Development of a Deep Learning Technique to Analyze Air Quality from Social Media Response

Thushara Sudheish Kumbalaparambil

SCMS School of Engineering and Technology

Ratish Menon (✉ ratishmenon@scmsgroup.org)

SCMS School of Engineering and Technology <https://orcid.org/0000-0002-8391-1672>

Vishnu P Radhakrishnan

SCMS School of Engineering and Technology

Vinod P Nair

Cochin University of Science and Technology

Research Article

Keywords: Twitter, air pollution, self attention network, Delhi, PM2.5

Posted Date: February 8th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1174813/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

One of the prominent new-age methods used by today's population in spreading awareness or drawing attention on an issue or concern is through social media platforms. In this study, the responses of general public to air quality that they share on a popular social media platform – Twitter, was taken as a virtual quantity that helps in measuring and analyzing the prevalent air quality. Machine learning technique based on self attention network was used to sort, clean and classify large amount of air pollution related Twitter responses extracted during 2019-2020 at Delhi in India. The temporal correlation of tweet response volumes with the ground monitored concentrations of air pollution and word cloud analysis were used to analyze the attitude of the public towards urban air quality. These analyses lead to the development of an air quality prediction model using 'spaCy' – similarity analysis method which attempts to predict the ground concentration range of pollutant $PM_{2.5}$ from tweets received on a particular day.

1. Introduction

Increasing air pollution has become a major concern for the environmental quality of life in the urban areas. Recent studies visualize the grim reality of air pollution and its health risks across the cities of the world (Dandona, 2018; Pant et al., 2019; IQAir, 2019; WHO, 2020). For strategizing any control or eradication program, first step would be to monitor air quality at possibly high temporal and spatial resolutions to identify the source, location, quality and quantity of pollutants. The existing regulatory monitoring networks, especially in developing countries such as India, however make sparse spatial and temporal measurements (CPCB, 2020). Any additional information in real time, even if qualitative, would be an advantage for air quality management initiatives in countries like India which has 6 out of the 10 most polluted cities in the world (IQAir, 2019).

With increased access to internet and emergence of social media platforms, people are now expressing their views and relation to the events around them like never before. Social media have grown tremendously in number and popularity ever since its emergence and now include multitude of platforms such as Twitter, Facebook, LinkedIn, YouTube channels, blogs, chat rooms, discussion forums etc. India with its population of ~1.3 billion people has over 560 million internet users and is the second largest online market in the world. The internet accessibility and use in the country largely varied based on factors like gender and socio-economic divide (Statista, 2020b). Analyzing public behavior on such platforms are not new to researchers and have been part of routine business analytics (Fan and Gordon, 2014). Many studies have also explored their potential for monitoring environmental events such as natural disasters in real time (Sakaki et al., 2010; Lindsay, 2011; Earle et al., 2011; Li et al., 2012; Kent and Capello Jr., 2013; Middleton et al., 2013, Singh et al., 2019).

Social media responses also have the potential to extract useful information on air pollution (Robinson and Fialkowski, 2010; Jiang et al., 2015; Gurajala and Matthews, 2018; Jiang et al. 2019). It has been seen that the exposed people tend to share pictures, videos, blogs and tweets about air quality events almost immediately on social media within minutes of occurrence of an event (Robinson and Fialkowski,

2010). These user generated contents (UGC) enable us to derive real time information on air pollution. The response of users or ‘human sensors’ may be direct in the form of immediate posts, tweets, blogs etc or indirect like re-sharing an already existing content or supporting it. This rise in popularity of social media has given people access to a humongous volume of information which can be used in a variety of valuable areas. Extracting and analyzing UGC to derive actionable knowledge is challenging task. However, recent approaches using machine learning (ML) techniques have been found promising for the purpose (Jackoway et al., 2011; Gurajala and Matthews, 2018). Such ML techniques are exploited curiously by new researchers are air quality prediction and forecasting as Xu et al. (2020), Chang et al. (2021) and Zhang et al. (2021).

Use of attention network in deep machine learning algorithms which is now widely recognised in sequence modeling as shown by Bahdanau et al. (2014) and Vaswani et al. (2017) to maximize performance in models using attention mechanism to connect between the input and output. Long - Short Term Memory (LSTM) are known to be effective for time series prediction tasks and multiple LSTM units are stacked to form the more effective Bi-directional Long - Short Term Memory (Bi-LSTM) models(Zhang et al., 2021). Wiedemann et al. (2018) and Wu et al. (2019) has also presented inspiring Bi-LSTM model architectures.

