IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Predicting Air Pollution in Delhi using Long Short-Term Memory Network

Shadab Ahmad Ghazali[1*], Raj Kumar[2]

[1,2] CSE, Rattan Institute of Technology and Management, JCB University of Science & Technology YMCA, Faridabad, India

[*]*Corresponding Author: shadab.ghazali@gmail.com, Tel: +91-9718441850*

DOI: https://doi.org/10.26438/ijcse/v7i5.482486 | Available online at: www.ijcseonline.org

Accepted: 15/May/2019, Published: 31/May/2019

***Abstract*—** Air pollution has become a great cause of concern nowadays. The worst affected areas are urban environments especially large metropolitan cities, like Delhi. It has adverse impact on the physical and mental health of human beings. In this context, predicting air pollution has become an urgent need of the hour. This would help people to take safety measures as well as government to enact policies to safeguard the citizens. Traditionally, climatologists and meteorologists have relied on physical simulations for weather forecasting. With the advancement in artificial neural network predicting the future values based on previously observed values has become quite popular. This paper focuses on Time series analysis to predict air pollution in Delhi using LSTM, an artificial recurrent neural network architecture. We use LSTM because it can work on sequences of arbitrary length. We have taken a data-centric approach to predict air pollution and used historical weather data of Delhi that includes several weather variables – atmospheric pressure, temperature, rain, wind direction and wind speed etc.

***Keywords*—** Air Pollution Prediction, RNN, LSTM, Deep Learning

I. INTRODUCTION

Air is vital to every living thing. It is one of the most crucial segments of our environment. Without air life would become impossible. Air, which we breathe nowadays has been polluted to alarming levels. With rapid industrialization, air has become polluted with harmful particles that are dangerous for human as well as for the environment. Air pollution has thus become a serious environmental problem, especially in the urban areas because of the impact on the physical and mental health. Air pollution is caused by harmful gases, particulates, and biological molecules in the atmosphere. Air pollution causes headache, dizziness, skin diseases, lung diseases to humans. It also causes damage to trees, fruits, vegetables and natural environment. For these reasons Governing bodies of large cities/states have tried several measures to reduce Air pollution as well as educating its citizens to take steps to safeguard themselves. Governments around the globe has formulated policies to tackle this menace. In India, the several large cities rank consistently among the most polluted cities of the world. Delhi is among the worst affected cities. The government has enacted several policies to cope this problem. For example, the odd-even formula where vehicles with registration number ending with odd numbers were only allowed on the road on odd days of the month and vice versa. Awareness among citizens for future air pollution level is very important, especially who suffer from various illnesses caused by it. It also helps in the success of policies framed by the government. If the citizens are well-informed in advance, they can take mitigation efforts and safety measures. Artificial Neural Networks are like biological neural networks, which functions same as the human brain. The human brain is a large interconnected network of neurons. The output of a single neuron may be result of inputs from several other neurons. The network learns by activating certain connections repeatedly. This reinforces the connection between neurons and it is more likely to produce a desired output based upon some specific inputs. Recurrent Neural Networks and its variant Long Short-Term Memory Networks work great on sequence prediction which is essentially predicting the next value for a given input sequence. The Long Short-Term Memory network has promising results of learning long sequences of observations. Since weather data can be represented as a sequence data, LSTMs can predict the future weather. LSTMs are also able to model problems which have multiple input variables. Air pollution is characterized by the concentration of air pollutants over certain area. From meteorological parameters, such as particulate matter concentration, temperature, pressure, wind speed, wind direction and rainfall, it is possible to predict air pollution in advance. Due to this feature LSTMs are great for time series prediction[1]. This paper is organized in the following sections. Section I provide the introduction of this paper. Related work is discussed in Section II. The theoretical details are provided in Section III followed by the proposed approach in Section IV. Experiments and results are provided in Section V. The

conclusion of this paper is provided in Section VI with future work direction.

