International Journal of
Information
and
Management
Sciences

http://ijims.ms.tku.edu.tw/main.php

International Journal of
Information and Management Sciences

33 (2022), 353-367. DOI:10.6186/IJIMS.202212_33(4).0005

CODEN IIMSEQ

ISSN 1017-1819
DOI : 10.6186/IJIMS

International Journal
of
Information and Management Sciences

Information and Management Sciences December, 2022 Volume 33, Number 4

IJIMS http://ijims.ms.tku.edu.tw

Volume 33, Number 4 December, 2022

Using Light Gradient Boosting Machine with Genetic Algorithms and Google Trends in Forecasting COVID-19 Confirmed Cases

Chia-Chi Fang, Ping-Feng Pai, Chi-Ju Lai and Ying-Lei Lin*
National Chi Nan University

Keywords

Google trends
Light gradient boosting machine
Genetic algorithms
COVID-19
Forecast

Abstract.

This study aims to employ the Google Trends searching volume to predict daily COVID-19 confirmed cases in the United States during the beginning and rapidly developing periods of COVID-19. Fourteen keywords related to the COVID-19 pandemic from Google trends in the United States were treated as independent variables. Daily confirmed cases served as dependent variables. This study developed a Light Gradient Boosting Machine with Genetic Algorithms and Google Trends (LGBMGAGT) model in forecasting COVID-19 confirmed cases. Forecasting horizons of one day to seven days ahead were employed to forecast confirmed cases. Furthermore, the Pearson correlation coefficient was used to select essential attributes and forecasting results with and without attributes selection were compared. Numerical results indicated that the proposed LGBMGAGT can obtain more accurate results than the other forecasting models. Therefore, the LGBMGAGT model is a feasible and promising alternative in forecasting COVID-19 confirmed cases.

1. Introduction

In March 2020, the World Health Organization announced that it was a world pandemic and named COVID-19. Millions of people have been diagnosed with COVID-19 and the number of deaths is increasing. The impact is not only blocking and collapsing medical systems, but also delaying orders due to the shutdown of factories. Because of COVID-19, the economic growth rate of many countries has become negative. The governments can deal with a new wave of peaks in advance by using accurate forecasting of confirmed cases. Nowadays websites have been used for searching for things that users

*corresponding author

concern about or are interested in. Google trend is a platform to collect search volume. Table 1 lists the Google Trends keywords used in the previous related literatures. Kurian et al. [17] used the daily confirmed data to investigate the feasibility of using Google Trends search intensity to forecast the number of new cases. Ten keywords were employed to analyze the relationship between Google Trends and new cases. Experiment results pointed out that a strong correlation between keywords and new cases existed. Lin et al. [18] used two Google Trends keywords, namely wash hand and face mask, as an alternative for national population health identifications to study whether or not the number of confirmed cases among 21 countries could be kept away from increasing. Statistical analyses concluded that the public awareness of COVID-19 was increasing to protect themselves when the pandemic expanded. Prasanth et al. [23] proposed a hybrid method, namely the long short-term memory with grey wolf optimizer, to forecast infection cases, cumulative cases, and death cases in India, the USA, and the UK by ten search keywords of Google Trends and con-firmed data collected from the European Centre for Disease Prevention and Control. The proposed approach can generate more accurate results in forecasting infection cases than the auto regression integrated moving average method. Springer et al. [26] utilized terms related to COVID-19 of website search and indicated that the top three keywords are COVID-19 symptoms, social distancing, and lockdown. Pascual and Pourmand [22] and Misiak et al. [20] employed Google Trends search terms to investigate associations between search volume and confirmed cases. Wang et al. [28] applied the autoregression model with Google Trends data to predict the future two weeks of national and state-level new COVID-19 hospitalization. Numerical results revealed that the proposed model is able to capture the new waves of infectious cases at both levels. Rabiolo et al. [24] used three time-series models, including error trend seasonality, autoregressive integrated moving average, and feed-forward neural network autoregression, to calculate the upcoming 14 days confirmed cases with eight countries' COVID-19 confirmed data and Google Trends data. The principal components analysis was employed to reduce data dimensions. This study reported that including Google Trends data in forecasting models can effectively improve the forecasting ability. Husnayain et al. [12] utilized generalized linear models and linear regression models to predict daily COVID-19 cases and deaths in Korea for both short-term and long-term ways. NAVER search volumes served as external factors in forecasting models to increase forecasting accuracy. Authors claimed that generalized linear models outperformed linear regression in most scenarios. In addition, the NAVER search volumes had essential influences on forecasting models. Izhar and Torabi [13] investigated online search behaviors during the COVID-19 pandemic by Google Trends. This study concluded that the decrease of public interest in searching COVID-19 information significantly increases the possibility of an outbreak. Ben et al. [4] examined the relationship between internet search trends and daily confirmed cases at both global and regional levels. Autoregressive integrated moving average models were utilized to forecast the search volumes of keywords before and after the COVID-19 outbreak. Numerical results pointed out that "fever" and "cough" have a high correlation globally and regionally. Furthermore, "diarrhea" and "loss of taste" were irregularly raised during the outbreak periods. Awijen et al. [3] investigated the influences of COVID-19 vaccination on panic and economic anxiety among 194 countries by Google Trends data. This study indicated that anxiety increased when the vaccine arrived. In addition, people lack confidence in the vaccine's efficiency to conquer the COVID-19 pandemic. Rao et al. [25] employed Google Trends data related to COVID-19 to analyze the relation between searching volume and confirmed cases by a bootstrapped Pearson correlation, a time-lead correlation, and a quantile regression. Numerical results revealed that a strong relationship between con-firmed cases and Google Trends searches. Khakimova et al. [16] utilized Google Trends data to gauge the global public interest in the COVID-19 vaccine during the pandemic period. Keywords related to vaccines were used to track users' interests. Numerical results showed that the public attitude toward the COVID-19 vaccine can be measured by the users' search volume. Brodeur et al. [7] employed using Google Trends data to explore whether lockdowns lead to changes in well-being by in Europe and America. This study indicated that pandemic and lockdown policy had significantly affected mental health. Alruily et al. [2] developed a stacked long short-term memory model to predict COVID-19 confirmed and death cases in the USA. In addition to historical data, keywords related to COVID-19 symptoms gathered from Google Trends were employed in this study. This study revealed that using a hybrid data including historical data and Google Trends data can provide more accurate results than only using historical data. Abbas et al. [1] presented that Google Trends are strongly associated with COVID-19 confirmed cases and mortality in the USA. The top nine COVID-19 related symptoms keywords were used to examine the relationship. This investigation indicated that the combination of COVID-19 historical data and Google Trends data can forecast COVID-19 spread and mortality up to three weeks ahead. Yousefinaghani et al. [31] investigated the relationship between COVID-19 confirmed cases and symptom related posts and search volume in Canada and the USA. Data from Twitter and Google Trends were collected to learn the lag between real outbreak waves and warning signals. Numerical results pointed out that social media data were helpful for conducting forecast.

