Pandemic Confirmed Cases Prediction using LSTM

CPE 695 Applied Machine Learning
Final Project Report

Daniel Gural
Stevens Institute of Technology
Hoboken, NJ
dgural@stevens.edu

Chloe Quinto
Stevens Institute of Technology
Hoboken, NJ
cquinto@stevens.edu

Jocelyn Ragukonis
Stevens Institute of Technology
Hoboken, NJ
jragukon@stevens.edu

Sam Alpaugh
Stevens Institute of Technology
Hoboken, NJ
salpaugh@stevens.edu

***Abstract*—Due to the recent novel COVID-19 pandemic, there is a strong motivation to forecast or predict the number of confirmed cases for each geographic location. Forecasts based on statistical and mathematical models are imperative in finding the best interventions to limit the spread of the COVID-19 from human to human. This group studied the accuracy of Long Short Term Memory (LSTM). LSTM constitute the best example of a successful learning of order dependence in sequence prediction problems. The group applied this model to not only COVID-19 but also Ebola, H1N1, MERS, and SARS pandemic and asked can we predict how many cases of each virus will spread within a respective country?**

I. Introduction

A. Background

The rise of computers in the modern era has in turn led to improved methods to record, process, and display that data. In order for that data to have the broadest impact, it must be comprehensive without being overly complicated, as well as able to make predictions about how future data will behave. These predictions for future data have gotten faster and more accurate due to increased processing power that allows for a program to take in a larger range of input data and update models as the data comes in. However, the data can lack dimensionality that provides details such as age range, ethnic background, etc. about patients who have contracted the disease.

The biggest challenge facing the modeling of the spread of infectious diseases is that no disease spreads in the same manner due to variables such as method of transmission (direct or indirect). In the last twenty years, several infectious diseases have spread throughout the globe with a range of symptoms, transmission methods, and fatality rates. These diseases include those such as COVID-19 (2019-present), Ebola (2014-2016), H1N1 (2009), MERS (2012-2019), and SARS (2003).

An ability to make accurate predictions of how a disease will spread through a country is extremely important to legislators and to the general public. Governing bodies will be able to see how a disease will impact their population in the near future unless they take action to slow the spread. Without this model, it is left up to the judgement of an individual person to predict what kind of policies they need to enact to protect the public health.

Certain models will be used to address this data including Occam's Razor. This problem-solving principle uses a preference of simplicity to solve complex problems. It is important for our own model to be as simple as we can make it because of the complexity of the problem that the group is trying to solve. Another aspect of the model that will be focused on is that ability for the model to work unsupervised, which is ideal in order to explore and model raw data. In addition, the group will use RNN and LSTM to make the model learn from data that it has already interpreted.

B. RNN and LSTM

A recurrent neural networks (RNNs), are networks with loops, which allow information to persist over time [1]. An RNN remembers its input, due to an internal memory. The memory helps it remember important things about the input they received. Therefore, RNNs are suited well for machine learning problems that involve sequential data [2]. When it makes a decision, it considers the current input and also what it has learned from the inputs it received previously [2]. The diagram below shows a chunk of an RNN. A looks at some input xt and outputs some value ht. The loop allows information to be passed from one step of a network to the next [1].

A basic RNN has short-term memory. RNNs fail to learn in the presence of time lags greater than 5 to 10 discrete time steps between relevant input events and target signals, LSTM is not affected by this problem [3]. When used in combination with a LSTM, RNNs can have a long-term memory [2]. Long short-term memory networks extend the memory of an RNN, which means LSTMs enable RNNs to remember inputs over a longer period of time [2]. They are able of learning order dependence in sequence prediction problems [3]. A standard RNN, as shown below, the repeating module has a simple structure that contains single layer [1].

While LSTMs also have a similar chainlike structure as shown above, their repeating module has four different layers, as shown below.

