Journal of Soft Computing Exploration

Homepage: https://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Predicting international tourist arrivals in north sumatra using machine learning and google trends keyword selection

Angga Putra Juledi [1], Fajar Maulana[2], Yofhanda Septi Eirlangga[3], Yomei Hendra[2,4], Qiuyun Cao[5]
[1]Study Program of Information System, Universitas Labuhanbatu, Indonesia
[2]Study Program of Informatics, Universitas Adzkia, Indonesia
[3]Study Program of Information System, Universitas Adzkia, Indonesia
[4]Graduate School of Environmental, Life, Natural Science and Technology, Okayama University, Japan

Article Info

Article history:

Received May 8, 2026
Revised June 3, 2026
Accepted July 4, 2026

Keywords:

Tourism forecasting
International tourist arrivals
Google trends
Random forest
Keyword selection

ABSTRACT

This study examined the prediction of monthly international tourist arrivals in North Sumatra by combining historical visitation data with Google Trends keyword selection in a machine learning framework. Monthly data from January 2010 to December 2025 were analyzed. The proposed approach constructed temporal features from past arrivals and search-interest features from tourism-related keywords grouped into destination, travel-intent, and attraction-specific segments. A seasonal naive model was used as the baseline, while Random Forest and Gradient Boosting were applied as the main prediction models. The results showed that the historical-only Random Forest model achieved the best performance among the main forecasting scenarios and clearly outperformed the seasonal baseline. The inclusion of all Google Trends keywords did not improve prediction accuracy consistently. However, the destination segment provided more useful predictive information than the other keyword groups. Further keyword selection revealed that Bukit Lawang was the most robust single keyword, while a compact subset consisting of Danau Toba, Medan, Sumatra Utara, and Bukit Lawang produced the best test performance. These findings indicated that Google Trends improved forecasting accuracy only when relevant keywords were selected carefully.

Corresponding Author:

Angga Putra Juledi,
Study Program of Information System,
Universitas Labuhanbatu,
Jl. S.M. Raja No. 126 A Aek Tapa, Rantauprapat, Kab. Labuhanbatu, Sumatera Utara, Indonesia
Email: anggapj19@ulb.ac.id
https://doi.org/10.52465/joscex.v7i3.107

1. INTRODUCTION

Tourism has contributed significantly to regional economic development through visitor spending, supporting services, and destination-related business activities [1], [2], [3]. For North Sumatra, the ability to predict international tourist arrivals has become increasingly important because it can support promotion strategies, destination capacity management, and evidence-based tourism planning [4]. However, tourism demand is dynamic, as it is influenced by historical visitation patterns, seasonality, and changing travel interest over time [5]. These characteristics make tourism forecasting a challenging problem, especially when the target series is affected by nonlinear temporal relationships and external shocks such as global health crises [6].

A growing body of research has examined machine learning methods for tourism demand forecasting, with considerable advances reported in recent years. Tree-based ensemble methods, particularly Random Forest and Gradient Boosting, have been widely applied to capture nonlinear patterns in tourist arrival series and have consistently outperformed traditional time-series benchmarks [7], [8]. Almotiri [9] applied Random Forest, Gradient Boosting, and ensemble voting models to predict city-level tourist arrivals in Saudi Arabia and demonstrated strong temporal holdout performance, highlighting the suitability of ensemble tree-based approaches for destination-specific forecasting tasks. Similarly, Wu et al. [10] reviewed big data forecasting methods in tourism and hospitality and found that tree-based ensemble models produced accurate and stable results across multiple destination contexts. In the Indonesian context, machine learning and time-series models have also been used to forecast international tourist arrivals, including studies incorporating structural break interventions to account for COVID-19 disruptions and post-pandemic recovery [11]. For North Sumatra specifically, research has examined the time-series structure of international arrivals during the pandemic period to improve data quality and modeling validity for subsequent forecasting analysis [12].

In parallel, the use of internet search data, particularly Google Trends, has gained substantial attention as a source of leading indicators for tourism demand [13]–[18]. Botha and Saayman [13] found that Google Trends indices provided useful predictive information for forecasting tourism demand recovery in South Africa following COVID-19 disruptions, while De Luca and Rosciano [14] demonstrated that transfer function models incorporating Google Trends data improved tourism demand predictions for United States tourists visiting Italy. Mikulić [15] also emphasized that the use of Google Trends and Baidu Index data in tourism forecasting needs careful assessment because their predictive value is not always consistent. Yang et al. [16] showed that search-query data can be useful during the pandemic, but its effectiveness depends on timing and destination context. In addition to search-query data, tourist-generated online reviews have also been used as internet-generated predictors in tourism demand forecasting [17]. Hu and Wu [18] further showed that Google Trends indices could contribute to forecast combinations in tourism. These findings suggest that internet-generated data can support tourism forecasting, but their usefulness depends on data type, keyword relevance, and modeling strategy.

Although studies on tourism forecasting with internet-generated data have continued to grow [13]–[18], research focusing specifically on North Sumatra remains limited [3], [11]. Existing studies in similar regional contexts generally used a restricted set of digital indicators or focused on different tourism targets, and they did not systematically investigate the role of Google Trends keyword segmentation and keyword subset selection for forecasting international tourist arrivals [3], [6], [11], [14], [17]. North Sumatra presents a distinct forecasting challenge because it hosts internationally recognized attractions including Lake Toba and Bukit Lawang that generate differentiated patterns of online search interest, yet empirical evidence on how these search signals relate to actual arrival dynamics is limited. This gap indicates that the predictive value of segmented Google Trends keywords in the North Sumatra context has not been sufficiently explored.

