Proceeding of International Conference on Digital Advance Tourism, Management and Technology 2023
Volume. 1 No. 1 2023 e-ISSN : 3031-5999, Hal 132-142
DOI: https://doi.org/10.56910/ictmt.v1i1.57

Can Google Trends(GT) be used to predict tourist arrivals?: FB Prophet Machine Learning(ML) for Predicting Tourist Arrivals

Indra Gunawan[1], Dwi Purnomo Putro[2], Adhika Pramita Widyassari[3]
[1-3] Sekolah Tinggi Teknologi Ronggolawe/Informatic Department, Cepu, Indonesia

Corresponding email: igunsttr@gmail.com

Abstract. *The big problem in tourism is how to provide appropriate preparations to serve tourists so that when the tourist season is low, resources can be saved and when the tourist season is busy, all resources can be provided effectively. Machine learning is a derivative branch of artificial intelligence, one of whose capabilities can be used to carry out data/dataset-based forecasting. This research uses a dataset obtained from GT from 2013-2023 with several keywords combining city names and tourist destination names in Yogyakarta Indonesia, then it will be compared with a dataset of tourist arrivals in the city of Yogyakarta obtained from the Central Statistics Agency. The Machine Learning model that will be used is Prophet Facebook.. This model uses a Bayessian as a backend algorithm. The results obtained from this research are that GTs can be used to predict tourist arrivals with some tweaks on the dataset. However, to get accurate results, various combinations of keywords are needed for the desired destination, and it is recommended to add some column namely max and mean to the dataset to prevent insufficiency of data of some keywords that make prediction result bad. In this research it can be concluded that the use of an additional max column can increase the COERR, MAPE and R2 values. Meanwhile, we found that the GT dataset can be used for forecasting best in time periods under 200 days. Also we found that using the GT dataset alone produces unstable COERR, MAPE and R2 values. Another finding is that the GT dataset that uses the YouTube filter is only suitable for use in Indonesia for the time period above 2018 considering that Indonesian people's access to YouTube has increased massively over that year and tends to decrease below that year. However, the trend shows that the use of searches on YouTube after 2018 tends to increase drastically, beating searches on the Google web.*

Keywords *Prophet Facebook; Machine Learning; Forecasting; Tourist Prediction; Google Trends*

INTRODUCTION

Tourism is a big contributor to foreign exchange for a country, why?, because the taxes generated from tourism are quite large so that it can increase state income. Another benefit of tourism is that it can reduce the income gap between upper class and lower class people, because the tourism industry can absorb a large workforce, the majority of which are people in the lower class. (Alam & Paramati, 2016; Nguyen dkk., 2021; Qin dkk., 2018).

Tourism does have great potential to increase the prosperity of a country. So, many countries are flocking to shift their business sectors to this sector, however, tourism also has big challenges that are of common concern. The challenge is how to prepare all resources appropriately according to the current tourism season. When the tourism season is in peak season, the entire tourism ecosystem can prepare all resources such as lodging, accommodation, consumption and sufficient workforce. Meanwhile, when the tourism season is in the low season, the ecosystem can provide all resources without excess, so that nothing is wasted (Anisa dkk., 2021).

This research aims to test GT's ability to predict tourist arrivals in Indonesia, especially in the city of Yogyakarta. Data from GT will be compared with actual data on tourist arrivals

Received November 19, 2023; Revised Desember 01 , 2023; Accepted Desember 31, 2023
* Indra Gunawan , *igunsttr@gmail.com*

in the city of Yogyakarta obtained from the Central Statistics Agency in the same period, namely 2013 to 2019, then a forecasting method using the FB prophet machine learning model will be used to test the accuracy of the dataset from GT.