Table 1 shows a tabulation of some of the previous works that attempted to predict air quality using different ML techniques. It is seen that the influence of social media for air quality prediction is lesser explored than using air quality parameters and other polluting factors. Also self attention networks have not been prominently used for air quality prediction using social sensors.

Table1

Comparison between various reference papers that performed air quality prediction

Reference papers	Monitored air quality parameter	Source/ Social media platform used for prediction	Country or city under study	Prediction Models/Techniques used, efficiency obtained
Jiang et al. (2015)	AQI	Sina Weibo(Chinese Twitter)	Beijing	Gradient tree boosting (GTB) - 59%
Gurajala and Matthews (2018)	PM _{2.5}	Twitter	Paris, Delhi, London	Air quality not predicted
Jiang et al. (2019)	AQI	Twitter	California, Idaho, Illinois, Indiana, Ohio	Natural Language Processing (NLP) - 6.9–17.7% improvement with social media intervention over base line method.
Xu et al. (2020)	PM _{2.5}	Historical meteorological data, road network data, administrative boundary vector data, POI data.	Beijing, Tianjin, Hebei	Temporal-spatial-regression-tree model, Grid prediction model - 90%
Chang et al. (2021)	PM _{2.5}	Local and neighboring station data, chimney and abroad pollution data	Taiwan	Aggregate LSTM - better than GTB, LSTM, SVR*
Zhang et al. (2021).	PM _{2.5}	PM _{2.5} - Hourly, daily, restructured multi hour data	Beijing	LSTM, Bi-LSTM, EMD**-BiLSTM, - More than 95%
*SVR - Support vector regression				
**EMD - Empirical mode decomposition				

In the present work, Twitter responses related to air quality in Delhi were analyzed during 2019-2020 to develop a deep learning model that can predict the concentration of PM_{2.5} (mass concentration of particulates having an aerodynamic diameter less than 2.5µm) from the tweets. A self attention network based classifier was used to characterize the tweets. The PM_{2.5} data from continuous ambient air quality monitoring stations (CAAQMS) within the city were used to train the deep learning algorithm. The aim of the study is to:

- analyze the fit and relevance of the physically monitored air quality parameter with the tweeting volume and behavior of people in an urban city of a developing country like Delhi, India and thus;
- evaluate the ability to predict pollution concentrations from this virtual medium;
- develop a new self attention deep learning classification model with high accuracy to classify tweets to those indicating poor air quality (Class I), good air quality (Class II) and neutral or noise tweets (Class 0) with minimal human intervention.
- develop a machine learning model to predict the PM_{2.5} concentration by analysing a tweet.

The further organization of the paper is as follows: methodologies and machine learning techniques followed for the accomplishment of the objectives are elaborated in Section 2. The results and discussions of the temporal and behavioral analysis of the study are discussed in Section 3 along with the efficiency on the classification and prediction models. Finally Section 4 wraps up the paper with the conclusions drawn, challenges faced by the study and the future scope of work.

2. Methodology

This section discusses in detail the extraction of the social media responses from the chosen area under study during 2019 - 2020 and the various analyses performed on them along with the descriptions of the machine learning models used for the classification of the extracted data and the prediction of $PM_{2.5}$ concentration using the classified data.

2.1 Selection of study area and the social media platform

Twitter was chosen as the social media platform for the present study. Tweet response volumes, defined as the sum of direct and indirect response of people to air quality, for the period of 12 months ranging from March 2019 to February 2020 was used as the data for the analysis. Indirect responses included likes and re-tweets for a particular tweet. There have been works like that of Jiang et al. (2015) who have considered the effects of direct and indirect responses separately but we chose to consider them as a collective response of the people of Delhi to their air quality issues. Twitter is peculiar for its short and crisp contents limited over 280 characters and its prominence at the time of an event. Twitter has also become an increasingly important tool in politics, mass media communications, and many more such that it touches almost every sector of life. Many world leaders, governments, ministries, influencers, institutions, news channels have their official Twitter accounts to make announcements, influence and engage with the general public. In India there were 13.15 million Twitter users as of April 2020 (Statista, 2020a).

Delhi, capital city of India, was selected as a representative urban area to demonstrate the methodology presented in this paper. A preliminary analysis of tweets related to air pollution from the 100 most populated cities in India as part of this study identified Delhi as the city with maximum number of air quality episodes (IQAir, 2019) and maximum number of tweets related to air quality. Also, a public health emergency was declared at Delhi in November 2019 as the air quality index exceeded more than 3 times the 'hazardous' level (CNN, 2020).

