Journal of Engineering Science and Technology
Vol. 19, No. 4 (2024) 1254 - 1267

PREDICTING COVID-19 CASES USING HYBRID ARIMAX-BIDIRECTIONAL LSTM MODEL

NUGROHO WIDYANTO[1,*], JIN-WHAN KIM[1], UMI NARIMAWATI[2]

[1]Department of Computer Engineering, Youngsan University, South Korea
[2]Department of Management, Universitas Komputer Indonesia, Indonesia
*Corresponding Author: nugrohowidyanto@office.ysu.ac.kr

Abstract

Currently, there has been improvement in the handling of the pandemic with the availability of vaccines, which has had an impact on improving the economy and starting to return to normal community activities. However, further studies are still needed. So that this pandemic condition can be anticipated in the future, there must be a model that can predict future COVID-19 cases, so that each country can anticipate it by implementing more appropriate policies and handling. In this research, we will propose 2 methods that are popularly used in the case of time-series forecasting involving other factors in the regression model, namely ARIMAX, which is very superior in predicting linear (stationary) models, and bidirectional LSTM, which is a better development than LSTM, which can predict non-linear (non-stationary) models. The hybrid approach used for these two models is expected to provide much better prediction results with a smaller error range, as well as reducing existing variations, compared to using both methods separately. The process of creating a hybrid ARIMAX-Bidirectional LSTM model involves six sub-processes: data transformation using ARIMAX, data normalization (min-max scaling), feature selection using a genetic algorithm, hyper-parameter tuning for ARIMA models and bidirectional LSTM models, prediction models with bidirectional LSTM, and lastly the creation of a hybrid model. The hybrid model was proven to provide much better prediction results than the models run individually. The ARIMAX-Bidirectional LSTM hybrid model has also been proven to be better than the ARIMAX-LSTM hybrid model, as well as other deep learning-based hybrid models. Even though the ARIMAX-Bidirectional LSTM hybrid model can provide very good predictions, its implementation has not been widely carried out, so it is an excellent opportunity to carry out research on its use and utilization.

Keywords: ARIMAX, Bidirectional LSTM, COVID-19, Deep learning, Forecasting, LSTM, Model hybrid, Time-series.

1. Introduction

The SARS-CoV-2 virus infection was first identified in Wuhan, China at around the end of 2019, which then spread throughout the world around June 2020. The World Health Organization (WHO) immediately declared pandemic status and then this virus was given the name COVID-19. Over time in almost all countries, the number of confirmed positive cases and deaths has increased. This has forced several countries to implement strict policies to limit the activities and mobility of their people. The consequences of this policy ultimately resulted in the collapse of the world economy [1-6].

Currently there has been improvement in the handling of the pandemic, with the availability of vaccines, which has had an impact on improving the economy and starting to return to normal community activities. However, further studies are still needed, so that this pandemic condition can be anticipated, there must be a model that can predict future COVID-19 cases, so that each country can anticipate it by implementing more appropriate policies and handling [1-6].

One prediction method that can be used is using time-series forecasting by considering other (external) influencing factors. However, by looking at data from the past (historical), it turns out that not all changes in the data are stationary, but there are also times when these changes are non-stationary. This could be due to the emergence of a new variant of the COVID-19 virus, where previously the number of cases had decreased drastically, but could suddenly increase, or this change could also be caused by other external factors, such as demographic factors, mobility factors, or public health factors, for example, such as the increase in comorbid diseases.

Therefore, the prediction model that will be applied must adopt a linear model and a non-linear model so that good prediction results can be obtained, namely prediction results with a relatively small error range [7]. In this research, we will propose two methods that are popularly used in the case of time-series forecasting, involving other factors in a regression model, namely ARIMAX which is very superior in predicting linear (stationary) models, and Bidirectional LSTM which is a better development from LSTM which can predict non-liner (non stationary) models. The hybrid approach used for these two models is expected to provide much better prediction results with a smaller error range, as well as reducing existing variations, compared to using these two methods separately [8].

2. Research method

The prediction model for COVID-19 cases is very interesting to discuss among researchers. These researchers typically use methodologies based on statistical and mathematical models as well as machine learning.

Since the COVID-19 virus has spread widely, scientists from a variety of fields have started searching for and viewing data on the evolution of COVID cases 19, such as the number of cases per day, the cumulative number of cases, the results of CT scans, MRI images of patients with lung disorders, and other data. The results of these data make it possible to conduct more in-depth research on the spread of the COVID-19 virus from various perspectives.

In order to predict the prevalence of COVID-19 in Egypt, the scientists used a nonlinear autoregressive neural network. They treated confirmed cases as a time series and contrasted their method with an Auto-Regressive Integrated Moving Average (ARIMA) model. Using confirmed instances that were reported in March, both methods were used to forecast the cumulative COVID-19 cases for ten days, from April 1 to April 10, 2020 [4].

