Air Pollution Forecasting using Multivariate Time Series and LSTM

Sushmashree S[1], Dr. S Satish Kumar[2]

[1]*Computer Science and Engineering, RNS Institute of Technology, Channasandra, Bangalore, India*

[2]*Associate Professor, Computer Science and Engineering, RNS Institute of Technology, Bangalore.*

Abstract-Day by Day Urbanization growth is increasing which is due to deforestation .Initially forest is called as lungs of a living beings but because of deforestation, now level of oxygen is decreasing in addition to that other poisonous gases and air pollutants increasing and directly affecting human health as well as earth atmosphere. To make aware of air pollution levels to people, it is necessary to study how air pollution spreading across the world, how much air pollution contents exist in an earth atmosphere. These studies help research fellows and industrialist as well as health organizations to keep track of air pollution to avoid its major effects on people. Mainly Air quality data is utilized for prediction purpose using Deep Learning and Machine Learning approaches which is needed for to make decisions and analysis of suitable data. A main aim is to provide users a effective prediction of air quality data on real time by Multivariate Time Series and Long Short Term Memory which allows timely prediction of Air quality using Deep Learning and Machine Learning.

Keywords: Deep Learning (DL), Machine Learning (ML), Multivariate Time Series, Long Short Term Memory (LSTM).

I. INTRODUCTION

Now a day's man upgrading his knowledge by increasing standard of Technologies in the field of Automobile industry, Factories and so on. As much Human Society becoming civilized some where they are becoming selfish too. Due to deforestation oxygen percentage in air reducing and air become polluted. Air pollution occurs usually in two ways either by humans or by nature. Aerosols are small liquid droplets or solid particles, because of increased levels of aerosols in an environment affects living beings lifecycles. Weather factors such as wind direction, wind speed, pressure, temperature, dew point, cumulative number of hours of rain and snow fall. Weather factor have impacts on an air pollution.. A main reason for Air pollution busy movements of vehicles, Acid rain, and work environments is far from residential places, Particulate contamination, Greenhouse effects, excessive Ultra violet radiations. A Particulate matters is an aerosol particles which are present in an atmosphere. Particulate matters are combination of liquid droplets and solid particles which are in suspended state in atmosphere. Particulate matters are indicated as PM2.5 and its unit is micrograms per cubic meter (µg/m3) .Particulate matter causes irritation in lungs and nose, lung cancer and heart problems.

Air pollution causes heart disease, lung cancer, bronchitis, respiratory problem and also affecting major body parts such as lungs, heart. Even for short time inhale of pollutants can affect the functionalities of a Lungs and Cardiovascular systems. To protect people from suffering due to air pollution it is very important to make people aware about air quality, air pollution and it is necessary to use technology in right manner for the betterment of living beings. It is very much required to carry out research on air pollution for analyzing and predicting air pollution to make suitable decision to take action against air pollution.

In Technology and Science an Air pollution forecasting is considered as an Application, its main job is for a given time and place it predicts Air pollution composition and provides an index of an air quality. Machine learning allows computer program to learn through an experience which is gained after performing a series of tasks and its performance also measured than enhances its experience. Machine learning is mainly used for air pollution forecasting that predicts timely concentrations of a pollutants which are present in an air. By using Large-scale optimization algorithms, machine learning can effectively trains a model in big data.

We have two types of Time series forecasting methods they are univariate time series and multivariate time series. Problem with respect to univariate time series contains two variables, one is fields of forecasting and another one is date-time. Multivariate time series considers other scenarios in addition to date-time variable, it holds multiple variables compared to univariate time series.

For time series forecasting deep learning techniques can be utilized. Whenever time series forecasting taken into account it is important to model problems with respect to multiple input variables, for this Neural networks an updated version of Recurrent neural networks(RNN) Long short term memory (LSTM) plays major role ,and it overcomes the problem associated with Recurrent neural network with limited amount of memory. LSTM model is developed for air pollution forecasting with respect to Multivariate time series. For air pollution forecast a deep learning library called Keras is used.