2.2 Data extraction

Tweets were extracted from Twitter by attaining access to Twitter streaming API. The tweets were fetched with the help of keywords pertaining to air pollution along with the name of the city. About 15 different keyword combinations with Delhi (as shown in Table 2) were used for tweet extraction for the study period. The keywords thus chosen were different combinations of 'air quality', 'air pollution' and 'smog'. The data set collected included attributes such as the day, date and time of the tweet, the tweet text and

the number of likes and retweets received by that particular tweet. In this manner, a data set of around 82,000 tweets were created. Tweets only in the English language were considered in this study as Twitter users in India are pre-dominantly English speakers (Poell and Rajagopalan, 2015). In order to reduce the noise in the data set, few words were used to filter out unnecessary tweets like those generated by automated websites on a periodic basis regardless of the change in pollution intensity. In order to relate Twitter activity with ambient air quality information, PM_{2.5} data from 38 CAAQMS stations within Delhi during the period of March 2019 to February 2020 were collected from the CPCB data repository (CCR-CPCB, 2020). The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Table 2

Keyword combinations used to extract tweets during study period

Keyword Combinations		
Air Quality + Delhi	Air Pollution + Delhi	DelhiChokes
AirQuality + Delhi	AirPollution + Delhi	Choke + Delhi
AirQualityDelhi	AirPollutionDelhi	Clean air + Delhi
DelhiAirQuality	DelhiAirPollution	DelhiEmergency
Delhi + Smog	DelhiSmog	Delhi + RightToBreathe

2.3 Data classification

Each tweet, within the allowed 280 characters contained words indicating good or bad air quality. Extracted dataset also contained some tweets that are either irrelevant or neutral with respect to the air quality. Segregating these tweets into various classes was an important step in the present study. The tweet data set was manually read and sorted by a team of 15 environmental engineers and the data set was sorted into 3 classes to segregate them into: poor air quality tweets (Class I), good air quality tweets (Class II) and noise or neutral tweets (Class 0). The sorted data set was then checked for variations in sorting and resorted (if needed) by an expert team of 3 environmental engineers.

The sorted set consisted of tweets in Class I and II which contributed as indicators of the ambient air quality whereas Class 0 consisted of tweets generated by bots, irrelevant tweets, etc. Examples of tweets in their respective classes is shown in Table 3. The classified set was then pre processed and used to train a self attention model that would automatically classify a sample tweet into its suitable classes.

Table 3

Examples of classified tweets

S. No.	Tweets	Classification
• 1.	New Delhi: Over 1.2 million people died in India due to air pollution in 2017, said a global report on air pollution	Class 0
• 2.	#DelhiAirEmergency #DelhiPollution #DelhiBachao #DelhiAirQuality #DelhiNCRPollution	Class 0
• 3.	Kab tak zindagi katoga bd or cigar mein, kuch din to gujaro delhi, in ncr	Class 0
• 4.	Just landed in #Delhi, the air here is just unbreathable	Class 1
• 5.	Amazing Air Quality today in Delhi! Enjoy the blue sky and clean air while it lasts.	Class 2

2.4 Data analysis

The following analyses were performed on the extracted and classified data set.

2.4.1 Temporal analysis of tweet response volume in relation to PM_{2.5} concentrations

The volume of tweet responses related to air quality was compared with the ambient concentrations of PM_{2.5} so as to analyze similarity in their trend. The comparison was done by graphical and statistical methods.

2.4.1.1 Graphical analysis

Semi log time series graphs for the tweet response volumes for each of Class I and Class II tweets were plotted against the ground average PM_{2.5} concentrations obtained from CPCB for the duration of the study. Twitter activity was plotted on a logarithmic scale considering their vast data range. The coherence of peaks and falls in the graph were analyzed. The graphical representation and the inferences drawn from it are discussed in Section 3.1.

2.4.1.2 Statistical analysis

In-order to understand the statistical relationship between the two data sets, box plots were made with PM_{2.5} on x-axis and tweet response volume on y-axis. The data was sorted on the basis of the classification of PM_{2.5} values into 6 categories as defined by the CPCB namely Good(0 - 30µg/m³),

Satisfactory(31 - 60 $\mu\text{g}/\text{m}^3$), Moderate(61 - 90 $\mu\text{g}/\text{m}^3$), Poor(91 - 120 $\mu\text{g}/\text{m}^3$), Very Poor(121 - 250 $\mu\text{g}/\text{m}^3$) and Severe (above 250 $\mu\text{g}/\text{m}^3$) for plotting the box plots. The relationship of tweet response volumes of each class (Class I and II) in these categories were analyzed and the results are given in section 3.2.

