



Assessment of urban air quality from Twitter communication using self-attention network and a multilayer classification model

Thushara Sudheish Kumbalaparambi¹ · Ratish Menon¹ · Vishnu P Radhakrishnan² · Vinod P Nair³

Received: 9 January 2022 / Accepted: 29 August 2022 / Published online: 8 September 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Social media platforms are one of the prominent new-age methods used by public for spreading awareness or drawing attention on an issue or concern. This study demonstrates how the twitter responses of public can be used for qualitative monitoring of air pollution in an urban area. Tweets discussing about air quality in Delhi, India, were extracted during 2019–2020 using a machine learning technique based on self-attention network. These tweets were cleaned, sorted, and classified into 3-class quality viz. poor air quality, good air quality, and noise or neutral tweets. The present study used a multilayer classification model with first layer as an embedding layer and second layer as bi-directional long-short term memory (BiLSTM) layer. A method was then devised for estimating PM_{2.5} concentration from the tweets using ‘spaCy’ similarity analysis of classified tweets and data extracted from Continuous Ambient Air Quality Monitoring Stations (CAAQMS) in Delhi for the study period. The accuracy of this estimation was found to be high (80–99%) for extreme air quality conditions (extremely good or severe) and lower during moderate variations in air quality. Application of this methodology depended on perceivable changes in air quality, twitter engagement, and environmental consciousness among public.

Keywords BiLSTM · Delhi · Air pollution · PM2.5 · spaCy · Deep learning

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 (Leffel et al. 2022; Castells-Quintana et al. 2021; WHO 2020; Pant et al. 2019). 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 (Gholami et al. 2021; Shanavas et al. 2020; Xu et al. 2018). 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, and discussion forums. 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 behaviour on such platforms are not new to researchers and have been part of routine business analytics (Hassani and Mosconi 2022; Fan and Gordon 2014). Many studies have also explored their potential for monitoring environmental

Responsible Editor: Baojing Gu

✉ Ratish Menon
ratishmenon@scmsgroup.org

¹ Department of Civil Engineering, SCMS School of Engineering & Technology, Kochi, Kerala, India

² Department of Computer Science & Engineering, SCMS School of Engineering & Technology, Kochi, Kerala, India

³ Department of Computer Applications, Cochin University of Science & Technology, Kochi, Kerala, India

events such as natural disasters in real time (Yigitcanlar et al. 2022; 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 (Mohammadifar et al. 2021; Gurajala and Matthews 2018; Jackoway et al. 2011;). Many ML techniques are exploited curiously by new researchers are air quality prediction and forecasting (Xu et al. 2020; Chang et al., 2020; Zhang et al. 2021; Al-Janabi et al. 2020, 2021).

Use of attention network in deep machine learning algorithms is now widely recognized in sequence modelling (Bahdanau et al. 2014; Vaswani et al. 2017) to maximize performance in models. 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) have 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. Comprehensive analysis of various classifiers for air pollution detection is also under represented in the prior literature.

In the present work, Twitter responses related to air quality in Delhi were analyzed during 2019–2020 using a deep learning model to estimate $PM_{2.5}$ (mass concentration of particulate matter having an aerodynamic particle diameter less than $2.5 \mu m$) from the tweet content. 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 compare with tweet content to find its relationship with the pollution level. The study analyzed the fit and relevance of the physically monitored air quality parameter with the tweeting volume and behaviour of people in Delhi and developed 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.

Table 1 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 to 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. (2020)	$PM_{2.5}$	Local and neighbouring 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

A method was then devised for estimating PM_{2.5} concentration from the tweets using ‘spaCy’ similarity analysis of classified tweets and data extracted from CAAQMS in Delhi for the study period.

Methodology

Figure 1 shows the methodology framework followed in this study. Each step is explained in detail in the following sections.

Selection of study area and the social media platform

Twitter was chosen as the social media platform for the present study. Twitter is peculiar for its short and crisp contents limited over 280 characters and its prominence at the time of an event. In India, there were 13.15 million twitter users as of April 2020 (Statista 2020a). Many world leaders, governments, ministries, influencers, institutions, and news channels have their official Twitter accounts to make announcements, influence, and engage with the general public. 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. Tweets for a period of 12 months ranging from March 2019 to February 2020 were used as data in this study. Tweet response volumes, defined as the sum of direct and indirect response of people with respect to air quality, were estimated. Indirect responses included likes and re-tweets for a particular tweet. There has been works like that of Jiang et al. (2015) who have considered the effects of direct and indirect responses separately. However, in this study, it was decided to consider