The contribution of this study is to investigate the relationship between keywords related to COVID-19 in Google Trends and daily COVID-19 confirmed cases in the United States for one day to seven days during the period of beginning and rapidly developing COVID-19. The rest of this study is organized as follows. Section 2 introduces light gradient boosting machine with genetic algorithms. Section 3 illustrates the proposed architecture of forecasting daily COVID-19 confirmed cases by light gradient boosting machine with genetic algorithms. Section 4 presents the numerical results. Section 5 depicts conclusions.

2. Light Gradient Boosting Machine with Genetic Algorithms

The Light Gradient Boosting Machine (LightGBM), proposed by Ke et al. [15], is a modified approach from the Gradient Boosting Decision Tree (GBDT) algorithm. Compared with the GBDT, the LightGBM can easily cope with high-dimensional data and significantly improve efficiency while not decreasing the accuracy of forecasting. Two auxiliary methods, namely the gradient-based one-side sampling and exclusive feature

Table 1: Literatures of Google Trends keywords for COVID-19.

Ref.	Keywords
Kurian et al.[17]	COVID-19 symptoms, Coronavirus symptoms, Sore throat AND shortness of breath AND fatigue AND cough, Coronavirus testing center, Loss of smell, Lysol, Antibody, Face Mask, Coronavirus vaccine, COVID-19 stimulus check.
Lin et al. [8]	Wash hands, Face mask.
Prasanth et al. [23]	Coronavirus symptoms, Coronavirus, Covid, Handwash, Healthcenter, Mask, Positive cases, Sanitizer, Coronavirus Vaccine.
Springer et al. [26]	Wash hands, Social distancing, Lock down, Panic buying, Covid19 symptoms.
Pascual and Pourmand [22]	Do I have coronavirus, How to get tested for coronavirus, Signs and symptoms of coronavirus, What is coronavirus, How is coronavirus spread.
Misiak et al. [20]	Suicide, Depression, Anxiety, Insomnia.
Wang et al. [28]	How long contagious, loss of smell, loss of taste, covid-19 vaccine, Cough, pneumonia, how long covid-19, sinus, symptoms of the covid-19, contagious coronavirus, coronavirus vaccine.
Rabiolo et al. [24]	Ageusia, Anosmia, Chills, Cough, Eye pain, Fever, Headache, Nasal congestion, Rhinorrhea, Shortness of breath, Sore throat.
Husnayain et al. [12]	Coronavirus, coronavirus test, Middle East respiratory syndrome, face mask, social distancing, Shinchoenji, kf94 mask, disposable mask, thermometer, hand sanitizer, mask strap, kf80 mask.
Izhar and Torabi [13]	Covid-19 Malaysia.
Ben et al. [4]	Nausea, vomiting, abdominal pain, diarrhea, anorexia, loss of taste, cough, fever.
Awijen et al. [3]	Recession, survivalism, conspiracy theory, stock market crash.
Rao et al. [25]	Coronavirus, Cough, COVID 19, Diabetes, Heart, COVID pneumonia, Temperature, Pneumonia, Symptoms.
Khakimova et al. [16]	2019 novel coronavirus disease, 2019 novel coronavirus infection, 2019 novel coronavirus, 2019nCov, Coronavirus disease 2019 virus, Coronavirus disease 2019, covid 19, COVID19 virus vaccine, COVID-19 virus vaccine, COVID19 virus, COVID-19virus, COVID19, COVID-2019, SARScov2, Vaccina, Vaccine, vaccination, Vaccines, Wuhan corona virus, Wuhan seafood market pneumonia virus, Severe acute respiratory syndrome coronavirus 2, AstraZeneca vaccine, China corona vaccine, Corona Chinese vaccine, Korea Corona Vaccine, Moderna vaccine, Pfizer Corona Vaccine, Russian corona vaccine, Sputnik v,UK corona vaccine, US Corona Vaccine.
Brodeur et al. [7]	Boredom, Contentment, Divorce, Impairment, Irritability, Loneliness, Panic, Sadness, Sleep, Stress, Suicide, Well-being, Worry.
Abbas et al. [1] and Alruily et al. [2]	Hypoxemia, ageusia, anosmia, dysgeusia, hypoxia, fever, pneumonia, chills, and shortness of breath (SOB).
Yousefinaghani et al. [31]	Fever, cough, tiredness, shortness of breath, loss of smell, sore throat.