Based on the above background, this research intended to develop a machine learning framework for predicting international tourist arrivals in North Sumatra by combining historical visitation data with Google Trends keyword selection. The objectives of this research were threefold: first, to compare the performance of historical-only and Google Trends-augmented models; second, to evaluate the predictive contribution of destination, travel-intent, and attraction-specific keyword segments; and third, to identify the most effective destination-related keyword subset for improving forecasting accuracy. The novelty of this study lies in its integration of a destination-specific keyword segmentation framework with exhaustive keyword subset selection within a machine learning forecasting pipeline, applied to a regional destination in Indonesia that has received limited attention in the tourism forecasting literature. Unlike previous studies that either used a single keyword or adopted all available search terms without systematic evaluation [6], [12]-[17], this study provides empirical evidence on which keyword categories and compact subsets most effectively augment historical forecasting models for an emerging international tourist destination.

2. METHOD

Research framework

This study employed a machine learning-based framework to forecast monthly international tourist arrivals in North Sumatra. As illustrated in Figure 1, the research procedure consisted of five stages: data collection, data preprocessing, feature extraction, model development, and model evaluation. The collected tourist-arrival and Google Trends data were aligned by month, transformed into historical and search-interest predictors, and evaluated through time-series forecasting experiments to determine the best model scenario and keyword subset.

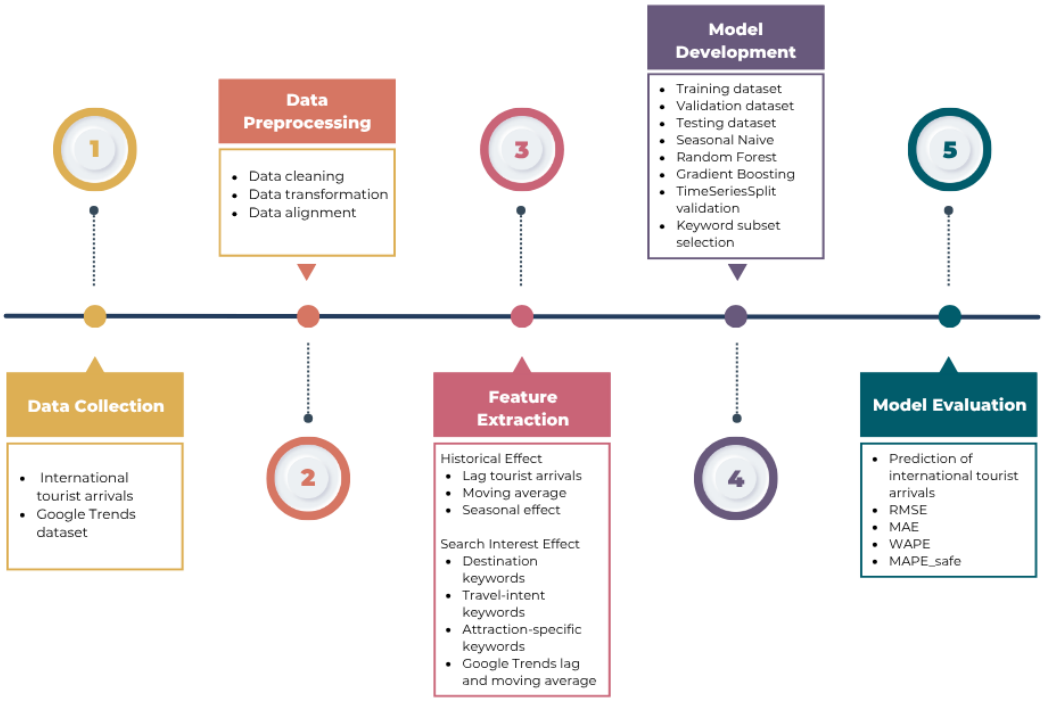

Figure 1. Proposed research methodology.

Data Colletion

This study used two main data sources: monthly international tourist arrivals to North Sumatra (2010-2025) from the Central Bureau of Statistics, and Google Trends data, representing online search interest related to tourism in the region. The Google Trends data, collected monthly for the same period, included keywords grouped into three segments: destination (e.g., Lake Toba, Medan), travel-intent (e.g., *flight to Medan, hotel in Medan*), and attraction-specific (e.g., *Lake Toba tour*). A sample of the raw dataset, including the target variable and Google Trends features, is shown in Table 1 before preprocessing and feature engineering.

Table 1. Sample of the raw monthly dataset

Year	Month	Tourist Arrivals	GT_Lake_Toba	GT_North_Sumatra	GT_Danau_Toba	GT_Medan	GT_Sumatra_Utara	GT_Bukit_Lawang
2010	Jan	12990	19	9	2	14	16	12
2010	Feb	13407	20	10	3	13	30	22
2010	Mar	17649	21	14	4	13	100	34
2010	Apr	14067	21	8	3	14	21	26
2010	May	15765	29	12	4	16	27	20
2010	Jun	17038	31	7	4	17	24	25

Data Preprocessing

The data were first aligned by month and year to ensure temporal consistency between the tourist arrival series and Google Trends series. The initial dataset consisted of 192 monthly observations covering the period from January 2010 to December 2025. The dataset was checked for duplicate records, format inconsistencies, non-numeric values, and missing observations. No missing values were found in the main monthly tourist-arrival and Google Trends variables. After cleaning, all observations were ordered chronologically to preserve the time-series structure.