II. LITERATURE SURVEY

Now a days in Developed and Developing countries like America, Russia, India, China, and Japan these countries governments considering examination with respect to regulation on air as a primary job. Air pollution mainly due to nature and by humans. Causes of Air pollution is because of busy traffics, excessive use of vehicles and burning of Plastics and so on. It is necessary to keep continuous track on air quality and its level which makes essential role for reducing air pollution. A statistical model called Regression always finds relation between variables. To check data sample is either polluted or not Logistic Regression is to be done [1]. Based on old values to predict future value of Particulate matter Auto Regression is used. Because of old existing records regarding Particulate matter allows peoples to reduce its hazardous level. A dataset holds atmospheric details such as temperature, dew point, pollution, pressure, wind speed and its direction cumulative hours of rain and snow fall.

Prediction is a process of inferring what will be happening based on analysis of past records. During air pollution forecasting prediction plays major role. Feature analysis and selection of feature is an advantages for prediction process, it only includes useful information's. A mathematical technique called Interpolation estimates value for an unknown function g(x) for argument x and sometime interval [a ,b].While computing urban air a finely grained air quality- Prediction, feature selection, Interpolation are considered major points.

All three points which are mentioned which are solved individually via various existing works through many model. One model among them is Deep Air Learning its main idea present in embedding semi-supervised learning and feature selection [2]. Beijing, China city real data source is taken into consideration foe extensive experiment. A literature shows the details such as when solving problems associated with interpolation, prediction and feature selection and analysis experiments indicates that Deep air learning is higher to the peer models. Especially in the domain of Quality of an air like prediction of weather it is also possible to define a model, it can be utilized for predicting future set of pollution with respect to atmosphere [3]. There are so many models are present, those models are referred as Atmospheric Dispersion Modelling. An air pollution forecasting is possible through Recurrent neural network and LSTM through an open source software library called Tensor Flow [4].

III. DETAILED DESIGN

Input Data

The dataset are collected from the k aggle website data service provider in order to train the data and for predicting the target value. Actually for air pollution forecasting a five years data collected in a dataset that mainly holds air pollution level and weather condition of an atmosphere of a Beijing city. Initially input dataset contains feature list such as row number ,year, month, day, hour, PM2.5, dewpoint, temperature, pressure, wind direction and wind speed, cumulative hours of snow and rain fall .Initially all these dataset are considered as a raw dataset which are used for further work.

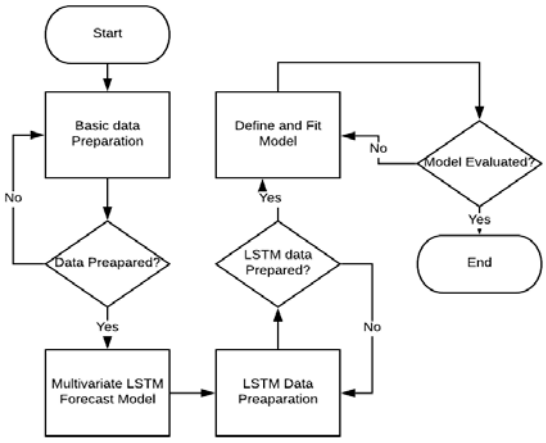

Figure 1 – Dataflow diagram

Data Preprocessing

Sometimes dataset contains incorrect or incomplete information which leads problem especially while generating training models; they may lead to poor quality of prediction which affects entire system performance. Data pre-processing ensures quality of data and also maintains flawless dataset and it can be performed before data cleaning takes place. There are two types of Data pre-processing they are automated pre-processing and manual pre-processing.

Basic Data Preparation

Initially datasets are not in a state to use for further work. Before starting any work dataset must be prepared first A very first step is to consolidate individual date time information into a single date-time, so in pandas these date-time are used as an index. For first 24 hours some fields of data filled as "NA" value .these "NA" values are marked as "0 "values. A python script is used for loading dataset that are raw and they are parsed into a suitable format. A column with no value is dropped.

Multivariate LSTM Forecast Model and LSTM Data Preparation.

Firstly it is necessary to prepare air pollution dataset for LSTM. In order to perform this it is important to frame a dataset into supervised learning problem and normalization of a input variables. In order to transform dataset into

supervised learning problem a series_to_supervised () function can be used. A pollution.csv dataset is loaded and feature with respect to wind speed is integer encoded next normalization taken place for all the features, and finally transformation of a dataset in to supervised learning problem is to done. Normalization is usually done to rescale the attributes according to required range.

