Air pollutant severity prediction using Bi-directional LSTM Network

Ishan Verma, Rahul Ahuja, Hardik Meisheri, Lipika Dey
TCS Research
Tata Consultancy Services
New Delhi, India
(ishan.verma, ahuja.rahul, hardik.meisheri, lipika.dey)@tcs.com

***Abstract*—Air pollution has emerged as a universal concern across the globe affecting human health. This increasing danger motivates the study of systems for predicting air pollutant severities ahead of time. In this paper, we have proposed the use of a bi-directional LSTM model to predict air pollutant severity levels ahead of time. We have shown that the predictions can be significantly improved using an ensemble of three Bi-Directional LSTMs (BiLSTM) that model the long-term, short-term and immediate effects of PM2.5 (the key air pollutant) severity levels. Further, weather information data has been taken into account while modelling, since they are found to boost prediction accuracies. Experimental results for multiple locations in New Delhi, India are presented to demonstrate model superiority over earlier techniques.**

***Index Terms*—Pollution severity prediction, Time-series analysis, Long-short term memory networks, ensemble learning**

I. Introduction

According to a report by WHO [1], air pollution presents the biggest environmental risk to health. In 2012, one out of every nine deaths reported globally was the result of air pollution-related conditions. Of those deaths, around 3 million are attributable solely to ambient (outdoor) air pollution, that affects all socio-economic groups irrespective of geography and age. An analysis of medical certifications reporting causes of deaths [2] in New Delhi, India shows that 8260 people died due to respiratory diseases in the year 2016. It is also reported that on an average 23 people die every day in the Indian capital due to respiratory diseases, a number that has doubled in the past four years. Though polluted air can have many constituents and compositions, particulate matters with diameter of $2.5\mu m$ or less i.e. PM2.5 is currently reported as a major global concern for human health.

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 the current work, we have presented a model based on Bi-directional Long Short-Term Memory networks (BiLSTM) [1] that outperforms traditional machine learning models. The proposed models are robust and have shown superiority over Artificial Neural Network model in predicting PM2.5 severity levels for multiple stations in New Delhi for prediction upto 6 hours, 12 hours and 24 hours ahead of time. We have also shown that an ensemble of such models further improves the accuracy of prediction.

II. Related Work

Researchers all around the globe have been working in the area of modeling, predicting or assessing pollution and its health impact [2]–[4]. In [5], authors have analysed the air pollution characteristics and their relation to multi-scale meteorological conditions during 2014-2015 for 31 provincial capital cities in China. They have also shown that PM2.5 was the major pollutant followed by PM10, O_3, NO_2, SO_2 and CO. Authors in [6] have proposed the use of multi-layer neural networks to predict PM2.5 concentrations at any hour of the day, by fitting a function of the 24 hourly average concentrations measured on the previous day. It also discussed the effects of explicit consideration of meteorological variables to improve prediction accuracy. In [7] authors presented an online air pollution forecasting system for Greater Istanbul Area. The system predicted three air pollution indicator (SO_2, $PM10$ and CO) levels for the next three days (+1, +2, and +3 days) using neural networks. It was shown that cumulative prediction using previous day's predicted values produced better results than independent predictions. In this work, the importance of day of week as an input parameter was also investigated. In [8] authors addressed the problem of the predicting ozone and PM10 for the city of Milan, Italy using feed-forward neural networks (FFNNs), pruned neural networks (PNNs) and lazy learning (LL). PNNs constitute a parameter-parsimonious approach, based on the removal of redundant parameters from fully connected neural networks; LL, on the other hand, is a local linear prediction algorithm, which performs a local learning procedure each time a prediction is required. No significant differences were found between the forecast accuracies of the different models.

To the best of our knowledge, the closest to our work is multiscale PM2.5 prediction using long short-term memory neural network extended (LSTME) model presented in [9]. Authors argue that existing statistical and machine learning models for predicting air pollutant concentration fail to effectively model long-term dependencies. Additionally, it was also shown that the use of auxiliary data improved model performance.

