

Deep Learning With Processing Algorithms for Forecasting Tourist Arrivals

Harun Mukhtar^{1,2}, Muhammad Akmal Remli^{2,3},
Khairul Nizar Syazwan Wan Salihin Wong², Mohd Saberi Mohamad⁴

¹ Faculty of Computer Science, Universitas Muhammadiyah Riau, Pekanbaru 28000, Riau, Indonesia

² Faculty of Data Science and Computing, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100 Kota Bharu, Kelantan, Malaysia

³ Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100 Kota Bharu, Kelantan, Malaysia

⁴ Department of Genetics and Genomics, College of Medical and Health Sciences, United Arab Emirates University, P.O. Box 17666, Al Ain, Abu Dhabi, United Arab Emirates

Abstract – The DL (Deep Learning) method is the standard for forecasting tourist arrivals. This method provides very good forecasting results but needs improvement if the data is small. Statistical data from the BPS (Central Bureau of Statistics) needs to be corrected, resulting in forecasts that tend to be invalid. This study uses statistical data and GT (Google Trends) as a solution so that the data is sufficient. GT data has a lot of noise because there is a shift between web searches and departures. This difference will produce noise that needs to be cleaned. We use monthly data from January 2008 to December 2021 from BPS sources combined with GT. Hilbert-Huang Transform (HHT) is proposed to clean data from various disturbances. The DL used in this study is long short-time memory (LSTM) and was evaluated using the root mean squared error RMSE and mean absolute percentage error (MAPE). The evaluation results show that the HHT-LSTM results are better than without data cleaning.

Keywords – Deep learning (DL), tourism arrivals, long short-time memory (LSTM), HHT, Google trends (GT), data.

1. Introduction

Tourism is an activity related to travel for recreation [23]. Tourism is one of the natural beauties and cultural. The existence of tourism will attract tourists. Tourists who come will increase regional income [11]. The number of visitors, the number of tourist objects, and the occupancy rate of hotels are factors that influence the tourism sector [2]. Tourism is the most dynamic industry in the world but it faces enormous challenges and requires good future planning. Forecasting is the only way to do the planning. This forecasting has the main purpose of imagining the future so that planning will be better. The arrival of tourists in large numbers will be able to improve the economy but there is also a very large risk [28]. The risk will be minimized with the forecast [16]. Tourism forecasts will be able to reduce the risk of failure in making decisions [4]. Tourism forecasting can also be used to improve the performance of employees or government employees [17]. Forecasting involves the following processes: 1. Determination of the purpose of the model; 2. Collection of past data; 3. Estimation of the initial model; 4. Doing forecasting; 5. Evaluate forecast of forecasting results.

First, determining the purpose of the model is a critical step in building a forecasting model. The modeler must identify the usefulness of the forecasting process. The modeler must also pay attention to the highly non-linear nature of the data. Second, the collection of past data from various sources.

DOI: 10.18421/TEM123-57

<https://doi.org/10.18421/TEM123-57>


Corresponding author: Harun Mukhtar
Faculty of Computer Science, Universitas Muhammadiyah
Riau, Pekanbaru 28000, Riau, Indonesia
Email: harunmukhtar@umri.ac.id

Received: 09 May 2023.

Revised: 29 July 2023.

Accepted: 14 August 2023.

Published: 28 August 2023.

 © 2023 Harun Mukhtar et al; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

This study uses two data sources, namely BPS and GT. BPS data is used as primary data, and GT data is used as additional data. Data is also processed using HHT to clean noise. Noise occurs because there are differences in search time and departure time. Third, estimate the model, namely estimating the appropriate forecasting model for the suitability of the volume of data and the type of data used. Fourth, evaluating forecasting results, evaluation is very important by considering how accurate the forecasting results are.

2. Problem Formulation

Forecasting tourist arrivals using time series data has four main components. First, Trend (T), the trend shows the number of tourist arrivals that have increased or decreased within a certain period. Second, Cycle (C), tourist arrival data can also be said to be this component because, at a certain time will experience the same event. Third, Seasonal (S), tourist arrivals usually have a certain season, and fourth, Irregular (I), natural events, for example, are unpredictable events that influence tourist arrivals. COVID-19 is an example of an event that affects tourist arrivals. Based on the components that have been described, forecasting tourist arrivals can be solved with two-equation models, namely the multiplication equation model and the addition equation model. See equation (1) as the multiplication model and equation (2) as the addition model.

$$Y = T * C * S * I \quad (1)$$

$$Y = T + C + S + I \quad (2)$$

Tourist arrival data is non-linear, so it is very difficult to make accurate predictions. In addition to statistical data, GT data is also used in this study. Google trends data has a noise that is getting harder and harder to resolve.

3. Fundamental Works

The discussion discussed includes three things: using Big Data for forecasting. HHT is used to clean data that has a lot of noise and DL uses LSTM for forecasting models.

3.1. Big Data

If not correctly estimated, the perishable nature of tourism products will result in considerable losses.

Tourism forecasting requires data timeliness, but in reality tourism data is always delayed by up to two months and also requires expensive costs.

We added data from GT. GT data can be analyzed to solve tourist arrival problems [26]. The use of trend data from big data analysis on the internet is able to improve forecasting performance compared to using only past arrival data and is able to improve prediction performance compared to using only past arrival data [36]. In fact, with the development of the times, there is big data that can be used as analysis [20]. Large collections of data from various types of sources, having different variations in data structures and spread across various technological devices are called big data [29]. The service provider's website to understand how frequently topics are covered is GT. The data generated by GT can be used as an explanatory variable in the tourist arrival forecasting model [10]. Although it has a lot of noise, this noise occurs due to many keyword searches [1]. However, the search for data on *Google Search* about tourism is always disturbed by noise [22]. Noise referred to is data disturbances that should be eliminated. The process of removing noise from the data needs to be done. Without noise, search engine data capabilities may be weak, even invalid.

Some researchers use web search data for their research. Among them Höpken et al [15], used web search data as an additional variable finding forecasting results that outperformed traditional autoregressive. Havranek et al [14], use GT's potential to predict more accurately on tourist arrival forecasting models. LSTM is able to handle complex time series problems in forecasting tourist arrivals [41].