Two types of methodology, namely deep learning techniques and statistical models, which were used to predict COVID-19 [9]. Using ARIMA and Support Vector Regression (SVR) as a basic machine learning approach, a prediction model is first designed. In the next stage, several deep learning methods used to predict COVID-19 are discussed, namely GRU, LSTM, and Bidirectional LSTM. Then in this research statistical performance was measured, namely MAE, RMSE, and R2 values which were then used in evaluating the performance measurements of each method. The purpose of this study was to forecast the number of COVID-19 instances in three categories: recovery cases, death cases, and positive confirmed cases [2].

Used a predictive strategy to assist decision makers in preventing the COVID-19 pandemic from spreading; precise forecasts of the disease's trajectory are crucial. This study uses a Bidirectional Long Short-Term Memory (Bi-LSTM) network applied to multivariate time series data to offer a deep learning method for predicting the cumulative number of COVID-19 cases. Unlike other prediction methods, our suggested strategy uses the K-means clustering algorithm to first group countries that have comparable socioeconomic and demographic characteristics as well as health sector data. To train the forecasting model, bidirectional long short-term memory (LSTM) is fed cumulative case data from grouped countries that has been augmented with information about lockout measures. We examine the illness outbreak in Qatar and the predicted dates of the proposed model from December 1, 2020, to verify the efficacy of the suggested methodology. Quantitative analysis demonstrates that the suggested method performs better than cutting-edge forecasting methods [1].

Suggested using lung X-ray pictures to diagnose and locate infected tissue from COVID-19 patients using a Convolutional Neural Network (CNN) based method. An artificial neural network trained on fractal properties of photographs was the first deep learning model. The second model uses images of lungs to train a CNN. The study's findings demonstrated that the CNN-based model outperformed the initial model, which had an accuracy of 83.4%, with a 93.2% accuracy rate [3].

Forecasted the peak of the epidemic in Japan from January 15 to February 29, 2020, using the SEIR model. A mathematical formulation for describing the spread of disease from one person to another is provided by SEIR. Four statuses are experienced by these people: susceptible (S), exposed (E), infectious (I), and recovered (R). The peak is expected to happen in early to mid-summer 2020, according to the SEIR model. Furthermore, some epidemiological inferences can be made, which can facilitate the implementation of measures that have a significant impact on postponing the epidemic's peak. To achieve an effective reduction in the extent of the epidemic, these treatments also need to be implemented over an extended period of time [6].

Used a data-driven methodology to examine the effects of the lockdown in India. According to data, the infection rate was three times lower after six weeks of

lockdown than it was before. Lockdown procedures are therefore crucial to controlling the epidemic. Nevertheless, when assessing daily or cumulative COVID-19 cases, these metrics are rarely taken into account. Furthermore, the majority of COVID-19 forecasting techniques usually rely on scant data from a single nation; yet pandemic trends may be similar among nations with comparable socioeconomic and demographic traits as well as comparable health sector indicators. Our contribution comprises of classifying nations based on comparable socioeconomic and demographic traits as well as health sector variables, and then use COVID-19 data from each group to construct a prediction model. Additionally, this study suggests a prediction method based on bidirectional LSTM and deep learning. This study uses a multivariate time series that includes the total number of daily cases and a time series that shows lockdown measures like closing schools, closing workplaces, imposing gathering restrictions, closing public transportation, and imposing restrictions on international travel in order to train a Bidirectional LSTM-based model. Multiple dependent time series can be modelled jointly to account for cross-correlation and within series that record factors that change simultaneously over time using the proposed Bidirectional LSTM on multivariate time series [5].

3. Research Method

This study discusses ARIMAX, Bidirectional LSTM, which is used to predict the number of COVID-19 positive confirmations, deaths, and recoveries. It also discusses the fundamental ideas and modelling procedure of the hybrid model that is created by logically combining these models, which will be explained in more detail below.

3.1. Feature selection using genetic algorithm

In the first sub-process, a selection process is carried out for important input features that have the most significant influence in the regression model, in this case ARIMAX, which is used to predict the number of COVID-19 cases. The initial stage is carried out using a Genetic Algorithm, starting with the creation of chromosomes with a number of features used in this research which are carried out in the population formation process. When a chromosome is represented by a 1, it indicates that the input feature is chosen, and when it is represented by a 0, it indicates that the input feature is not. As the objective function, the largest R-squared (R2) estimate is selected. In this study, the crossover probability is 0.5, the number of generations is 10, the population is 100, and the mutation probability is 0.002. Because a genetic algorithm selects just the most significant features, feature selection can enhance the predictive performance of the regression model [10].