An LSTM has three gates: input, forget, and output gate. These gates determine whether or not to let new input in (input gate), delete the information because it isn't important (forget

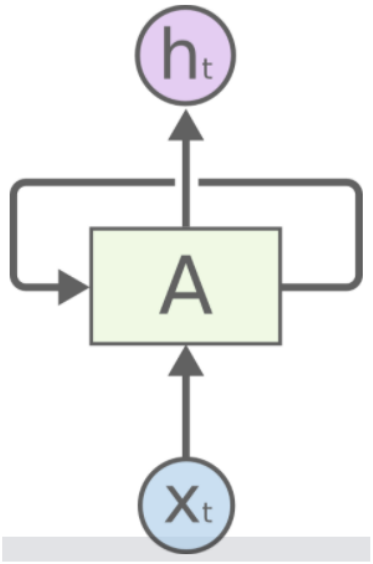

Fig. 1. Recurrent Neural Network

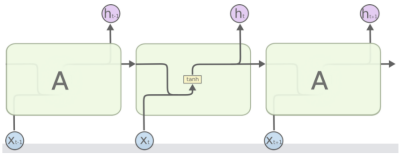

Fig. 2. Repeating Module in a Standard RNN [1]

gate), or let it impact the output at the current timestep (output gate) [2].

The first step of an LSTM, the forget gate, decides what information is going to be thrown away from the cell state. This layer looks at ht-1 and xt and outputs a number between 0, completely get rid of this information, and 1, definitely keep this information, for each number in the cell state Ct-1 [1].

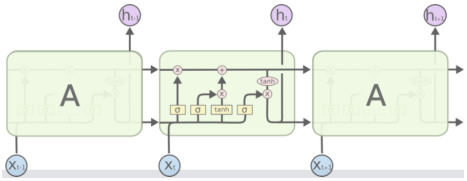

Fig. 3. Repeating Module in an LSTM[1]

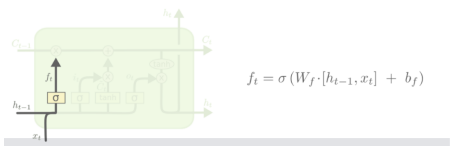

Fig. 4. LSTM Forget Gate Layer [1]

The second step of an LSTM, as shown below, decides which values to update. This layer is comprised of a sigmoid and tanh layer. The sigmoid layer, also known as the input gate layer, decides which value to update. After the sigmoid layer is a tanh layer, which creates a vector of new candidate values that could be added to the state [1].

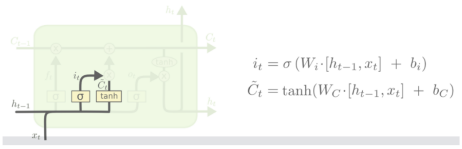

Fig. 5. LSTM Input Layer [1]

The third step of an LSTM updates the old cell into the new cell state. The new candidate values are scaled by how much was decided to update each state value in the previous layers.

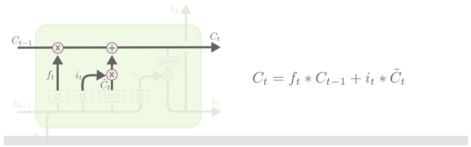

Fig. 6. LSTM Update Cell State[1]

The forth and last step of an LSTM decides what will be output, also known as the output gate. The output is based on the cell state, but is filtered. A sigmoid layer decides what parts of the cell will be output. The cell is then put through a tanh layer, which puts the values between -1 and 1 and then multiplied by the output of the sigmoid gate.