There have been several studies on forecasting using GT datasets. Yang (Yang et al., 2015) researched the comparison of forecasting using the GT dataset compared to Baidu in China, the result was that Baidu was more dominant as a search website than GT. Havranek (Havranek & Zeynalov, 2021) conducted research to compare the effectiveness of the GT dataset compared to the mixed data sampling (MIDAS) dataset. The result was that the GT dataset was able to outperform the MIDAS dataset. Other research was conducted by Claud (Claud, 2020) who examined the effectiveness of forecasting using the GT dataset with a hidden Markov algorithm. Another research was conducted by Li (Li et al., 2021) who examined the combination of GT datasets with social media. The results were that the combination of GT and datasets from social media gave good results. Another research was conducted by Bokelmann (Bokelmann & Lessmann, 2019) who examined the use of the GT dataset for tourist forecasting in Germany, the results shows that the GT dataset gave good results and was able to be used in short-range forecasting. Lastly is research from Patandung (Patandung & Jatnika, 2021) which examined the use of the Prophet FB model for forecasting international tourist arrivals in Indonesia during the Covid pandemic. The results found that international tourist arrivals in Indonesia experienced a significant decline and will return to normal after March of 2022.

METHOD

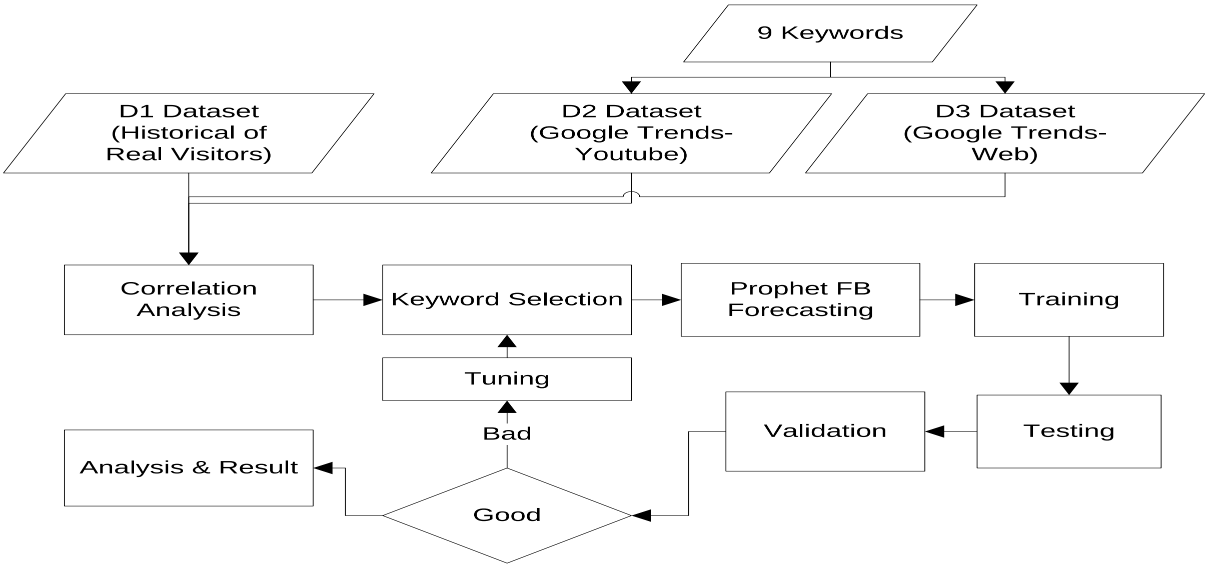
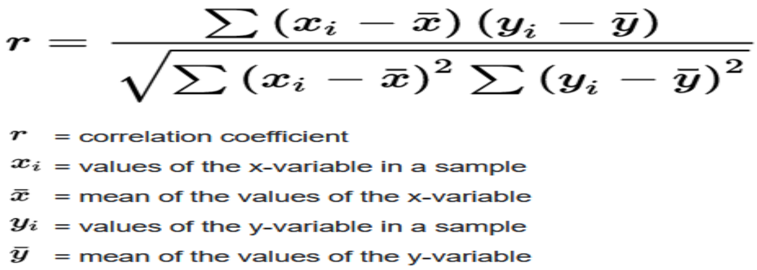