Evaluation of a Model.

When model becomes fit for entire test dataset forecast is to be done. Further forecast is to be combined with the test dataset and invert scaling also performed for a test dataset with pollution numbers. Calculation of Root Mean Square Error also performed for a model. To find the difference among predicted and observed values RMSE is used. When RMSE value is small there will be more accuracy in prediction system.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(g_j - x_j)^2}$$

IV. WORK FLOW

For air pollution forecasting it is important to know about LSTM networks and Multivariate time series which plays an important role for performing work with respect to forecasting of air pollution.

Long short term memory (LSTM)

LSTM as an updated version of recurrent neural network. It is Deep Learning model. LSTM contains memory so it can hold information for long period of time. It contains three gates such as Input gate, Forgot gate and output gate. An input gate allows input data through input gate, forgot gate acts as a limiting factor, it deletes an unwanted data through forgot gate. Output data is travelled out through output gate.

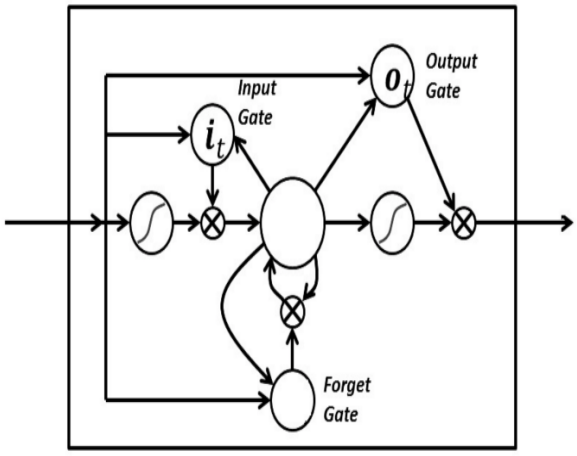

Figure 2 – Structure of a LSTM

An LSTM contains memory unit, these units are capable to take decisions based on current input and previous memory, previous output. LSTM alters its memory only when new output is generated. It can seamlessly model problems which are associated with the multiple input variables, it is an added advantage for time series forecasting.

Multivariate forecasting methods

Multivariate time series mainly allows multiple observations for each time stamp. There are two types of models that use data with respect to multiple time series. They are Multiple Input series and Multiple Parallel series. In multiple Input series problems contains two or more input time series, but always output time series is depended on input time series. In Multiple Parallel time series a value is to be predicted for each. It is very important in economics.

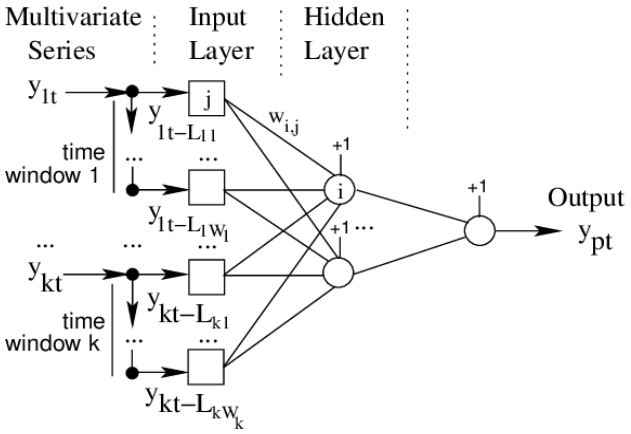

Figure 3 – Architecture of Multivariate time series

V. IMPLEMENTATION ANDRESULTS

During data preparation a consolidation of a date time information into single date-time, so it can be used as an Index in pandas .Column with value no is dropped and “NA“ value replaced with “0”.

