 21, 2021, doi: 10.3390/app112110291.

[14] A. Setiawan and F. H. Sukmana, "MENGURAI PENGALAMAN POSITIF TAMU SAAT MENGINAP DI SHERATON SENGGIGI BEACH RESORT: BUKTI DARI ULASAN TRIPADVISOR Unravelling Positive Experiences of Guests Staying at Sheraton Senggigi Beach Resort: Evidence from TripAdvisor," vol. 17, no. 1, pp. 64–84, 2023.

[15] I. Ihwandi and F. Sukmana, "Hotel Tua dan Ulasan Online Negatif: Apa Yang Dikatakan Pelanggan?," *Jurnal Master Pariwisata (JUMPA)*, p. 354, 2022, doi: 10.24843/JUMPA.2022.v09.i01.p16.

[16] K. Kozlovskis, Y. Liu, N. Lace, and Y. Meng, "Application of Machine Learning Algorithms To Predict Hotel Occupancy," *Journal of Business Economics and Management*, vol. 24, no. 3, pp. 594–613, 2023, doi: 10.3846/jbem.2023.19775.

[17] J. W. Bi, Y. Liu, and H. Li, "Daily tourism volume forecasting for tourist attractions," *Annals of Tourism Research*, vol. 83, no. September 2019, p. 102923, 2020, doi: 10.1016/j.annals.2020.102923.

[18] R. Law, G. Li, D. K. C. Fong, and X. Han, "Tourism demand forecasting: A deep learning approach," *Annals of Tourism Research*, vol. 75, no. October 2018, pp. 410–423, 2019, doi: 10.1016/j.annals.2019.01.014.

[19] S. Sun, Y. Wei, K. L. Tsui, and S. Wang, "Forecasting tourist arrivals with machine learning and internet search index," *Tourism Management*, vol. 70, no. February 2018, pp. 1–10, 2019, doi: 10.1016/j.tourman.2018.07.010.

[20] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[21] L. Wiranda and M. Sadikin, "Penerapan Long Short Term Memory Pada Data Time Series Untuk Memprediksi Penjualan Produk Pt. Metiska Farma," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 8, no. 3, pp. 184–196, 2019.

[22] M. Davi and E. Winarko, "Rancang Bangun Aplikasi Peramalan Jumlah Penumpang Menggunakan Long Short-Term Memory (LSTM)," *Infotekmesin*, vol. 14, no. 2, pp. 303–310, 2023, doi: 10.35970/infotekmesin.v14i2.1911.

[23] J. Zhang, L. Ye, and Y. Lai, "Stock Price Prediction Using CNN-BiLSTM-Attention Model," pp. 1–18, 2023.

[24] J. Zhang, Y. Peng, B. Ren, and T. Li, "PM2.5 Concentration Prediction Based on CNN-BiLSTM and Attention Mechanism," *Algorithms*, vol. 14, no. 7, 2021, doi: 10.3390/a14070208.

[25] Q. Nie, D. Wan, and R. Wang, "CNN-BiLSTM water level prediction method with attention mechanism," *Journal of Physics: Conference Series*, vol. 2078, no. 1, 2021, doi: 10.1088/1742-6596/2078/1/012032.

[26] B. S. Prasetyo, D. Adytia, and I. A. Aditya, "Time Series Forecasting of Electricity Load Using Hybrid CNN-BiLSTM with an Attention Approach: A Case Study in Bali, Indonesia," *2023 International Conference on Data Science and Its Applications, ICoDSA 2023*, pp. 460–464, 2023, doi: 10.1109/ICoDSA58501.2023.10277176.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.

[28] P. Gajewski, B. Čule, and N. Rankovic, "Unveiling the Power of ARIMA, Support Vector and Random Forest Regressors for the Future of the Dutch Employment Market," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 18, no. 3, pp. 1365–1403, 2023, doi: 10.3390/jtaer18030069.

[29] P. Swetha, A. H. K. P. Rasheed, and V. P. Harigovindan, "Random Forest Regression based Water Quality Prediction for Smart Aquaculture," *2023 4th International Conference on Computing and Communication Systems, I3CS 2023*, pp. 1–5, 2023, doi: 10.1109/I3CS58314.2023.10127488.

[30] T. Petukhova, D. Ojkic, B. McEwen, R. Deardon, and Z. Poljak, "Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario, Canada," *PLoS ONE*, vol. 13, no. 6, pp. 1–17, 2018, doi: 10.1371/journal.pone.0198313.