Figure 1 Research Methodology

This research is experimental research. This research began by obtaining a historical dataset of tourist arrivals in the city of Yogyakarta which was obtained from the Central Bureau of Statistics for the city of Yogyakarta Indonesia, this dataset was given the code D1. This D1 dataset is a monthly timeseries type with columns for date and number of visitors. This D1 dataset has visitor data from 2013 to 2019. This D1 dataset is the basic dataset used as a comparison and testing standard. The next dataset is D2, which is a dataset that comes from GT which is then shortened to GT. This dataset is specifically for searches on GT with search filters on YouTube only. Dataset D2 contains 10 columns consisting of date columns, and 9 search keyword columns which contain the number of searches on certain keywords for the monthly period from 2013 to 2019. Dataset D3 is a GT dataset containing the same 9 keywords as D2, but with filter for searches only on the web.

The next step is to carry out a COERR correlation analysis for keywords D2 and D3 against the number of tourists coming to D1 to determine the right keywords to use in machine learning. After finding the right keywords, a training and testing process is carried out on datasets D1 and D2. The next process is validation of the machine learning forecasting results. The validation carried out is the assessment of the MAPE score (Mean Absolute Percentage Error) and the assessment of the R2 coefficient (Coefficient of Determination). If the resulting score is bad then the forecasting process is carried out again using different keywords by optimizing the D2 and D3 datasets. If the resulting score is good, then the keywords and optimization process will be selected as the final recommendation result of the research. The research flow can be seen in Figure 1.

The COERR score is a number between –1 and 1 which measures the strength and direction of the relationship between two variables. When one variable changes, other variables also change in the same direction. The COERR score used in this research is Pearson correlation. The Pearson score is formulated as follows:



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Figure 2 Pearson Score

The MAPE score is a function of assessing the value of measuring the average absolute percentage error of the average error produced by a model, or how far the average prediction is off. The MAPE score is formulated as follows:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\, y_i - \hat{y}_i \,|}{y_i} \;.\; 100\%$$

Figure 3. MAPE Score

The R2 score is a function to measure the average value of the absolute percentage error produced by a model, or how far the average prediction is off. The R2 score is formulated as follows:

$$\begin{aligned} \text{r2\_score} &= 1 - \frac{\text{total\_error\_model}}{\text{total\_error\_baseline}} \\ &= 1 - \frac{\sum_{i=1}^{N} (\text{predicted}_i - \text{actual}_i)^2}{\sum_{i=1}^{N} (\text{average\_value} - \text{actual}_i)^2} \end{aligned}$$

Figure 4. R2 Score

RESULTS AND DISCUSSION

Analysis of Raw data

The first process carried out in this research was to analyze the raw data into graphic form. This process is used to recognize data and record anomalies that occur in the data so that they can be used as a reference in the dataset optimization process in the next step. In D2, a data gap was found in the keyword 'Yogyakarta', causing the graph to drop to 0, this causes the forecasting process in machine learning to be inaccurate if the keyword 'Yogyakarta' is used. Further analysis can be seen in the graph that only the keywords 'Jogja' and 'Jogja Tourism' have sufficient data to be processed. Furthermore, by analyzing the graphs on the D3 dataset, it can be concluded that only the keywords "Malioboro", "Parangtritis", "Borobudur", "Yogyakarta Tourism" do not have sufficient data to be processed.

