Comparison Between Two Systems for Forecasting Covid-19 Infected Cases

Tatiana Makarovskikh, Mostafa Abotaleb

Comparison between two systems for forecasting Covid-19 infected cases*

Tatiana Makarovskikh[1][0000−0002−3656−9632] and Mostafa Abotaleb[2][0000−0002−3442−6865]

[1] Department of System Programming, South Ural State University, Chelyabinsk 454080, Russia Makarovskikh.T.A@susu.ru
[2] Department of System Programming, South Ural State University, Chelyabinsk 454080, Russia abotalebmostafa@bk.ru

Abstract. *Building a system to forecast Covid-19 infected cases is of great importance at the present time, so in this article, we present two systems to forecast cumulative Covid-19 infected cases. The first system (DLM-System) is based on deep learning models, which include both long short-term memory (LSTM), bidirectional long short-term memory (Bi-LSTM), and Gated recurrent unit (GRU). The second system is a (TS-System) based on time series models and neural networks, with a Prioritizer for models and weights for time series models acting as an ensemble between them. We did a comparison between them in order to choose the best system to forecast cumulative Covid-19 infected cases, using the example of 7 countries. As some of them have finished the second wave and others have finished the third wave of infections (Russia, the United States of America, France, Poland, Turkey, Italy, and Spain). The criterion for choosing the best model is MAPE. It is a percentage, not an absolute value. It was concluded that an ensemble method gave the smallest errors compared to the errors of the models in the (TS-System).*

Keywords: Covid-19 · LSTM · BiLSTM · GRU · Deep learning models · Time series models · ARIMA · BATS · TBATS · Holt's Linear trend · NNAR · Forecasting system.

1 Introduction

After more than a year since the spread of the Covid-19 virus in the Chinese city of Wuhan and its spread around the world, a problem appeared in modeling and forecasting cases of Covid-19, so there was a need to develop a system that has the ability to model and predict cases of Covid-19 in addition to obtaining accurate predictions with the least possible error. At the same time, the scientific pandemic of Covid-articles began. An enormous number of articles dedicated to

* The work was supported by Act 211 Government of the Russian Federation, contract No. 02.A03.21.0011. The work was supported by the Ministry of Science and Higher Education of the Russian Federation (government order FENU-2020-0022)

different fields dealing with Covid have appeared. Among them, there are thousands of articles devoted to the forecasting of Covid-19 in different countries and regions (especially in South-East Asia and some other countries mostly affected by pandemics). Most of these articles consider one or two models for the fixed data and for the fixed region and look like a news digest. There are two methods for predicting the spread of a pandemic. The first one is based on mathematical models (machine learning, neural networks, time series). The second one is based on data analysis from social media (e.g., tweets, Facebook). For example, In [9] by using a multivariate time series associated with a geographic region, obtained by quantifying indicators from massive online surveys on COVID symptoms, it is offered. Through the Facebook platform, they show how a neural ODE is able to learn the dynamics that connect these variables and detect virus outbreaks in the region. We show that the neural ODE can predict up to sixty days into the future in a virus-spreading environment by analyzing data from US states.Our work focus will be on the mathematical method for forecasting cumulative daily Covid-19 infection cases.

In our article, we consider two forecasting systems that allow us to choose the best model to analyse the time series appearing as input data. The first system (DLM-system) is based on deep learning models that include LSTM, BiLSTM, and GRU. The second system (TS-system) is based on time series and neural network models, which include models (NNAR, BATS, TBATS, Holt Linear trend, and ARIMA). We also designed a weight-through Prioritizer for models and gave weights where it assigns weights to the time series models and neural network model and gets the Ensembling model and compares its errors with the errors of each model of time series and neural networks in the second system (TS-system), where the Ensembling model very accurate results were given in five countries (Russian federation- France- Poland- Turkey- Italy) out of seven countries (Russia, the United States of America, France, Poland, Turkey, Italy, and Spain). As some of them have finished the second wave and others have finished the third wave of infections. The Prioritizer idea is due to giving a higher weight to the best model and the remaining models give it constant weights. In [1] It is shown that Holt's linear trend model is better than the ARIMA model for China, Italy, and the USA. In [4] used RNN, LSTM, (SARIMA) Seasonal Autoregressive Integrated Moving Average, and Holt winter's exponential smoothing and moving average methods to forecast Covid-19 cases in Iran. Their comparative study on these methods showed that the LSTM model outperformed other models in terms of the least error values for infection development in Iran. In [8] We concluded that it is difficult to obtain a highly accurate forecast without periodically updating the model's parameters.As a result, the development of a system to automatically select the best forecasting model and its best parameters is critical.In [10] Models ranked from good performance to the lowest in all scenarios are Bi-LSTM, LSTM, GRU, SVR, and ARIMA. Bi-LSTM generates the lowest MAE and RMSE values of 0.0070 and 0.0077, respectively, for deaths in China. The best R squared score value is 0.9997 for recovered cases in China. On the