3.2. Hyper-parameter tuning for ARIMAX and bidirectional LSTM models

The next stage is to find the best parameters for the ARIMAX and Bidirectional LSTM models which will be used in the hybrid model. After completing selecting important features using a genetic algorithm [11]. Then these features are used in the ARIMAX model and Bidirectional LSTM model [12]. In the ARIMAX model, parameters (p, d, q) will be selected which are considered to provide the smallest

error results through checking using RMSE and MAE evaluation [13]. Determining the best ARIMAX model by parameter search will use the Particle Swarm Organization (PSO) Optimizer algorithm [14].

Determining the best Bidirectional LSTM model to be used for the hybrid model in the next sub-process, focuses on the parameters of the number of hidden nodes and the optimization algorithm used [15, 16]. The goal of this research is to determine the optimal number of concealed nodes by setting an optimization algorithm that makes use of Adaptive Moment Estimation (ADAM). In increments of 10, the number of nodes in the hidden layer is experimented with between 10 and 50. In order to assess model performance using the estimated R-squared (R2) value, the K-Fold Cross Validation method will also be employed throughout the cross-validation process for these models. which is acquired [17, 18].

3.3. ARIMAX

We will first go over the ARIMA sub-process before moving on to the third sub-process, which is ARIMAX. When developing prediction models using time series data, ARIMA is frequently used to estimate future data by applying trend cycles from prior data [19-21]. There are three components to ARIMA (p, d, q). The first section, integrated as d, uses differential testing and Augmented Dickey-Fuller (ADF) to transform non-stationary data into stationary data. The second component is autoregressive (AR), which forecasts the next timestamp based on past data. The moving average (MA) as q, which determines the average error or noise from historical data, is the final component [7]. The ARIMA's output is computed using:

$$y_{t=}\sum_{i=1}^{p}(\varphi_i x_{t-1})-\sum_{j=1}^{q}(\theta_j \varepsilon_{t-j})+\varepsilon_t \quad (1)$$

In Eq. (1), y_t shows the predicted value at time t, φ_t and θ_t shows the coefficient values at time t, x_t shows the historical value at time t, y_t shows the predicted value at time t, ε_t shows the error and noise terms at time t. The moving average series is represented by q, and the autoregressive sequence by p.

ARIMAX is simply ARIMA with additional input features. To generate new input features, this study transforms data using ARIMAX (p, d, q). To build an ARIMAX model, input features selected from the preceding procedure are employed in a training dataset. Subsequently, every piece of data from the training dataset is fed back into the ARIMAX model, yielding fresh input features in the form of output data.

$$y_{t=}\sum_{i=1}^{p}(\varphi_i x_{t-1})-\sum_{j=1}^{q}(\theta_j \varepsilon_{t-j})+\sum_{k=1}^{r}(\beta_k x_{tk})+\varepsilon_t \quad (2)$$

where: x_kt denotes the k input features at time t, β_t and θ_t shows the coefficient values at time t, and r indicates the number of input features.

3.4. Bidirectional LSTM

The normalization of data is the next subprocess. The new input characteristics from the first sub-process are normalized using min-max normalization (mix-max scaling) in this process. This study selects a range of 0 to 1 for the new input feature range based on the activation function of the input for the Bidirectional LSTM model, which is the next sub-process.

While unfolded RNN can process current data using past data, RNN struggles to maintain long-term dependency. RNN has been utilized for sequential time series applications with temporal dependency. One of the RNN variations developed as a result of this work, LSTM, which is patented by Hochreiter and Schmid Huber, can solve this. With the addition of hidden layer components called memory cells, LSTM, a new breed of RNN networks, has been able to overcome RNN's drawbacks. Memory cells are governed by gates called input, output, and forget gates. These gates are connected to the memory cells and store the temporal state of the network [14, 20, 22, 23].

The input and output flow of memory cells in other areas of the network is managed by the input and output gates. Furthermore, the memory cell now has a forget gate, which eliminates high-weight information from the preceding neuron. The results of high activation determine what is stored in memory; if the input of the unit is highly activated, the information is stored in the memory cell; if the output of the unit is highly activated, the information is stored in the subsequent neuron; if not, the input information with a high weight is stored in the memory cell [22, 23].