II. RELATED WORK

Scientific methods for weather forecasting are in place for several decades now. There are many studies focusing on physical deterministic models. However, forecasting air pollution using machine learning techniques have come into picture quite recently with the advent of sophisticated algorithms as well as powerful computer systems to simulate those algorithms. Machine learning models apply statistics, applied mathematics, and probability theory for the data driven approach. Of late, this approach has been widely used to emulate the real-world problems. As a result, there are models based on statistical methods that use neural networks to get higher prediction performance. In [2], authors used big data analysis techniques using distributed approach on weather forecasting data. In [3], efficient monitoring of benzene gas has been done by using the ensemble approach and random forest method. In [4], the authors developed models for forecasting Nitrogen Dioxide at Taj Mahal in Agra using Artificial Neural Networks. Simple machine learning algorithms like k-nearest neighbour, support vector machines and decision tree were used by the authors in air quality index prediction in [5]. In [6], authors have used deep learning methods for prediction of IoT data. The results of the proposed model by the authors are found as promising.

III. THEORETICAL DETAILS

A. LONG-SHORT TERM MEMORY NETWORK

Long Short-Term Memory Network[7] abbreviated as LSTM were first proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997. Currently, it is one of the most popular models used in Deep Learning. Recurrent Networks suffer from what is called the "Vanishing Gradient Problem". During backpropagation the Gradient exponentially decays. The neural networks weight is updated by a value known as gradient. The vanishing gradient problem is when gradient shrinks as it back propagates through time of a gradient value becomes extremely small. If gradient value becomes extremely small, it doesn't contribute to much learning. so, in RNN, layers that get a small gradient update doesn't learn. Because these layers don't learn RNNs can forget what is seen in longer sequences does have short term memory. LSTMs were created as the solution to short-term memory. They have internal mechanisms known as gates which can regulate the flow of information.

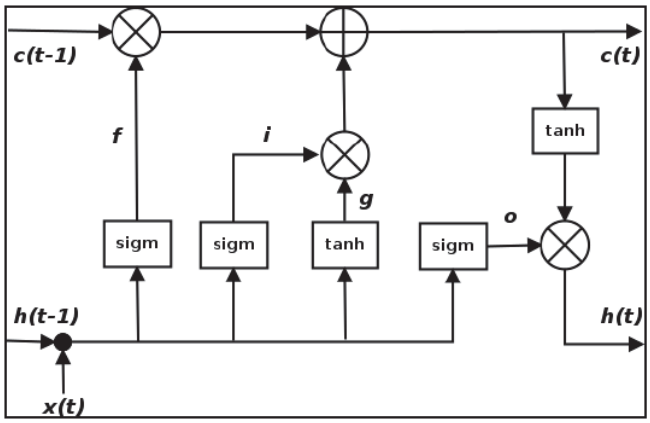

Figure 1. Block Diagram of Long-Short Term Memory Network[8].

There are three different gates that regulate the information flow in an LSTM cell "forget gate", "input gate" and "output gate". The "forget gate", decides what information should be thrown or kept away. Information from a previous hidden State and information from the current input is passed through the sigmoid function. The LSTM model used in this research is given by the following equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (3)$$

$$c_t = f_t c_{t1} + i_t \tanh(W_{xg}x_t + b_g) \quad (4)$$

$$h_t = o_t \tanh(c_t) \quad (5)$$

where σ is the logistic sigmoid function, and i, f, o and c denote the input gate, forget gate, output gate, and cell activation vectors respectively, having same dimensionality as the hidden vector h. The weight matrices from the cell to gate vectors denoted by W_{xi} are diagonal, so element m in each gate vector only receives input from element m of the cell vector. We are going to implement a time series analysis using LSTMs to predict the air pollution in Delhi. Time series analysis is essentially use of historical data to analyze existing data patterns and use them in predicting a future outcome.

IV. PROPOSED APPROACH

A. Data Collection

The air pollutant data was collected from Central Pollution Control Board website (http://cpcb.nic.in/). The air pollutant data contains various parameters such as PM2.5, PM10, NO2, SO2 etc. The meteorological data was downloaded from Weather Underground (https://www.wunderground.com/). It contains various

features such as temperature, pressure, wind direction, speed, humidity, precipitation etc. The data was cleaned as the obtained data contained many missing and junk values. The data was preprocessed to fit the LSTM model. Missing values were imputed by using median strategy.