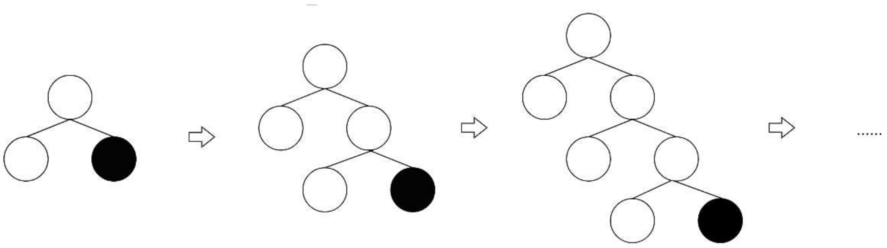

Figure 1: The leaf-wise tree growth (see Ke et al. [15], Sun et al. [27]).

bundling, are included in developing the LightGBM algorithm. Additionally, when growing the leaves, the leaf-wise tree generation strategy (Figure 1) which improves performance of the LightBGM and avoids overfitting is conducted.

The main purpose of the LightGBM is to find an appropriate objective function, expressed as Equation (2.1), by a given training set $\{(X_i, Y_i)\}_{i=1}^{n}$ to minimize the loss function $LS(Y, f(X))$. A number of T regression trees $\sum_{t=1}^{T} g_t(X)$ are combined with LightGBM to approximate the final model and illustrated as Equation (2.2). (see Ke et al. [15], Sun et al. [27])

$$\hat{g}(X) = \text{MIN}\, LS(Y, g(X)) \tag{2.1}$$

$$g_T(X) = \sum_{t=1}^{T} g_t(X). \tag{2.2}$$

The weight of the node W_j, where $j \in \{1, 2, \ldots, s\}$, is calculated by the regression tree, where s is the number of leaves, j represents the decision rules of the tree, and W depicts the sample weight of leaf nodes. The objective function can be estimated by Newton's technique. Equation (2.3) shows the transformed formulation by deleting the constant element.

$$\delta_t = \sum_{i=1}^{n} L(Y_i, g_{t-1}(X_i) + g_t(X_i)) \cong \sum_{i=1}^{n} \Big(p_i g_t(X_i) + \frac{1}{2} q_i g_t^2(X_i)\Big) \tag{2.3}$$

where p_i and q_i illustrate the first-order and second-order of the loss function, respectively. Given H_s represents the sample set of leaves, Equation (2.3) is converted into Equation (2.4).

$$\delta_t = \sum_{s=1}^{s} \bigg(\Big(\sum_{i \in H_s} p_i\Big) W_s + \frac{1}{2}\Big(\sum_{i \in H_s} q_i + \beta\Big) W_s^2\bigg). \tag{2.4}$$

The weight of each leaf node W_s^* and evaluation function of the tree δ_t^* can be generated by Equations (2.5)-(2.6). Finally, the objective function is expressed as Equation (2.7).

$$W_s^* = -\frac{\sum_{i \in H_s} p_i}{\sum_{i \in H_s} q_i + \beta} \tag{2.5}$$

$$\delta_t^* = -\frac{1}{2}\sum_{s=1}^{s}\frac{(\sum_{i\in H_s} p_i)^2}{\sum_{i\in H_s} q_i + \beta} \tag{2.6}$$

$$\text{Gain} = \frac{1}{2}\Big(\frac{(\sum_{i\in H_l} p_i)^2}{\sum_{i\in H_l} q_i + \beta} + \frac{(\sum_{i\in H_r} p_i)^2}{\sum_{i\in H_r} q_i + \beta} - \frac{(\sum_{i\in R} p_i)^2}{\sum_{i\in R} q_i + \beta}\Big) \tag{2.7}$$

where H_l and H_r are the left subtree and the right subtree correspondingly.

3. The Proposed Architecture for Forecasting Daily COVID-19 Confirmed Cases

In this study, light gradient boosting machine models with genetic algorithms and Google Trends were presented to forecast the daily COVID-19 confirmed cases in the United States. The other three forecasting models, including Back Propagation Neural Network (BPNN), General Regression Neural Network (GRNN), and Classification and Regression Tree (CART), were utilized to perform forecasting tasks with the same data to compare performance with the presented Light Gradient Boosting Machine with Genetic Algorithms and Google Trends (LGBMGAGT) models. Figure 2 illustrates the flowchart of this study. Three parts are included in this section, namely data collection, data preprocessing, and parameters selection by genetic algorithms. First, Google Trends data and the daily confirmed cases in the United States were gathered. Subsequently, Google Trends data were normalized and arranged into seven datasets with time lags of daily confirmed cases numbers gathered from the World Health Organization (WHO). Finally, data were divided into training data sets and testing data sets for each time lag data and parameters of four models were provided by genetic algorithms. The training data sets were employed to determine forecasting models and testing data sets were used to evaluate performances of forecasting models.

During the beginning of the pandemic, people immensely use the internet to search the information about COVID-19. Google Trends [8] displays the volume of web searches for specified keywords and provides the relative search volume for each keyword. In this study, 14 Google Trends keywords were collected from March 25, 2020 to September 30, 2020 in the United States. The keywords used in this study were referred to literature reviewed in Section 1. Keywords included “COVID symptoms”, “coronavirus symptoms”, “sore throat AND shortness of breath AND fatigue AND cough”, “coronavirus testing center”, “loss of smell”, “antiseptic”, “antibody”, “face mask”, “vaccine”, “COVID stimulus check”, “social distancing”, “lock down”, “panic buying”, and “wash hands”. The gathered data respectively present the relatively searching volume of keywords with the strength from 100 to 0, where 100 indicates the popular apex of keywords, and the value of 0 is the opposite. The daily confirmed cases in the United States were gathered from March 25, 2020, to September 30, 2020, in the World Health Organization (WHO) COVID-19 dashboard [29]. Pearson correlation coefficient expressed as Equation (3.1) was employed to select essential attributes by predetermined threshold values (see Benesty et al. [5], Pan et al. [21], Wu et al. [30], Hu et al. [11]). The threshold values of

Pearson correlation coefficient used in this study are +0.3 and -0.3, respectively.

$$r = \frac{\mathrm{Cov}\,(X,Y)}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{3.1}$$

where X indicates variables of 14 keywords and Y represents the daily confirmed cases.