The mapping between the input and output sequences, X = (X1, X2,..., Xn) and Y = (y1, y2,..., yn), is computed by the LSTM network. Use the following equation to calculate

$$f_{orget}g_{ate} = sigmoid\ (W_{fg}X_t + W_{hfg}h_{t-1} + b_{fg}) \quad (3)$$

$$i_{nput}g_{ate} = sigmoid\ \left(W_{ig}X_t + W_{hig}h_{t-1} + b_{ig}\right) \quad (4)$$

$$O_{utput}g_{ate} = sigmoid\ (W_{og}X_t + W_{hog}h_{t-1} + b_{og}) \quad (5)$$

$$(C)_t = (C)_{t-1} \otimes (f_{orget}g_{ate)_t} + (i_{nput}g_{ate)_t} \otimes (\tanh\ (W_cX_t + W_{hc}h_{t-1+}\text{bc})) \quad (6)$$

$$h_t = O_{utput}g_{ate} \otimes tanh\ ((C)_{t-1}) \quad (7)$$

In Eqs. (3) to (7), Wig, Wog, WhC, Wfg and bfg, big, bog, bCr - represent the weight and bias variables of each of the three levels and memory cells in the aforementioned equations. Here, ht-1 represents the previous hidden coating unit to which the element is added with three levels of weight. This implies that it is the result of multiplying the output of the previous memory unit by the output of the prior hidden unit after processing Equation (6). (C)t modifications to the current memory unit cell. Equation (3-7) illustrates how to add linearity across all levels using a hand sigmoid activation function. Here, t-1, as well as the past and present periods. to get around LSTM cells' restrictions, which prevent them from being applied to new content but allow them to function on older stuff. Two separate long short-term memory neurons (LSTMs) with the same output going in opposing directions make up the directed artificial neural network (BRNN) that Schuster and Paliwal developed [18].

This architecture makes use of information from the past and present in the output layer. The input sequence is represented as X = (X1,...,...,...,...,..., hn) and is represented in reverse as ↞ht = (↞h1,↞h2,...,↞hn). The complete sequence of outcomes for this cell, which is made up of both →ht and ↞ht, is y = (y1,y2,...,yt...,yn). shows both the bidirectional and LSTM displays.

3.5. Hybrid

For modelling non-linear interactions in time series, the ARIMAX model is insufficient, despite its proficiency at simulating linear relationships. Both linear and non-linear interactions can be modelled by bidirectional LSTM models, albeit they can't produce the same outcomes for all data sets. Therefore, a hybrid model based on the idea of separating the linear and non-linear components of the time series is utilized to get the best prediction results [12, 14, 19, 23].

Despite the fact that the hybrid model and the results produced by applying the models separately have no relationship, it can be observed that they can reduce variance or error in general [12, 17]. Because of this, it is acknowledged that the hybrid model works well for tasks involving forecasting and prediction.

This study develops a hybrid ARIMAX-Bidirectional LSTM model for COVID-19 case time-series prediction. Each time series in this model is assumed to be the mathematical sum of two linear and non-linear model components. On the basis of this, a hybrid model was developed to generate predictions utilizing past data from a time series. This model is shown as follows in the block diagram in the Fig. 1.

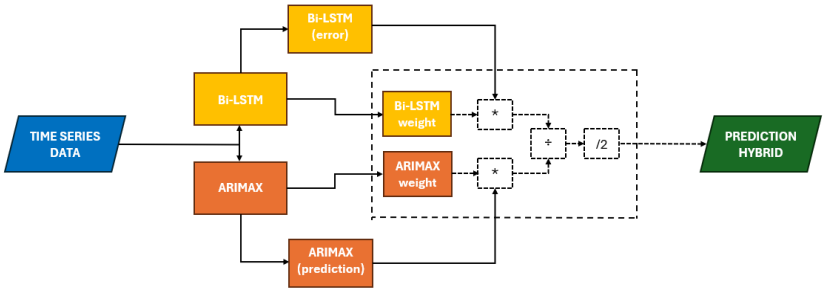

Fig. 1. The research builds a hybrid ARIMAX-Bidirectional LSTM model for time-series prediction of COVID-19 cases.

The final model's time series prediction formula is typically represented as the product of its linear and nonlinear components, as the following equation demonstrates:

$$y_t = L_t + N_t \tag{8}$$

The time series' linear component is represented by the symbol Lt, and its non-linear component is symbolized by Nt. In the hybrid model, the LSTM model is used to estimate Nt after the linear component Lt of the time series is first estimated using the ARIMAX model. The two models' error values are then computed. The following formula can be used to determine this error:

$$BiLSTM_{error} = BiLSTM\_average[error] \tag{9}$$

$$ARIMAX_{error} = ARIMAX\_average[error] \tag{10}$$

The model weights are calculated using the error values obtained in the equation below. The normalization process is carried out in calculating the weight values.

$$BiLSTM_{error} = 2\left(1 - \frac{BiLSTM_{error}}{BiLSTM_{error} + ARIMAX_{error}}\right) \tag{11}$$

The weight values from the model and finally each predicted value from the hybrid model is obtained using the equation below.