[1] www.who.int/phe/publications/air-pollution-global-assessment/en/
[2] www.delhi.gov.in

978-1-5386-7325-6/18/$31.00 ©2018 IEEE
DOI 10.1109/WI.2018.00-19

The key differentiator in our work is in the fact that we are predicting severity rather than actual values. while it is sufficient to predict severity indices for pollution state analysis and using it for pollution based planning, the accuracy of prediction improves significantly while working with discretized values. Another differentiating factor is the use of Bi-directional LSTM model that learns to utilize both past and future influence on prediction.

III. Pollution Data Description

Air pollution is determined by the presence of various pollutants like PM2.5, PM10, NO_2, NO, SO_2 etc. PM10 and PM2.5 refer to particulate matters that are present in the environment as solid fine particles with diameters less than $10\mu m$ and $2.5\mu m$ respectively. Air pollution also depends on meteorological parameters like temperature, rainfall, wind speed etc. The meteorological parameters indirectly impact air quality index by increasing or decreasing the concentrations of pollutants in the air.

Air Quality Index (AQI) is a single index used by Environmental agencies to report daily air quality. The higher the AQI value, the greater the level of air pollution and the greater the health concern. Table I shows how Indian National Ambient Air Quality categorizes AQI into different categories along with their potential impacts on health. It also shows the corresponding PM2.5 level ranges for each category. The severity levels shown corresponding to the ranges are used for prediction task.

Since AQI is computed as a 24-hour average, it is not a suitable variable for hourly prediction. Rather predicting PM2.5 or PM10 serves as a useful factor for predicting pollution levels, since one of them is mandatory for computing AQI. In this work, we have presented a deep-learning based model to predict the severity category of PM2.5 as good, moderate or severe corresponding to 0,1 and 2 respectively as shown in Figure 1.

A. Data Pre-processing

Pollutant data like any other sensor data is not free from missing data and abnormal values. The irregularities may occur due to instrumental error or some other external factors like power-shutdown or severance of connectivity etc. There were instances where pollutant data was not reported by a source monitoring station. These missing values were interpolated using rolling average of available data values of past three time instances. A value lying outside the permissible range for a parameter is treated as an abnormal value. Abnormal values are also replaced by rolling average of past three instances.

IV. Prediction Model

The most popular neural models used earlier were Multilayer perceptions. While they were effective in predicting the next value, these models are incapable of capturing both long and short-term dependencies. Pollution data typically exhibits both. There are long-term static patterns that are exhibited seasonally or weekly as well as dynamic patterns that show the sudden build-up of a spike over a few hours. The relationships among all the factors are quite complex and cannot be captured easily without considering both.

Recurrent Neural Networks (RNN) have proved to be very efficient in processing temporal data. However, future input information coming up later than the current time instance is also useful for prediction. RNNs can partially achieve this by delaying the output by a certain number of time frames to include future information. Theoretically, a large delay could be used but in practice, it is found that prediction results drop if the delay is too large. While delaying the output by some frames has been used successfully to improve results for sequential data as shown in [10], the optimal delay is task dependent and need to be obtained by the trial and error error method. Also, two separate networks, one for each direction could be trained on all input information and then the results could be merged using arithmetic or geometric averaging for final prediction. However, it is difficult to obtain optimal merging since different networks trained on the same data can no longer be regarded as independent. To overcome these limitations, [1] proposed bidirectional recurrent neural network (BRNN) that can be trained using all available input information in the past and future of a specific time frame. It was shown that since the network concentrates on minimizing the objective function for both time directions simultaneously, the problem of merging outputs and optimal delay become insignificant as all future and past information around the currently evaluated time point is theoretically available and does not depend on a predefined delay parameter.