In this study, the Google Trends data were normalized between 0 to 1 first. Then, the data were arranged into seven datasets according to time lags of daily confirmed cases numbers from one day to seven days ahead, namely $(T+1)$, $(T+2)$, $(T+3)$, $(T+4)$, $(T+5)$, $(T+6)$, and $(T+7)$. The keywords from Google Trends served as independent variables, and the daily confirmed cases in the United States were treated as the dependent variable. A training data set and a testing data set were divided from seven data sets with a proportion around 80the training dataset includes 146 days, whereas the testing data consisted of 37 days. Three other forecasting models, Back Propagation Neural Networks (BPNN), Generalized Regression Neural Networks (GRNN), Classification and Regression Trees (CART), were employed to deal with the seven datasets in forecasting confirmed cases.

The parameter selection of forecasting models influences forecasting accuracy a lot. This study employed genetic algorithms (see Holland [10]) to determine parameters of four forecasting models, namely LGBM, BPNN, GRNN, and CART. The integrated LGBM forecasting model with the genetic algorithm was illustrated in Figure 3. First, parameters of LGBM model was initialized randomly and represented by genes codes. Then, the forecasting task was performed and MAPE value was calculated. The MAPE served as the fitness of the genetic algorithm. If the stop criterion was reached, then the procedure stopped and a finalized parameter set of the LGBM model was generated. Otherwise, operations of crossover and mutation were performed individually or successively. Population with the best fitness was selected. Finally, a new parameter set of LGBM was provided. For LGBM models, six parameters including the number of boost rounds, the number of leaves, learning rates, the maximum of depth, the feature fraction, and the subsample, are determined. Learning rates and momentum are tuned for the BPNN models. The spread is selected for the GRNN models. The maximum samples to split the minimum leaf samples, the number of variables to sample, are determined for CART models. A single-point crossover with binary coding was applied for genetic algorithms in this study. Referring to a survey paper (Hassanat et al. [9]), the iterations, the population size, the crossover rate, the mutation rate, and binary bits of genetic algorithms were set to be 100, 40, 0.8, 0.6, and 25 respectively. Table 2 lists searching ranges by genetic algorithms for parameters of four forecasting models. Table 3 presents the selected values of parameters for forecasting models.

4. Numerical Results

Five models, including LightGBM, BPNN, GRNN, and CART, were used to forecast the daily United States confirmed cases with search volume of Google Trends keywords. Three indices, namely mean absolute percentage error (MAPE), root mean square error

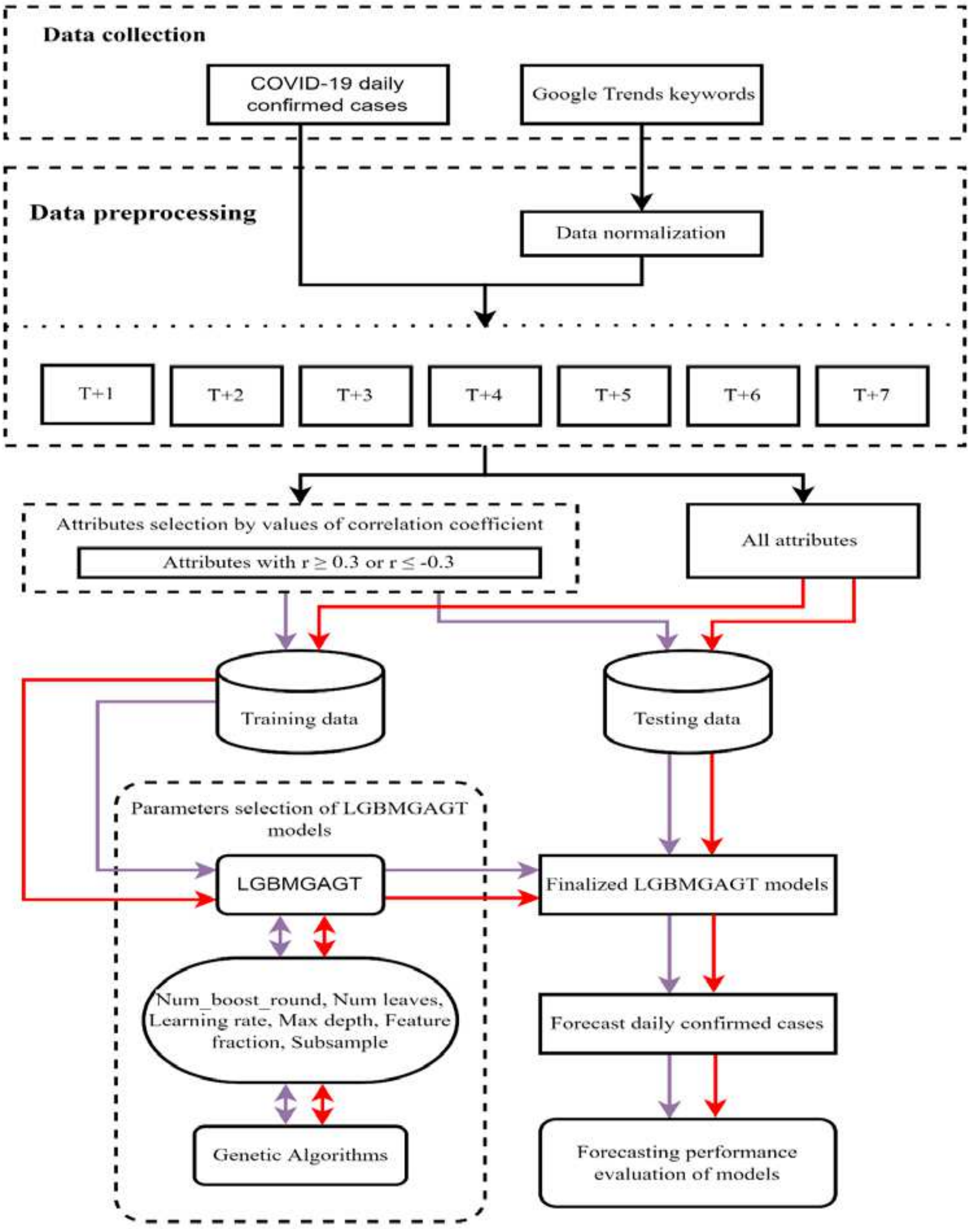
Data collection
COVID-19 daily confirmed cases
Google Trends keywords
Data preprocessing
Data normalization
T+1
T+2
T+3
T+4
T+5
T+6
T+7
Attributes selection by values of correlation coefficient
Attributes with r ≥ 0.3 or r ≤ -0.3
All attributes
Training data
Testing data
Parameters selection of LGBMGAGT models
LGBMGAGT
Finalized LGBMGAGT models
Num_boost_round, Num leaves, Learning rate, Max depth, Feature fraction, Subsample
Forecast daily confirmed cases
Genetic Algorithms
Forecasting performance evaluation of models