$$Hybrid_{predict}[i] = \left(BiLSTM_{weight}[i] * BiLSTM_{error}[i] + ARIMAX_{weight}[i] * ARIMAX_{error}[i]\right)/2 \quad (12)$$

3.6. Evaluation

This study's assessment centers on how accurate the hybrid model's prediction results are; specifically, this is done by computing the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Residual Mass (CRM), and Determination Coefficient R2, with the resulting equation being as follows [14]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \quad (13)$$

$$MAE = \frac{1}{n}\sum_{i-1}^{n}\left|\frac{x_i - y_i}{x_i}\right| \quad (14)$$

$$CRM = \frac{\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} \quad (15)$$

$$R^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})\right)^2}{\sum_{i=1}^{n}(x_i - \hat{x})^2 \sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (16)$$

where: x_i is actual data on COVID-19 cases, y_i is the prediction result for COVID-19 cases, xˆ is the average value of actual data on COVID-19 cases, yˆ is the average value of the prediction results for COVID-19 cases.

The model that achieves the lowest RMSE, lowest MAE, greatest R2 value, and CRM value closest to 0 (zero) is the best prediction model.

4. Result and Discussion

This section describes the results of experiments with the proposed model. Exhibitions presented relate to data sets used such as, Exploratory Data Analysis (EDA), modelling, and visualization.

4.1. Data set

The dataset used in this research is data on COVID-19 cases starting from February 2020 to November 2021. The data covers regionally including Indonesia and as a comparison, data from South Korea is used.

The data that will be used to predict in time series units is data on the number of confirmed positive cases, data on the number of deaths, and data on the amount of recovery. Meanwhile, the data involves external factors, namely demographic factors, sector health indicators, and community socio-economic factors.

For demographic factors, the data used is data on population size, population density, median age, while the health indicator sector includes data on total vaccinations, data on the number of deaths due to heart disease, data on the number of people with diabetes, data on the percentage of female smokers, data on the percentage of male smokers. Then socio-economic factors, the data used in this research, are GDP per capita data, life expectancy data and human resource development index data.

All of this data was assumed at the start to have an influence on the COVID-19 case data. These data will be processed through the methodological stages that have been determined in this research. This dataset will be used to predict up to 3 months or 90 days into the future

4.2. Exploratory data analysis (EDA)

The results of applying the time series prediction algorithm with the Python 3.10 programming language, running on the Google Colab platform, are discussed next. Other equipment when running this algorithm software program is using the Tensor flow 2.14 and Keras 2.14 libraries. Tensor Flow is a broad and versatile machine learning library, while Keras provides an easy-to-use, high-level interface for building neural network models on top of Tensor Flow.

These two libraries are really needed for algorithm implementation, both the ARIMAX model and the Bidirectional LSTM model, or those combined in the Hybrid model. In order to compare, testing is carried out, and the predicted results of COVID-19 case data for the next 3 months (90 days) can be seen. These results will be used as a basic source for conducting analysis and determining conclusions.

The initial stage when implementing the algorithm is carried out by reading the COVID-19 Dataset that has been provided in this research. This dataset is saved in CSV (Comma-Separated Values) format, which is used to represent data in table form, where each row of the table is a row in the file, and each column is separated by a comma (,). This format is used to store data used in this research, especially in the context of data processing and analysis. From the results of reading the Dataset, it is obtained that the amount of power that can be processed consists of 134,020 rows x 67 columns, but this data still comes from many countries.

Then the next stage is to only take data originating from South Korea, from the results of sorting based on country location, finally the amount of data that can be processed for the next stage is 667 rows x 67 columns. Furthermore, from the sorted dataset, only the dataset for 2021 is taken.

The dataset for 2021 that will be used will be separated into past data and future data. The benchmark for taking the dataset for past data is the data available before October and not October itself, while for future data the data will be taken after October. After that, the data that has been separated is prepared to be formed as variable X and variable Y.

For variable X, the data column "Total Cases" will be taken. Past data for this variable Then for variable Y the data columns "Population", "GDP Per Capita", "New Cases", "Total Deaths", "Total Vaccinations" will be taken, the same as the data for variable X, future data for variable Y will be used as test data and past data will be used as training data

4.3. ARIMAX algorithm implementation

The implementation of the ARIMAX algorithm is carried out using the Pmdarima (Python Auto-Regressive Integrated Moving Average) library which provides an automatic algorithm for ARIMA (Auto Regressive Integrated Moving Average) modelling. In the pmdarima library there is an auto_ARIMA function which is used to estimate the best parameters for the ARIMAX model (ARIMA with exogenous

variables) used in this research. The exogenous variable used is the variable X in the training data [15].

After the ARIMAX model can be run using the auto_ARIMA function, then the model results can be used to predict data in the future within a specified period of the amount of test data for variable Y, and the test data for variable X is used as an exogenous variable.