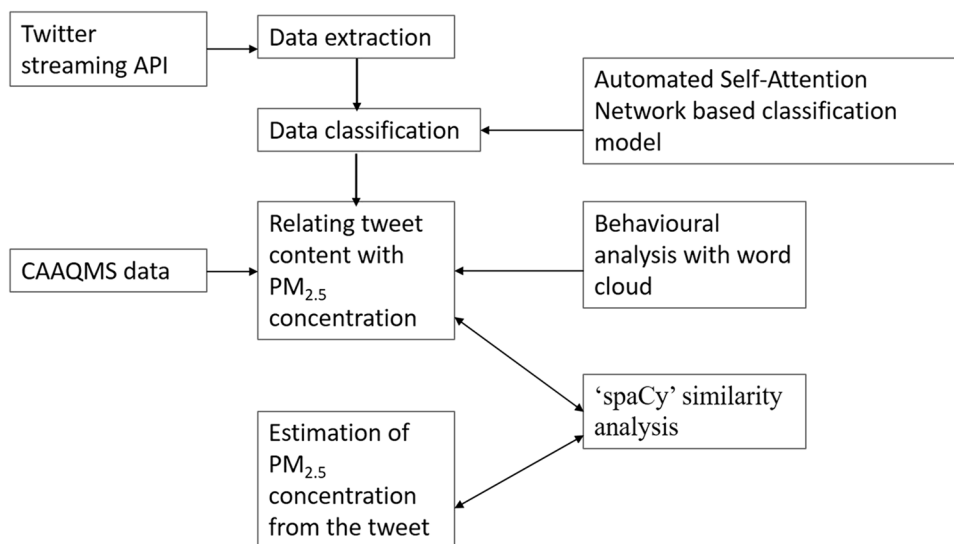
them as a collective response of the people of Delhi to the cities air quality issues.

Delhi, capital city of India, was selected as a representative urban area to demonstrate the methodology presented in this paper (Fig. 2). Delhi is bounded by the Indo-Gangetic alluvial plains in the North and East, by Thar desert in the West and by Aravalli hill ranges in the South and its land use comprises of residential, industrial, and commercial areas (Hang and Rahman 2018). Delhi has been witnessing tremendous population growth from 1.7 million in 1951 to over 16 million in 2011 (Census 2011). 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).

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

Fig. 1 Framework of the study



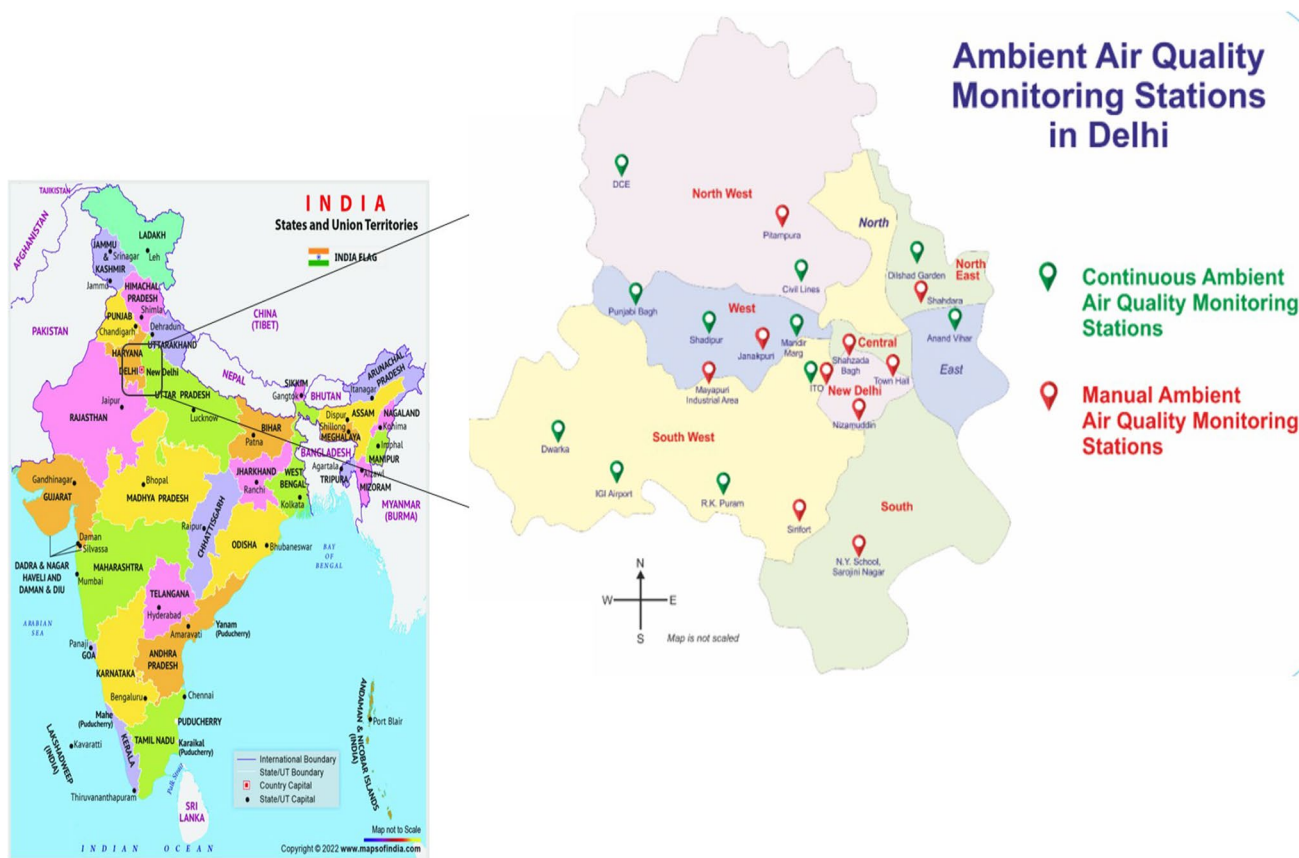


Fig. 2 Study area and air pollution monitoring stations in Delhi, India (Source: CPCB)