Since the feature engineering stage generated lagged variables and moving-average features, several observations at the beginning of the series did not have complete predictor values. In particular, the use of a 12-month lag required the removal of the first 12 observations before modeling. Therefore, the model-ready dataset started from January 2011 and ended in December 2025, resulting in 180 monthly observations. This procedure ensured that all lag and moving-average predictors were constructed only from previous observations, so that no current or future information was used in the forecasting process.

Feature Extraction

Feature extraction was performed using two groups of predictors, namely historical features and Google Trends-derived features. The historical features were constructed from the monthly international tourist arrival series to capture temporal dependence and seasonal behavior [10], [11]. Specifically, lagged arrival features at 1, 2, 3, 6, and 12 months were generated. In addition, two moving-average features were constructed, namely a 3-month moving average and a 6-month moving average, both calculated from prior observations only.

The moving-average features were defined as:

$$MA3_t = \frac{y_{t-1} + y_{t-2} + y_{t-3}}{3} \quad (1)$$

$$MA6_t = \frac{y_{t-1} + y_{t-2} + y_{t-3} + y_{t-4} + y_{t-5} + y_{t-6}}{6} \quad (2)$$

where y_t denotes the number of international tourist arrivals at month t.

To represent annual seasonality, two cyclical variables were generated from the month index using sine and cosine transformations:

$$month_{sin_t} = \sin\left(\frac{2\pi m_t}{12}\right) \quad (3)$$

$$month_{cos_t} = \cos\left(\frac{2\pi m_t}{12}\right) \quad (4)$$

where m_t is the month number from 1 to 12.

Google Trends-derived features were constructed for each keyword included in a given scenario. For every selected keyword, three lag features and one short-term moving-average feature were generated, namely *GT_lag1, GT_lag2, GT_lag3*, and *GT_MA3*. The lagged Google Trends variables were used to capture delayed search effects, while the moving average was intended to smooth short-term fluctuations in online search interest [17]-[20]. The use of diverse feature extraction techniques for internet search query data has also been emphasized in tourism demand forecasting research [21].

Thus, the feature engineering process combined (1) historical arrival lags, (2) historical moving averages, (3) seasonal cyclical features, and (4) Google Trends lag and moving-average features. Since some engineered variables required previous observations, the initial rows with incomplete lag or moving-average values were removed before model training and testing.

Table 2 presents a sample of the engineered features used in the forecasting models. The table illustrates how the raw monthly variables were transformed into temporal and search-interest predictors before model development.

Table 2. Example of feature engineering results

Month	Tourist Arrivals	ta_lag_1	ta_lag_12	ta_MA3	month_sin	month_cos	GT_Bukit_Lawang lag1	GT_Bukit_Lawang_MA3
2011/01	15757	14782	12990	14858	0,5	0,866025404	23	25
2011/02	17034	15757	13407	14708	0,866025404	0,5	22	23
2011/03	17138	17034	17649	15857,66667	1	6,12323E-17	32	25,6667
2011/04	17192	17138	14067	16643	0,866025404	-0,5	27	27
2011/05	19133	17192	15765	17121,33333	0,5	-0,866025404	27	28,6667
2011/06	18888	19133	17038	17821	1,22465E-16	-1	25	26,3333

Model Development

Three forecasting approaches were developed in this study, namely Seasonal Naive, Random Forest, and Gradient Boosting. These models were selected to provide a comparison between a simple seasonal benchmark and two tree-based machine learning models capable of capturing nonlinear relationships in time-series and search-interest data, as commonly applied in recent tourism forecasting studies [19]–[21].

The first approach was Seasonal Naive, which served as the baseline model. In this approach, the prediction for a given month was set equal to the observed tourist-arrival value from the same month in the previous year. Formally, the Seasonal Naive prediction can be written as:

$$\hat{y}_t = y_t - 12 \tag{5}$$

where y_t denotes the actual number of international tourist arrivals at month t, and $\hat{y}_t$ denotes the predicted value. This baseline was used to evaluate whether the machine learning models provided meaningful gains beyond a simple yearly seasonal repetition [22].

The second approach was Random Forest, which is an ensemble learning method based on multiple decision trees constructed using the bagging principle [8], [23]. Each tree was trained on a bootstrap sample of the training data, and the final prediction was obtained by averaging the predictions of all trees. Random Forest was selected because it is able to capture nonlinear dependencies, handle interactions among predictors, and remain relatively robust in the presence of noisy or redundant variables, which is particularly relevant when using multiple Google Trends-derived features [19]–[21]. In this study, the Random Forest model was implemented with 300 trees, a maximum tree depth of 8, and a fixed random state of 42 to ensure reproducibility.

The third approach was Gradient Boosting, which is another tree-based ensemble method, but unlike Random Forest, it constructs decision trees sequentially. Each new tree was trained to minimize the residual errors generated by the previous ensemble, thereby gradually improving the overall predictive performance [23]–[25]. Gradient Boosting was included to compare a boosting-based approach with a bagging-based approach in the same forecasting setting. In this study, the Gradient Boosting model was implemented using the experimental configuration applied in the forecasting pipeline, with a fixed random state of 42.