3.2. The Hilbert Huang Transform Algorithm (HHT)

Previous research conducted to clean or reduce noise [12] which discussed reducing noise in datasets for Multi Class Classification with Decision Trees to detect rotor rod damage in induction motors [32]. One technique that is often used to clean noise is HHT [5]. HHT uses Empirical Mode Decomposition (EMD) to analyze and decipher signals [8] and to obtain frequency and amplitude information by utilizing the Intrinsic Mode Function (IMF) [39].

Xiaoxuan et al [37], proposed a denoising and forecasting model using search engine data. This denoising model is used to clean up data interference in tourism data retrieved from search engines. They predict that the forecasting ability will be weak or even invalid when data from search engines is processed without denoising.

Forecasting tourist arrivals is very important in tourism decision making. New strategies need to be developed, denoising factors are considered to improve forecasting accuracy [19] Denoising can increase the accuracy of the less-than-optimal Neural Network Autoregressive (NNAR) algorithm with seasonal tourism demand data [33].

3.3. The Deep Learning (DL) Algorithm

Method DL provides insights to improve forecasting accuracy. Forecasting tourist arrivals has challenges in using highly non-linear data. Broad insight to improve accuracy is very important to solve these forecasting problems. Non-linear data is complicated to solve, but DL solves the problem well [43]. Forecasting problems that use non-linear data require the right solution. The DL method can be a solution for forecasting and promises to achieve maximum accuracy [18]. Tourism forecasting using highly non-linear time series data is very suitable to be solved with LSTM [7].

This study uses LSTM, one of the DL methods, as an experimental method. LSTM is very popular among researchers for long-term and short-term forecasting. The arrival of tourists to Indonesia is used as the object of research. Monthly tourist arrival data is used and combined with search data in GT. The DL method is currently popularly used in all fields, including medicine and nursing [24], the field of forecasting the COVID-19 that has recently hit the world [25], and the areas of smart city development that are currently popular are discussed in [40].

4. A Deep Learning Method with Processing Algorithms

The main algorithm proposed in this study is the strategy used to produce near-optimal accuracy; in particular, this study suggests incorporating Google Trends data as additional data which is believed to reduce overfitting. GT data has a lot of noise caused by the shift between search and departure. The noise coping strategy is carried out using the HHT algorithm. These strategies can improve forecasting performance so that it is close to the maximum in terms of accuracy.

4.1. The Process of Forming Datasets

This study uses data from BPS Indonesia. The dataset comprises monthly entry data to Indonesia spanning from January 2008 to December 2021. An example of this data is taken from the website: <https://www.bps.go.id> in June 2022. Data is presented in tabular form in excel format. The contents of the table in this data consist of the number of foreign tourist visits per month to Indonesia by entry point. Other data in this study uses a dataset obtained directly from GT in an excel file with seven queries: Indonesia, tour to Indonesia and Bali. The query recommendations are: Bali Indonesia, Bali Indonesia hotels, the best hotels in Bali Indonesia using data from 01-01-2008 to 31-12-2021. Table 1 describes data sourced from BPS and then combined with Table 2 sourced from GT. The results of the merger are in Table 3.

Table 3 consists of 16 columns and 168 rows of data. The data in the first column (A) is for the month, namely from January 2008 to December 2021. The data in the second column (B) is taken from BPS Indonesia. The third column (C) data is taken from GT. The fourth column (D) to the sixteenth column (P) is filled with delays or shifts in arrival estimates after browsing Google Trends. To fill in the delay, formula (3) is needed.

$$D = ((x * y) * (x/y)) * y \quad (3)$$

Where D is the delay consisting of delay₀ to delay₁₂ (columns D to P from Table 3). X is the percentage of GT popularity, namely the trend data table (column C), and y is the number of queries used to obtain data from google trends (this study uses 7 queries (so y is 7). The delay is determined to be 12 because it is assumed that there will be a shift between one to twelve months. The resulting data is as shown in Table 3 and 7.

Table 7 is a dataset that was formed after the value 0 was deleted. By deleting the value 0 the data can be used properly. Table 7 is additional data with an available class. This dataset is derived through a distinct process, transforming it into GT like data format. With the formation of data sourced from Google Trends, this research uses two data, namely the main data from BPS and additional data from GT. Table 4 shows equation (3) in pseudocode form.

Table 1. Monthly tourist arrival data for 2008 to 2021 (source: BPS)

Month	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
January	437966	473165	493799	548821	652692	614328	753079	724698	814303	1107968	1097839	1201735	1290411	137230
February	465449	421555	523135	568057	592502	678415	702666	794302	888309	1023388	1197503	1243996	872765	115765
March	502041	511314	594242	598068	658602	725316	765607	792804	915019	1059777	1363426	1311911	486155	130933
April	459129	487121	555915	608093	626100	646117	726332	750999	901095	1171386	1302321	1274231	158066	125001
May	508955	521735	600031	600191	650883	700708	752363	794294	915206	1148588	1242705	1249536	161842	152604
June	529064	550582	613422	674402	695531	789594	851475	815307	857651	1144001	1322674	1434103	156561	137247
July	567364	593415	658476	745451	701200	717784	777210	815351	1032741	1370591	1547231	1468173	155742	135438
Aug	599506	566797	586530	621084	634194	771009	826821	853244	1031986	1393243	1511021	1530268	161549	124751
September	501018	493799	560367	650071	683584	770878	791296	870351	1006653	1250231	1370943	1388719	148984	124071
October	529391	547159	594654	656006	6688341	719903	808767	826196	1040651	1161565	1291605	1346434	152293	148645
November	524162	531669	578152	654948	693867	807422	764461	777976	1002333	1062030	1157483	1280781	144476	153199
December	610452	625419	644221	724539	766966	860655	915334	913828	1113328	1147031	1405554	1377067	164079	163619