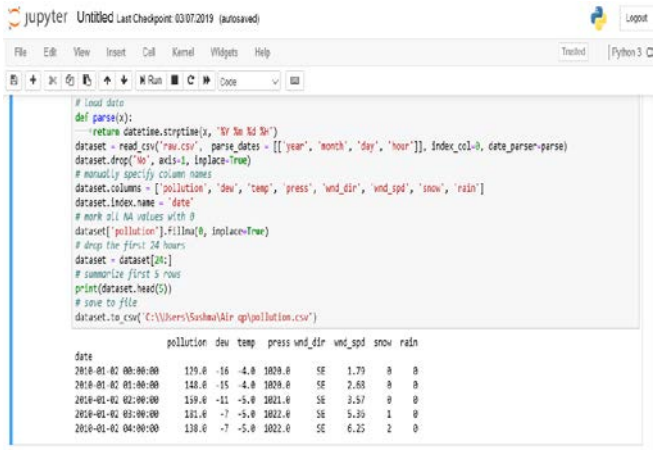

Figure 4 – Snippet 1

Snippet 1 shows the transformed dataset.

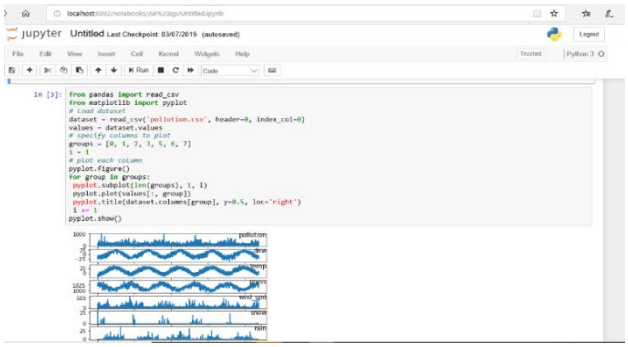

Figure 5 – Snippet 2

A transformed dataset are stored inpollution.csv which is in the form of ready to use. Snippet 2 shows plot diagram of seven feature variables. Each variable has its own time series. A 7 subplot indicates 5years of data for each feature variable. A first step is to prepare the pollution dataset for the LSTM, mainly during LSTM Data Preparation. Further framing of dataset into a supervised learning problem and normalization of an input variables are carried out. Snippet 3 shows first 5 rows of data that are transformed and it shows one output variable that indicates current hour pollution level and 8 input variables.

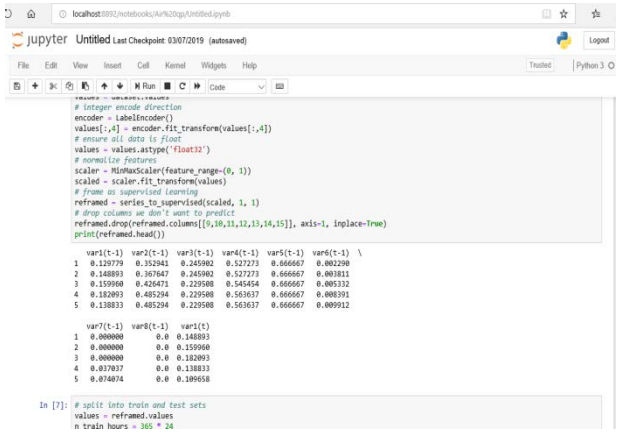

Figure 6 – Snippet 3

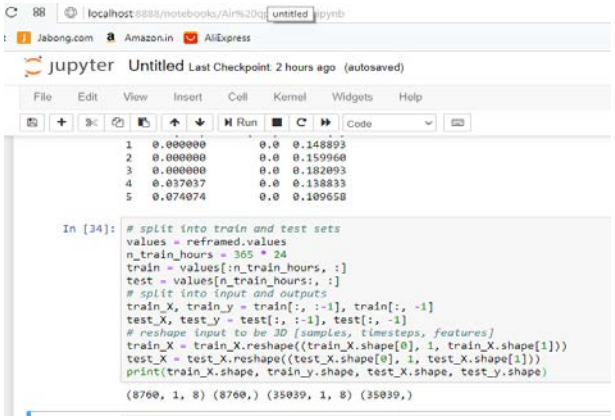

Figure 7 – Snippet 4

Snippet 4 shows train and test input sets shapes.