2.4.2 Behavioral analysis using word clouds

Word clouds were generated corresponding to each season to understand how people express the changes in the air quality through words using a python program utilizing necessary supporting libraries. India has 4 meteorological seasons namely Winter (January, February), Pre Monsoon (March – May), Southwest Monsoon (June – September) and Post Monsoon (October – December) as per Indian meteorological Department (IMD). A prominent transition in the air quality levels in Delhi was observed during post monsoon and winter seasons having the worst air quality which improved over pre monsoon and southwest monsoon seasons.

2.5 Machine Learning Techniques

In this study, machine learning techniques were used for (i) automated classification of the tweet data set and (ii) for the prediction of PM_{2.5} concentration range from tweet response. The program codes constructed for the classification, training and prediction of the data in this work is available from the corresponding author on reasonable request.

2.5.1 Automated data classification model

The tweets extracted over a year along with the likes and retweets appended to them comprised the dataset for the study. The dataset was a highly imbalanced multiclass data set with around 62,000 tweets in Class 0, 18000 tweets in Class I and 1700 tweets in Class II. As initial step, the manually classified dataset was tokenized into words. Preprocessing was performed in order to remove special symbols, stop words, punctuations, Twitter handles, etc. Natural Language Toolkit(NLTK) library was used for the same. The data set was further groomed by performing processes like word indexing, integer encoding and assigning them with pad descriptions before using the data for supervised learning. The respective classes of the tweets were label encoded. The processed data set was split into a training set and test set as 90% and 10% respectively for a 10 fold cross validation analysis to evaluate the overall performance of the model and a test-train split of 80%-20% for result analysis.

The present study uses a multilayer classification model. The first layer is an embedding layer in which an embedding matrix is constructed from the data set using pertained global vector (GloVe) embedding for creating weights for the embedding layer (Pennington et al., 2014). The second layer is a BiLSTM layer (Units: 100) where the input sequences are analyzed in the forward and backward direction; that is, if a tweet has 'n' number of words ($w_1, w_2, \dots, w_n$) then the tweet is first analyzed to from w_1 to w_n by forward LSTM and w_n to w_1 by backward LSTM and forms a word feature (W_f) for each word on the basis of both the forward analysis (f_f) and backward analysis (f_b) such that

$$Wf = f_F \parallel f_B$$

where $\parallel$ stands for concatenation function.

Similarly a word feature is formed for every word of the data file. After this the data flows into a self attention layer (SeqSelfAttention) layer where words are assigned weights and added to the resultant word features of the previous layer based on the relative importance of that word to the entire tweet. This layer improves the efficiency of the classification as 'attention' is given to words based on their importance. For further detailed reference on Bi-LSTM and attention neural networks Bahdanau et al, 2014; Vaswani et al., 2017; Wiedemann et al, 2018; Zhang et al., 2018; Wu et al, 2019; Xu et al., 2020; Chang et al., 2021 and Zhang et al., 2021 may be used. Python libraries like Keras (<https://keras.io/>) and Keras_self_attention (<https://pypi.org/project/keras-self-attention/>) was used to support this layer.

The context vectors or sentence feature vector that are obtained as the output of the attention layer using weighted sum function is then given to two dense layers or the fully connected layers where all the inputs and outputs are connected to all the neurons in each layer. The both dense layer contains 50 units and is activated with Rectified Linear Unit (ReLU) layer activation function. It is then flattened in the next layer, the flatten layer. The final output is obtained in a dense layer which has units equal to the number of output classes and is then given output activation by softmax function which provides the probabilities of the potential outcomes. Other parameters used in this modelling architecture includes: Epochs=5, Batch size:128, Optimizer: Adam. The model architecture is shown in Fig. 1.

2.5.2 PM_{2.5} prediction model

The prediction model used similarity analysis to find the most similar tweets to a test tweet to predict the probable ground PM_{2.5} concentration that would have been recorded on that day. The flow chart of the model is shown in Fig. 2. This model used 'spaCy' library available in python in order to find the top 10% of the most similar tweets of the test tweet from the season wise training set. A test tweet first undergoes model classification as mentioned in the previous section and then from the date of the tweet identify respective season's training set. The test tweet was then analyzed for similarity with each tweet of its corresponding training set and a similarity index was assigned to each training tweet. The training set tweets are then arranged in descending of the cosine similarity index and the first 10% of the training set in this sorted set was taken as the resultant 'most similar tweets set'. The arithmetic mean of PM_{2.5} concentrations corresponding to that similarity set was then reported as the output result of the model.

The usefulness of the model was analyzed by measuring the correctly predicted concentrations as the percentage of the total predictions under each category specified by CPCB (CPCB, 2014). This was also represented as a histogram plot to understand the social media behavior with respect to change in air quality.