Table 2. Tourist search data on Google Trends

Month	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
January	204	217	374	356	379	283	342	457	427	432	506	452	518	235
February	177	208	333	317	308	321	426	358	393	415	489	426	428	306
March	287	256	341	257	338	336	395	381	446	415	458	457	377	280
April	198	333	313	304	302	314	397	390	398	454	439	420	264	252
May	238	312	419	337	338	419	364	411	421	484	491	448	247	209
June	212	291	278	297	339	385	357	391	376	424	419	478	289	276
July	212	315	269	396	342	348	353	339	380	450	559	459	221	206
Aug	213	283	318	259	310	354	448	418	406	510	539	470	260	244
September	198	305	275	349	305	360	377	352	379	483	492	477	285	279
October	207	350	287	256	340	341	322	367	408	441	460	429	249	224
November	195	291	241	318	322	330	433	371	355	526	387	442	247	224
December	209	220	351	313	278	362	388	388	450	396	410	451	253	242

Table 3. Dataset after being combined

Month	BPS	Trend	Delay 0	Delay 1	Delay 2	Delay 3	Delay 4	Delay 5	Delay 6	Delay 7	Delay 8	Delay 9	Delay 10	Delay 11	Delay 12
2008-01	437966	204	2913120	0	0	0	0	0	0	0	0	0	0	0	0
2008-02	465449	177	219303291312	0	0	0	0	0	0	0	0	0	0	0	0
2008-03	502041	287	576583219303	291312	0	0	0	0	0	0	0	0	0	0	0
2008-04	459129	198	274428576583	219303	291312	0	0	0	0	0	0	0	0	0	0
2008-05	508955	238	396508274428	576583	219303	291312	0	0	0	0	0	0	0	0	0
...
2021-08	124751	244	416751297052	533232	305767	444528	548800	655452	386575	448063	427063	434007	568575	473200	
2021-09	124071	279	544887416751	297052	533232	45767	444528	548800	655452	386575	448063	427063	434007	568575	
2021-10	148645	224	351232544887	416751	297052	533232	305767	444528	548800	655452	386575	448063	427063	434007	
2021-11	153199	224	35123235123257	544887	41236751	297054	55232	38636751	444528	548800	655452	386575	448063	427063	
2021-12	163619	242	409948351232	351	232	544887	416751	297052	533232	305767	444528	548800	655452	386575	448063

4.2. Noise reduction with HHT

$$h1k = h1(k-1) - m1k \tag{5}$$

It is believe that the GT dataset still has noise that will interfere with accuracy if applied immediately. This dataset needs to be cleaned before use [30] The steps for cleaning a dataset using HTTP can be seen from the following mathematical formula:

$$h1 = x(t) - m1 \tag{4}$$

where h1 is the data (initial signal), x(t) is the extreme value of the movement, which includes the upper envelope and lower envelope, m1 is the mean of the different data between the upper envelope and lower envelope, h1k is the process of shifting up to k times, k is iteration which is done up to k times.

Table 4. Algorithm 1: Pseudocode to the process of forming data based on GT.

Input: data stream x, y <----- $y = 7$; queries used (seven queries)
Output: New X = D0, D1, D2, D3, D4, D5, D6, D7, D8, D9, D10, D11, D12
 For New X <----- $((x * y) * (x/y)) * y$ do
 input <----- $(x * y), (x/y), y$
 target <----- New X
 end

Table 5. Algorithm 2: Pseudocode to clean data noise using HHT.

Input: data stream $X = \{x1, x2, \dots, xi, \dots\}$; New X; New X1.
Output: Predicted result O
 X <----- $x1, x2, \dots, xi, \dots$ //Input data
 $O \{ \}$ //Output data
 New X <----- $y1 + delay$
 New X1 <----- $New X + New X1$ Combination
 $x(t)$ <----- $h1$
 For h_1 <----- $x(t) - m1$ do
 input <----- $h1(k-1), m1k$
 target <----- $h1k$
 end
 For $r1$ <----- $x(t) - h1k$ do
 input <----- $x(t), h1k$
 target <----- $r1k$
 end

Equation (4) functions to convert data into an initial signal. Then the signal is carried out by a second shifting process using equation (5). H1k is then separated from the residue (r). The mathematical formula for separating residues is given in equation (6). Rn is the repetition from the first residue to the last written in equation (7).

$$r1 = x(t) - h1k \tag{6}$$

$$x(t) = \sum_1^n h1k + rn \tag{7}$$

Formulas (1) to (7) are used to parse a signal called (EMD). If done repeatedly, it will produce IMF. Which produces a fixed residual value. To ensure that the results still have a physical meaning of modulation, amplitude, and frequency, it is necessary to limit the size of the standard deviation using equation (8).

$$SD = \sum_t^T 0 \left[\frac{|h1(k-1)(t) - h1(k)(t)|^2}{h2(k1)(t)} \right] \tag{8}$$

$h_{1(k-1)}$ is the initial signal and h_{1k} is the result of the initial signal minus the average of the initial signal. The formula is translated into pseudo-code. The HHT pseudo-code to clean data from noise is shown in Table 5.

4.3. LSTM Architecture Modeling

The DL model used in this study is the LSTM which is a development of the Recurrent Neural Network (RNN). to reduce errors during training using backpropagation [31]. LSTM can overcome long-term dependence on its inputs. This method provides predictions using step-by-step sequence data. The advantage of LSTM is that it is suitable for time series prediction. It contains special units called memory blocks in hidden layers that repeat. The memory block contains memory cells temporarily storing network state in addition to special copying units called information flow control gates. This is a very remarkable DL performance and has recently attracted the attention of researchers in the field of forecasting [27]. The following equation is used to predict:

$$ft = \sigma(Wf * [ht - 1 * xt] + bf) \tag{9}$$

$$it = \sigma(Wi * [ht - 1 * xt] + bi) \tag{10}$$

$$Ct = \tanh(Wc * [ht - 1 * xt] + bc) \tag{11}$$

$$Ct = ft * Ct - 1 + it * Ct \tag{12}$$

$$ot = \sigma(Wo * [ht - 1 * xt] + bo) \tag{13}$$

$$ht = ot * \tanh * Ct \tag{14}$$

All LSTM outer layers are connected with the cell state in the form of a horizontal line which is the primary key. Cell state can also be added or subtracted by LSTM. The LSTM has a sigmoid layer, and a dot multiplication operation called a gate. The sigmoid layer produces two outputs namely 1 and 0; 1 if the information is forwarded and 0 if the information is stopped.

$$\sigma(x) = 1/(1+e^x) \tag{15}$$

where, X = Input data; E = mathematical constant (2.71828 18284 59045 23536 02874 71352). The advantages of LSTM are found in three gates; the first is the forget gate function to determine which information will be deleted from the cell. Second, the input gate functions to determine the input value to be updated by processing ht-1 and xt in memory.