Figure 2: The flowchart of this study.

(RMSE), and mean absolute error (MAE), were employed to measure the performance of the forecasting models and expressed as Equations (4.1)-(4.3).

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{Y}_i - Y_i|}{Y_i} \times 100\% \tag{4.1}$$

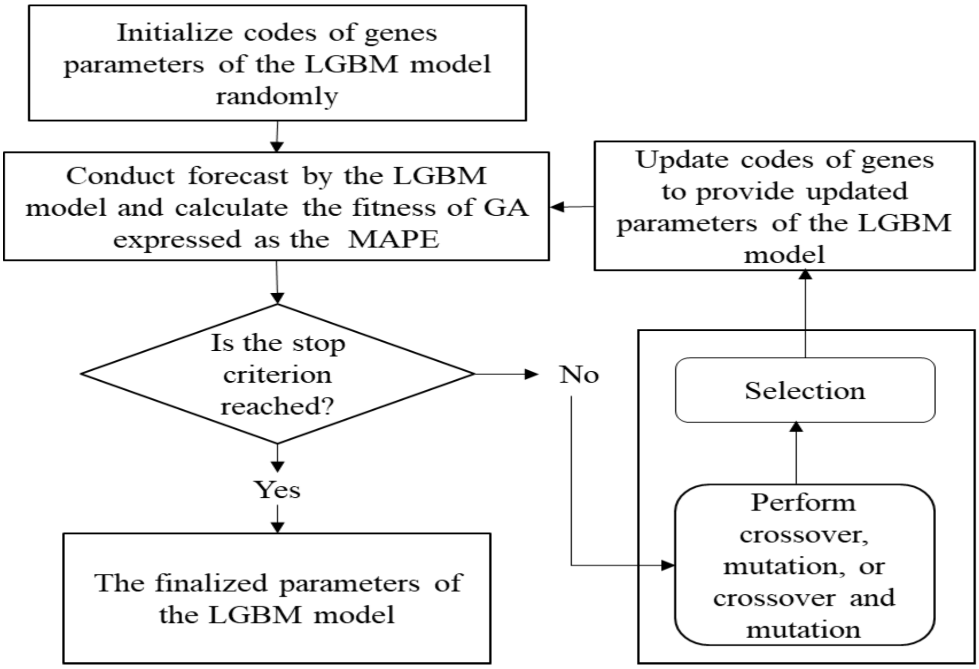
Initialize codes of genes parameters of the LGBM model randomly
Conduct forecast by the LGBM model and calculate the fitness of GA expressed as the MAPE
Update codes of genes to provide updated parameters of the LGBM model
Is the stop criterion reached?
No
Selection
Yes
Perform crossover, mutation, or crossover and mutation
The finalized parameters of the LGBM model

Figure 3: The integrated LGBM forecasting model with the genetic algorithm.

Table 2: Searching ranges for parameters of forecasting models.

Models	Parameters	Searching ranges
LightGBM	The number of boost rounds	[70,1120]
	The number of leaves	[31,38]
	Learning rates	[0.01,0.9]
	The maximum of depth	[2,9]
	The feature fraction	[0.01,0.9]
	The subsample	[0.01,0.9]
BPNN	Learning rates	[0.01,0.9]
	Momentum	[0.01,0.9]
GRNN	The spread	[0.01,500]
CART	The maximum samples to split	[1,500]
	The minimum leaf samples	[1,100]
	The number of variables to sample	[1,40]

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{4.2}$$

Table 3: The determined parameter values by genetic algorithms of forecasting models.