3. Results And Discussions

The data classification model sorted the tweets into their corresponding Class (Class 0, I or II) with an overall accuracy of 96.7% from the 10 fold cross validation testing and an accuracy of 87.4% for the model using test-train split of 80% - 20%. On the classified data further analyses were done, the results of which are discussed in the following sections.

3.1 Temporal variation of tweet responses with PM_{2.5} concentrations: Graphical analysis

A time series analysis for tweet responses of Class I and II along with corresponding PM_{2.5} concentration was conducted for the duration of study (Fig. 3). It was observed that there is rise in peak of the tweet response volume along with rise in pollution in the air and also a corresponding dip in tweet response volume with a fall in air pollution. It was noticed that there is more response in Class I than Class II. The larger response volume in Class I might be due to the alertness of the public to poor quality of air that brings distress to their daily lives and their eagerness to spread the information to others, draw attention of government or other related or powerful individuals in the society, apart from the fact that the quality of air remained poor for a major part of the study period. The low volume in Class II is primarily due to lesser number of good air quality days in Delhi and also might be because good air quality days are appreciated mostly after a spike in pollution and not much while quality of air has been staying good for a period of time.

During pre-monsoon season (March – May 2019), Delhi had its worst air quality towards the middle of May and that was the time when the highest tweet response (around 200) in Class I was recorded. The day with best air quality of the season reported no Class I tweets. High volume in Class II tweet response was found during the early days of March when the air quality used to improve to moderate category after a poor air quality period. Most of the times when the PM_{2.5} concentrations fell within the desirable limit there was some Class II tweet response activity. A sudden fall in pollutant level resulted in more tweet responses than a gradual fall.

During the period June – September 2019 (southwest monsoon season), the average PM_{2.5} concentration levels in Delhi were mostly within the desirable limit. Therefore, Class I tweet response activity was comparatively lesser (below 100) compared to other seasons. There were more number of days with minimal or no Class I tweet response volume. Whereas in the case of Class II, the maximum tweet response volumes were recorded on the day with best air quality of the year. Even though the quality of air in this season was mostly good, there wasn't a tweet response for each of those days. Class II responses were mostly observed on good air quality days after an increase in pollution or if the pollutant concentrations were well below the desirable limits.

In the post monsoon season the average $PM_{2.5}$ concentration in Delhi was seldom within the desirable limits that is below $60\mu g/m^3$ as specified in National Ambient Air Quality Standards by CPCB. This was the infamous times when the stubble burning in northern parts of India highly polluted the air in the northern states and public health emergency was declared in Delhi (CNN, 2020). This season also coincides with the Indian festival Diwali which is mostly celebrated with fireworks, crackers, etc. Hence, for almost every day of this season, there was Class I tweet response. Class I tweet response volume was varied from 100 units and ranges as high as above 50000 on the most polluted day of the year when average $PM_{2.5}$ concentration over Delhi was around $550\mu g/m^3$. The time series of the tweet response volume of this season resembles much with its $PM_{2.5}$ concentrations. There was a high response in Class II as well. This was due to the fact that even a small fall in pollutant concentration was a relief to the citizen that they readily responded, increase the volume in Class II.

In the winter season, Class I response was observed almost every day of the season owing to the fact that the average $PM_{2.5}$ concentration in Delhi during this season was seldom within the desirable limits. As the air pollution was not bad as in the previous season, the tweet response volume range has reduced. The low volume of response on 1st January 2020 in contrary to the high ground $PM_{2.5}$ concentration on the same day might have been a result of deviation of attention of public from air pollution to New Year celebrations and activities. As the air quality in this season was generally poor, the volume of tweet responses in Class II was less. However, Class II tweets were prominent with sharp fall in pollution as seen in the previous cases.

3.2 Statistical relationship of tweet response volume with $PM_{2.5}$

To analyze the statistical relationship between the tweet response volumes and the average ground $PM_{2.5}$ concentrations, box plots were made for each class as shown in Fig. 4. For this analysis the data was classified according to the different $PM_{2.5}$ concentration categories as specified by CPCB.

It was seen from Fig. 4(a) that the tweet response volumes for Class I were the largest when the $PM_{2.5}$ concentrations were in the poor - severe category and the lowest when it was in the good – moderate category. This proves that the people were more responsive on Twitter and posted tweets indicating poor air quality during the times when the quality of air was in the poor – severe conditions and the Class I tweet response volumes were proportional to the severity of air pollution. The small volume of Class I tweets even on days with good – moderate quality of air could be because the air quality of a day was poorer than the previous day though the concentrations were not above the desired limits. It could also be a response of people visiting the city for the first time and finding the air quality to be bad compared to the place from which they arrived. From the present dataset, it was difficult to verify the behavior of the floating population.

