A Project Report

On

**Detecting Air Pollution Levels using LSTMs and other Deep Learning Architectures**

BY

**Anshuman Pati : 2016B4A70470H**

Under the supervision of

**Dr. Manik Gupta**

**SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF**

**CS F366: LABORATORY ORIENTED PROJECT**



**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)**

**HYDERABAD CAMPUS**

**(MARCH 2020**)

**ACKNOWLEDGMENTS**

I would like to express my sincere thanks and deepest gratitude to Dr. Manik Gupta and the institute BITS Pilani Hyderabad Campus for giving me the wonderful opportunity to work on the lab oriented project on **Detecting Air Pollution Levels using LSTMs and other Deep Learning Architectures**.

I would like to thank Dr. Manik Gupta for her constant guidance during the project right from defining the problem statement to exploring different deep learning architectures to solve the problem and making a comparative analysis on these techniques. This has helped me understand what research is about and also helped me do a lot of research, explore and learn many new things in the process. I would also like to thank her for her constant support at each and every stage of the project.

****

**Birla Institute of Technology and Science-Pilani,**

**Hyderabad Campus**

**Certificate**

This is to certify that the project report entitled “**Detecting Air Pollution Levels using LSTMs and other Deep Learning Architectures”** submitted by Mr. ANSHUMAN PATI (ID No. 2016B4A70470H) in partial fulfillment of the requirements of the course CS F366, Lab Project Course, embodies the work done by him under my supervision and guidance.

**Date: 15/03/2020 (Dr. Manik Gupta)**

BITS- Pilani, Hyderabad Campus

**ABSTRACT**

In this study we take a look at making hourly predictions for air pollution levels of Hyderabad with live data collected from several pollution sensors placed across different zones in Hyderabad. Based on the data collected we construct an LSTM model which makes multivariate predictions (PM2.5, dew, temperature, pressure, NO2, PM10, CO, Ozone, NOX, AQI) on an hourly basis so as to predict the level of air pollution in a locality and accurately determine the source of the emissions contributing to the pollution via sensor data. Moreover, the source of emissions can be validated using the satellite data of the emissions in that locality. We classify a given city into different zones based on the signature of the pollutant(s) it emits which is obtained from the satellite data as well as the predictions obtained about Air Quality Index (AQI) from deep learning architectures.

**CONTENTS**

Title page…………..………………………………………………….1

Acknowledgements…………………………………………………...2

Certificate……………………………………………………………..3

Abstract……………………………………………………………….4

Motivation....………………………………………….…………........6

Literature Survey…………..………………………………………….7

Methodology………………………………………………………...12

Future Scope………………....……………………………………...17

Conclusion……..…………………………....………………...……..18

References….……………………………………………….…….....19

**MOTIVATION**

Over the past few decades, air pollution has risen as a global concern adhering to the rapid industrialization of the world, with more people being exposed to the risks associated with this modernisation. Recent advancements in the area of prediction using deep learning algorithms have shown promise in terms of achieving better prediction accuracy for multiple domains. Air pollution has been a major concern in India since last few years. Growing air pollution has led to an increase in health related ailments like stroke, heart disease, lung cancer etc. To monitor air pollution, particularly in major pollution prone cities, the government of India as well as several private companies have installed pollutant measuring sensors. The sensors regularly monitor the level of air pollutants like PM2.5, PM10, CO, NO2, SO2, O3. The data from these sensors can be used to effectively predict the air pollution level of a city, categorize pollution-prone cities and pinpoint the sources of emissions from a particular zone. This, in turn, can effectively reduce the problem of air pollution by identifying the sources of cities with bad Air Quality Index (AQI).

**LITERATURE SURVEY**

Over the past few decades, air pollution has risen as a global concern adhering to the rapid industrialization of the world, with more people being exposed to the risks associated with this modernisation. Recent advancements in the area of prediction using deep learning algorithms have shown promise in terms of achieving better prediction accuracy for multiple domains.