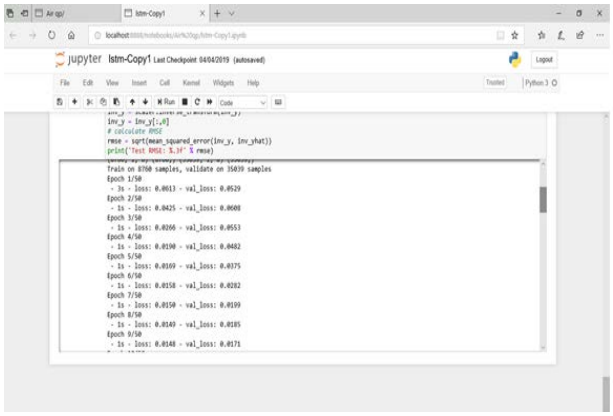

Figure 8 – Snippet 5

Snippet 5 shows RMSE value for test dataset and indicates that model achieved 26.496 of RMSE.

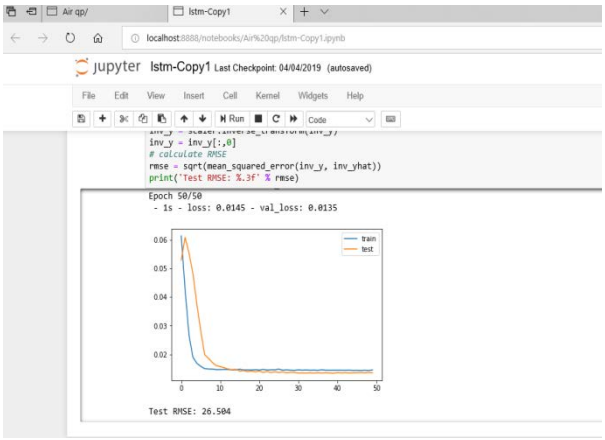

Figure 9– Snippet 6

Snippet 6 shows line plot of train and test loss during training from Multivariate LSTM.

VI. CONCLUSION

Quality of air is main issue .A polluted air causes many health problems such as Heart and Lungs problem and so on, It is necessary is make people aware about quality of air. Platforms such as Deep learning and Machine learning plays major role for Prediction and forecasting of air pollution. Multivariate time series are used; it reduces error and increases prediction accuracy. Before forecasting a data with respect to time series are given to auto encoders. An input feature vector is given to a model that performs forecasting it makes prediction based on input variables. An LSTM is used for multivariate time series for forecasting a data regarding air pollution.

Further the system can be enhanced with IOS and Android platform where it is easy to make people alert about atmospheric air quality level. This system can also applied to wearable devices, so it can be very helpful for people to protect nature and Living beings from Air pollution and its effects.

ACKNOWLEDGEMENT

I would like to thank Dr. S. Satish Kumar (Dept. of CSE, RNSIT, Bangalore) for his valuable suggestions and comments that helped improving this works, this support is greatly appreciated.

REFERENCES

[1] Aditya C R , Chandana R Deshmukh , Nayana D K, Praveen Gandhi Vidyavastu “Detection and Prediction of Air Pollution using Machine Learning Models“, Department of Computer Science and Engineering, VidyaVikas Institute of Engineering and Technology,Mysuru, Karnataka, India 570028.

[2] Zhongang Qi, Tianchun Wang, Guojie Song, Weisong Hu, Xi Li , Zhongfei (Mark) Zhang“Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-grained Air Quality “.

[3] Xia Xi, Zhao Wei, RuiXiaoguang, Wang Yijie, Bai Xinxin, Yin Wenjun, Don Jin ,“A Comprehensive Evaluation of Air Pollution Prediction Improvement by a Machine Learning Method “, IBM CRL Beijing, China {xiaxi, wzhaow,ruixg,wyijie,baixx,yinwenj,dongjin}@cn.ibm.com .

[4] Yi-Ting Tsai, Yu-Ren,Zeng,Yue-Shan Chang ,“Air pollution forecasting using RNN with LSTM, Dept. of Computer Science and Information Engineering National Taipei University New Taipei, Taiwan, R.O.C s7106831@webmail.ntpu.edu.tw and Dept. of Computer Science and Information Engineering National Taipei University New Taipei, Taiwan, R.O.C ysc@mail.ntpu.edu.tw.

[5] Shi, J.P., Harrison, R. M., “Regression modelling of hourly NOX and NO2 concentrations in urban air in London”. Atmospheric Environment 31, 4081ü4094. , 1997

[6] Sousa, S.I.V., Martins, F.G., Alvimü Ferraz, M.C.M., Pereira, M.C.” Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations”. Environmental Modelling and Software 22, 97ü103. 2007.