The model development process was conducted using the engineered features described in the previous subsection. For each experimental scenario, the corresponding feature matrix was generated and then used to train the prediction models. The historical-only scenario used lagged arrivals, moving averages, and seasonal cyclical features. The Google Trends-augmented scenarios extended the historical feature set by adding segmented search-interest variables and their derived lag and moving-average features.

To evaluate model stability, the training data were further validated using *TimeSeriesSplit* with five folds, so that each validation fold occurred after its corresponding training fold. This design preserved the chronological structure of the time series and avoided information leakage from future observations. The final models were then trained on the complete training set and evaluated on the last 24 months of unseen data.

After the main forecasting comparison was completed, an additional keyword subset selection stage was applied within the destination-keyword segment. Because this segment showed the strongest potential in the earlier experiments, all non-empty combinations of the eight destination-related keywords were tested. This resulted in 255 keyword subsets, each of which was evaluated using the same Random Forest modeling framework. The purpose of this stage was to determine whether a compact and relevant subset of Google Trends keywords could improve forecasting accuracy beyond the historical-only benchmark.

Overall, the model development stage was designed as a sequential forecasting comparison. First, the Seasonal Naive model was used as a baseline to represent a simple yearly seasonal forecasting approach. The Random Forest and Gradient Boosting models were then developed using historical features to evaluate whether machine learning models could capture nonlinear temporal patterns more effectively than the baseline model. After that, Google Trends-derived features were added to the historical feature set to examine whether online search-interest data could improve forecasting accuracy. The Google Trends variables were evaluated based on keyword segments, namely destination-related, travel-intent, and attraction-specific keywords. Finally, because the destination segment showed the strongest forecasting potential, an exhaustive keyword subset-selection process was conducted within this segment to identify a compact group of keywords that provided the best forecasting performance.

Experimental Design

The experiments were conducted using a time-ordered forecasting design. The last 24 months of the dataset were used as the testing set, while the earlier observations were used as the training set. This setting allowed the models to be evaluated on the most recent unseen data.

Within the training data, validation was performed using *TimeSeriesSplit* with five folds. This procedure preserved the temporal order by ensuring that the validation observations always occurred after the corresponding training observations.

The experiments were organized into several scenarios. The first scenario used only historical features. The second scenario added Google Trends-derived features. The Google Trends variables were further tested in three segments: destination, travel-intent, and attraction-specific. Another scenario used all Google Trends variables together. After the destination segment showed the best potential, an exhaustive subset-selection experiment was conducted within this segment. Since the destination segment contained eight keywords, all non-empty combinations were evaluated, resulting in 255 keyword subsets.

Evaluation Metrics

Model performance was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Weighted Absolute Percentage Error (WAPE), and safe Mean Absolute Percentage Error (MAPE_safe). These metrics were selected to evaluate prediction error from different perspectives: RMSE penalizes large errors, MAE measures the average absolute error in the original unit, WAPE provides a scale-independent aggregate percentage error, and MAPE_safe avoids division by zero by calculating percentage error only for observations with non-zero actual values [26]–[33]. Let y_i denote the actual number of international tourist arrivals, $\hat{y}_i$ denote the predicted value, and n denote the number of observations, and S denote the set of observations with non-zero actual values. The metrics were calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

$$WAPE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i|} \times 100\% \tag{8}$$

$$MAPE\_safe = \frac{1}{n^*}\sum_{i \epsilon S}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{9}$$

3. RESULTS AND DISCUSSIONS

Descriptive Analysis

The monthly series of international tourist arrivals to North Sumatra from 2010 to 2025 showed substantial temporal variation. In general, the series reflected recurring seasonal patterns, but it also contained periods of sharp decline and recovery. This indicated that tourist arrivals were influenced not only by historical patterns but also by external shocks and changing travel interest over time. These characteristics justified the use of forecasting models capable of capturing nonlinear patterns and temporal dependence.

The Google Trends variables also showed different patterns across keyword groups. Destination-related keywords tended to have more stable fluctuations over time, while several travel-intent and attraction-specific keywords appeared more volatile. This suggested that not all search-query groups provided the same level of predictive information for international tourist arrivals.

Figure 2 illustrates the monthly pattern of international tourist arrivals in North Sumatra during the observation period.

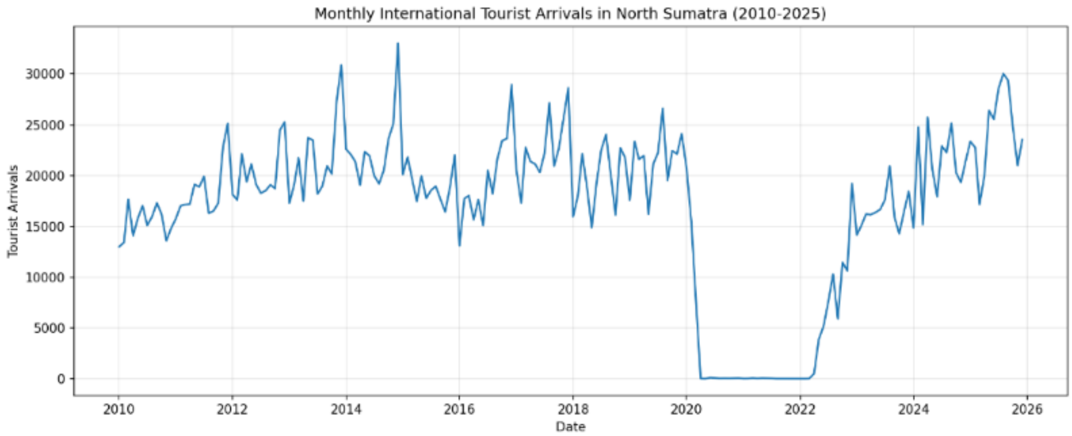

Figure 2. Monthly international tourist arrivals in North Sumatra (2010-2025)

Baseline Comparison

A Seasonal Naive model was used as the baseline. In this approach, the predicted value for a given month was set equal to the observed value from the same month in the previous year. This baseline was selected because the target series was monthly and potentially contained yearly seasonality.