[31] D. T. Andariesta and M. Wasesa, "Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach," *Journal of Tourism Futures*, pp. 1–17, 2022, doi: 10.1108/JTF-10-2021-0239.

[32] V. Leoni, P. Figini, and J. O. W. Nilsson, "It's not all about price. The determinants of occupancy rates in peer-to-peer accommodation: A methodological contribution," *International Journal of Contemporary Hospitality Management*, vol. 32, no. 4, pp. 1693–1711, 2020, doi: 10.1108/IJCHM-05-2019-0464.

[33] R. Law, "Room occupancy rate forecasting: A neural network approach," *International Journal of Contemporary Hospitality Management*, vol. 10, no. 6, pp. 234–239, 1998, doi: 10.1108/09596119810232301.

[34] G. Aji, Z. Arfani, A. Mila Sari, and R. Seprtiani, "Dampak Pemindahan Ibukota Negara Baru terhadap Ekonomi dan Sosial di Provinsi Kalimantan Timur," *Kultura: Jurnal Ilmu Hukum, Sosial, Dan Humaniora*, vol. 1, no. 5, pp. 1–8, 2023, doi: https://doi.org/10.572349/kultura.v1i5.474.

[35] D. K. Kim, S. K. Shyn, D. Kim, S. Jang, and K. Kim, "A Daily Tourism Demand Prediction Framework Based on Multi-head Attention CNN: The Case of the Foreign Entrant in South Korea," *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings*, pp. 1–10, 2021, doi: 10.1109/SSCI50451.2021.9659950.

[36] S. K. Prilistya, A. E. Permanasari, and S. Fauziati, "The Effect of the COVID-19 Pandemic and Google Trends on the Forecasting of International Tourist Arrivals in Indonesia," *TENSYMP 2021 - 2021 IEEE Region 10 Symposium*, pp. 1–8, 2021, doi: 10.1109/TENSYMP52854.2021.9550838.

[37] N. K. Sutrisnawati, "Dampak Bencana Alam Bagi Sektor Pariwisata Di Bali," *Jurnal Ilmiah Hospitality Management*, vol. 9, no. 1, pp. 57–66, 2018, doi: 10.22334/jihm.v9i1.144.

[38] M. Nowacki and J. Kowalczyk-Anioł, "Experiencing islands: is sustainability reported in tourists' online reviews?," *Journal of Ecotourism*, vol. 22, no. 1, pp. 59–79, 2023, doi: 10.1080/14724049.2022.2041648.

[39] S. Mujilahwati, M. Sholihin, R. Wardhani, and M. R. Zamroni, "Python Based Machine Learning Text Classification," *Journal of Physics: Conference Series*, vol. 2394, no. 1, 2022, doi: 10.1088/1742-6596/2394/1/012015.

[40] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for," *Eighth International AAAI Conference on Weblogs and Social Media*, pp. 216–225, 2014.

[41] S. Loria, "TextBlob: Simplified Text Processing." [Online]. Available: https://textblob.readthedocs.io/

[42] Kementerian Kesehatan Republik Indonesia, "Peran Ditjen Kesmas Daman Pandemi COVID 19 2020-2021," *Jakarta*, vol. 15, no. 2, pp. 1–23, 2021.

[43] Humas Sekretariat Kabinet Republik Indonesia, "Inilah Keppres Penetapan Berakhirnya Status Pandemi COVID-19 di Indonesia." Accessed: Jan. 11, 2024. [Online]. Available: https://setkab.go.id/inilah-keppres-penetapan-berakhirnya-status-pandemi-covid-19-di-indonesia/

[44] A. Salamanis, G. Xanthopoulou, D. Kehagias, and D. Tzovaras, "LSTM-Based Deep Learning Models for Long-Term Tourism Demand Forecasting," *Electronics (Switzerland)*, vol. 11, no. 22, 2022, doi: 10.3390/electronics11223681.

[45] S. Siregar, *Statistik Parametrik untuk Penelitian Kuantitatif Dilengkapi dengan Perhitungan Manual & Aplikasi SPSS Versi 17*. Jakarta: PT. Bumi Perkasa, 2017.

[46] C. D. Lewis, *Industrial and Business Forecasting Methods*. London: Butterworth, 1982.

[47] Dr. P. R. Motivation, "Neural Networks vs. Random Forests – Does it always have to be Deep Learning?," 2018.