In [1], the authors have proposed models which are robust and have shown superiority over Artificial Neural Network model in predicting PM2.5 severity levels for multiple stations in New Delhi (www.who.int/phe/publications/air-pollution-global-assessment/en/, www.delhi.gov.in) for predicting upto 6 hours, 12 hours and 24 hours ahead of time. The authors have implemented Bi-directional LSTM network to model air pollutant severity for multiple locations in New Delhi, India. Their data incorporates the PM2.5 concentration along with several meteorological parameters like temperature, rainfall, wind speed, etc. on an hourly basis. The authors have made a comparative study on how the implemented model was superior to the ANN implemented framework for the same. However, the authors have predicted *severity* in their model rather than the actual values. This provides us with categorical attributes to identify the severity of air pollution in a locality but does not quantify this impact thereby making a comparative study (with several other models that can be implemented) implausible. The authors have carefully analyzed the advantages of using Bi-LSTMs owing to its structure to use both past and future influences for prediction. They have also explained how Air Quality Index (AQI) cannot be considered as a parameter for input since it is taken at 24 hours average time interval. Hence, the authors have considered PM2.5 and PM10 data for hourly data to estimate. Another drawback of this implementation is that the authors have provided rolling averages for missing values and abnormal values which could have been better filled up with regression data. The authors have stressed on how the multilayer perceptrons existing earlier could just predict the next value but could not influence the short and long term (seasonal or weekly) data. However, this method provides another drawback where the optimal delay for RNN architecture on temporal data is measured by trial and error which is not very efficient. Two separate networks, one in each direction could be obtained and trained on all inputs and averages. The authors have also provided a comparative analysis of BiLSTM and Ensemble (which gives the best results) against multilayer perceptron.

In [2], the authors have used Keras, which is a high-level neural networks API written in Python and capable of running on top of Tensorflow, to build a neural network and run RNN with LSTM through Tensorflow. The training data used in the network is retrieved from the EPA (Environmental Protection Administration) of Taiwan from year 2012 to 2016 and is combined into 20-dimensions data; and the forecasting test data is the year 2017. The authors have conducted experiments to evaluate the forecasting value of PM2.5 concentration for the next four hours at 66 stations around Taiwan. The result shows that the proposed approach can effectively forecast the value of PM2.5. They authors have discarded any station which had lack of enough data for the model to act upon. The authors have filled in the missing values with rolling averages and used min-max normalization to properly scale the data. The missing values could have been better filled with regression outputs. The authors have compared the accuracy of their model to a model implementing Artificial Neural Network on the basis of RMSE values. One of the major drawbacks of this model is that LSTM can be used to forecast PM2.5 properly if the effect of climate or air pollution from abroad is not taken into account. More often than not this is not the case. Moreover, not enough data is present on the air pollutants emitted by transportation and industry. In the prediction of PM2.5 values in the more distant future through different atmospheric altitude airflow data, the fugitive or cumulative area of air pollutants can be learned more accurately. In the prediction of more distant PM2.5 values through different atmospheric height airflow data, the flow or cumulative characteristics of air pollutants can be learned more accurately. And it can provide accurate PM2.5 prediction data for a longer time.

In [3], the authors used LSTM model for multivariate time series forecasting in Keras and TensorFlow deep learning library in a Python SciPy environment with Machine Learning scikit-learn, Pandas, NumPy and Matplotlib libraries. The Raw pollution dataset of this work consists of many columns of CSV for many environmental factors concentrations merged to one CSV file with python merge script and minimized to pollution concentrations of each hour for the following: SO2, NO\_2, NO, CO with the datetime. Later the dataset was minimized for two years (2015 and 2016) for best fitting the expected model in order to get the best forecasting for Madrid’s Air Quality. This paper framed the dataset to forecast the pollution concentrations at the next hour depending on the pollution concentrations of prior hours. The difficulty of predicting multivariate time series and the development of intelligent systems that can design predictive complex environmental dataset automatically requires a good Machine Learning model. To tackle such problems, the authors have proposed an approach based on Neural networks like Long Short-Term Memory (LSTM) in keras and TensorFlow environment. The proposed approach can process multivariate environmental time series data and is robust to the number of samples, numeric ranges of data etc. Running the model many times with increasing the number of memory cells in a hidden layer validates the effectiveness of the approach in accomplishing multivariate time series forecasting. This model can be expanded to predict new environmental observations and also recommend varying the time step index of the sequence output and training epochs to see if there is a relationship between the index and how hard the problem is to learn. Records involving missing data were dropped during the preprocessing stage. This can be detrimental if the dataset is a sparse dataset.