The Seasonal Naive model produced a test Root Mean Squared Error of 5472.72, a test Mean Absolute Error of 4663.71, a safe Mean Absolute Percentage Error of 19.70%, and a Weighted Absolute Percentage Error of 20.62%. These results served as the benchmark for evaluating whether the machine learning models provided meaningful improvements. The baseline errors were substantially higher than those of the machine learning models. This indicated that simple yearly repetition was insufficient to capture the dynamics of international tourist arrivals in North Sumatra.

Main Forecasting Performance

The first main experiment compared historical-only models with models augmented by Google Trends-derived features. The results are presented in Table 2.

Table 2. Comparison of the main forecasting scenarios

No	Scenario	Model	Test RMSE	Test MAE	Test WAPE
1	Historical Only	Random Forest	**3913.07**	3236.96	14.31
2	Historical Only	Gradient Boosting	4413.95	3438.94	15.20
3	Historical + Destination Keywords	Random Forest	4056.00	**3234.20**	**14.30**
4	Historical + Destination Keywords	Gradient Boosting	4410.63	3300.05	14.59
5	Historical + Travel-Intent Keywords	Random Forest	4302.65	3368.57	14.89
6	Historical + Travel-Intent Keywords	Gradient Boosting	4460.44	3505.26	15.50
7	Historical + Attraction-Specific Keywords	Random Forest	4057.18	3351.00	14.81
8	Historical + Attraction-Specific Keywords	Gradient Boosting	4387.40	3525.34	15.58
9	Historical + All Google Trends	Random Forest	4255.22	3285.81	14.53
10	Historical + All Google Trends	Gradient Boosting	4259.21	3263.52	14.43

The results showed that the historical-only Random Forest model achieved the best overall performance among the main scenarios, with a test Root Mean Squared Error of 3913.07. This model also clearly outperformed the Seasonal Naive baseline, reducing the Root Mean Squared Error by 28.50%. The historical-only Gradient Boosting model also outperformed the baseline, but its performance remained below that of Random Forest.

When Google Trends-derived features were included, the improvement was not consistent across all scenarios. Among the segmented Google Trends experiments, the destination keyword segment produced the most promising results. In this scenario, the Random Forest model yielded slightly lower Mean Absolute Error and Weighted Absolute Percentage Error than the historical-only model, although its Root Mean Squared Error remained slightly higher. By contrast, the travel-intent, attraction-specific, and all-keywords scenarios did not outperform the historical-only Random Forest model in a consistent manner.

Segment-Based Evaluation of Google Trends

The segment-based analysis showed that the usefulness of Google Trends depended strongly on the type of keyword being used. Among the three keyword groups, destination-related queries provided the most informative external signal. This was likely because searches involving destination names such as *Lake Toba, Medan, North Sumatra,* and *Bukit Lawang* were more closely related to destination awareness and travel consideration.

In contrast, travel-intent keywords such as *flight to Medan, hotel in Medan,* and *Kualanamu airport* did not improve the prediction results. One possible explanation was that these queries were too broad or not sufficiently specific to actual international arrivals in North Sumatra. Similarly, the attraction-specific segment was not strong enough to improve performance consistently.

These findings indicated that the predictive value of Google Trends could not be assumed simply from the availability of search data. Because the destination segment showed the strongest potential, it was further examined through keyword subset selection.

Keyword Selection within the Destination Segment

Since the destination segment showed the best potential, an additional keyword-selection experiment was conducted within this group. All non-empty combinations of the eight destination-related keywords were evaluated, resulting in 255 subsets. The purpose of this experiment was to determine whether a compact subset of keywords could provide better predictive performance than the full destination segment. The best results are presented in Table 3.

Table 3. Keyword selection within the destination segment

Model Setting	Keywords	Test RMSE	Test MAE	Test WAPE	RMSE Improvement vs Historical Only
Historical Only + Random Forest	–	3913.07	3236.96	14.31	–
Historical + Random Forest	GT_Bukit_Lawang	3859.43	3187.31	14.09	+1.37%
Historical + Random Forest	GT_Danau_Toba, GT_Medan, GT_Sumatra_Utara, GT_Bukit_Lawang	**3788.96**	**3131.16**	**13.84**	**+3.17%**

The keyword-selection results demonstrated that the predictive value of Google Trends depended strongly on the selected keywords. The single keyword *GT_Bukit_Lawang* reduced the Root Mean Squared Error to 3859.43, improving the historical-only Random Forest benchmark by 1.37%. This result showed that even one carefully selected keyword could provide useful external information. The visual comparison of actual and predicted values for the selected models is shown in Figure 3.