basis of demonstrated robustness and enhanced prediction accuracy, Bi-LSTM can be exploited for pandemic prediction.

Given the similarity in the characteristics of the models in the United States and Italy, it was suggested that in [11] that the corresponding forecasting tools can be applied to other countries facing the Covid-19 pandemic, as well as to any pandemics that may arise in the future. However, a general principle for choosing models for forecasting the spread of Covid-19 has not yet been formulated. Moreover, for different states and different conditions for the spread of the epidemic, it is advisable to build a forecast using different models. For example, in [6] The LSTM model was shown that had consistently the lowest rates of forecast errors for tracking the dynamics of infection cases in the four countries considered. There are also studies that show that the ARIMA model and cubic smoothing spline models had lower forecast errors and narrower forecast intervals compared to Holt's and TBATS models.

As for the SIR model, even at the beginning of the pandemic, it was shown to be ineffective in predicting cases of coronavirus infection. For example, using this model, it was found that the peak of the second wave of infection cases in Pakistan should have occurred on August 25, 2020. However, in fact, the peak of infections in this country was in December 2020 [7]. The "covid19.analytics" package, developed by using the R language for programming, has the same drawbacks. This is evidenced by the results of the SIR model and the prediction of the time of occurrence of the second (and subsequent) wave cycles.Because of these drawbacks to epidemiological models in dealing with Covid-19, there was a need to rely on time series models and deep learning models for their accuracy in detecting the pattern of the spread of Covid-19 and predicting cases of infection. As a result, two systems were created: one that relies on deep learning models and the other that relies on time series and neural network models and combines them and gives weights to time series models through the Prioritizer.

2 The Review of Two System For forecasting Covid-19

The purpose of our work is to create an algorithm that allows for the available initial data on the spread of coronavirus infection in a certain region for a given period of time to determine the best model for making a forecast for a given period. The algorithm analyses forecasts from time series models (TS-system) (ARIMA, Holt's linear model, BATS, and TBATS), and the neural networks model (NNAR) and selects a model that produces a forecast with a minimum mean absolute percentage error (MAPE). The article describes a R program (TS-system) that generates a forecast using the above-mentioned models and combines them with them and weights for each time series model using the Prioritizer.On the other hand, we apply (DLM-system) deep learning model systems and compare both systems' errors to obtain accurate forecasts with the least MAPE errors.

Fig.1 shows global variables for running the second system (TS-system) that is based on time series models and a neural network model.

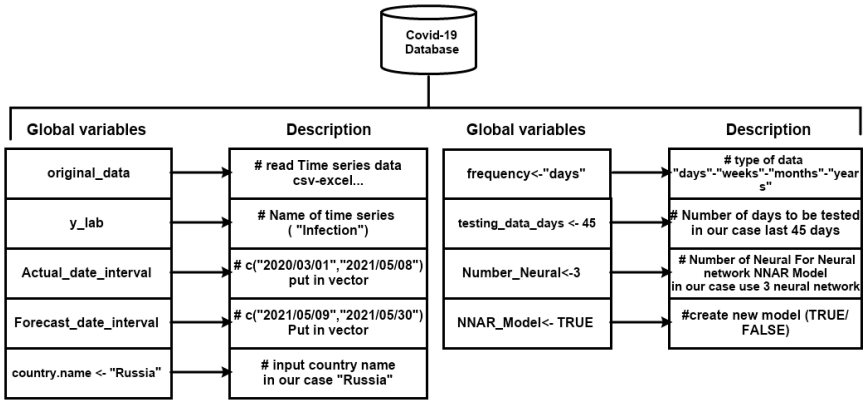

Fig. 1. Global variables for running the second system (TS-system) that is based on time series models and a neural network model.