Third, the output gate functions to determine the output to be produced in the form of 0 to 1 in the Ct1 state.

Forget gate uses output at time t-1 and input at time t; if the result is 0 then it is forgotten and if the result is 1 then the state does not change.. The forget gate equation is as follows.

$$ft = \sigma(Wf * [ht - 1 * xt] + bf) \quad (16)$$

where ft is forget gate; σ is forgetgate; Wf is forgetgate; $ht - 1$ is the output value before t; xt is input value on order t; bf is bias value at forget gate. The weight value is described in equation (17).

$$W = \left(\frac{-1}{\sqrt{d}} \cdot \frac{1}{\sqrt{d}} \right) \quad (17)$$

where: W is weight; d is number of data. Storage of information into cells consists of two parts; first, the input gate decides which value needs to be updated, and the tanh layer creates a candidate with the new value added to the cell state. Second, the output from the input and the tanh gate is combined to fix the cell state. The final result of this process is between 0 and 1.

$$it = \sigma(Wi * [ht - 1 * xt] + bi) \quad (18)$$

where: it is the input gate, σ is the sigmoid function, Wi is the weight value for the input gate, $ht - 1$ is the output value before order t, x is the input value at sequence t, bi is the bias value at the input gate. Then the input is multiplied by the candidate output. The new candidate equations are described in equation (19).

$$Ct = \tanh(Wc * [ht - 1 * xt] + bc) \quad (19)$$

where: t is the new value that can be added to the cell, \tanh is the tanh function, Wc is the weight value for the cell, $ht - 1$ is the output value before order t, xt is the input value at sequence t, bc is the bias value for the state of the cell.

Table 6. Algorithm 2: Pseudo-code for forecasting tourist arrivals using the DL method

<p>Input: Xt-1 and Input order t: Xt+1 <----- range 1 to 0.</p> <p>Output: $ht - 1$ <----- range 0 to 1</p> <p>Clean data are taken from Algorithm 1, to Input</p> <p>Output: Predicted result O</p> <p>if $ft = \sigma(Wf * [ht - 1 * xt] + bf)$ then 1; else if $it = \sigma(Wi * [ht - 1 * xt] + bi)$ then 1; else if $Ct = \tanh(Wc * [ht - 1 * xt] + bc)$ then 1; else if $Ct = ft * Ct - 1 + it * Ct$ then 1; else if $ot = \sigma(Wo * [ht - 1 * xt] + bo)$ then 1; else if $ht = ot * \tanh * Ct$ then 1; end if</p> <p>If Sigmoid = $\sigma(x) = 1/(1+\epsilon)^{-x}$ then ht-1 and xt if else $ft = \sigma(Wf * [ht - 1 * xt] + bf)$ then 1; if else $C\sim t + \text{cell state} <----$ then 1; If else $it = \sigma(Wi * [ht - 1, xt] + bi)W =$ $\left(\frac{-1}{\sqrt{d}} \cdot \frac{1}{\sqrt{d}} \right) <----$ then 1; // input multiplied if else $it = \sigma(Wi * [ht - 1 * xt] + bi) <----$ it * $C\sim t$ then 1; if else $Ct = \tanh(Wc * [ht - 1 * xt] + bc)$ then 1; if else $ot = \sigma(Wo * [ht - 1 * xt] + bo)$ then 1; if else $ht = ot * \tanh(Ct)$ then 0 end if</p>
--

Equation (20) describes the formation of a new cell state by multiplying the old cell state by ft, then adding a unique value used to update the status.

$$ct = ft * (ct - 1) + it * t \quad (20)$$

where: ct is the cell state; ft is the forget gate; $ct - 1$ is the cell state before order t; it is the input gate. Equation (21) is used to determine the output on the LSTM. The resulting work must match the input received. The sigmoid layer decides which part will be the result, then the output is fed into the tanh layer to change the value between -1 to 1 and multiplied by the sigmoid gate. The output gate functions to control how many states will be issued.

$$ot = \sigma(Wo * [ht - 1 * xt] + bo) \quad (21)$$

where: ot is the output gate; σ the is output gates; Wo is the weight value for output gate; $ht - 1$ is the weight value for output gate; xt is the input value on order t; bo is the input value on order t.

The output value equation of order t is described in equation (22).

$$ht = ot * \tanh(Ct) \tag{22}$$

where: ht is the output value of order t , ot is the output gate, \tanh is the output value of order t , Ct is the output gate. Equations above are described in a problem solving algorithm using pseudocode. Table 6 explains pseudocode for forecasting tourist arrivals using the DL method.

4.4. Hybrid HHT – DL

The research methodology involves employing the DL model or specifically the DNN (Deep Neural Network) to forecast foreign tourist arrivals.

The models used are often very complex and overfitting. This overfitting is caused by two things, namely there is little data and the requirements for explanatory variables are needed. Big data analysis using web search can be used as additional data and if processed properly will increase accuracy in forecasting tourism visits [41].

Predicting tourist arrivals using a combination of denoising and DL algorithms is proposed. HHT is the best choice for the denoising algorithm used on the discussed trend data. Hybrid HHT-DL is predicted to be able to increase the accuracy of the proposed forecasting. In this work, hybrid is intended to combine data processing algorithms with DL algorithms. The output of the processing algorithm is used as input to the DL.