In [4], the authors have proposed a Bayesian LSTM machine-learning model that utilizes proxy data including Aerosol Optical Depth (AOD) and meteorology which can explain 80% of the PM2.5 variability. The preliminary results show that air pollution control regulatory measures introduced in China and Beijing have reduced PM2.5 pollution in Beijing by 23% on average. The dataset comprises official hourly station-level PM2.5 concentration data from 1 January 2013 to 31 December 2017. In addition, the data also constitutes hourly PM2.5 concentration data recorded at the US Embassy, Beijing from 9 April 2008 to 30 June 2017. AOD data was collected from the NASA MODIS satellite database from 26 March 2008 to 21 May 2017. Eight features were selected based on data availability during the period of study. In addition, meteorology data, including temperature, pressure, humidity, visibility, precipitation, and wind speed, measured at the Beijing Capital International Airport, from 1 January 2008 to 31 December 2017, were collected. The pre-processed data is fed into a Bayesian deep learning model for training. In this study, the authors have focused on the Bayesian RNN, which is a particular type of Bayesian deep learning model capable of modelling time-series data. The proposed model was developed based on DeepMind Sonnet, a TensorFlow-based neural network library, and multiple linear regression, which is commonly used in econometric modelling, was selected as the baseline model. The experimental results were in coherence with the theory the authors have proposed. The R2 value of the final fitted model was 80% when using the mean of the posterior over the network weight parameters for prediction; while the R2 for the baseline model was 70% only. However, the proposed model has some drawbacks. Causal inference from observed data can be biased due to the lack of counterfactual outcomes. Moreover, there exists some selection bias with or without the regulatory interventions. The authors’ work only highlights the aggregate effect of APCR (Air Pollution Control Regulations) but not the individual effects. Additionally, limited data is another drawback.

**Bibliography**

[1] Ishan Verma, Rahul Ahuja, Hardik Meisheri, Lipika Dey, “Air pollutant severity prediction using Bi-directional LSTM Network”, *IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2018.*

[2] Yi-Ting Tsai, Yu-Ren,Zeng, Yue-Shan Chang, “Air pollution forecasting using RNN with LSTM”,  *IEEE 16th Int. Conf. on Dependable, Autonomic & Secure Comp., 16th Int. Conf. on Pervasive Intelligence & Comp., 4th Int. Conf. on Big Data Intelligence & Comp., and 3rd Cyber Sci. & Tech. Cong., 2018.*

[3] Shaheen Alhirmizy, Banaz Qader, “Multivariate Time Series Forecasting with LSTM for Madrid, Spain pollution”, *IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2018.*

[4] Yang Han, Jacqueline C.K. Lam, Victor O.K. Li, “A Bayesian LSTM Model to Evaluate the Effects of Air Pollution Control Regulations in China”, *IEEE International Conference on Big Data (Big Data), 2018.*

[5] Wei Song, Jinkun Han, Junwei Xie, Yanning Gao, Liangliang Song, “System for Detecting and Forecasting PM2.5 Concentration Levels Using Long Short-Term Memory and LoRa”, *International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2019.*

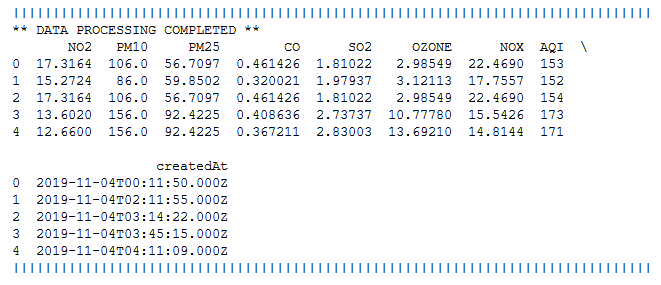
[6] Pratyush Singh, Lakshmi Narasimhan T, Chandra Shekar Lakshminarayanan, “DeepAir: Air Quality Prediction using Deep Neural Network”, *IEEE Region 10 Conference (TENCON 2019), 2019.*