Class II tweets were made not only based on present day air quality, but also when there was a decrease in pollution levels from previous days. Therefore, for the statistical analysis, the change in $PM_{2.5}$ concentration every 3 days (Δ_3) was calculated and the cumulative Class II tweet response volume during that period was considered and plotted to a box plot. From Fig. 4b, it was observed that days with good - satisfactory $PM_{2.5}$ concentrations had smaller Δ_3 and therefore had lesser tweet response volumes. During the very poor – severe air pollution days, the tweets response volumes were more in Class I category than Class II. The pollution was too high to have anything positive to tweet about. It was only while the air quality improved from severe to moderate the people were more responsive to tweet in a positive manner which explains the larger tweet response volume observed under moderate-poor category.

3.3 Behavioral analysis from word clouds

Word clouds were generated from the model classified & pre processed data segregated season wise in order to study the responsive nature of public about air quality over various situations and seasons. Word clouds helps in identifying the most frequent words in a season, the nature of emotions expressed through these words & the focus of people. The word cloud for each season of the study period is shown in Fig. 5.

The word clouds for the different season in the study period collectively suggest that the citizens of Delhi were much interested in following the air quality index(AQI) of the city regularly. People seems to pay continuous attention to the news updates which are a major source of the AQI indices apart from dedicated apps, Twitter handles, etc and like and retweet tweets put up by news channel in regard with the varying air quality conditions. During poor air quality seasons, the tweeting pattern seem to be dependent on how distressed they feel about air pollution along with concern for health and calling for help and attention from the government and related departments. During good air quality season, people were found to tweet about the content and happiness in having a clear sky and asking the government to maintain the good conditions. The reason for an improvement or deterioration of the air quality is also of interest to the people.

From the word cloud for pre monsoon season, it was evident that the air quality was mostly in the 'moderate' to 'severe' zone. The social media responses have reported when the air quality turned 'poor' or when it is 'deteriorate'-ing or when it has shown an 'improve'-ment. This was also the summer season in Delhi and there was tweets correlating air quality with 'temperature' and 'dust' and how it was an 'unhealthy' condition and difficult to 'breathe'. In the monsoon season, tweets were mostly about 'improve'-d air quality and tweets about 'rain', 'wind', and 'storm' with an 'AQI' mostly in the 'good', 'satisfactory' or 'moderate' ranges and asking the 'government' to maintain the 'clean' 'sky'. The post monsoon season had large number of tweets in comparison to the other seasons and the wordcloud had the maximum words. The public mostly spoke about how the condition was 'severe', 'poor', 'toxic', 'hazardous', and an 'emergency' situation which 'choked' and affected the daily life of the 'people' and led to 'school' to shut and 'flights' delayed or redirected. They reported about the effect of 'diwali' too in

turning Delhi into a 'gas chamber'. Even the slightest 'improve'-ment brought a 'relief' causing people to tweet about it. In the winter season, tweets were mainly about 'poor' quality of air still 'remain'-ing in the city with its share of 'improve'-ments. But the condition was mostly 'toxic' and 'bad' as 'hell'. The 'fog' contributed much to the 'smog'.

3.4 Prediction model results

Machine learning based prediction model that can estimate $PM_{2.5}$ concentrations from the content of tweets was developed in the present study. Fig. 6 shows the percentage accuracy of these predictions for each $PM_{2.5}$ concentration category (CPCB, 2014).

Prediction accuracy was found to be high for extreme conditions of air quality such as good air quality (80%) and severe air quality (99%) categories. This was because the public could clearly experience these conditions without ambiguity, who then reflected these experiences with appropriate words in their tweet and thus making it easily predictable with our prediction model. As the $PM_{2.5}$ concentration moved into other categories, mixed tweeting behavior was observed thereby reducing the prediction accuracy for the model. Also, for a city like Delhi, the citizens were used to moderate to poor air quality conditions for the major part of the year and thus their reactivity to moderate air pollution situations were not so prominent. It was the deviations from the moderate conditions that made the people more alert and produced larger specific tweets suitable for predicting accurately.