Table 7. Description of dataset

Month	Delay1	Delay2	Delay3	Delay4	Delay5	Delay6	Delay7	Delay8	Delay9	Delay10	Delay11	Delay12
2009-01	305767	266175	299943	274428	317583	314608	314608	396508	274428	576583	219303	291312
2009-02	329623	305767	266175	299943	274428	317583	314608	314608	396508	274428	576583	219303
2009-03	302848	329623	305767	266175	299943	274428	317583	314608	314608	396508	274428	576583
2009-04	458752	302848	329623	305767	266175	299943	274428	317583	314608	314608	396508	274428
2009-05	776223	458752	302848	329623	305767	266175	299943	274428	317583	314608	314608	396508
...
2021-08	297052	533232	305767	444528	548800	655452	386575	448063	427063	434007	568575	473200
2021-09	416752	297052	533232	305767	444528	548800	655452	386575	448063	427063	434007	568575
2021-10	544887	416752	297052	533232	305767	444528	548800	655452	386575	448063	427063	434007
2021-11	351232	544887	416752	297052	533232	305767	444528	548800	655452	386575	448063	427063
2021-12	351232	351232	544887	416752	297052	533232	305767	444528	548800	655452	386575	448063

156 rows x 12 columns

5. Experiments

Experiments were carried out to get the best results. The investigation starts from the dataset, the experimental results, and the evaluation.

5.1. Dataset and experimental setup

The datasets used in the study are summarized in Table 7. This dataset consists of two datasets with different sources. They were obtained after combining the two datasets discussed. This dataset consists of 156 rows and 12 columns. This data line contains monthly data from January 2009 to December 2021. The 2008 data line was deleted because it has a value of 0. All data used for this research is stored at: <https://github.com/harunmukhtar/Prediksi-Visits-Tourist-Ke-Indonesia/tree/main/data>.

The program script for analysis and testing can be downloaded at:

<https://github.com/harunmukhtar/HTT---LSTM/tree/main/project%20hht-lstm>.

Comparisons were made using three data, namely data from BPS, GT that had not been cleaned, and GT that had been cleaned with HHT. All experiments were performed using google colab on a Dell Precision 3640 workstation with Intel(R) Xeon(R) W-1250P RAM @4.10GHz (12 CPU), 4.1Ghz 32768MB on Windows 10 Pro for Workstations 64-bit (10.0, build 19044). To achieve statistically significant results, the experiment was carried out ten times. The first experiment uses BPS data. This second experiment uses GT, which still needs to be cleaned. The third uses a GT that has been cleaned with HHT. The use of computation time for each experiment is 2750s 4ms or the equivalent of 2 hours 25 minutes.

5.2. Experimental result

The models used are often very complex and overfitting. This overfitting is caused by two things, namely there is little data and the requirements for explanatory variables are needed. Big data analysis using web search can be used as additional data if processed properly will increase accuracy in forecasting tourism visits [41]. Recently, research on DL for tourist arrivals has been mostly done by adding data from web queries as an example of data [42]. This new data source can be used to increase the volume so that it is sufficient for forecasting [34]. This data is also a solution to improve forecasting accuracy [21]. Google trends have also been widely used to add tourist visit data such as [3], [9], [14]. However, this data has many disturbances such as noise which causes the forecasting to be more inaccurate [35], [37]. This study discusses data cleaning techniques using HHT which are then used for forecasting using LSTM. Comparison of forecasting results between two data namely clean data and dirty data is also carried out. The results show a significant difference between the two data. Data that has been processed using HHT cleaning is more accurate than data that has not been processed with HHT.

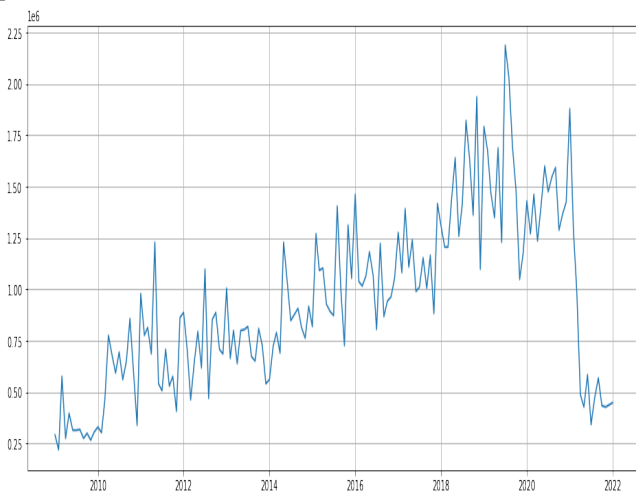


Figure 1. Converted data in the form signal

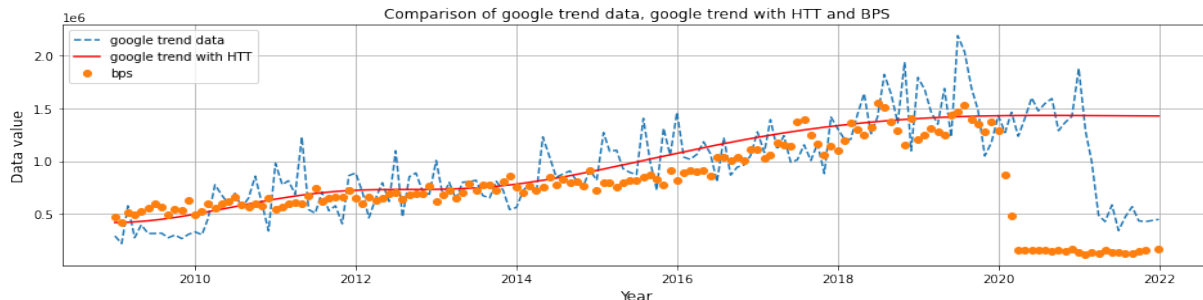


Figure 2. Graph of results after data is cleaned with HHT

5.2.1. Data cleaning using HHT

As shown in Table 4, the dataset is converted into signals for processing. This data signal is then processed to remove noise. These movements break down into components and residues of the IMF. Figure 1 shows the data cleaning process in the first stage, consisting of four iterations until clean data is found, its use for forecasting.