Models	Parameters	T+1	T+2	T+3	T+4	T+5	T+6	T+7
LightGBM	The number of boost rounds	70	140	70	700	770	840	210
	The number of leaves	34	32	34	33	37	34	34
	Learning rates	0.01	0.01	0.24	0.04	0.01	0.01	0.01
	The maximum of depth	4	4	6	3	4	6	9
	The feature fraction	0.24	0.27	0.21	0.84	0.47	0.76	0.81
	The subsample	0.64	0.41	0.33	0.87	0.33	0.41	0.38
BPNN	Learning rates	0.26	0.78	0.21	0.51	0.55	0.03	0.25
	Momentum	0.66	0.04	0.42	0.66	0.43	0.11	0.46
GRNN	The spread	0.04	0.05	0.23	0.07	0.02	0.03	0.23
CART	The maximum samples to split	179	303	81	109	364	323	260
	The minimum leaf samples	2	27	5	5	7	36	2
	The number of variables to sample	11	3	10	3	10	3	20

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{4.3}$$

where Y_i and $\hat{Y}_i$ are the actual value and predicted value at time i, respectively. n is the number of the forecasting periods. In addition, the Pearson correlation coefficient (see Benesty et al. [5], Pan et al. [21], Wu et al. [30], Hu et al. [11]) was used to select the essential attributes. Thus, two scenarios, forecasting with all attributes and forecasting with selected attributes, were provided for further analysis in this study.

4.1. Forecasting with all attributes

In this scenario, all attributes were employed to train four forecasting models. The performances of four forecasting models with seven datasets and all attributes were presented in Table 4-6 in terms of MAPE, RMSE, and MAE correspondingly. Numerical results revealed that the LightGBM outperforms the other three forecasting models according to average values of three measurements correspondingly.

4.2. Forecasting with selected attributes

As illustrated in Table 7, the raw data were used to attribute selection by Pearson's correlation analysis (Jebli et al. [14] and Liu et al. [19]). Correlation coefficients for selected variables with r values less than -0.3 or greater than 0.3 were marked in bold, and the number of variables was presented. The performances of four forecasting models with seven datasets and selected attributes were expressed in Tables 8-10 in terms of MAPE, RMSE, and MAE, respectively. Numerical results revealed that LightGBM models with selected variables are able to provide the more accurate average forecasting

Table 4. MAPE values with all attributes.

Datasets	LightGBM	BPNN	GRNN	CART
T+1	15.06	18.99	19.88	26.98
T+2	14.55	17.39	20.56	15.10
T+3	14.78	14.28	13.83	18.49
T+4	13.53	12.82	13.91	28.99
T+5	13.64	14.58	16.91	14.12
T+6	14.02	17.21	17.55	35.03
T+7	14.33	18.26	22.06	19.27
Average	**14.28**	**16.22**	**17.81**	**22.57**

Table 5. RMSE values with all attributes.

Datasets	LightGBM	BPNN	GRNN	CART
T+1	6694	8590	9254	12371
T+2	7095	7973	9311	6688
T+3	6905	6628	6252	9367
T+4	6496	6114	6135	12787
T+5	6277	6845	7287	6734
T+6	6244	7974	7693	15227
T+7	6456	7958	9227	9123
Average	**6595**	**7440**	**7880**	**10328**

Table 6: MAE values with all attributes.

Datasets	LightGBM	BPNN	GRNN	CART
T+1	5771	6640	7123	11198
T+2	5674	6594	7327	5298
T+3	5568	5373	4920	6908
T+4	5327	4585	4968	10365
T+5	4960	5270	6076	5499
T+6	5139	5899	6172	14013
T+7	5353	6493	7912	7543
Average	**5399**	**5836**	**6357**	**8689**

results than the other three models with attributes selection. Besides, all forecasting models with attributes selection can obtain smaller MAPE values than models without attributes selection.

5. Conclusions

The outbreak of COVID-19 has spread rapidly in the world until now. The pandemic also changed our daily life and caused a global economic recession. Therefore, the governments have announced new policies for the purposes of decreasing people contact and controlling inflection. The government can make earlier decisions by forecasting confirmed cases. Accurate prediction of daily COVID-19 confirmed cases is crucial for allocating medical resources effectively and efficiently. In this study, search volumes of Google Trends keywords were used to predict the daily confirmed cases in the United States by four forecasting models, including LGBMGAGT, BPNN, GRNN, and CART. Datasets with seven time lags of daily confirmed cases and volumes of Google Trends keywords search were employed to examine forecasting performances of different models. Genetic algorithms were employed to determine parameters of forecasting models. In ad-

Table 7: Correlation of coefficients.

Variables	Keywords	T+1	T+2	T+3	T+4	T+5	T+6	T+7
x1	COVID symptoms	**0.638**	**0.664**	**0.700**	**0.712**	**0.694**	**0.676**	**0.650**
x2	Coronavirus symptoms	-0.171	-0.163	-0.154	-0.146	-0.150	-0.151	-0.158
x3	Sore throat AND shortness of breath AND fatigue AND cough	0.091	0.082	0.074	0.071	0.034	0.015	-0.001
x4	Coronavirus testing center	-0.015	0.048	0.054	0.096	0.019	0.010	0.018
x5	Loss of smell	**0.384**	**0.408**	**0.412**	**0.402**	**0.377**	**0.336**	0.262
x6	Antiseptic	-0.247	-0.232	-0.188	-0.216	-0.240	-0.218	-0.268
x7	Antibody	**-0.417**	**-0.376**	**-0.318**	**-0.325**	**-0.344**	**-0.364**	**-0.372**
x8	Face Mask	-0.105	-0.098	-0.106	-0.101	-0.118	-0.108	-0.087
x9	Vaccine	**0.423**	**0.537**	**0.609**	**0.573**	**0.479**	**0.382**	0.296
x10	COVID stimulus check	-0.233	-0.217	-0.200	-0.204	-0.207	-0.258	-0.288
x11	Social distancing	**-0.475**	**-0.430**	**-0.424**	**-0.425**	**-0.435**	**-0.438**	**-0.445**
x12	Lock down	-0.306	-0.278	-0.284	-0.278	-0.287	-0.292	-0.274
x13	Panic buying	-0.221	-0.223	-0.234	-0.196	-0.209	-0.219	-0.240
x14	Wash hands	**-0.365**	-0.298	-0.278	-0.291	**-0.311**	**-0.323**	**-0.348**
Numbers of variables		**7**	**5**	**5**	**5**	**6**	**6**	**4**

Table 8. MAPE values with selected attributes.