Fig.2 It shows the scheme of the developed second system (DLM-system), which allows us to choose the best model with the available initial data.

This software module works according to the following algorithm.

Figure3 shows the idea of the Prioritizer for models and giving weights, where after obtaining expectations from the time-series models and neural networks, they are Ensembling and given weights. It was found that giving a weight of 0.9 to the best model of the time series and neural networks, and distributing (1-0.9) equally over the other models gives accurate results and fewer errors.

Algorithm Covid-19 Forecasting

Step 1. Insert time series data, Covid-19, and global variables. (see fig. 2).

Step 2. Split the data into training and testing.

Step 3. Transform time series to be stationary, and supervised. Using the first system (DLM-system) for deep learning models.

Step 4. Run deep learning models by using the first system (DLM-system).

Step 5. Using the second system,(TS-system) we run time series models and neural Network (NNAR) model, and ensembling them between them using a prioritizer for models and giving weights for each model.

Step 6. Calculate the accuracy of the training data (ME-MAE-RMSE-MPE-MAPE-MASE-ACF).

Step 6. Calculate the accuracy of the testing data (MAPE), and obtain the summary tables for forecasting by using each model for each system (DLM-system and TS-system).

Step 7. Select the best model for forecasting with the least error MAPE.

The source code for two systems for forecasting Covid-19 using this algorithm is published on GitHub. [3].