Bidirectional Long Short Term memory (BiLSTM) networks are BRNNs using LSTM hidden layers which were proposed in [11]. BiLSTMs are two LSTM stacked over each other. Figure 1 shows our pollution severity prediction model. We have stacked two BiLSTM layers to better capture the hierarchical features in the temporal domain. Between these two layers we have used fully connected Time-distributed layer, which helps in weighting each time-step output before feeding it to the next BiLSTM layer and in decoupling the noise effects while propagating. Dropout has been used to reduce over-fitting in the BiLSTM layer.

Hourly data is fed into the first BiLSTM layer which is denoted as vt_i, where *i* ranges from $0-k$, *k* being maximum number of hourly data taken as a sequence. At each time-step (t_i), BiLSTM layer calculate the output based on current and previous output. Output of each time-step from BiLSTM (O_{t_i}) is taken and fed into a fully connected single layer neural network based Time distributed layer. This dense representation, denoted by D_{t_i} is then fed into second BiLSTM layer. Output of the final time step (F_{t_i}) results into a compact representation of the temporal domain. This output is then merged with inputs from daily parameters, so that the granularity of both the input data is same. The combined representation is then fed to 3 layer fully connected network which act as a classifier.

Normalized outputs seem to be really helpful in sta-

TABLE I
AQI RANGE AND HEALTH IMPACTS

AQI Range and PM 2.5 Range (μm)	Severity	Associated Health Impacts
Good (AQI : 0-50; PM 2.5 : 0-30)	0	Minimal impact
Satisfactory (AQI : 51-100; PM 2.5 : 31-60)		Minor breathing discomfort to sensitive people.
Moderately Polluted (AQI : 101-200; PM 2.5 : 61-90)	1	Breathing discomfort to people with lung and heart disease, children and older adults.
Poor (AQI : 201-300; PM 2.5 : 91-120)		Breathing discomfort to anyone on prolonged exposure.
Very Poor (AQI : 301-400; PM 2.5 : 121-250)	2	Respiratory illness to anyone on prolonged exposure.
Severe (AQI : 401-500; PM 2.5 : 250+)		Respiratory impact even on healthy people, and serious health impacts on people with lung/heart disease.

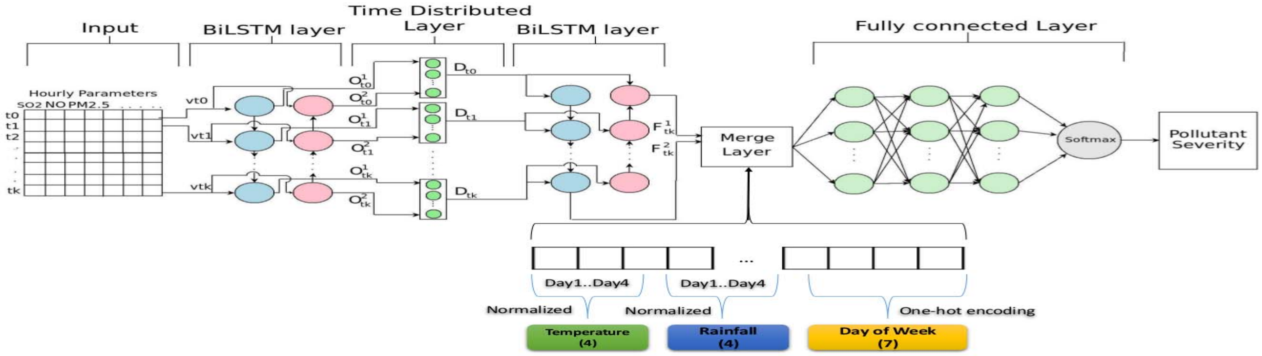

Fig. 1. Model Architecture

bilizing the training process. Scaled Exponential Linear Units(*selu*) [12] produces the normalized output by centering its mean towards zero and constraining its variance to unity. We have used *selu* as activation function in the fully connected layers except in last layer where *softmax* is used.

We have used ADAM [13] as optimizer with default momentum as presented in the paper with a decay rate of 0.00001. We have utilized randomized search for obtaining hyperparameters such as the number of hidden units, learning rate, and its decay. These parameters are tuned for each station separately to better tune model to that specific station. Selected values for the number of hidden units are shown in Table II.