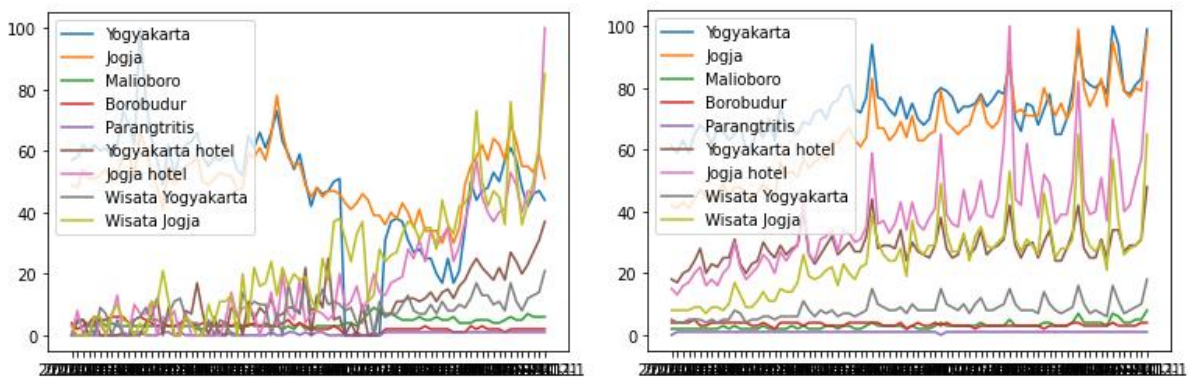


Figure 5. D2 and D3 Dataset

Below in Table 1 are some sample data from the D2 dataset obtained from GTs with filters only on YouTube. This D2 dataset has 90 rows of data starting from 2013 to 2019. Added Max and Mean columns.

Table 1 Dataset D2 – GT Youtube

Month	Yogyaka ta	Jogja	Malioboro	Borobudur	Parangtritis	Yogyakarta hotel	Jogja hotel	Wisata Yogyakarta	Wisata jogja	Max	Mean
01/01/2013	57	49	3	4	1	0	0	0	3	57	13
01/02/2013	58	48	4	2	0	5	8	0	0	58	13,89
01/03/2013	62	54	5	3	0	5	0	0	0	62	14,33
01/04/2013	60	51	3	4	0	0	4	4	0	60	14

…

In Table 2 are some examples of data in the D3 Dataset obtained from GTs with a Web only filter. This dataset has 90 rows of data starting from 2013 to 2019. Then the Max and Min columns are added.

Table 2 Dataset D3 – GT Web.

Month	Yogyakarta	Jogja	Malioboro	Borobudur	Parangtritis	Yogyakarta hotel	Jogja hotel	Wisata Yogyakarta	Wisata jogja	Max	Mean
01/01/2013	61	42	2	4	0	18	15	5	8	61	17,22
01/02/2013	59	41	2	4	1	17	13	4	8	59	16,556
01/03/2013	63	43	2	4	1	20	16	4	8	63	17,889
01/04/2013	59	41	2	4	1	21	17	5	8	59	17,556

…

In Table 3 are some examples of data in Dataset D1 obtained from the Yogyakarta City Central Statistics Agency. This dataset has 90 rows of data starting from 2013 to 2019.

Table 3 Dataset D1 - Historical of Visitor Data.

Month	Visitor
01/01/2013	35
01/02/2013	50
01/03/2013	58
01/04/2013	21
01/05/2013	48
01/06/2013	36
…	

Keyword Testing

The second process carried out was testing the relationship between keywords in D2 and D3 and the number of visitors in D1. At this stage new columns "Max" and "Mean" are added to datasets D2 and D3. The "Max" column is obtained by searching for the highest number from columns one to nine in each row. Meanwhile, the "Mean" column is the average value of all columns one to nine. The addition of these four columns is to fill deficiencies in other columns, some of which have blank search data, which can reduce forecasting accuracy. Another goal is to test the comparison of columns one to nine with the new column resulting from a combination of other columns. Testing the relationship between columns one to eleven uses the COERR (Pearson Coefficient Correlation) function. The results of the scores can be seen in Table 4.