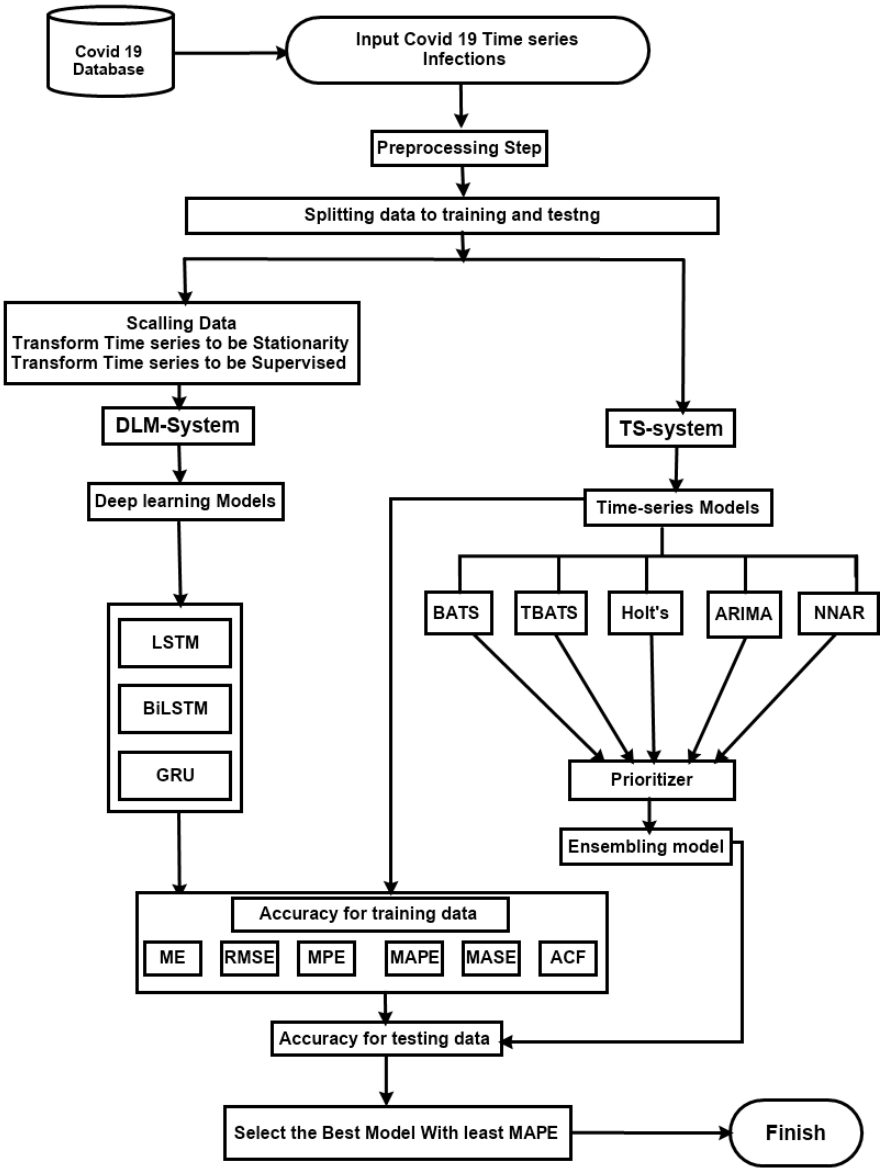

Fig. 2. For forecasting daily cumulative Covid-19 infection cases, the structural scheme consists of two systems (DLM-system and TS-system).

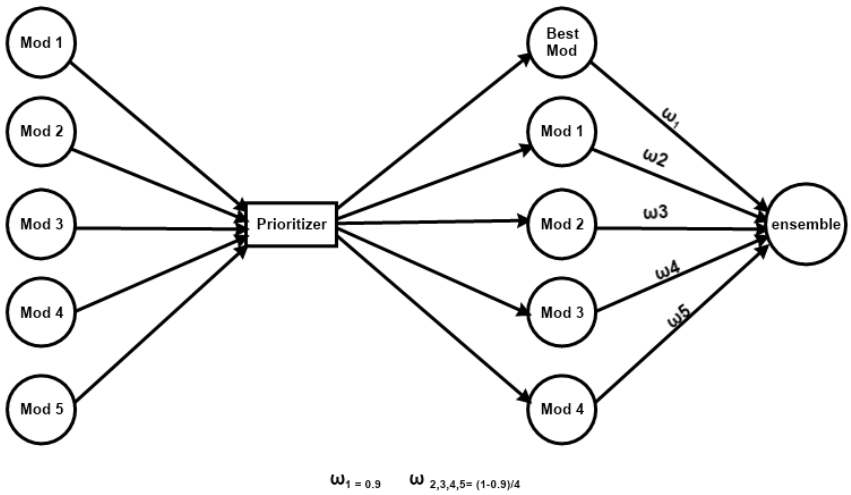

Fig. 3. Ensembling time series models by prioritizing models and assigning weights to time series models for the second system (TS-system).

The TS-system selects the best model from five time-series models forecasting Covid-19 with the least error in the MAPE. Note that the considered system can be used to forecast not only the time series associated with the spread of the epidemic. The study of this system implementation for other time series (for example, to forecast the production volume, the prices of goods, etc.) is a topic for further research.

3 Computational Experiments

Let us consider the results of using the two systems for forecasting cumulative infection cases. [2].

3.1 Covid-19 Datasets

The system uses Covid-19 data from the World Health Organization [5] about Covid-19 infection in the Russian Federation, the United States of America, France, Poland, Turkey, Italy, and Spain for the period from March, 1, 2020 to May, 8, 2021. We use them for our computational experiments by using two systems, the following in [2]:

3.2 Analysing the Obtained Results by using the two different systems

From table 1, it is clear that the best model in the first system is the LSTM model, which achieved the least errors in the testing data for the last 45 days

Table 1. MAPE % for Model selection for forecasting cumulative daily Covid-19 infection cases by using the (DLM-system) Deep learning system.

Country	LSTM	BiLSTM	GRU	Best Model
Russian federation	0.042257	0.094228	3.634369	LSTM
USA	1.037883	1.23026	1.622161	LSTM
France	0.231087	3.588404	2.273541	LSTM
Poland	1.324627	7.303683	6.399338	LSTM
Turkey	3.117724	4.556783	21.827662	LSTM
Italy	0.246362	2.550975	7.110975	LSTM
Spain	0.888655	1.558458	8.811285	LSTM

Table 2. MAPE % for Model selection for forecasting cumulative daily Covid-19 infection cases by using the (TS-system) time series models and the neural Network (NNAR) model.