TABLE II
LAYER PARAMETERS AND ACTIVATION FUNCTIONS

Layer	Hidden Units	Activation Function
Bi - Directional	60	tanh
Time Distributed Dense	10	linear
Bi - Directional	40	tanh
Dense	100	selu
Dense	20	selu
Dense	3	softmax

V. EXPERIMENTS AND RESULTS

We have collected pollutant data for 3 stations viz. Punjabi Bagh, Anand Vihar and RK Puram of New Delhi area from Central Pollution Control Board (CPCB) [3] as shown in Table III. The meteorological data has been collected from Weather Underground [4]. Historic data for daily temperature and rainfall was collected as daily aggregate for entire New Delhi region.

TABLE III
PARAMETERS USED FOR PM2.5 SEVERITY PREDICTION

Type	Parameter	Temporality
Pollutant	Particulate Matter (PM2.5)	Hourly
	Sulphur Dioxide (SO_2)	
	Nitrogen Dioxide (NO_2)	
	Nitric Oxide (NO)	
	Ozone (O_3)	
Meteorological	Wind Speed (WS)	
	Wind Direction (WD)	
	Temperature	Daily
	Rainfall	

The collected data ranges from April 15 to November 17. The data is split into training and testing set where training data is from April '15 to June '17 and test data is from July '17 to November '17. The predictions are made for three different time-scale i.e. 6hours(6h), 12hours(12h) and 24hours(24h) ahead for each station separately.

Ensemble prediction model: Apart from independent 6h, 12h and 24h predictions, We have also used a three layer neural network based ensemble model with *relu* as activation for hidden layers. For 6h prediction we use outputs of 6h, 12h and 24h BiLSTM models as inputs. In case of 12h predictions, the outputs of 12h and 24h BiLSTM models are used as input to the ensemble. The resultant ensemble model shows improvement over existing models in both the cases.

[3] www.cpcb.nic.in/

[4] www.wunderground.com

A. Model Input parameters

The hourly input data is N*T*F three-dimensional vector where, N is the number of samples, T is the length of sequence considered and F is the number of features. Here, F is 38, that can be broken down into 5 pollutant values, wind speed, wind direction, one hot encoding of the hour of the day and one hot encoding of the day of the week. In addition, we have considered 72 as value for T i.e. past 72 hours hourly data. All the values extracted from sensor are normalized. Daily input consist of N*P dimensions where, P denotes the temperature and rainfall data of past 4 days normalized along with one hot encoding vector for week of the day.

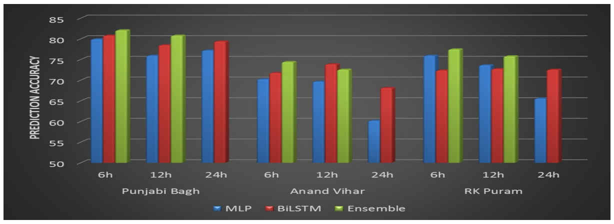

Fig. 2. Overall Performance comparison

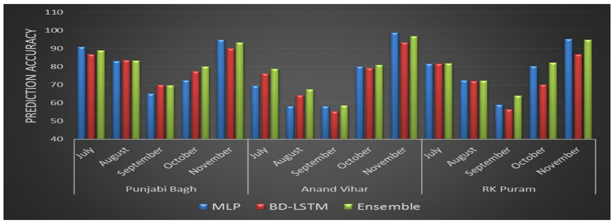

Fig. 3. Month-wise performance comparison for 6hour prediction

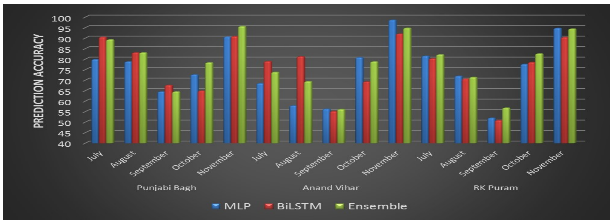

Fig. 4. Month-wise Performance comparison for 12hour prediction

B. Results

Figure 2 shows overall performance comparison of BiLSTM and Ensemble model against Multi-layer perceptron (MLP) model for different monitoring stations for 6h, 12h and 24h prediction scale in terms of prediction accuracy. It can be observed that Ensemble model performs the best under most of the circumstances except the exception of 12h Anand vihar where its results are comparable with BiLSTMs. BiLSTM model shows superiority over MLP in 24h severity prediction.