4. Conclusion

Present study explored the social media behavior of urban dwellers towards air pollution and developed a deep learning based model to predict the concentration of $PM_{2.5}$ in the city based on the air pollution related social media responses. The population of Delhi expressed their agony and disappointment on urban air quality through various modes and present study analyzed their response to air quality through popular social media forum - Twitter. The Twitter generated data was sorted before analysis using a well efficient self attention network. The relationship of the Twitter responses to actual pollution level was inferred from temporal and statistical analysis of average $PM_{2.5}$ concentrations with tweet response volumes received per day over various season of a year and categories of $PM_{2.5}$ concentration ranges. The emotions and attitudes that the Delhi public hold towards air quality are depicted through the words they use in a tweet was analyzed using word clouds which showed that the Delhi public closely follows the AQI in Delhi through news streams or dedicated apps or through social media. It is evident from these analyses that social media is a powerful tool to monitor air quality variations in an urban city like Delhi. The PM concentration prediction results showed that there was high prediction efficiency (above 80%) for extreme air quality conditions and lower prediction efficiency for moderate air quality conditions.

It was observed that limited people in India geo tag their tweets and therefore it was difficult to find location of specific tweets unless they mention the location in the tweet. The present study has only

considered the text in tweets for analysis here. There were Twitter responses in other forms such as pictures, re-directing internet links, etc. which were not considered in this study. This may be overcome by developing a system to identify multiple languages, image interpretations, etc. Also tweets in regional languages were also overlooked in this study and could further develop the models to include more vernacular languages of the area and mixed scripts using cross lingual knowledge transfer and analysis. The nature of response of people on Twitter in terms of likes or retweets may be influenced by tweets of influential persons, popular organisations, bots, etc. This influence maybe taken into attention and the current study could be further advanced with the analysis of account origins of the tweets. It was seen during the pilot study of this study that many India cities face air quality issues but most of them do not have an active Twitter response culture. Metro cities showed relatively better responses and Delhi with its severe air quality issues had high response volume. Though the study focused on the social media behavior of people in Delhi and depends on several factors such as access to social media, environmental awareness among public, range of changes in air quality experienced by public etc., the methodology is replicable to any urban area in the world. Twitter was also not very popular in India as compared to other social media platforms like Facebook, YouTube, Instagram, etc. Such other social media platforms may also be explored in future to analyze if they hold higher relevance to real time air pollution related responses monitoring in India than Twitter for similar studies. In case of prediction of pollutant concentrations, similarity analysis might be insufficient for peculiar case of air quality conditions unfamiliar to the current training data set such as the unduly fall in air pollution levels in India during Covid-19 pandemic lockdown.

Declarations

Data Availability

PM_{2.5} data is available from CPCB at <https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing>. Twitter data can be extracted using Twitter streaming API [<https://developer.twitter.com/en/docs/twitter-api>]. Data are also available from the corresponding author on reasonable request.

Funding

No funding was received to assist with the preparation of this manuscript.

Ethical Approval

Not applicable.

Consent to Participate

Not applicable

Consent to Publish

Not applicable

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Thushara Sudheish Kumbalaparambil: Formal analysis, Writing -original draft. **Ratish Menon**: Supervision, Writing- original draft, Writing- review & editing. **Vishnu P Radhakrishnan**: Programming, Data curation. **Vinod P Nair**: Formal analysis, Writing-review & editing

References

1. Bahdanau D, Cho K, Bengio Y (2014). Neural machine translations by jointly learning to align and translate. ICLR 2015, arXiv:1409.0473
2. CCR-CPCB (2020). <https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing>, last accessed on 01-12-2021
3. Chang Y, Chiao H, Abimannan S, Huang Y, Tsai Y, and Lin K (2020). An LSTM-based aggregated model for air pollution forecasting. Atmospheric Pollution Research, Volume 11, Issue 8, August 2020, Pages 1451-1463
4. CNN (2020). <https://edition.cnn.com/2020/02/25/health/most-polluted-cities-india-pakistan-intl-hnk/index.html>, last accessed on 01-12-2021
5. CPCB (Central Pollution Control Board), 2014. National Air Quality Index. Control of Urban Pollution Series, CUPS/82/2014-15
6. CPCB (2020). <https://cpcb.nic.in/monitoring-network-3/>, last accessed on 01-12-2021
7. Dandona L (2018). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The Global Burden of Disease Study 2017. Lancet Planet Health; 3: e26–39
8. Earle P S, Bowden D C, and Guy M (2011). Twitter earthquake detection: Earthquake monitoring in a social world. Annals of Geophysics, 54(6), 708–715
9. Fan W and Gordon M (2014). The Power of Social Media Analytics. Communications of the ACM. 57. 74-81. 10.1145/2602574
10. Gurajala S and Matthews J N (2018). Twitter data analysis to understand societal response to air quality. Proceedings of the 9th International Conference on Social Media and Society ,July 2018 ,Pages 82–90, <https://doi.org/10.1145/3217804.3217900>
11. IQAir (2019). World Air Quality Report - Region & City PM2.5 Ranking. <https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2019-en.pdf>, last accessed on 01-12-2021
12. Jackoway A, Samet H and Sankaranarayanan J (2011). Identification of Live News Events using Twitter. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social