Country	NNAR	BATS	TBATS	Holt's	ARIMA	ARIMA Model	Ensembling	Best Model
Russia	2.766	0.071	0.079	0.099	0.316	ARIMA(1,2,4)	0.018	Ensembling
USA	3.465	0.627	0.633	0.612	0.795	ARIMA(3,2,2)	0.689	Holt Model
France	9.37	2.291	2.32	1.229	0.909	ARIMA(2,2,3)	0.825	Ensembling
Poland	2.897	3.462	3.43	5.188	4.194	ARIMA(3,2,2)	2.305	Ensembling
Turkey	1.793	9.036	9.094	8.986	9.334	ARIMA(2,2,2)	1.402	Ensembling
Italy	3.265	3.897	2.021	2.37	3.066	ARIMA(2,2,2)	1.969	Ensembling
Spain	5.651	2.007	1.439	0.746	1.355	ARIMA(5,2,0)	0.931	Holt Model

Table 3. MAPE % for System selection for forecasting cumulative daily Covid-19 infection cases.

Country	Least error for the (DLM-system)	Least error for the (TS-system)	Best System
Russian federation	0.042257	0.018	TS-system
USA	1.037883	0.612	TS-system
France	0.2310873	0.825	DLM-system
Poland	1.324627	2.305	DLM-system
Turkey	3.117724	1.402	TS-system
Italy	0.246362	1.969	DLM-system
Spain	0.888655	0.746	TS-system

in all seven countries. We can see that the LSTM model is the best one for forecasting cumulative daily Covid-19 infection cases.

From table 2, it is clear that the best model in the (TS-system) is the Ensembling model, which achieved the least errors in the testing data for the last 45 days in five countries. We can see that the Ensembling model is the best one for forecasting cumulative daily Covid-19 infection cases.

Table 3, shows the subdivision of the considered countries into two groups according to the best forecasting system used (DLM-system and TS-system).

4 Conclusion

So, we compared the (TS-system) with (DLM-system) deep learning models (LSTM-BI-LSTM-GRU) and compared their errors. When comparing (DLM-syatem) models' errors and Ensembling model errors, it was found that Ensembling models yielded fewer errors at the level of 4 countries, so we found that the second system was able to outperform (DLM-system) deep learning models at the level of four countries through Ensembling between them by using a Prioritizer for models and giving weights for time series that were added in the second system. Thus, we conclude that expectations can be obtained. By Ensembling models in the (TS-system), errors can be reduced.

The open task is testing the (TS-system) for epidemic data for different countries and different ways of Covid-19 infections spreading to get the low MAPE forecasting of infection cases and to define the optimal criteria for choosing the best model.

References

1. Abotaleb, M.S.A.: Predicting covid-19 cases using some statistical models: An application to the cases reported in china italy and usa. Academic Journal of Applied Mathematical Sciences **6**(4), 32–40 (2020)
2. Abotaleb M., M.T.: Comparison between two systems for forecasting covid 19 cumulative infected case. `https://rpubs.com/abotalebmostafa/771031` (2021)
3. Abotaleb M., M.T.: Two systems for forecasting covid-19. `https://github.com/abotalebmostafa11/2-systems-for-forecasting-covid-19` (2021)
4. Azarafza, M., Azarafza, M., Tanha, J.: Covid-19 infection forecasting based on deep learning in iran. medRxiv (2020)
5. in Data, O.W.: Daily covid 19 vaccine doses administrated. `https://ourworldindata.org/grapher/daily-covid-19-vaccination-doses` (2021)
6. Gecili, E., Ziady, A., Szczesniak, R.D.: Forecasting covid-19 confirmed cases, deaths and recoveries: revisiting established time series modeling through novel applications for the usa and italy. Plos one **16**(1), e0244173 (2021)
7. Hussain, N., Li, B.: Using r-studio to examine the covid-19 patients in pakistan implementation of sir model on cases. Int. J. Sci. Res. in Multidisciplinary Studies Vol **6**(8) (2020)
8. Makarovskikh, T.A., Abotaleb, M.S.: Automatic selection of arima model parameters to forecast covid-19 infection and death cases. Vestnik Yuzhno-Ural'skogo

Gosudarstvennogo Universiteta. Seriya Vychislitelnaya Matematika i Informatika" **10**(2), 20–37 (2021)
9. Núñez, M., Barreiro, N., Barrio, R., Rackauckas, C.: Forecasting virus outbreaks with social media data via neural ordinary differential equations. medRxiv (2021)
10. Shahid, F., Zameer, A., Muneeb, M.: Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. Chaos, Solitons & Fractals **140**, 110212 (2020)
11. Tian, Y., Luthra, I., Zhang, X.: Forecasting covid-19 cases using machine learning models. MedRxiv (2020)