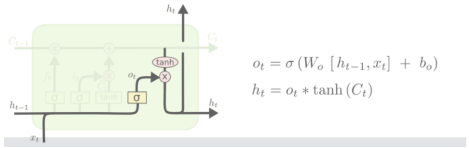

Fig. 7. LSTM Output Gate Layer[1]

RNN contain cycles that feed the network activations from a previous time step as inputs to the network to influence predictions at the current time step. These activations are stored in the internal states of the network which can in principle hold long-term temporal contextual information. This mechanism allows RNNs to exploit a dynamically changing contextual window over the input sequence history. [3]

II. Related Work

There are numerous publications that attempt to predict time series data sets. A popular application is in stock market price predictions. For example, see [6] where deep learning architectures such as RNN and LSTM were used for predicting stock prices in the National Stock Exchange of India and New York. The following [Tab. 1] shows the MAPE incurred during the prediction of Bank of America and Chesapeak Energy NYSE values for the Deep Learning Network

TABLE I
MAPPED INCURRED DURING PREDICTION OF BOA AND CHESAPEAK ENERGY STOCK VALUES

Company	RNN	LSTM	CNN	MLP
Bank of America	5.38	6.01	5.31	4.82
Chesapeak Energy	8.94	8.98	9.18	7.85

Their conclusion found that deep learning models are capable of identifying the patterns existing in both the stock markets. Furthermore, they found that that neural networks, specifically, CNN, outperformed RNN, LSTM, and MLP models for prediction.

During the process of the group's study, very few machine learning models were published as the COVID-19 pandemic was taking effect.

III. SOLUTIONS

A. Data Collection

The data obtained in the project was taken from six sources found within Kaggle. The data was background checked to validate that the number were coming from official sources. For all diseases, the data was obtained originally from the World Health Organization (WHO). Due to the ongoing outbreak, COVID-19 data was taken from multiple sources such as John Hopkins, the CDC, and other national and international agencies. The group decided that the data would only be taken up March 23rd for COVID-19 as it was the furthest recorded data set at the time. While some of the data sets contained many dimensions and columns of data, the group was looking for country, date, and confirmed cases per day. The country would be used to track the three most infected regions.

B. Pre-Processing

Date and confirmed cases would be used to create sequences for the LSTM network to forecast. Since countries infections start on different dates, an arbitrary number of cases confirmed was used as a starting point for each country. This number of cases varied based off the rate of infection for the disease. This report used days after the 10th for Ebola and COVID-19, days after 2nd case for H1N1 and SARS, and weeks after first confirmed case for MERS. After all the pre-processing mentioned above took place, the following graphs were obtained.

It can be observed how there are many differences in the distribution of the cases and infections in each disease. COVID-19 has yet to reach the peak. MERS is not only graphed by week, but is also very seasonal, having large spikes about every year. H1N1 is also described by confirmed cases per day and not cumulative. The final noteworthy detail of the data is the fact that due to the nature of infectious diseases and how we track them, the data is limited by day and by the total number of days the infection lasted. This leaves us with rigid data that can be limited to less than 100 samples.