Table 4 Uji Nilai COERR Kata Kunci

Keyword	D2 Dataset	D3 Dataset
Yogyakarta	-0,368	0,15
Jogja	-0,3	0,26
Malioboro	-0,002	-0,086
Borobudur	-0,34	-0,16
Parangtritis	-0,046	-0,03
Yogyakarta Hotel	-0,09	0,23
Jogja Hotel	-0,21	0,27
Wisata Yogyakarta	0,106	0,29
Wisata Jogja	0,0042	0,33
Max	-0,45	0,179
Mean	-0,3	0,272

Furthermore, the results of the analysis in Table 4 can be concluded that in D2, the highest Pearson scores are for the keywords "Yogyakarta Tourism" and "Jogja Tourism". In D3, it was found that the best keywords were "Jogja Tourism", "Yogyakarta Tourism", "Jogja hotel", "Jogja". A combination of first stage analysis (data adequacy) and second stage analysis (Pearson value) is used to decide which keywords to choose. The result was that the keywords "Yogyakarta", "Jogja", "Yogyakarta Tourism" and "Jogja Tourism" were selected.

Prediction

At this stage, a prediction/forecasting process is carried out using the Prophet FB algorithm and the Python programming language. The dataset is divided into training data covering the years 2013 to 2018 and the remaining 2019 data is used as testing data. The results of the forecasting carried out can be seen in Figure 6.