4.4. Bidirectional LSTM algorithm implementation

The implementation of the Bidirectional LSTM algorithm was carried out using the Tensor flow and Keras libraries. The Bi-direction LSTM model starts by using Sequential model objects. This object is used to create a neural network model sequentially (layer by layer). Each layer is added one by one in the order of execution.

Embedding parameters, namely (50, 128, input_length=3) are used in the model used. This is the embedding layer, which is used to convert the word representation into a numeric vector. The first parameter (50) is the size of the vocabulary (the number of possible words), the second parameter (128) is the dimension of the embedding vector for each word, and the third parameter (input_length=3) is the length of each input sequence.

After the Embedding parameters above, then proceed with determining the Bidirectional LSTM layer, which uses two sets of LSTM cells to process the input sequence in two directions (forward and backward). Parameter 64 is the number of units or cells in the LSTM layer.

Next is the Dropout parameter, namely (0.5). This is a dropout layer, which helps prevent overfitting by randomly ignoring some nodes during training. Parameter 0.5 indicates that half of the nodes will be dropped during each training epoch.

Lastly is the Dense parameter, namely (1, activation='sigmoid'): This is a dense (fully connected) layer with one neuron (output) and a sigmoid activation function. This layer is used to produce the output of the model.

After all the parameters of the Bidirectional LSTM model have been determined, it is then compiled using the Adam Optimizer technique, and focuses on calculating the loss function using the MSE (Mean Squared Error) method and measuring accuracy during the training process.

4.5. Hybrid ARIMAX-Bidirectional LSTM algorithm implementation

This hybrid algorithm combines 2 ARIMAX and Bidirectional LSTM models together. The use of parameters in the model in both algorithms is still used as in the previous implementation of the ARIMAX and Bidirectional algorithms, including the use of Python libraries, namely Pmdarima, Tensor flow and Keras, the same as before.

However, in this hybrid algorithm, the Bidirectional LSTM model is used to evaluate the X and Y variables used in the hybrid model. The results of the Bidirectional model will choose which data column I am from variable "GDP Per Capita", "New Cases", "Total Deaths", and "Total Vaccinations", the data column for variable the new Y variable, as well as the test data, which is the evaluation result of the Bidirectional LSTM model.

The next ARIMAX process still uses the same parameters and libraries as the previous ARIMAX algorithm implementation

4.6. ARIMAX algorithm result

Figure 2 shows the results of implementing the ARIMAX algorithm. Next, the prediction results of the ARIMAX algorithm were compared with the actual values (ground-truth) to be evaluated using the MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) methods (see Fig. 2). The respective results were obtained, namely MAE = 4877.7647, and MSE = 5702.2962.

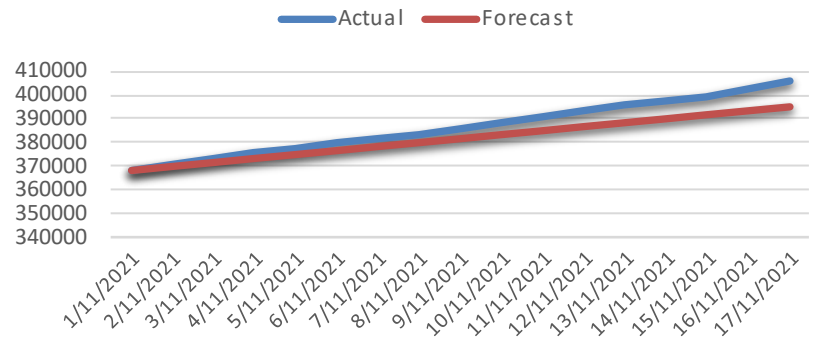

Fig. 2. Prediction results for COVID-19 Cases using the ARIMAX algorithm. A comparison between actual and forecast results.

4.7. Bidirectional LSTM algorithm result

Figure 3 show the results of implementing the Bidirectional LSTM algorithm. Next, the prediction results of the Bidirectional LSTM algorithm were compared with the actual values (ground-truth) to be evaluated using the MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) methods (see Fig. 3).. The respective results were obtained, namely MAE = 10015.5294, and MSE = 11953.1689.

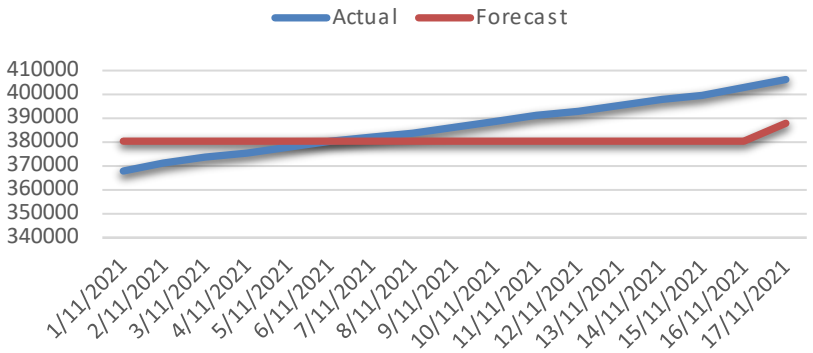

Fig. 3. Prediction results for COVID-19 cases using the bidirectional LSTM algorithm. A comparison between actual and forecast results.