Datasets	LightGBM	BPNN	GRNN	CART
T+1	14.07	17.29	14.62	15.98
T+2	13.96	16.88	17.76	28.00
T+3	14.03	17.26	15.52	17.12
T+4	14.00	13.82	15.26	14.67
T+5	13.37	17.35	15.37	13.33
T+6	14.25	15.05	15.92	31.60
T+7	14.29	15.10	15.36	15.81
Average	**13.99**	**16.11**	**15.68**	**19.50**

Table 9. RMSE values with selected attributes.

Datasets	LightGBM	BPNN	GRNN	CART
T+1	6415	8294	6697	7637
T+2	6749	7494	7811	13178
T+3	6563	8412	6906	7491
T+4	6778	6186	6894	6464
T+5	6165	8163	7263	6024
T+6	6440	7678	6547	7852
T+7	6456	7958	9227	9123
Average	**6502**	**7567**	**7029**	**8775**

dition, data with and without attributes selection were employed to observe performances of forecasting models. This study showed that attributes selection can improve forecasting accuracy for all forecasting models in terms of average values of measurements. The main contribution of this study is to examine the feasibility and effectiveness of the developed LGBMGAGT (Light Gradient Boosting Machine with Genetic Algorithms and Google Trends) model in forecasting COVID-19 confirmed cases with google trend search data during the beginning and rapidly developing periods. The empirical results revealed

Table 10: MAE values with selected attributes.

Datasets	LightGBM	BPNN	GRNN	CART
$T+1$	5127	6232	5303	6353
$T+2$	5482	5848	6299	11548
$T+3$	5438	7171	5448	5949
$T+4$	5514	4900	5344	5163
$T+5$	4853	6668	5482	5034
$T+6$	4987	5763	5767	11279
$T+7$	5101	6187	5617	6033
Average	**5215**	**6110**	**5609**	**7337**

that the designed LGBMGAGT model was a feasible as well as promising alternative in forecasting daily COVID-19 confirmed in the USA with forecasting MAPE values less than 15% when feature selection was applied.

The maximum number of Google Trends daily data can be gathered is 270 days (Bleher and Dimpfl [6]). Thus, data collection for a longer time frame is a major limitation of this study. For future study, integrating periods of Google Trends data into a longer time frame to performing forecast COVID-19 confirmed cases is a challenging direction for future study. In addition, keywords in different languages can be used for collecting the search volume of Google Trends. Thus, the designed LGBMGAGT can be used to predict confirmed cases in other countries. Thirdly, deep learning models can be used to deal with the same datasets and compare results with the LGBMGAGT model. Finally, the latest data collected from 2022 could be a possible direction for future study.

Acknowledgements

This research was funded by the Ministry of Science and Technology, Taiwan under the Contract Number MOST 109-2410-H-260-023.

References

[1] Abbas, M., Morland, T. B., Hall, E. S. and El-Manzalawy, Y. (2021). *Associations between google search trends for symptoms and covid-19 confirmed and death cases in the United States*, International Journal of Environmental Research and Public Health, Vol.18, 4560.

[2] Alruily, M., Ezz, M., Mostafa, A. M., Yanes, N., Abbas, M. and El-Manzalawy, Y. (2022). *Prediction of covid-19 transmission in the united states using google search trends*, Computers, Materials and Continua, Vol.71, 1751-1768.

[3] Awijen, H., Zaied, Y. B. and Nguyen, D. K. (2022). *Covid-19 vaccination, fear and anxiety: Evidence from Google search trends*, Social Science & Medicine, Vol.297, 114820.

[4] Ben, S., Xin, J., Chen, S., Jiang, Y., Yuan, Q., Su, L., Christiani, D. C., Zhang, Z., Du, M. and Wang, M. (2022). *Global internet search trends related to gastrointestinal symptoms predict regional COVID-19 outbreaks*, Journal of Infection, Vol.84, 56-63.

[5] Benesty, J., Chen, J., Huang, Y. and Cohen, I. (2009). *Pearson correlation coefficient*, In Noise reduction in speech processing (pp.1-4). Springer, Berlin, Heidelberg.

[6] Bleher, J., and Dimpfl, T. (2022). *Knitting Multi-Annual High-Frequency Google Trends to Predict Inflation and Consumption*, Econometrics and Statistics, Vol.24, 1-26.

[7] Brodeur, A., Clark, A. E., Fleche, S. and Powdthavee, N. (2021). *COVID-19, lockdowns and well-being: Evidence from Google Trends*, Journal of Public Economics, Vol.193, 104346.

[8] Google Trends website. Available online: https://trends.google.com.tw/trends/?geo=US

[9] Hassanat, A., Almohammadi, K., Alkafaween, E. A., Abunawas, E., Hammouri, A. and Prasath, V. S. (2019). *Choosing mutation and crossover ratios for genetic algorithms — a review with a new dynamic approach*, Information, Vol.10, 390.

[10] Holland J.H. (1975). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, MIT Press, 439-444.

[11] Hu, C., Liu, B., Wang, S., Zhu, Z., Adcock, A., Simpkins, J. and Li, X. (2022). *Spatiotemporal Correlation Analysis of Hydraulic Fracturing and Stroke in the United States*, International Journal of Environmental Research and Public Health, Vol.19, 10817.