13. Jiang J, Sun X, Wang W and Young S (2019). Enhancing Air Quality Prediction with Social Media and Natural Language Processing. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 28 - August 2, 2019, Pages 2627–2632
14. Jiang W, Wang Y, Tsou M-H, and Fu X (2015). Using Social Media to Detect Outdoor Air Pollution and Monitor Air Quality Index (AQI): A Geo-Targeted Spatiotemporal Analysis Framework with Sina Weibo (Chinese Twitter). PLoS ONE 10(10): e0141185.
15. Kent J D and Capello Jr. H T (2013). Spatial patterns and demographic indicators of effective social media content during the Horseshoe Canyon fire of 2012. Cartography and Geographic Information Science, 40:2,78-89
16. Li R, Lei K H, Khadiwala R and Chang K C (2012). TEDAS: A Twitter-based Event Detection and Analysis System. IEEE 28th International Conference on Data Engineering, 1084-4627/12
17. Lindsay B R (2011). Social Media and Disasters: Current Uses, Future Options, and Policy Considerations. Congressional Research Service, Washington, DC, 7-5700
18. Middleton S E, Middleton L and Modafferi S (2013). Real-Time Crisis Mapping of Natural Disasters Using Social Media. IEEE Intelligent Systems 29 (2), 9-17
19. Pant P, Lal R M, Guttikunda, S K, Russell AG, Nagpure A S, Ramaswami A and Peltier R E (2019). Monitoring particulate matter in India: recent trends and future outlook. Air Quality, Atmosphere & Health.12, 45-58
20. Pennington J, Socher R and Manning C D (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532–1543
21. Poell, T and Rajagopalan S (2015). Connecting Activists and Journalists: Twitter communication in the aftermath of the 2012 Delhi rape. Journalism Studies, 16:5, 719-733
22. Robinson E M, and Fialkowski W E (2010). Air Twitter: Using Social Media And Scientific Data To Sense Air Quality Events. 2010 IEEE International Geoscience and Remote Sensing Symposium, July 25-30, 2010, Honolulu, Hawaii, USA
23. Sakaki T, Okazaki M and Matsuo Y (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World Wide Web, Pages 851–860
24. Singh J P, Dwivedi Y K, Rana N P, Kumar A and Kapoor K K (2019). Event classification and location prediction from tweets during disasters. Annals of Operations Research, 283,737-757
25. Statista (2020a). <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>, last accessed on 01-12-2021
26. Statista (2020b). <https://www.statista.com/statistics/255146/number-of-internet-users-in-india/>, last accessed on 01-12-2021

27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I (2017). Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

28. WHO (2020). https://www.who.int/health-topics/air-pollution#tab=tab_1, last accessed on 01-12-2021.

29. Wiedemann G, Ruppert E, Jindal R, and Biemann C (2018). Transfer Learning from LDA to BiLSTM-CNN for Offensive Language Detection in Twitter. Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018). arXiv:1811.02906

30. Wu G, Tang G, Wang Z, Zhang Z, and Wang Z (2019). An Attention-Based BiLSTM-CRF Model for Chinese Clinic Named Entity Recognition. Special Section On Data-Enabled Intelligence For Digital Health, Vol. 7, pages 113942-113949

31. Xu C, Tong T, Zhang W and Meng M (2020). Fine-grained prediction of PM2.5 concentration based on multisource data and deep learning. Atmospheric Pollution Research, Volume 11, Issue 10, October 2020, Pages 1728-1737

32. Zhang L, Liu P, Zhao L, Wang G, Zhang W and Liu J (2021). Air quality predictions with a semi-supervised bidirectional LSTM neural network. Atmospheric Pollution Research, Volume 12, Issue 1, January 2021, Pages 328-339

33. Zhang Y, Wang J and Zhang X (2018). YNU-HPCC at SemEval-2018 Task 1: BiLSTM with Attention Based Sentiment Analysis for Affect in Tweets. Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), Pages 273–278

Figures

Figure 1

Classification model architecture

Figure 2

Flowchart of the PM_{2.5} concentration prediction model

Figure 3

(a) Time series of Class I tweet response and (b) Time series of Class II tweet response

Figure 4

Statistical variations of the tweet volume responses with PM_{2.5} concentration ranges for (a) Class I & (b) Class II.

Figure 5

Word clouds for (a) pre monsoon, (b) southwest monsoon, (c) post monsoon and (d) winter season.

Figure 6

Prediction accuracy of the model within different PM_{2.5} concentration categories