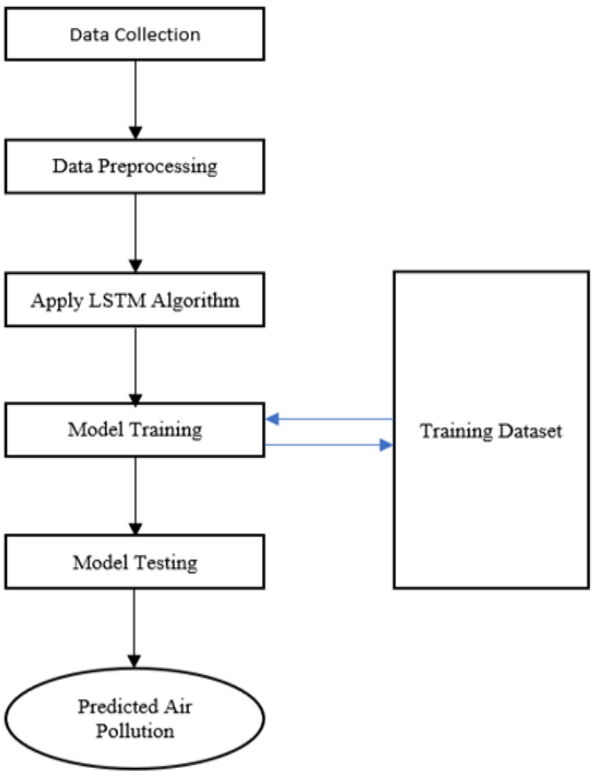

Figure 2. Block Diagram for the Air Prediction Model

B. *Evaluation Metrics*

We have used Root Mean Square Error (RMSE) as the evaluation metric. It is the measure of how good our Long Short-Term Memory model performed. It does so by measuring the difference between predicted values of the air pollution and the actual values of the air pollution. RMSE is the difference between air pollutant values predicted by a model or an estimator and the values observed. Since the model is trained on past data, we report the RMSE for future pollutant prediction value. Following equation represents the equation used to calculate RMSE.

$$RMSE = \sqrt{1/n\sum_{i=1}^{n}(f-o)}\text{^}2 \qquad (6)$$

Where,

n is the test sample size

f = Predicted values

o = observed values.

C. *Ground truth*

The data was split. 70% of data was used for the training and 30% for the test. We used 20% of the training dataset for validation. The LSTSM network is built using Keras and Tensorflow as the backend. For training Adam Optimizer was used with batch size of 32. The pollutant concentration observed at that time is used as the ground truth to measure the RMSE.

V. EXPERIMENTS AND RESULTS

There are multiple parameters (number of epochs, hidden layers, hidden units, learning rate etc.) on which LSTM model work. Tuning all these parameters results in different training time or RMSE. We perform many experiments to find optimal value of parameters to get least RMSE and training/prediction time. The correlation chart shows correlation coefficient between various features. In Graph the coefficient of correlation between T and AQI is negative that means if temperature drops AQI increases and vice versa. Graph 2 shows that average AQI is higher in winter seasons.

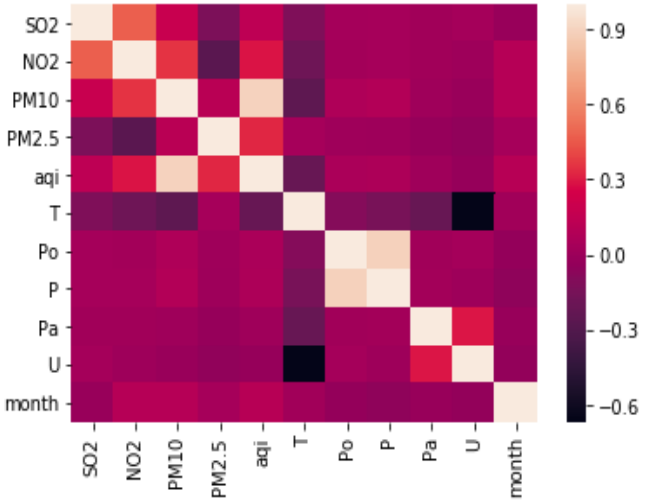

Figure 3. Correlation Chart between Meteorological Parameter and Pollution Parameter