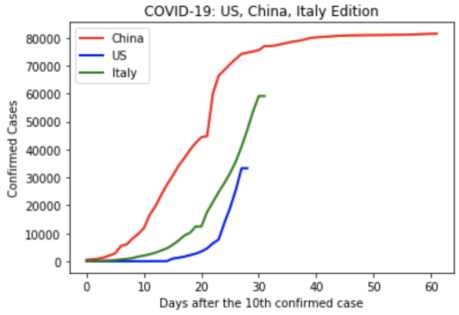

Fig. 8. COVID-19

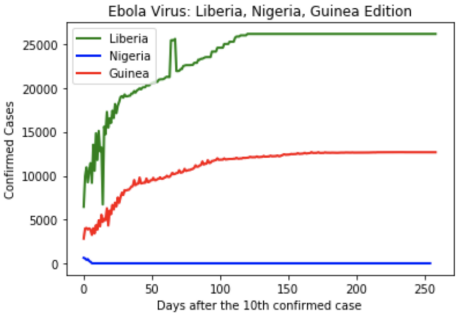

Fig. 9. Ebola

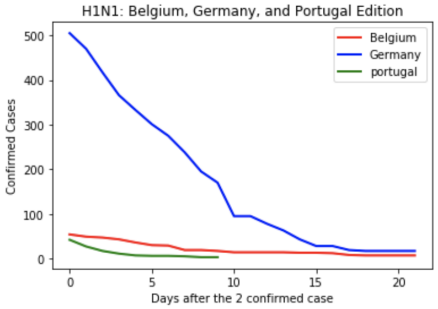

Fig. 10. H1N1

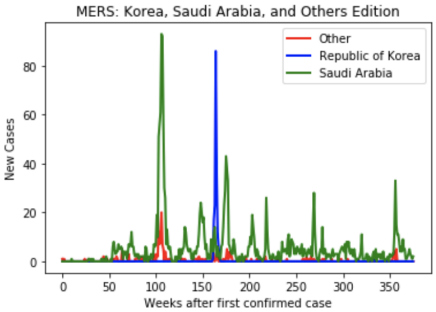

Fig. 11. MERS

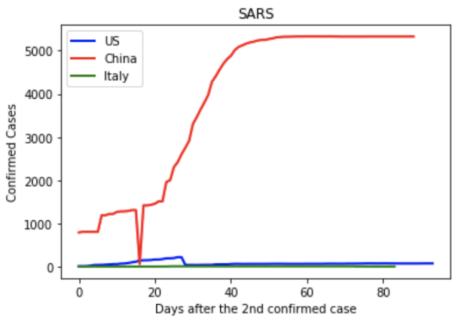

Fig. 12. SARS

C. Choosing a Model

The data set for this study is a univariate time series. The group analyzed which type of LSTM model works best for the data set. There were three types of LSTM models that were analyzed. Namely, Vanilla LSTM, Stacked LSTM, and Bidirectional LSTM. Vanilla LSTM is a model that has a single hidden layer of LSTM units and an output layer used to make a prediction. Stacked LSTM is a model with multiple hidden LSTM layers on top of each other. Finally, Bidirectional LSTM trains two instead of one LSTM - one input from the past to the future and the other from the future to the past.

The loss function of the model was calculated using mean squared error (MSE) as shown in Eq.(1)and root mean squared error (RMSE) Eq. (2)

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}\Big(y_i - \widehat{y_i})^2 \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(y_i - \widehat{y_i})^2} \tag{2}$$

The analysis found that Vanilla LSTM had a lower MSE and RMSE score than Stacked LSTM and Bidirectional LSTM on the COVID-19 data set as seen in [Tab. 2]. This could be due in part the data set for COVID-19 is small. However, all pandemic data sets are relatively small due to the short-lived nature of pandemics. Moving forward, the group will use Vanilla LSTM to predict the confirmed cases for each pandemic.

TABLE II
TIME STEPS

Model	MSE	RMSE
Vanilla LSTM	107970	328.589
Stacked LSTM	1.07524e+07	327.09
Bidirectional LSTM	1.77851e+06	1333.61

D. Training And Evaluation

Vanilla LSTM were trained on each of the pandemic data sets. Time steps were set based on each pandemic as seen in [Tab. 3]

TABLE III
TIME STEPS

Model	Time Steps
COVID-19	3
Ebola	7
H1N1	7
MERS	7
SARS	3

The reasoning behind choosing the time steps for COVID-19 and SARS were 3 days instead of 7 days was due to the exponential growth seen in both data sets.

The data points were transformed by scaling each feature between zero and one.

The model was hypertuned by evaluating the validation curves if it was over fitting or under fitting.

The following figure is the model summary including the layer, shape, and parameters. We define our activation as ReLU and optimizer algorithm as adam and the loss as the mean squared error.

```
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 200)               161600
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 201
=================================================================
Total params: 161,801
Trainable params: 161,801
Non-trainable params: 0
_________________________________________________________________
```

Fig. 13. Model Summary

The following are graphs [Fig. 14-18] that depict the LSTM forecasts for the next x time steps for a given model.