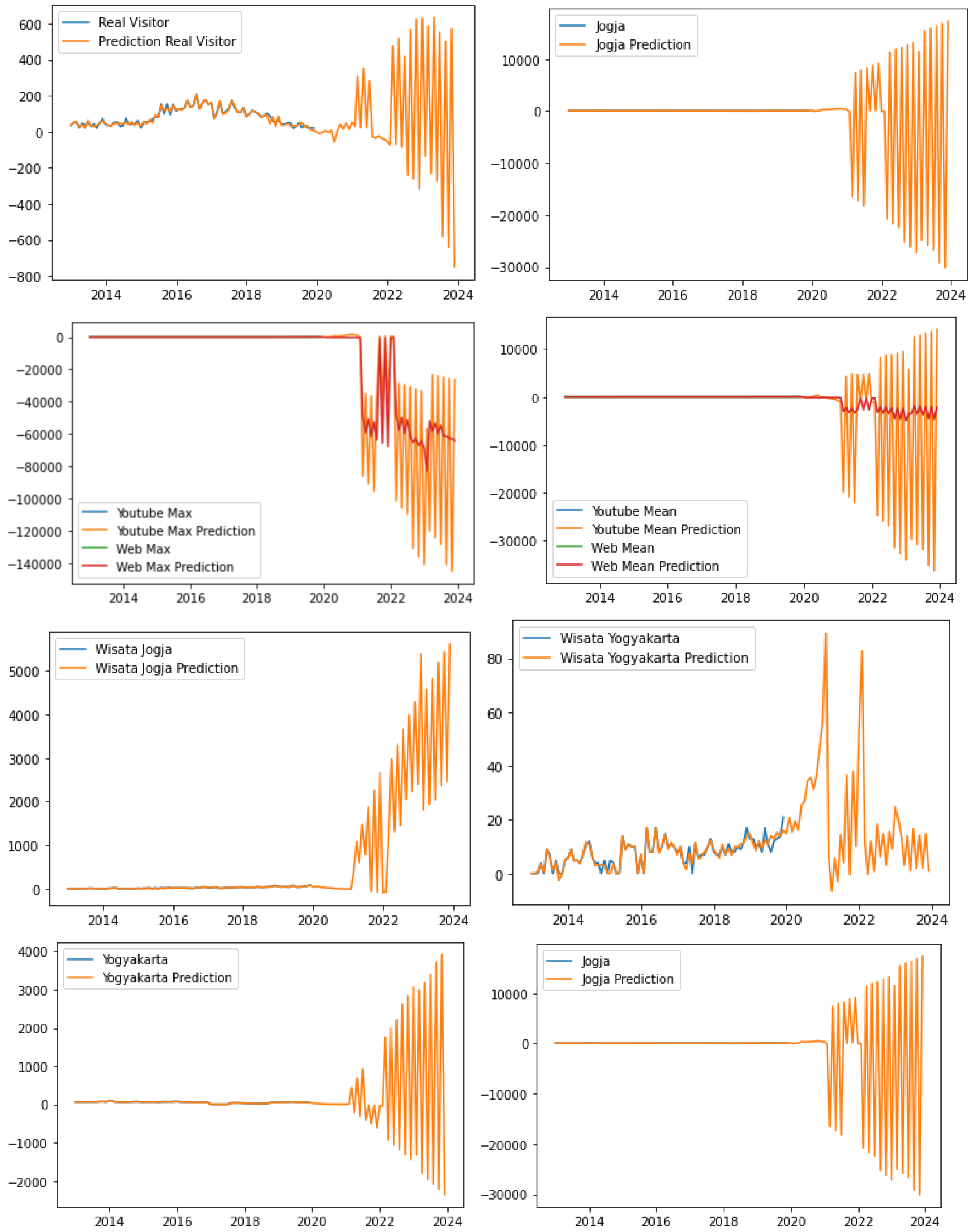


Figure 6. Selected keyword forecasting graph.

From Figure 6 it can be concluded that the keywords "Jogja Tourism" and "Yogyakarta Tourism" have a negative relationship to the data on the number of tourist arrivals in dataset D1.

Testing and Validation

MAPE Scoring

At this stage, the process of validation and testing of prediction results is carried out using the MAPE (Mean Absolute Percentage Error) and R2 Score functions. MAPE is used to see the percentage comparison of prediction results for each keyword in the D2 and D3 datasets against visitor data in the D1 dataset. The results of the MAPE score can be seen in Table 5.

Table 5 Mean Absolute Percentage Error

Column	Dataset	Score
Visitor	D1	0.17
jogja	D3	0.01
wisata jogja	D3	774918096165179.0
wisata yogyakarta	D3	696753804273722.1
yogyakarta	D3	168682868020812.94
jogja	D2	0.015
wisata jogja	D2	774918096165179.0
wisata yogyakarta	D2	696753804273722.1
yogyakarta	D2	168682868020812.94
max	D2	0.019
max	D3	0.010
mean	D3	0.018
mean	D2	0.016

From Table 5 it can be concluded that the best MAPE value is for the keyword "Jogja" and the values "Max" and "Min". Other keywords have very high scores, which indicates that the prediction results deviate too far from the original data.

Next, graphs are made from the MAPE score data to facilitate the analysis process. The results of the MAPE score graph can be seen in Figure 6.