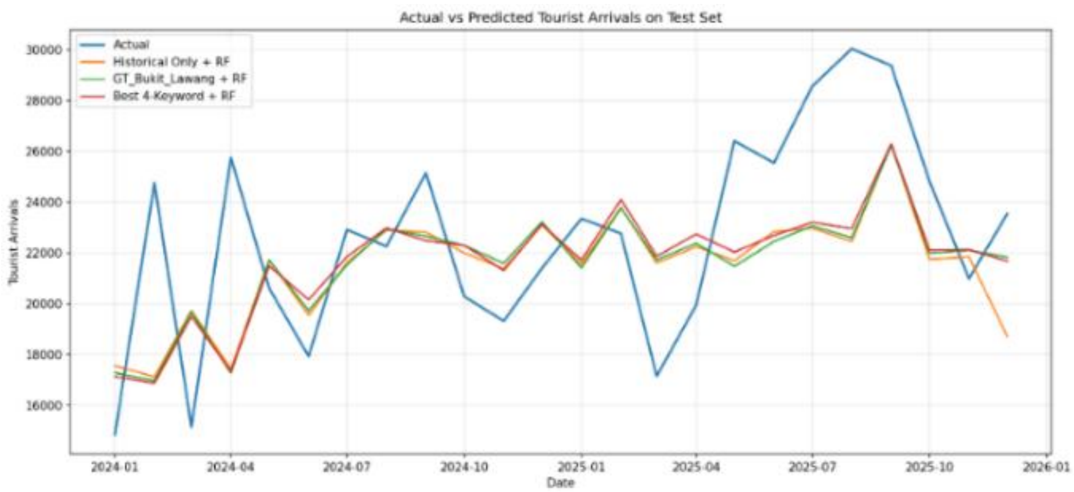

Figure 3. Actual and predicted tourist arrivals on the test set for the selected Random Forest models.

The best test performance was obtained from a four-keyword subset consisting of *GT_Danau_Toba, GT_Medan, GT_Sumatra_Utara,* and *GT_Bukit_Lawang*, which reduced the Root Mean Squared Error to 3788.96. This corresponded to an improvement of 3.17% over the historical-only Random Forest model. The same subset also produced the lowest Mean Absolute Error and Weighted Absolute Percentage Error among the evaluated models.

These results suggested that keyword selection was more important than simply increasing the number of Google Trends variables. When all keywords were used simultaneously, the prediction accuracy did not improve. However, when a compact and relevant subset was selected, measurable forecasting gains were achieved.

Discussion

The findings of this study indicate that historical visitation dynamics remained the dominant predictor of international tourist arrivals in North Sumatra. This can be seen from the strong performance of the historical-only Random Forest model, which outperformed most of the Google Trends-augmented scenarios. The result suggests that past arrival patterns, seasonal behavior, and short-term movement captured through lag and moving-average features already contained substantial predictive information The result suggests that past arrival patterns, seasonal behavior, and short-term movement captured through lag and moving-average features already contained substantial predictive information [34]–[36]. Therefore, although external digital indicators can provide additional information, historical tourist-arrival data still played the central role in forecasting monthly international arrivals in this study.

The contribution of Google Trends was found to be conditional rather than automatic. Including all Google Trends variables did not consistently improve forecasting accuracy, indicating that a broad set of search keywords may introduce redundancy, irrelevant variation, or noise into the model. This finding supports the importance of keyword selection before using search-interest data as forecasting predictors. In this study, destination-related keywords provided more useful information than travel-intent and attraction-specific keywords because they were more directly associated with destination awareness and general interest in North Sumatra. In contrast, travel-intent and attraction-specific keywords may have been too broad, too narrow, or less directly connected to aggregate international tourist arrivals.

The subset-selection results further confirmed that the predictive value of Google Trends depended on the relevance of the selected keywords. *Bukit Lawang* emerged as the most robust individual keyword, while the compact subset consisting of *Danau Toba, Medan, Sumatra Utara,* and *Bukit Lawang* produced the best forecasting performance. This result shows that using more search terms does not necessarily improve model accuracy. Instead, a smaller but more relevant keyword subset can provide additional predictive value beyond historical features. Overall, these findings demonstrate that Google Trends can be a valuable complementary predictor for tourism forecasting, but its effectiveness depends strongly on keyword segmentation and careful keyword selection.

4. CONCLUSION

This study investigated the prediction of monthly international tourist arrivals in North Sumatra by combining historical visitation data with Google Trends keyword selection in a machine learning framework. In line with the research objectives stated in the Introduction, the results and discussion confirmed that the predictive value of Google Trends depends on keyword segmentation and careful keyword selection rather than on the simple inclusion of all available search terms. The results showed that the historical-only Random Forest model achieved the best performance among the main forecasting scenarios and clearly outperformed the Seasonal Naive baseline. This indicated that historical visitation dynamics remained the strongest predictor in the current setting.

The experiments also showed that the inclusion of Google Trends did not improve forecasting accuracy consistently when all keywords were used together. However, the segment-based analysis revealed that destination-related keywords provided more useful predictive information than travel-intent and attraction-specific keywords. Further keyword selection demonstrated that *Bukit Lawang* was the most robust single keyword, while a compact subset consisting of *Danau Toba, Medan, Sumatra Utara*, and *Bukit Lawang* produced the best test performance. These findings indicated that Google Trends was useful as a complementary predictor only when relevant keywords were selected carefully.