The frequency in the initial data is very high. The decrease in frequency for all IMF functions alternates and does not overlap, while the residuals are closed for monotonous functions. The horizontal axis represents the time span, and the vertical axis shows the function value of each IMF. After the process is complete, Figure 2 shows that the data with the blue line is GT, the orange line is HHT, and the orange dotted line is BPS.

5.2.2. Forecasting tourist arrivals

Experiments in this study with three experimental models including: The first experiment made predictions using data sourced from BPS. The second experiment uses GT; The third uses cleaned GT-HHT. The experimental step begins with data processing. The practical step was carried out using several steps; first, the data is formed into a fit model by dividing per month into one column. Second, framing sequences as a supervised learning problem. Third, is the division of the dataset to be evaluated. Fourth is determination of the appropriate test scale. The fifth forms an inverted scale to determine the estimated value. Sixth is adapting the LSTM network to the training data used and conducting step-by-step tests for forecasting and transforming the data into stationary. Seventh is turning the data into supervised learning and continuing by dividing the data between training data and test data. Eighth, data is transformed with a certain scale.

Table 8. LSTM architecture for predicting tourist arrivals

Layer (type)	Output shape	Parameter
embedding_2 (Embedding)	(none, 2, 100)	14323230 0
lstm_2 (LSTM)	(none, 2, 100)	80400
dense_4	(none, 2, 1)	101
flatten 2 (Flatten)	(none, 2)	0
dense_5 (Dense)	(none, 1)	3
Total parameter : 143,312,804		
Trainable parameter : 80,504		
Non-trainable parameter : 143,232,300		

Table 8 shows that the total parameters used were 143,312,804, the parameters that could be trained were 80,504 and the parameters that could not be trained were 143,232,300. The results of LSTM performance can be seen using training by reading the history that has been done before. The experimental results were observed based on the specified number of epochs. Tables 9, 10, and 11 are the results of experiments conducted using 10 epochs, namely epochs 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500. Based on Table 9, the best results were obtained at epoch 350, with the lowest loss value of 6.58 and the highest loss of 11.86. using a time of 350s 4ms. The test ends at epoch 500 with a value of nan. While the best deal is at 300 with average time duration per step of 1s 4ms, the lowest loss and MAPE value is 6.77 and the highest is 10.32, as shown in Table 10.

Table 9. Comparison of trains on epochs with BPS data and Google Trends

Epoch	50	100	150	200	250	300	350	400	450	500
Lowest time	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms
Highest time	4s 4ms	1s 7ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms
Lowest loss	21.9 2	19.3 8	12.71 5	10.2 5	7.53 7.57	7.57 6.58	6.58 16.6	16.6 4	14.0 4	19.18
Highest loss	27.8 9	24.5 5	21.36 7	14.2 5	84.1 89	425. 6	11.8 9	39.6 9	119. 09	nan
Lowest MAPE	21.9 5	19.3 8	12.71 5	10.2 5	7.53 7.57	7.57 6.58	6.58 16.6	16.6 4	14.0 4	19.18
Highest MAPE	27.8 9	24.5 5	21.36 7	14.2 5	84.1 89	425. 6	11.8 9	39.6 9	119. 09	nan

Table 10. Comparison of trains on epoch

Epoch	50	100	150	200	250	300	350	400	450	500
Lowest time	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	1s 4ms	0s 3ms	0s 3ms	0s 3ms	0s 3ms
Highest time	2s 4ms	1s 4ms	1s 7ms	1s 4ms	1s 4ms	1s 4ms	2s 4ms	1s 4ms	1s 4ms	1s 4ms
Lowest loss	20.4 4	21.0 4	19.80 1	13.4 1	8.97 0	6.77 2	12.8 6	9.02 4	14.2 6	nan
Highest loss	33.4 2	25.7 1	24.36 9	23.0 0	15.4 0	10.3 2	85.2 6	14.2 4	nan	nan
Lowest MAPE	20.4 4	21.0 4	19.80 1	13.4 1	8.97 0	6.77 1	12.8 1	9.02 6	14.2 6	nan
Highest MAPE	33.4 2	25.7 1	24.36 9	23.0 0	15.4 0	10.3 2	85.2 6	14.2 4	nan	nan

5.2.3. Forecasting result

This study compares BPS, GT, and GT-HHT data. The results showed that the cleaning process using HHT gave better results.

5.3. Evaluation

The accuracy of the forecasting model is assessed based on forecasting errors. The smaller the value of the forecasting error, the more accurate the forecasting results. Determining the level of this error requires measurement tools such as Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). These two measuring tools are used in [6], [13], [38].

Equation (23) calculates the forecast error as a percentage. This evaluation is known as MAPE. The way MAPE works determines the error divided by the actual value in each period. The error value obtained is written in absolute percentage form. Equation (24), called RMSE, is used to find the concentration of data around the linear regression line or determine the distribution of point deviations on the linear regression line. The smaller the value generated, the higher the accuracy of the model.

$$MAPE = \frac{100}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{y_i} \tag{23}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y - \hat{y})^2}{N}} \tag{24}$$

where N is the number of observed data, $y_1, y_2, y_3, \dots, y_n$ is the observed value, while $\hat{y}_1, \hat{y}_2, \hat{y}_3, \dots, \hat{y}_n$ is the predicted value.

Table 11. Comparison of LSTM and HHT-LSTM accuracy measured using MAPE and RMSE for epoch 300

MODEL	MAPE	RMSE
LSTM	157,00	217628,75
HHT – LSTM	93,71	123882,20

Table 12. Comparison of accuracy of LSTM and HHT-LSTM as measured using MAPE and RMSE for epoch 300

MODEL	MAPE	RMSE
LSTM	475.93	696105.49
HHT-LSTM	96.59	124520.30

After all the experiments were carried out with various presentation patterns, then evaluated using MAPE and RMSE, the results were obtained as in Tables 11 and 12. Table 11 received the MAPE value for forecasting without a data cleaning algorithm worth 157.00 and RMSE 217628.75, meaning this value is greater than forecasting using LSTM with HHT. Table 12 shows that in, the results of the evaluation with MAPE there is a very significant difference in value, namely 475.93 for data that is not cleaned compared to 96.59 for data that has been cleaned using data processing algorithms. The RMSE value is also very far away, which is 696105.49 compared to 124520.30.