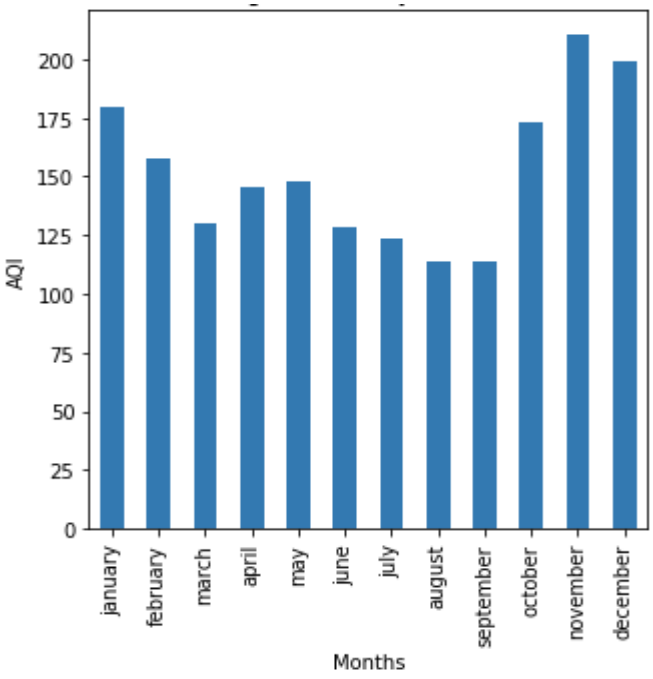

Figure 4. Average Monthly AQI of Delhi

The graph in Figure 4 shows the average monthly Air Quality Index of Delhi. It was observed that AQI was high in the months of October, November, December, and January. This indicates that thee months have higher air pollution concentration than the remaining months.

Training and test results are shown in the following graph.

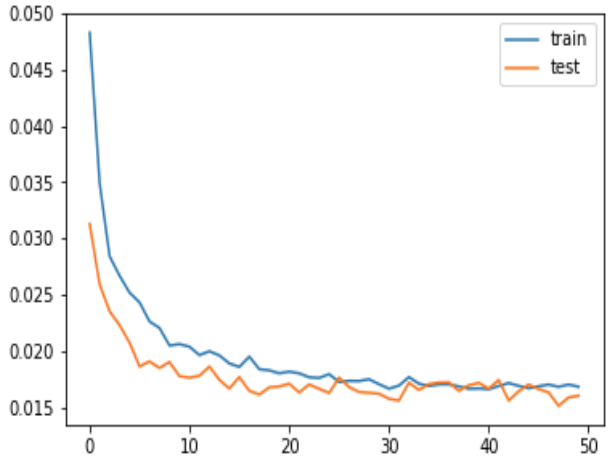

Figure 5. Training and Testing of LSTM Network

We were able to achieve Test RMSE of 28.414 with Variance score of 0.871334 on best run. The RMSE value after multiple experiments is approximately 29, which is quite good. A good machine learning model should have RMSE value lesser than 180. The model was able to predict the future air pollution quite accurately. Following table depicts various results based on different number epochs, and batch size. It is found that the by varying the number of epochs and batch size it results in different RMSE which help is fine-tuning the RMSE.

Table 1. Number of Epochs, Batch Size, RMSE and Variance

No. Of Epochs	Batch Size	RMSE	Variance Score
50	32	28.884	0.867038
100	24	28.414	0.871334
100	32	30.211	0.854540
100	72	29.652	0.859875

The actual predicted air pollution is shown in the following graph against the actual pollution data