This study contributed by showing that the predictive value of Google Trends depended on keyword segmentation and subset selection rather than on the inclusion of a large number of search terms. From an application perspective, the proposed forecasting framework can support tourism planners and local tourism authorities in monitoring destination-related online search interest as an additional signal for tourism demand planning, which is relevant to tourism economic planning and destination-related business activities [35]. This study was limited to two machine learning models and a restricted set of external variables. Future research may extend this work by incorporating additional external variables, evaluating more forecasting models, and applying interpretable machine learning techniques to better understand feature contributions. The framework may also be applied to other regional tourism destinations to examine whether similar keyword-selection patterns occur in different destination contexts.

ACKNOWLEDGEMENTS

The authors would like to express their sincere appreciation to Universitas Labuhanbatu, the Institute for Research and Community Service (LPPM) of Universitas Labuhanbatu, and all individuals and institutions who supported the completion of this research.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

[1] X. Li, R. Law, G. Xie, dan S. Wang, “Review of tourism forecasting research with internet data,” *Tour. Manag.*, vol. 81, 2021, doi: https://doi.org/10.1016/j.tourman.2020.104245.

[2] Y. Yang, J. Guo, dan S. Sun, “Tourism demand forecasting and tourists ’ search behavior : evidence from segmented Baidu search volume,” *Data Sci. Manag.*, vol. 4, no. September, hal. 1–9, 2021, doi: 10.1016/j.dsm.2021.10.002.

[3] E. Setiawati *et al.*, “Planning an effective communication strategy for tourism information centers (TICs) in managing product information,” *J. Gastro Tour.*, vol. 4, no. 1, hal. 19–30, 2026, doi: https://doi.org/10.52465/jogasto.v4i1.643.

[4] E. Nurmawati, T. Sugiyarto, N. Artiari, dan A. Rahmawati, “Analyzing Instagram Engagement to Forecast Domestic Tourist Trips in Lake Toba and North Sumatra : A Dual Approach with Conventional Statistics and Machine Learning Techniques,” *J. Kepariwisataan Destin. Hosp. dan Perjalanan*, vol. 8, hal. 178–196, 2024, doi: 10.34013/jk.v8i2.1619.

[5] A. Guizzardi, F. M. E. Pons, G. Angelini, dan R. Ranieri, “Big data from dynamic pricing: A smart approach to tourism demand forecasting,” *Int. J. Forecast.*, vol. 37, no. 3, hal. 1049–1060, 2021, doi: 10.1016/j.ijforecast.2020.11.006.

[6] J. Bi, H. Li, dan Z. Fan, “Annals of Tourism Research Tourism demand forecasting with time series imaging : A deep learning model,” *Ann. Tour. Res.*, vol. 90, hal. 103255, 2021, doi: 10.1016/j.annals.2021.103255.

[7] S. Sun, M. Hu, S. Wang, dan C. Zhang, “How to capture tourists ’ search behavior in tourism forecasts ? A two-stage feature selection approach How to capture tourists ’ search behavior in tourism forecasts ? A two-stage feature selection approach,” *Expert Syst. Appl.*, vol. 213, 2023, doi: 10.1016/j.eswa.2022.118895.

[8] E. Park, J. Park, dan M. Hu, “Tourism demand forecasting with online news data mining,” *Ann. Tour. Res.*, vol. 90, 2021, doi: https://doi.org/10.1016/j.annals.2021.103273.

[9] A. G. Alsulami, A. Alharbi, U. A. Khan, A. M. Al-hejri, S. S. Alshamrani, dan J. Almotiri, “Predicting tourism growth in Saudi Arabia with machine learning models for vision 2030 perspective,” *Nature*, vol. 16, hal. 1–23, 2026.

[10] D. C. Wu, S. Zhong, dan J. Wu, “Tourism and Hospitality Forecasting With Big Data : A Systematic Review of the Literature,” *J. Hosp. Tour. Res.*, 2025, doi: 10.1177/10963480231223151.

[11] M. Hu, R. T. R. Qiu, D. C. Wu, dan H. Song, “Hierarchical pattern recognition for tourism demand forecasting,” *Tour. Manag.*, vol. 84, 2021, doi: 10.1016/j.tourman.2020.104263.Abstract.

[12] T. Eddy, “Adjusting Anomalies in International Tourist Arrivals to North Sumatra During the Peak COVID-19 Period (April 2020 to June 2022) to Enhance the Validity of Time Series Modeling,” *J. Res. Math. Trends Technol.*, vol. 7, no. 2, hal. 72–77, 2025, doi: 10.32734/jormtt.v7i2.21718.

[13] I. Botha dan A. Saayman, “Does Google Analytics Improve the Prediction of Tourism Demand Recovery ?,” *forecasting*, vol. 6, no. 4, hal. 908–924, 2024, doi: https://doi.org/10.3390/forecast6040045.

[14] G. De Luca dan M. Rosciano, “Google Trends data and transfer function models to predict tourism demand in Italy,” *J. Tour. Futur.*, hal. 1–17, 2023, doi: 10.1108/JTF-01-2023-0018.

[15] J. Mikulić dan R. M. Baumgärtne, “Google Trends and Baidu index data in tourism demand forecasting: A critical assessment of recent applications,” *Tour. Manag.*, vol. 110, 2025, doi: https://doi.org/10.1016/j.tourman.2025.105164.