**METHODOLOGY**

#### **Tools Used:**

TensorFlow (Backend), Keras, Pandas, NumPy, Scikit-learn Libraries on Jupyter Notebook

#### **Steps followed:**

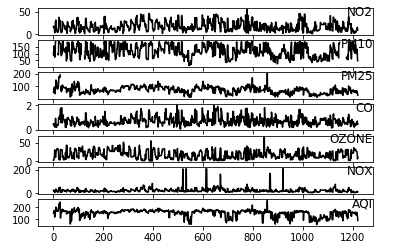
1. Data preparation
2. Data visualization  
   
3. LSTM data preparation
4. Fit model along with regularization term
5. Evaluate model (From RMSE)

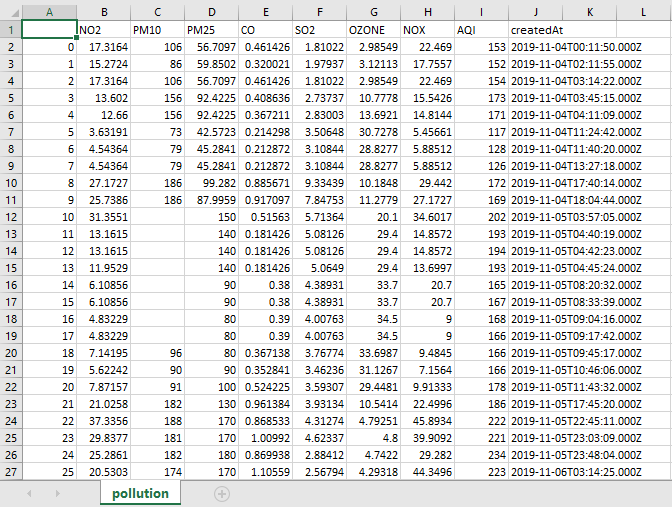
**Data preparation:**

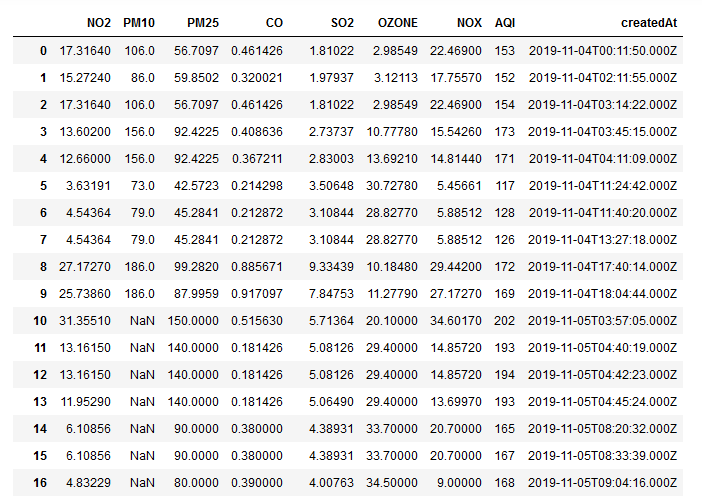
* Replace NA values
* Parse date-time into pandas dataframe index
* Specified clear names for each columns

**Data visualization**

* Used matplotlib to plot

*****Corresponding to parameters obtained from BITS Pilani Hyderabad Campus*

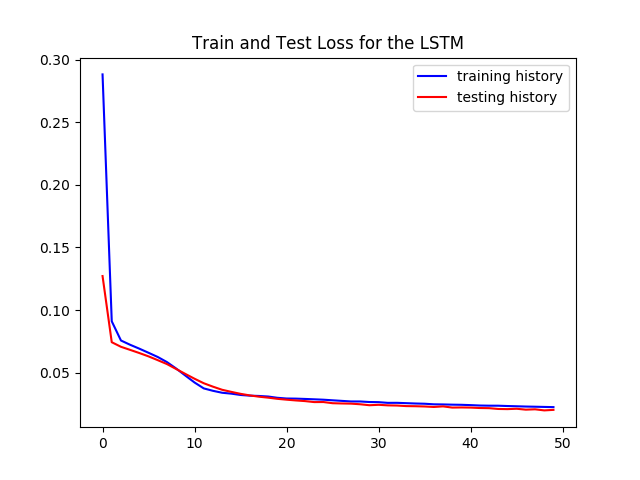
**LSTM data preparation***Stored on pollution.csv file  
*