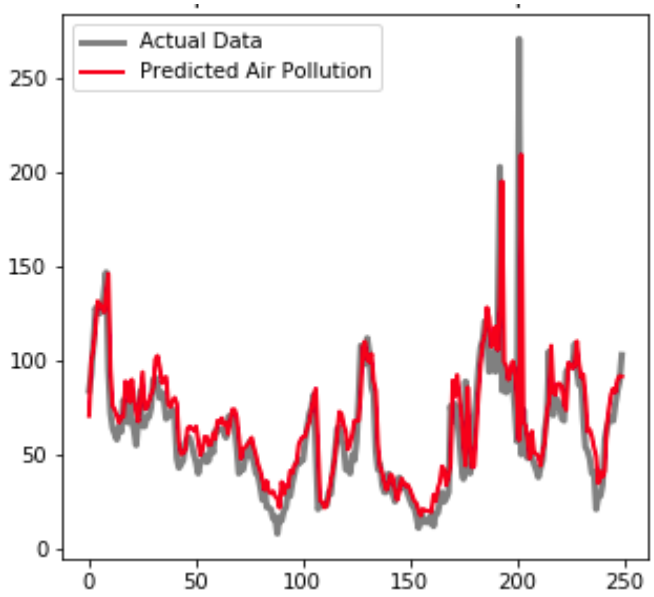

Figure 6. Correct Data VS predicted Data after 50 epochs of training

We measured the accuracy of proposed model based on limited data. It is observed that predicted air pollution levels are quite comparable to the actual air pollution levels on a specific day. However, if more air pollution related data is used the accuracy of the system may increase. From the graphical analysis we observe a trend of Air Pollution in Delhi, that is pollution levels increase in winter season and decrease in monsoon and summer season.

VI. CONCLUSION

Delhi is constantly rated as one of the world's most polluted city. Prediction of air pollution is very crucial for tackling this issue. In this paper we have used Long-Short Term Memory network to predict the air pollution in future. It has been observed that the predicted results were often comparable with the actual data. The experiment shows promising results. However, much work needs to be done to predict the air pollution for larger number of hours, say 24 hours. Predicting air pollution for large number of hours in advance could really help mitigate the issue. The real challenge lies in the weather data. The weather data may not be large enough to predict future air pollution levels. Often these data have missing/junk values which make it more difficult. Further, other machine learning approaches can be combined to get the best results. Predication accuracy can be improved using different algorithmic approaches and ANN models, such as using Bi-directional LSTM[9].

REFERENCES

[1] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, *"Forecasting fine-grained air quality based on big data,"* In the Proceedings of The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **USA** pp**. 2267-2276, 2015**

[2] Amit Palve, Ajit Patil, Amol Potgantwar *"Big Data Analysis Using Distributed Approach on Weather Forecasting Data"*, International Journal of Scientific Research in Network Security and Communication , Vol.**5,** Issue.**3,** pp. **39-43, 2017.**

[3] Gagandeep Kaur, Harmanpreet Kaur *"Ensemble based J48 and random forest based C6H6 air pollution detection"*, International Journal of Scientific Research in Computer Science and Engineering, Vol.**6,** Issue.**2,** pp. **41-50, 2018**

[4] D. Mishra, P. Goyal*," Development of artificial intelligence based NO2 forecasting models at Taj Mahal, Agra",* Atmospheric Pollution Research, Vol.6**,** Issue.**1,** pp**. 99-106, 2015**

[5] Kostandina Veljanovska and Angel Dimoski, *"Air Quality Index Prediction Using Simple Machine Learning Algorithms"*, International Journal of Emerging Trends & Technology in Computer Science, Vol.**7,** Issue.**1,** pp. **25-30,2018**

[6] Ibrahim Kok, Mehmet Ulvi S¸ and Suat Ozdemir. *"A deep learning model for air quality prediction in smart cities"*. International Conference on Big Data, **USA,** pp.**1983–1990. IEEE, 2017**

[7] S. Hochreiter, J. Schmidhuber *"Long Short-Term Memory"*, Neural Computation, Vol.9**,** Issue.**8,** pp.**1735-1780, 1997.**

[8] Antonio Gulli, Sujit Pal, *"Deep Learning with Keras"*, Packt Publishing **UK**, pp. **187-188, 2017**

[9] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Transactions on Signal Processing, Vol.**45,** Issue.**11,** pp.**2673–2681, 1997**

AUTHORS PROFILE

Shadab Ahmad Ghazali completed his Bachelor of Engineering in Computer Science Engineering from Jamia Millia Islamia, New Delhi, India in 2013. He is currently pursuing M. Tech in Computer Science Engineering from Rattan Institute of Technology & Management, JCB University of Science & Technology YMCA, Faridabad. He has worked with HCL Technologies, Noida, UP, India.

His main research work focuses on Machine Learning, Deep Learning, Data Science, and Cloud Computing. He has more than 5 years of Industry Experience.

Raj Kumar pursued B.E (CSE) from Maharshi Dayanand University, Rohtak (HR), India in the year 2007 and M. Tech (CSE) from Maharshi Dayanand University, Rohtak (HR), India in the year 2013. He is currently working as Assistant Professor in Department of Computer Science & Engineering, Rattan Institute of Technology & Management, JCB University of Science & Technology YMCA, Faridabad, Haryana, India since 2013. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. He has 7 years of teaching experience.