6. Conclusion

This paper discusses deep learning methods for forecasting tourist arrivals. The DL method used in this paper is LSTM. LSTM is a popular method today used as a forecasting tool. But unfortunately, data on tourist arrivals are often late and very non-linear. In addition, data on tourist arrivals is also minimal. LSTM also has many parameters that require a lot of data. Little data will result in a low level of forecasting accuracy which tends to be invalid. To overcome these difficulties, this research is proposed to improve the accuracy of using big data. The data contained in the GT is used as an explanatory variable. However, GT data is considered to have high noise; this is due to a shift in search time and departure time. Processing algorithm is proposed to be a solution to overcome this problem. Based on the research that has been done, it turns out that processing algorithms can improve accuracy. HHT combines very well with the DL method.

Acknowledgement

This work was supported by the Universitas Muhammadiyah Riau under the Institute for Research and Community Service with a Research Contract Agreement number: 27/PRJ/II.3.AU/F/7/2021.

References:

- [1]. Andariesta, D. T., & Wasesa, M. (2022). Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach. *Journal of Tourism Futures*, 1–17. Doi:10.1108/JTF-10-2021-0239
- [2]. Anisa, M. P., Irawan, H., & Widiyanesti, S. (2021). Forecasting demand factors of tourist arrivals in Indonesia's tourism industry using recurrent neural network. *IOP Conference Series: Materials Science and Engineering*, 1077(1), 012035. Doi: 10.1088/1757-899x/1077/1/012035
- [3]. Antolini, F., & Grassini, L. (2019). Foreign arrivals nowcasting in Italy with Google Trends data. *Quality and Quantity*, 53(5), 2385–2401. Doi: 10.1007/s11135-018-0748-z
- [4]. Assaf, A. G., Li, G., Song, H., & Tsionas, M. G. (2019). Modeling and Forecasting Regional Tourism Demand Using the Bayesian Global Vector Autoregressive (BGVAR) Model. *Journal of Travel Research*, 58(3), 383–397. Doi: 10.1177/0047287518759226
- [5]. Atbi, A., Debbal, S. M., Meziani, F., & Meziane, A. (2013). Separation of heart sounds and heart murmurs by Hilbert transform envelopogram. *Journal of Medical Engineering and Technology*, 37(6), 375–387. Doi: 10.3109/03091902.2013.816379
- [6]. Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2018). Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, 11(7). Doi: 10.3390/en11071636
- [7]. Chandra, R., Goyal, S., & Gupta, R. (2021). Evaluation of Deep Learning Models for Multi-Step Ahead Time Series Prediction. *IEEE Access*, 9, 83105–83123. Doi: 10.1109/ACCESS.2021.3085085
- [8]. Chen, B., Zhao, S. L., & Li, P. Y. (2014). Application of Hilbert-Huang Transform in Structural Health Monitoring: A State-of-the-Art Review. *Mathematical Problems in Engineering*. Doi: 10.1155/2014/317954
- [9]. Claude, U. (2020). Predicting Tourism Demands by Google Trends: A Hidden Markov Models Based Study. *Journal of System and Management Sciences*, 10(1), 106–120. Doi: 10.33168/JSMS.2020.0108
- [10]. Dergiades, T., Mavragani, E., & Pan, B. (2018). Google Trends and tourists' arrivals: Emerging biases and proposed corrections. *Tourism Management*, 66, 108–120. Doi: 10.1016/j.tourman.2017.10.014