* Normalized data
* Transformed dataset into supervised learning problem  
  

**Model Fitting**

* Split data into train and test
* Split into i/p and o/p
* Reshape into 3D
* Define a 50 neuron followed by 1 neuron LSTM
* Add dropout at 30%
* Plot history of training and testing loss



****

**Evaluate model**

* Make prediction
* Invert scalings
* Calculate RMSE

**FUTURE SCOPE**

The model can make accurate hourly predictions on the data for PIN code 500078 from the data obtained by the sensors around BITS Pilani Hyderabad Campus. With this verification, this model can be expanded to analyze and predict the air pollution levels of the entire city of Hyderabad.

Moreover, the accuracy can be significantly improved with special modifications to the LSTM architecture. For example, in the paper on *Air pollutant severity prediction using Bi-directional LSTM Network* by Ishan Verma, Rahul Ahuja, Hardik Meisheri, Lipika Dey; the authors provide an introduction to the use of Bi-directional LSTMs to improve the accuracy further owing to its feature to use both past and future influences to make accurate predictions. Further, a comparative analysis can be established between the GRU (Gated Recurrent Unit) which uses two gates in its architecture and the LSTM which uses three gates in its architecture.

Additionally, this model will be aided with the satellite aided clustering and classification of polluted zones to identify the sources of pollution. The satellite data analysis gives faster and clearer visualizations of a zone but the analysis is restricted by the coarser temporal resolution. Here, our deep learning model can provide the required finer temporal resolution. Both the models can be integrated to provide a tool to determine the air pollution level of a given zone and pinpoint the source(s) of pollutants quickly and accurately.

**CONCLUSION**

The comparative analysis with vivid visualizations will provide a clearer picture as to which model gives us the best accuracy. The ensembled model, along with the satellite data analysis can provide a tool to quickly as well as accurately determine the pollution threat level and the causes of pollution of various metropolitan cities. Such a tool can assist various municipal corporations governing these metropolitan cities to counter the threat posed by the growing air pollution levels across various cosmopolitan cities.

**CONCLUSION**

[1] Ishan Verma, Rahul Ahuja, Hardik Meisheri, Lipika Dey, “Air pollutant severity prediction using Bi-directional LSTM Network”, *IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2018.*

[2] Yi-Ting Tsai, Yu-Ren,Zeng, Yue-Shan Chang, “Air pollution forecasting using RNN with LSTM”,  *IEEE 16th Int. Conf. on Dependable, Autonomic & Secure Comp., 16th Int. Conf. on Pervasive Intelligence & Comp., 4th Int. Conf. on Big Data Intelligence & Comp., and 3rd Cyber Sci. & Tech. Cong., 2018.*

[3] Shaheen Alhirmizy, Banaz Qader, “Multivariate Time Series Forecasting with LSTM for Madrid, Spain pollution”, *IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2018.*

[4] Yang Han, Jacqueline C.K. Lam, Victor O.K. Li, “A Bayesian LSTM Model to Evaluate the Effects of Air Pollution Control Regulations in China”, *IEEE International Conference on Big Data (Big Data), 2018.*

[5] Wei Song, Jinkun Han, Junwei Xie, Yanning Gao, Liangliang Song, “System for Detecting and Forecasting PM2.5 Concentration Levels Using Long Short-Term Memory and LoRa”, *International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2019.*

[6] Pratyush Singh, Lakshmi Narasimhan T, Chandra Shekar Lakshminarayanan, “DeepAir: Air Quality Prediction using Deep Neural Network”, *IEEE Region 10 Conference (TENCON 2019), 2019.*