4.8. Hybrid ARIMAX-Bidirectional LSTM algorithm results

Figure 4 show the results of implementing the hybrid ARIMAX-Bidirectional LSTM algorithm. Next, the prediction results of the Bidirectional LSTM algorithm were compared with the actual values (ground-truth) to be evaluated using the MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) methods (see Fig. 4). The respective results were obtained, namely MAE = 2492.0988, and MSE = 2908.1411.

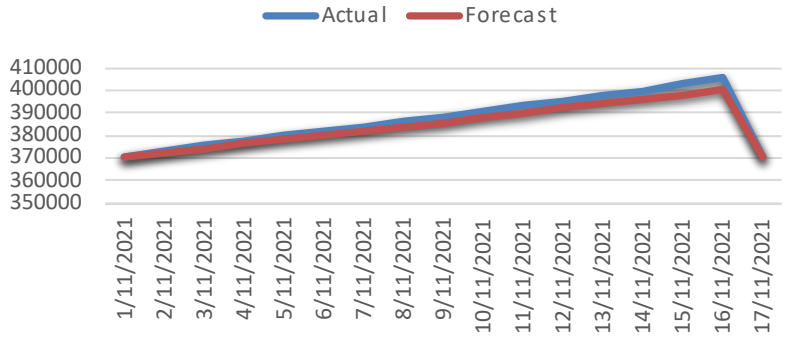

Fig. 4. COVID-19 case prediction results using the hybrid ARIMAX-bidirectional LSTM algorithm. A comparison between actual and forecast results.

From the results of the comparison of the three algorithms above, both ARIMAX, Bidirectional LSTM, and the hybrid ARIMAX-Bidirectional LSTM algorithm, it can be seen that the one with the lowest MAE and MSE values is the hybrid model, and as seen in the graph pattern, the prediction results using the hybrid model have similar patterns. almost similar to the actual value.

5. Conclusion

Two popular time series prediction algorithms are combined to form an optimal model. The combined ARIMAX model and Bidirectional LSTM model are able to provide the lowest MAE and RMSE values compared to the individual models. The application of the feature selection and hyper-parameter tuning stages with the highest R-squared value using the genetic algorithm method and the PSO algorithm contributes to obtaining a prediction model with a small error range. The hybrid model in this research has been successful and can be used to predict COVID-19 case data for the next 3 months (90 days), using datasets for the past 20 months, starting from February 2020 to November 2021.

External factors have a significant influence on the ability of the hybrid model to make predictions that are more appropriate to actual conditions. It can be seen that the hybrid model developed in this research is able to overcome the differences in time-series data in the past which consisted of stationary data and non-stationary data.

The Bidirectional LSTM method can analyse time series and use two directions, namely forward and backward so that it can increase the accuracy of prediction results for COVID-19 cases, however, the weakness of Bidirectional LSTM is that it requires a large amount of input data, and relatively longer computation time than LSTM, and only can be used to predict data over a small number of time units in the future.

The hybrid model was proven to provide much better prediction results than the models run individually. The ARIMAX-Bidirectional LSTM hybrid model has also been proven to be better than the ARIMAX-LSTM hybrid model, as well as other deep learning-based hybrid models. Even though the ARIMAX-Bidirectional LSTM hybrid model can provide very good predictions, its implementation has not been widely carried out, so it is an opportunity great to carry out research on its use and utilization. The COVID-19 pandemic requires a prediction model to help anticipate a spike in cases in the future. The hybrid ARIMAX-Bidirectional LSTM model has great potential to be used as a predictive model for multivariate data, both linear and non-linear in time series involving external factors.