- [11]. Ete, A., Fitriawati, M., & Arifin, M. (2019). Forecasting the Number of Tourist Arrivals to Batam by applying the Singular Spectrum Analysis and the Arima Method. In *First International Conference on Progressive Civil Society (ICONPROCS 2019)*, 119–126. Atlantis Press.
Doi: 10.2991/iconprocs-19.2019.24
- [12]. Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41(4), 1937–1946. Doi: 10.1016/j.eswa.2013.08.089
- [13]. Han, Y., Wang, C., Ren, Y., Wang, S., Zheng, H., & Chen, G. (2019). Short-term prediction of bus passenger flow based on a hybrid optimized LSTM network. *ISPRS International Journal of Geo-Information*, 8(9).
Doi: 10.3390/ijgi8090366
- [14]. Havranek, T., & Zeynalov, A. (2019). Forecasting tourist arrivals: Google Trends meets mixed-frequency data. *Tourism Economics*, 27(1).
Doi: 10.1177/1354816619879584
- [15]. Höpken, W., Eberle, T., Fuchs, M., & Lexhagen, M. (2019). Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of Åre, Sweden. *Information Technology and Tourism*, 21(1), 45–62.
Doi: 10.1007/s40558-018-0129-4
- [16]. Höpken, W., Eberle, T., Fuchs, M., & Lexhagen, M. (2020). Improving Tourist Arrival Prediction: A Big Data and Artificial Neural Network Approach. *Journal of Travel Research*, 60(5).
Doi: 10.1177/0047287520921244
- [17]. Jiao, X., Li, G., & Chen, J. L. (2020). Forecasting international tourism demand: a local spatiotemporal model. *Annals of Tourism Research*, 83, 102937.
Doi: 10.1016/j.annals.2020.102937
- [18]. Laaroussi, H., Guerouate, F., & Sbihi, M. (2020). Deep Learning Framework for Forecasting Tourism Demand. *2020 IEEE International Conference on Technology Management, Operations and Decisions, ICTMOD 2020*, 1–4. Doi: 10.1109/ICTMOD49425.2020.9380612
- [19]. Li, C., Ge, P., Liu, Z., & Zheng, W. (2020). Forecasting tourist arrivals using denoising and potential factors. *Annals of Tourism Research*, 83, 102943. Doi: 10.1016/j.annals.2020.102943
- [20]. Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, 68, 301–323.
Doi: 10.1016/j.tourman.2018.03.009
- [21]. Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management*, 68, 116–126.
Doi: 10.1016/j.tourman.2018.03.006
- [22]. Li, X., Li, H., Pan, B., & Law, R. (2021). Machine Learning in Internet Search Query Selection for Tourism Forecasting. *Journal of Travel Research*, 60(6), 1213–1231. Doi: 10.1177/0047287520934871
- [23]. Mariyono, J. (2017). Determinants of Demand for Foreign Tourism in Indonesia. *Jurnal Ekonomi Pembangunan*, 18(1), 82.
Doi: 10.23917/jep.v18i1.2042
- [24]. Mazlan, A. U., Sahabudin, N. A., Remli, M. A., Ismail, N. S. N., Mohamad, M. S., Nies, H. W., & Abd Warif, N. B. (2021). A Review on Recent Progress in Machine Learning and Deep Learning Methods for Cancer Classification on Gene Expression Data. *Processes*, 9(8), 1466. Doi: 10.3390/pr9081466
- [25]. Mukhtar, H., Taufiq, R. M., Herwinanda, I., Winarso, D., & Hayami, R. (2022). Forecasting Covid-19 Time Series Data using the Long Short-Term Memory (LSTM). *International Journal of Advanced Computer Science and Applications*, 13(10), 211–217.
Doi: 10.14569/IJACSA.2022.0131026
- [26]. Önder, I. (2017). Forecasting tourism demand with Google trends: Accuracy comparison of countries versus cities. *International Journal of Tourism Research*, 19(6), 648–660. Doi: 10.1002/jtr.2137
- [27]. Peng, L., Wang, L., Ai, X. Y., & Zeng, Y. R. (2021). Forecasting Tourist Arrivals via Random Forest and Long Short-term Memory. *Cognitive Computation*, 13(1), 125–138. Doi: 10.1007/s12559-020-09747-z
- [28]. Petrevska, B. (2012). Forecasting International Tourism Demand: the Evidence of Macedonia. *UTMS Journal of Economics*, 3(1), 45–55.
- [29]. Praveena, M. D. A., & Bharathi, B. (2017). A Survey Paper on Big Data Analytics. *International Conference on Information, Communication & Embedded Systems (ICICES 2017)*, *Icices*.
- [30]. Qin, S. R., & Zhong, Y. M. (2006). A new envelope algorithm of Hilbert-Huang Transform. *Mechanical Systems and Signal Processing*, 20(8), 1941–1952.
Doi: 10.1016/j.ymsp.2005.07.002
- [31]. Rizal, A. A., Soraya, S., & Tajuddin, M. (2019). Sequence to sequence analysis with long short term memory for tourist arrivals prediction. *Journal of Physics: Conference Series*, 1211(1).
Doi: 10.1088/1742-6596/1211/1/012024
- [32]. Saddam, B., Aissa, A., Ahmed, B. S., & Abdellatif, S. (2017). Detection of rotor faults based on Hilbert Transform and neural network for an induction machine. *2017 5th International Conference on Electrical Engineering - Boumerdes, ICEE-B 2017*, 1–6.
Doi: 10.1109/ICEE-B.2017.8192029
- [33]. Silva, E. S., Hassani, H., Heravi, S., & Huang, X. (2019). Forecasting tourism demand with denoised neural networks. *Annals of Tourism Research*, 74, 134–154. Doi: 10.1016/j.annals.2018.11.006
- [34]. Sun, S., Wei, Y., Tsui, K., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, 70, 1–10.
Doi: 10.1016/j.tourman.2018.07.010
- [35]. Volchek, K., Song, H., Law, R., & Buhalis, D. (2018). Forecasting London Museum Visitors Using Google Trends Data. *E-Review of Tourism Research*.

- [36]. Höpken, W., Ernesti, D., Fuchs, M., Kronenberg, K., & Lexhagen, M. (2017). Big data as input for predicting tourist arrivals. In *Information and Communication Technologies in Tourism 2017: Proceedings of the International Conference in Rome, Italy, January 24-26, 2017*, 187-199. Springer International Publishing.
- [37]. Xiaoxuan, L., Qi, W., Geng, P., & Benfu, L. (2016). Tourism forecasting by search engine data with noise-processing. *African Journal of Business Management*, 10(6), 114–130. Doi: 10.5897/ajbm2015.7945
- [38]. Xie, G., Qian, Y., & Wang, S. (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management*, 82, 104208. Doi: 10.1016/j.tourman.2020.104208
- [39]. Xu, T. Y., Zhen, H. F., & Wang, L. D. (2013). Detection of Flicker Caused by Wind Farm Based on Mathematical Morphology Filter and Hilbert-Huang Transform. *Advanced Materials Research*, 724, 555-560. Doi: 10.4028/www.scientific.net/AMR.724-725.555
- [40]. Zanury, N. A., Remli, M. A., Adli, H. K., & Wong, K. N. S. W. S. (2022). Recent Developments of Deep Learning in Future Smart Cities: A Review. *Machine Learning for Smart Environments/Cities. Intelligent Systems Reference Library*, 121, 199–212. Doi: 10.1007/978-3-030-97516-6_11
- [41]. Zhang, B., Li, N., Shi, F., & Law, R. (2020). A deep learning approach for daily tourist flow forecasting with consumer search data. *Asia Pacific Journal of Tourism Research*, 25(3), 323–339. Doi: 10.1080/10941665.2019.1709876
- [42]. Zhang, B., Pu, Y., Wang, Y., & Li, J. (2019). Forecasting hotel accommodation demand based on LSTM model incorporating internet search index. *Sustainability*, 11(17), 4708.
- [43]. Zhang, J., Chen, F., Cui, Z., Guo, Y., & Zhu, Y. (2021). Deep Learning Architecture for Short-Term Passenger Flow Forecasting in Urban Rail Transit. *IEEE Transactions on Intelligent Transportation Systems*, 22(11), 7004–7014. Doi: 10.1109/TITS.2020.3000761