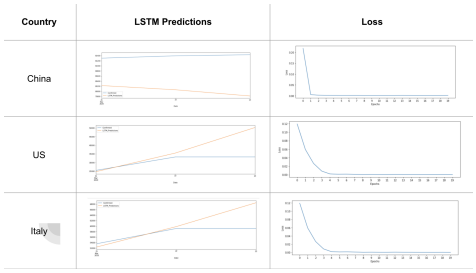

Fig. 14. COVID-19

IV. COMPARISON

The following are the calculated loss functions for each of the pandemic models. The group wanted to focus on 3 main countries for the COVID-19 virus. Namely, China, US, and Italy as they were the hardest hit countries. The data was compiled in order from the days after the first 10th confirmed case. The results show that the Vanilla LSTM model performed well on China's dataset. The model also shows that it did not predict US dataset well. The group believes this is due to the

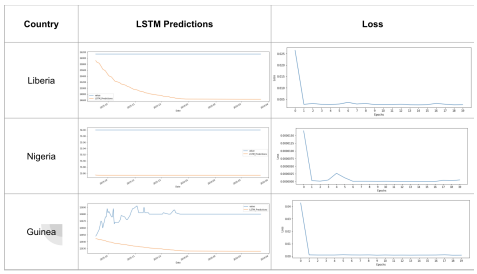

Fig. 15. Ebola

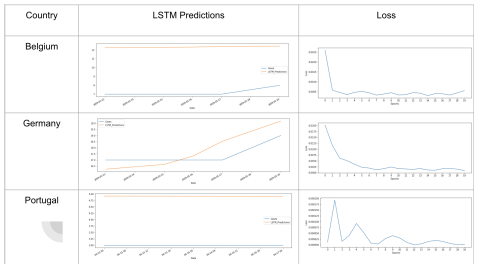

Fig. 16. H1N1

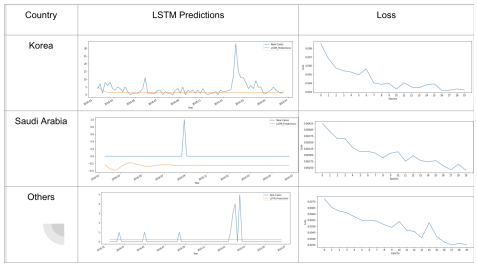

Fig. 17. MERS

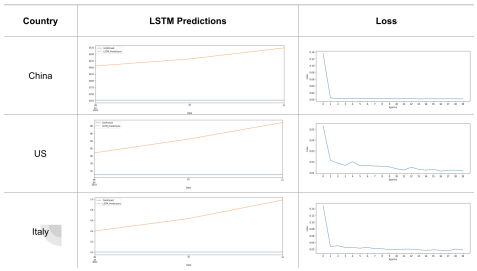

Fig. 18. SARS

TABLE IV
COVID-19

Model	MSE	RMSE
China LSTM Errors	26424.2	162.555
USA LSTM Errors	106.097	10.3004
Italy LSTM Errors	0.512954	0.716208

fact that there were less data on the US in comparison to Italy and China.

The group selected the 3 main countries affected by the Ebola in the dataset. Those were Liberia, Nigeria, and Guinea The data was compiled in order from the days after the first 10th confirmed case. The dataset shows three unique curves of cases. This data set is also useful as it had many cases over a long span of time. The model performed poorly over the data set. One possible reason why this occurred could be that the data set included much of a flat contained cases slope

TABLE V
EBOLA

Model	MSE	RMSE
Liberia LSTM Errors	1.45472e+06	1206.12
Nigeria LSTM Errors	0.00166147	0.0407611
Guinea LSTM Errors	2005.42	44.7819

The group wanted to focus on 3 main countries affected by the H1N1 in the dataset. Those were Belgium, Germany, and Portugal The data was compiled in order from the days after the first 2nd confirmed case. The dataset shows great parity between the countries The results show that the Vanilla LSTM model performed well on Germany's dataset and did well to predict. This is reflected in it's low RMSE.