References

1. Said, A.B.; Erradi, A.; Aly, H.A.; and Mohamed, A. (2021). Predicting COVID-19 cases using bidirectional LSTM on multivariate time series. *Environmental Science and Pollution Research*, 28(40), 56043-56052.
2. Shahid, F.; Zameer, A.; and Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos*, *Solitons & Fractals*, 140(1), 110212-110224.
3. Hassantabar, S.; Ahmadi, M.; and Sharifi, A. (2020). Diagnosis and detection of infected tissue of COVID-19 patients based on lung X-ray image using convolutional neural network approaches. *Chaos*, *Solitons & Fractals*, 140(1), 110170-110182.
4. Saba, A.I.; and Elsheikh, A.H. (2020). Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. *Process safety and environmental protection*, 141(1), 1-8.
5. Sahoo, B.K.; and Sapra, B.K. (2020). A data driven epidemic model to analyse the lockdown effect and predict the course of COVID-19 progress in India. *Chaos*, *Solitons & Fractals*, 139(1), 110034-110045.
6. Kuniya, T. (2020). Prediction of the epidemic peak of coronavirus disease in Japan, 2020. *Journal of clinical medicine*, 9(3), 789-790.
7. Dave, E.; Leonardo, A.; Jeanice, M.; and Hanafiah, N. (2021). Forecasting Indonesia exports using a hybrid model ARIMA-LSTM. *Procedia Computer Science*, 179(1), 480-487.
8. Farooqi, A. (2014). ARIMA model building and forecasting on imports and exports of Pakistan. *Pakistan Journal of Statistics and Operation Research*, 10(2), 157-168.
9. Fathi, O. (2019). Time series forecasting using a hybrid ARIMA and LSTM model. *Velvet consulting*, 15(1), 1-7.
10. Zhang, J.; Qu, S.; Zhang, Z.; and Cheng, S. (2022). Improved genetic algorithm optimized LSTM model and its application in short-term traffic flow prediction. *PeerJ Computer Science*, 8(1), 1048-1069.
11. Ohyver, M.; and Pudjihastuti, H. (2018). Arima model for forecasting the price of medium quality rice to anticipate price fluctuations. *Procedia Computer Science*, 135(1), 707-711.
12. Mahawan, A.; Jaiteang, S.; Srijiranon, K.; and Eiamkanitchat, N. (2022). Hybrid ARIMAX and LSTM model to predict rice export price in Thailand.

Proceedings of the 2022 *International Conference on Cybernetics and Innovations* (*ICCI*), Ratchaburi, Thailand, 1-6.

13. Peter, Ď.; and Silvia, P. (2012). ARIMA vs. ARIMAX-which approach is better to analyse and forecast macroeconomic time series. *Proceedings of the* 30th *International Conference Mathematical Methods in Economics*, Karvina, Czech Republic, 136-140.
14. Joseph, R.V.; Mohanty, A.; Tyagi, S.; Mishra, S.; Satapathy, S.K.; and Mohanty, S.N. (2022). A hybrid deep learning framework with CNN and bi-directional LSTM for store item demand forecasting. *Computers and Electrical Engineering*, 103(1), 108358-108367.
15. Faouzi, J.; and Janati, H. (2020). Pyts: A python package for time series classification. *Journal of Machine Learning Research*, 21(46), 1-6.
16. Supriya, K. (2021). A study on the performance of the arimax-ann hybrid forecasting model over the other time series forecasting models arimax and ann in forecasting the rice yield. *International Journal of Current Microbiology and Applied Sciences*, 10(1), 3421-3428.
17. Khashei, M.; Hejazi, S.R.; and Bijari, M. (2008). A new hybrid artificial neural networks and fuzzy regression model for time series forecasting. *Fuzzy sets and systems*, 159(7), 769-786.
18. Schuster, M.; and Paliwal, K.K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), 2673-2681.
19. Handhayani, T.; Lewenusa, I.; Herwindiati, D.E.; and Hendryli, J. (2022). A comparison of LSTM and BiLSTM for forecasting the air pollution index and meteorological conditions in jakarta. *Proceedings of the* 2022 5th *International Seminar on Research of Information Technology and Intelligent Systems* (*ISRITI*), Yogyakarta, Indonesia, 334-339.
20. Deng, Y.; Fan, H.; and Wu, S. (2023). A hybrid ARIMA-LSTM model optimized by BP in the forecast of outpatient visits. *Journal of ambient intelligence and humanized computing*, 14(1), 1-11.
21. Wu, Z. (2021). The comparison of forecasting analysis based on the ARIMA-LSTM hybrid models. *Proceedings of the* 2021 *International Conference on E-Commerce and E-Management* (*ICECEM*), Dalian, China, 185-188.
22. Sunny, M.A.I.; Maswood, M.M.S.; and Alharbi, A.G. (2020). Deep learning-based stock price prediction using LSTM and bi-directional LSTM model. *Proceedings of the* 2020 2*nd novel intelligent and leading emerging sciences conference* (*NILES*), Giza, Egypt, 87-92.
23. Siami-Namini, S.; Tavakoli, N.; and Namin, A.S. (2019). The performance of LSTM and BiLSTM in forecasting time series. *Proceedings of the* 2019 *IEEE International Conference on Big Data* (*Big Data*), Los Angeles, USA, 3285-3292.