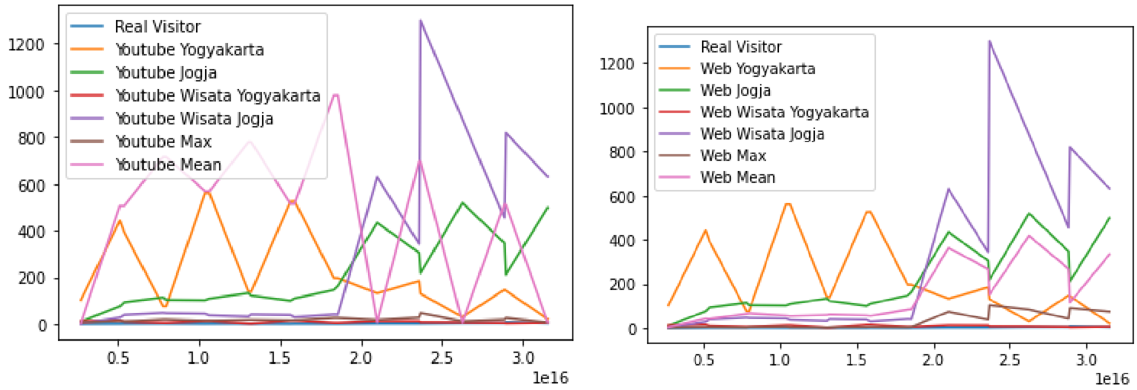


Figure 6. MAPE Score D2 Dataset vs D3 Dataset

From the graph in Figure 6, it can be concluded that in dataset D2, the best columns or keywords are "Youtube Max" and "Youtube Wisata Yogyakarta" in the forecasting window under 200 days. Meanwhile, in the D3 dataset, the best columns or keywords are "Web Max" and "Yogyakarta Tourism Web" in the forecasting window under 200 days.

R2 Scoring

The next stage is testing the R2 score which aims to determine the coefficient of determination of the prediction results against the original data. The R2 score results can be seen in Table 6 below.

Table 6 R2 Score

Column	Dataset	Score
Yogyakarta	D3	0.9948
Wisata Jogja	D2	0.9948
Mean	D2	0.9922
Mean	D3	0.9892
Max	D2	0.9860
Jogja	D3	0.9843
Jogja	D2	0.9843
Max	D3	0.9798
Wisata Jogja	D3	0.9722
Wisata Yogyakarta	D2	0.9722
Visitor	D1	0.9420
Wisata Yogyakarta	D3	0.8902
Yogyakarta	D2	0.8902

From Table 6 it can be concluded that the keywords/columns "Yogyakarta", "Jogja Tourism", "Mean" and "Max" have the highest R2 scores.

Analysis

At this stage, an analysis of the scoring results of all keywords in the D2 and D3 datasets is carried out. It can be concluded that the additional column "Max" has the best and most stable results in all Pearson, MAPE and R2 tests and while other keywords have unstable scores due to missing data in varying time periods, causing poor prediction results.

Another finding obtained is that searches on YouTube have increased since 2018 and the trend will always increase compared to searches on the Google web.