TABLE VI
H1N1

Model	MSE	RMSE
Belgium LSTM Errors	24.8203	4.982
Nigeria LSTM Errors	1.70265	1.30486
Guinea LSTM Errors	3.70391	1.92455

The group concentrated on the two most affected groups of the MERS virus in the data set and others. These were Republic of Korea, Saudi Arabia, and all other countries The data was compiled in order from the days after the first confirmed case. The overall dataset was small for MERS as it was a well contained pandemic. The results show that the Vanilla LSTM model performed well on all but Saudi Arabia's dataset. Most confirmed cases used in the test set were from when the disease well contained, leading to some misleading results.

TABLE VII
MERS

Model	MSE	RMSE
Korea LSTM Errors	0.0132744	0.115215
Saudi Arabia LSTM Errors	24.5149	4.95126
Other LSTM Errors	0.742677	0.861787

The group wanted to focus on 3 main countries for the SARS virus. Namely, China, US, and Italy as they were the hardest hit countries. The data was compiled in order from the days after the first 2nd confirmed case. The overall dataset was small for SARS as it was a contained pandemic. The

results show that the Vanilla LSTM model performed well on Italy's dataset. This was due to the fact that the confirmed cases narrowed down to 4 cases at the tail end of the pandemic.

TABLE VIII
SARS

Model	MSE	RMSE
China LSTM Errors	26424.2	162.555
Saudi Arabia LSTM Errors	106.097	10.3004
Other LSTM Errors	0.512954	0.716208

V. Future Research Directions

In order to extend this study, there are a few things that could be improved upon. To begin with, the dataset used in this research paper contains data up to March 23rd, 2020. While this encompasses most of the confirmed cases in China, it does not give a full representation of the other countries' confirmed cases. Public datasets such as Johns Hopkins Center for Systems Science and Engineering (JHU CSSE) and the Gobal Health Data from the World Bank can be used for up-to-date cases.

Furthermore, the model can be more robust by adding noise for larger datasets.

VI. Conclusion

The question teh group posed is: Can we predict how many cases of each virus willl spread within a respective country?

For each pandemic, the model was able to, in general, sucessfully predict the next few days of confirmed cases. The MSE and RMSE scores ranged from 0-10000 for Ebola, H1N1, MERS, and SARS. The MSE and RMSE scores of COVID were significantly higher since the dataset for COVID was small.

Overall, the model would have performed better with larger datasets of each pandemic. However, due to the nature of pandemics (short lived and contained), there is not much data to learn from. Possible applications for this model could be for future pandemics, but hopefully that never happens.

References

[1] C. Olah, "Understanding LSTM Networks," 27 August 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[2] N. Donges, "Recurrnet Neural Networks 101: Understanding The Basics of RNNSs and LTSM," Built In, 4 September 2019. [Online]. Available: https://builtin.com/data-science/recurrent-neural-networks-and-lstm. [Accessed May 2020].

[3] J. Brownlee, "A Gentle Introduction to Long Short-Term Memory Network by the Experts," Machine Learning Mastery, 24 May 2017. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/. [Accessed May 2020].

[4] Yang, C., Chen, Y., Chan, Y. et al. Influenza-like illness prediction using a long short-term memory deep learning model with multiple open data sources. J Supercomput (2020). https://doi.org/10.1007/s11227-020-03182-5

[5] Chae, Sangwon et al. "Predicting Infectious Disease Using Deep Learning and Big Data." International journal of environmental research and public health vol. 15,8 1596. 27 Jul. 2018, doi:10.3390/ijerph15081596

[6] M, H., E.A., G., Menon, V. and K.P., S., 2018. NSE Stock Market Prediction Using Deep-Learning Models. Procedia Computer Science, 132, pp.1351-1362.