[12] Husnayain, A., Shim, E., Fuad, A. and Su, E. C. Y. (2021). *Predicting New Daily COVID-19 Cases and Deaths Utilizing Search Engine Query Data in South Korea from 2020 to 2021: Infodemiology Study*, Journal of Medical Internet Research, Vol.23. DOI: 10.2196/34178

[13] Izhar, T. A. T. and Torabi, T. (2022). *Online searching trend on Covid-19 using Google trend: infodemiological study in Malaysia*, International Journal of Information Technology, Vol.14, 675-680.

[14] Jebli, I., Belouadha, F. Z., Kabbaj, M. I. and Tilioua, A. (2021). *Prediction of solar energy guided by pearson correlation using machine learning*, Energy, Vol.224, 120109, 1-20.

[15] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. Y. (2017). *Lightgbm: A highly efficient gradient boosting decision tree*, Advances in Neural Information Processing Systems, Vol.30. 1-9.

[16] Khakimova, A., Abdollahi, L., Zolotarev, O. and Rahimd, F. (2022). Global interest in vaccines during the COVID-19 pandemic: Evidence from Google Trends, Vaccine: X, Vol.10, 10015.

[17] Kurian, S. J., Bhatti, A. R., Alvi, M. A., Ting, H. H., Storlie, C., Wilson, P. M., Shah, N. D., Liu, H. and Bydon, M. (2020). *Correlations Between COVID-19 Cases and Google Trends Data in the United States: A State by State Analysis*, Mayo Clinic Proceedings, Vol.95, 2370-2381.

[18] Lin, Y. H., Liu, C. H. and Chiu, Y. C. (2020). *Google searches for the keywords of "wash hands" predict the speed of national spread of COVID-19 outbreak among 21 countries*, Brain, Behavior, and Immunity, Vol.87, 30-32.

[19] Liu, J., Yu, C., Hu, Z., Zhao, Y., Bai, Y., Xie, M., and Luo, J. (2020). *Accurate prediction scheme of water quality in smart mariculture with deep Bi-S-SRU learning network*, IEEE Access, Vol.8, 24784-24798.

[20] Misiak, B., Szcześniak, D., Koczanowicz, L. and Rymaszewska, J. (2020). *The COVID-19 outbreak and Google searches: Is it really the time to worry about global mental health*?, Brain, Behavior, and Immunity, Vol.87, 126-127.

[21] Pan, Y., Zhang, L., Unwin, J. and Skibniewski, M. J. (2022). *Discovering spatial-temporal patterns via complex networks in investigating COVID-19 pandemic in the United States*, Sustainable Cities and Society, Vol.77, 103508.

[22] Pascual, K. J. and Pourmand, A. (2020). *Using Google Trends to Determine Perceived Viral Exposure during the Early Phase of the COVID-19 Pandemic in the United States*, Annals of Emergency Medicine, Vol.76, Supplement, S108-S109.

[23] Prasanth, S., Singh, U., Kumar, A., Tikkiwal, V. A. and Chong, P. H. J. (2021). *Forecasting spread of COVID-19 using Google Trends: A hybrid GWO-Deep learning approach*, Chaos, Solitons and Fractals, Vol.142, 110336.

[24] Rabiolo, A., Alladio, E., Morales, E., McNaught, A. I., Bandello, F., Afifi, A. A. and Marchese, A. (2021). *Forecasting the COVID-19 Epidemic by Integrating Symptom Search Behavior Into Predictive Models: Infoveillance Study*, Journal of Medical Internet Research, Vol.8, 23.

[25] Rao, A., Sharma, G. D., Pereira, V., Shahzad, U. and Jabeen, F. (2022). *Analyzing Cyberchondriac Google Trends Data to Forecast Waves and Avoid Friction: Lessons From COVID-19 in India*, IEEE Transactions on Engineering Management, Vol.0, 1-14.

[26] Springer, S., Menzelb, L. M. and Ziegerc, M. (2020). *Google Trends provides a tool to monitor population concerns and information needs during COVID-19 pandemic*, Brain, Behavior, and Immunity, Vol.87, 109-110.

[27] Sun, X., Liu, M. and Sima, Z. (2022). *A novel cryptocurrency price trend forecasting model based on LightGBM*, Finance Research Letters, Vol.32, 101084.

[28] Wang, T., Ma, S., Baek, S. and Yang, S. (2022). *COVID-19 Hospitalizations Forecasts Using Internet Search Data*, arXiv, Vol.181, 940-947.

[29] World Health Organization. COVID 19 dashboard. Available online: https://covid19.who.int/info

[30] Wu, Y. K., Lai, Y. H., Huang, C. L., Phuong, N. T. B. and Tan, W. S. (2022). *Artificial Intelligence Applications in estimating invisible solar power generation*, Energies, Vol.15, 1312.

[31] Yousefinaghani, S., Dara, R., Mubareka, S. and Sharif, S. (2021). *Prediction of COVID-19 waves using social media and Google search: a case study of the US and Canada*, Frontiers in Public Health, Vol.9, 656635.

Department of Information Management, National Chi Nan University, Taiwan.

E-mail: amy087226g@gmail.com

Major area (s): Data analysis, forecasting, and machine learning.

Department of Information Management, National Chi Nan University, Taiwan.

Department of Information Management, Ph.D. Program in Strategy and Development of Emerging Industries, National Chi Nan University, Taiwan.

E-mail: paipf@ncnu.edu.tw

Major area (s): Data analysis, forecasting, and machine learning.

Department of Information Management, National Chi Nan University, Taiwan.

E-mail: s107213029@mail1.ncnu.edu.tw

Major area(s): Data analysis, forecasting, and machine learning.

Ph.D. Program in Strategy and Development of Emerging Industries, National Chi Nan University, Taiwan.

E-mail: s109245911@ncnu.edu.tw

Major area(s): Data analysis, forecasting, and machine learning.

(Received September 2022; accepted November 2022)