Figure 3 and Figure 4 shows performance comparison charts for 6h and 12h severity predictions for all three methods for different stations across different months. For 6h prediction the improvement in accuracies of BiLSTM over MLP and Ensemble is as expected except the case of November where contradictory results are there. It was found out in our data analysis that November month has highly sporadic pollution severity level due to a very popular Indian festival Diwali which is a known pollution contributor due to burning of firecrackers. Accuracy results are fairly consistent for 12h prediction where both BiLSTM and Ensemble models show improved accuracy over MLP except November for 2 stations, the reason for which are same as that for 6h predictions.

VI. Conclusion

In this paper, we have presented an effective way of predicting the severity of pollutants by leveraging various sensor data in 6,12 and 24 hour in advance using Deep learning models. We have validated this claim over the pollution data from New Delhi, India by predicting the severity of PM2.5 pollutant. We present our experimentations with comparison to baseline system over different stations and over different time periods. In addition, we have presented a Ensemble system which performs better in most of the cases, and also proves to be more robust.

References

[1] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[2] N. Künzli, R. Kaiser, S. Medina, M. Studnicka, O. Chanel, P. Filliger, M. Herry, F. Horak Jr, V. Puybonnieux-Texier, P. Quénel *et al.*, “Public-health impact of outdoor and traffic-related air pollution: a european assessment,” *The Lancet*, vol. 356, no. 9232, pp. 795–801, 2000.

[3] M. Jerrett, R. T. Burnett, R. Ma, C. A. Pope III, D. Krewski, K. B. Newbold, G. Thurston, Y. Shi, N. Finkelstein, E. E. Calle *et al.*, “Spatial analysis of air pollution and mortality in los angeles,” *Epidemiology*, vol. 16, no. 6, pp. 727–736, 2005.

[4] P. Goyal, A. T. Chan, and N. Jaiswal, “Statistical models for the prediction of respirable suspended particulate matter in urban cities,” *Atmospheric Environment*, vol. 40, no. 11, pp. 2068–2077, 2006.

[5] J. He, S. Gong, Y. Yu, L. Yu, L. Wu, H. Mao, C. Song, S. Zhao, H. Liu, X. Li *et al.*, “Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major chinese cities,” *Environmental pollution*, vol. 223, pp. 484–496, 2017.

[6] P. Pérez, A. Trier, and J. Reyes, “Prediction of pm2. 5 concentrations several hours in advance using neural networks in santiago, chile,” *Atmospheric Environment*, vol. 34, no. 8, pp. 1189–1196, 2000.

[7] A. Kurt, B. Gulbagci, F. Karaca, and O. Alagha, “An online air pollution forecasting system using neural networks,” *Environment International*, vol. 34, no. 5, pp. 592–598, 2008.

[8] G. Corani, “Air quality prediction in milan: feed-forward neural networks, pruned neural networks and lazy learning,” *Ecological Modelling*, vol. 185, no. 2-4, pp. 513–529, 2005.

[9] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi, “Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation,” *Environmental Pollution*, vol. 231, pp. 997–1004, 2017.

[10] A. J. Robinson, “An application of recurrent nets to phone probability estimation,” *IEEE transactions on Neural Networks*, vol. 5, no. 2, pp. 298–305, 1994.

[11] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[12] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, “Self-normalizing neural networks,” *CoRR*, vol. abs/1706.02515, 2017.

[13] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.