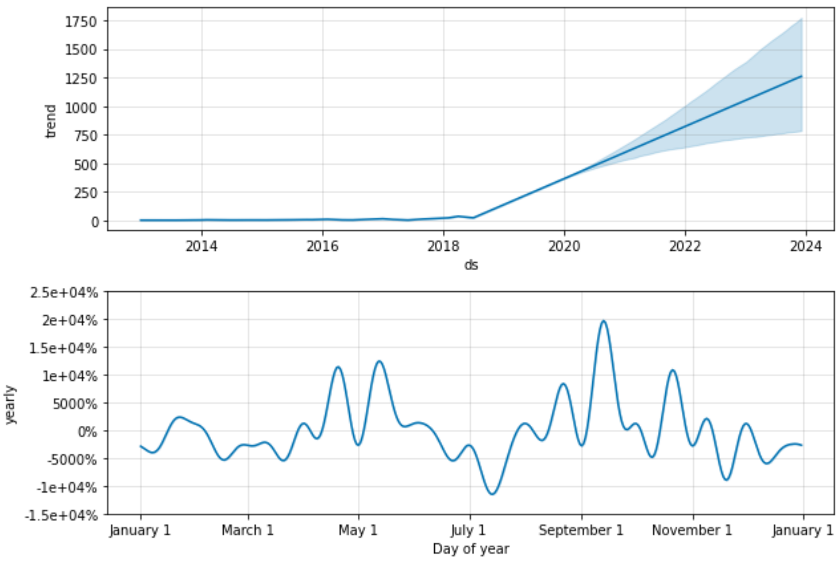


Figure 7 Youtube search trend

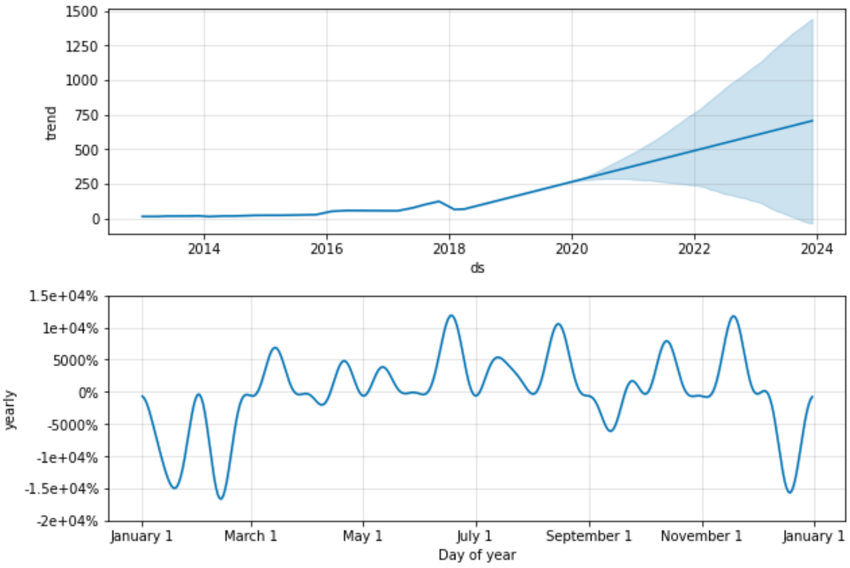


Figure 8 Google search trend

CONCLUSION

It can be concluded that the additional column "Max" has the best and most stable results in all Pearson, MAPE and R2 tests and while other keywords have unstable scores due to missing data in varying time periods, causing poor prediction results. Another finding is that the D3(GT-Web) dataset has a better score than the D2(GT-Youtube) dataset. This is due to the lack of YouTube search data in the early years of testing considering that YouTube was not very popular in Indonesia compared to now. so that searches on the Google website are relatively more stable when used as a data base if a long time span is used. However, based on data analysis, searches on YouTube have increased sharply since 2018, which is linear with the increase in internet quality in Indonesia. So it is recommended that if you want to use the GT dataset with YouTube filters, it is only for 2018 and above.

REFERENCES

Alam, Md. S., & Paramati, S. R. (2016). The impact of tourism on income inequality in developing economies: Does Kuznets curve hypothesis exist? *Annals of Tourism Research*, *61*, 111–126. https://doi.org/10.1016/j.annals.2016.09.008

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. https://doi.org/10.1088/1757-899X/1077/1/012035

Bokelmann, B., & Lessmann, S. (2019). Spurious patterns in Google Trends data—An analysis of the effects on tourism demand forecasting in Germany. *Tourism Management*, *75*, 1–12. https://doi.org/10.1016/j.tourman.2019.04.015

Claud, U. (2020). Predicting Tourism Demands by Google Trends: A Hidden Markov Models Based Study. *Journal of System and Management Sciences*. https://doi.org/10.33168/JSMS.2020.0108

Havranek, T., & Zeynalov, A. (2021). Forecasting tourist arrivals: Google Trends meets mixed-frequency data. *Tourism Economics*, *27*(1), 129–148. https://doi.org/10.1177/1354816619879584

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

Nguyen, C. P., Schinckus, C., Su, T. D., & Chong, F. H. L. (2021). The Influence of Tourism on Income Inequality. *Journal of Travel Research*, *60*(7), 1426–1444. https://doi.org/10.1177/0047287520954538

Patandung, S., & Jatnika, I. (2021). *The FB Prophet Model Application to the Growth Prediction of International Tourists in Indonesia during the COVID-19 Pandemic*. *6*(2).

Qin, Y., Luo, Y., Zhao, Y., & Zhang, J. (2018). Research on relationship between tourism income and economic growth based on meta-analysis. *Applied Mathematics and Nonlinear Sciences*, *3*(1), 105–114. https://doi.org/10.21042/AMNS.2018.1.00008

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, *46*, 386–397. https://doi.org/10.1016/j.tourman.2014.07.019