[16] Y. Yang dan D. Ph, “Search query and tourism forecasting during the pandemic : When and where can digital footprints be helpful as predictors ?,” *Ann. Tour. Res.*, vol. 93, no. 1, 2022, doi: 10.1016/j.annals.2022.103365.

[17] M. Hu, H. Li, H. Song, X. Li, dan R. Law, “Tourism demand forecasting using tourist-generated online review data,” *Tour. Manag.*, vol. 90, 2022, doi: https://doi.org/10.1016/j.tourman.2022.104490.

[18] Y.-C. Hu dan G. Wu, “The impact of Google Trends index and encompassing tests on forecast combinations in tourism,” *Tour. Rev.*, vol. 77, no. 5, hal. 1276–1298, 2022, doi: https://doi.org/10.1108/TR-02-2022-0088.

[19] H. Li, H. Gao, dan H. Song, “Tourism forecasting with granular sentiment analysis,” *Ann. Tour. Res.*, vol. 103, no. 6, 2023, doi: 10.1016/j.annals.2023.103667.

[20] S. Xu, Y. Liu, dan C. Jin, “Forecasting daily tourism demand with multiple factors,” *Ann. Tour. Res.*, vol. 103, 2023, doi: https://doi.org/10.1016/j.annals.2023.103675.

[21] G. Xue, S. Liu, L. Ren, dan D. Gong, “Forecasting hourly attraction tourist volume with search engine and social media data for decision support,” *Inf. Process. Manag.*, 2023, doi: 10.1016/j.ipm.2023.103399.

[22] Y. Li, Z. Lin, dan S. Xiao, “Using social media big data for tourist demand forecasting : A new machine learning analytical approach,” *J. Digit. Econ.*, vol. 1, no. 1, hal. 32–43, 2022, doi: 10.1016/j.jdec.2022.08.006.

[23] L. He, H. Li, J. Bi, J. Yang, dan Q. Zhou, “Annals of Tourism Research The impact of public health emergencies on hotel demand - Estimation from a new foresight perspective on the COVID-19,” *Ann. Tour. Res.*, vol. 94, no. 38, hal. 103402, 2022, doi: 10.1016/j.annals.2022.103402.

[24] B. Wu dan L. Wang, “Interpretable tourism volume forecasting with multivariate time series under the impact of COVID-19,” *Neural Comput. Appl.*, vol. 35, no. 7, hal. 5437–5463, 2023, doi: 10.1007/s00521-022-07967-y.

[25] T. Hu, H. Wang, R. Law, dan J. Geng, “Diverse feature extraction techniques in internet search query to forecast tourism demand: An in-depth comparison,” *Tour. Manag. Perspect.*, vol. 47, 2023, doi: 10.1016/j.tmp.2023.101116.

[26] B. Wu, L. Wang, dan Y. Zeng, “Interpretable tourism demand forecasting with temporal fusion transformers amid COVID-19,” *Appl. Intell.*, vol. 53, hal. 14493–14514, 2023, doi: https://doi.org/10.1007/s10489-022-04254-0 Interpretable.

[27] J. Chen, Z. Ying, C. Zhang, dan T. Balezentis, “Forecasting tourism demand with search engine data: A hybrid CNN-BiLSTM model based on Boruta feature selection,” *Inf. Process. Manag.*, vol. 61, no. 3, 2024, doi: 10.1016/j.ipm.2024.103699.

[28] M. Hu, N. Dong, dan F. Hu, “Tourism demand forecasting using short video information,” *Ann. Tour. Res.*, vol. 109, 2024, doi: https://doi.org/10.1016/j.annals.2024.103838.

[29] X. Li, Y. Xu, R. Law, dan S. Wang, “Enhancing tourism demand forecasting with a transformer-based framework,” *Ann. Tour. Res.*, vol. 107, 2024, doi: https://doi.org/10.1016/j.annals.2024.103791.

[30] M. Crispino dan V. Mariani, “A Tool to Nowcast Tourist Overnight Stays with Payment Data and Complementary Indicators,” *SSRN Electron. J.*, 2023, doi: 10.2139/ssrn.4464052.

[31] M. Hu, M. Li, Y. Chen, dan H. Liu, “Tourism forecasting by mixed-frequency machine learning,” *Tour. Manag.*, vol. 106, 2025, doi: https://doi.org/10.1016/j.tourman.2024.105004.

[32] F. Qin, J.-W. Bi, H. Li, dan H. Xu, “Tourism demand forecasting using social media data: A deep learning–based ensemble model with social media communication conversion rates,” *Ann. Tour. Res.*, vol. 115, 2025, doi: https://doi.org/10.1016/j.annals.2025.104058.

[33] P. Fuleky dan C. S. Bonham, “Forecasting with Mixed Frequency Samples : The Case of Multiple Common Trends,” *Ann. Tour. Res.*, vol. 110, 2012.

[34] R. J. Hyndman dan G. Athanasopoulos, “Forecasting: principles and practice,” 2018.

[35] C. Li, W. Zheng, dan P. Ge, “Tourism demand forecasting with spatiotemporal features,” *Ann. Tour. Res.*, vol. 94, 2022, doi: https://doi.org/10.1016/j.annals.2022.103384.

[36] C. A. Tisdell, “Handbook on tourism economics,” in *World Scientific*, 2013.

.