Table 2 Keyword combinations used to extract tweets during study period

	Keyword combinations	
Air Quality + Delhi	AirPollution + Delhi	DelhiChokes
AirQuality + Delhi	AirPollution + Delhi	Choke + Delhi
AirQualityDelhi	AirPollutionDelhi	Clean air + Delhi
DelhiAirQuality	DelhiAirPollution	DelhiEmergency
Delhi + Smog	DelhiSmog	Delhi + Right-ToBreathe

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. The tweets extracted over a year along with the likes and retweets appended to them comprised the dataset for the study. As initial step, the manually classified dataset was tokenized into words. Preprocessing was performed using Natural Language Toolkit (NLTK) library to remove special symbols, stop words, punctuations, Twitter handles, etc. 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.

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 analyzed during the current study are available from the corresponding author on reasonable request.

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. 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 are shown in Table 3. The classified set was then preprocessed and used to train a self-attention model that would automatically classify a sample tweet into its suitable classes.

Data analysis

Relationship between tweet responses and PM_{2.5} mass concentration

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. 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 ‘Temporal variation of tweet responses with PM_{2.5} concentrations’.

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 (0–30 µg/m³), satisfactory (31–60 µg/m³), moderate (61–90 µg/m³), poor (91–120 µg/m³), very poor (121–250 µg/m³), and severe (above 250 µg/m³) for plotting the box plots. The relationship of tweet response volumes of each class (Class I and II) in these categories was analyzed, and the results are given in ‘Statistical relationship of tweet response with PM_{2.5}’.

Behavioural 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. A python program with necessary

supporting libraries was utilized for the purpose. 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.

Automated data classification model

The dataset was a highly imbalanced multiclass data set with around 62,000 tweets in Class 0, 18,000 tweets in Class I, and 1700 tweets in Class II. 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 used a multilayer classification model. The network architecture is shown in Table 4. The first layer was an embedding layer in which an embedding matrix was 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 was a BiLSTM layer (units: 100) where the input sequences were 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 was 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 (Wf) for each word on the basis of both the forward analysis (f_F) and backward analysis (f_B) such that

$$Wf = f_F || f_B \quad (1)$$

where $||$ in Eq. (1) stands for concatenation function.

Similarly, a word feature was formed for every word of the data file. After this, the data flowed into a self-attention layer (SeqSelfAttention) layer where words were assigned weights and added to the resultant word features of the previous layer based on the relative importance of that word

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

Table 4 Network architecture (deep learning model)

Layer
Embedding: weight matrix using Glove used
BiLSTM: units: 100
SeqSelfAttention
Dense: (units: 50, activation: Relu)
Dense: (units: 50, activation: Relu)
Flatten
Dense: (3 (for class 0, class 1, class 2), activation: softmax)
Epochs = 5, batch size: 128, Optimizer: Adam.

to the entire tweet. This layer improved the efficiency of the classification as ‘attention’ was 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. (2020), 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/>) were used to support this layer.

The context vectors or sentence feature vector that were obtained as the output of the attention layer using weighted sum function was 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. Each 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 was obtained in a dense layer which had units equal to the number of output classes and was then given output activation by softmax function which provided the probabilities of the potential outcomes.

PM_{2.5} estimation from the tweet content

The methodology used similarity analysis to find the most similar tweets from a selected set of tweets to estimate the probable ground PM_{2.5} concentration that would have been recorded on that day. This model used ‘spaCy’ library available in python in order to find the top 10% of the most similar tweets out of a season wise selected set of tweets. A test tweet first undergoes model classification as mentioned in the previous section and then from the date of the tweet identify respective season for comparing with right set of tweets. The test tweet was then analyzed for similarity with each of the tweets in the selected set and a similarity index was assigned to each tweet in the selected set. The tweets are then arranged in descending of the cosine similarity index and the first 10% of this sorted set was taken as the resultant ‘most similar tweets

set’. The arithmetic mean of PM_{2.5} concentration as measured by CAAQMS stations on days corresponding to the tweets in the ‘most similar tweets set’ was then reported as the PM_{2.5} concentration which influenced the test tweet behaviour and its content.

The usefulness of this method was analyzed by measuring the correctly estimated concentrations as the percentage of the total estimations under each air quality category specified by CPCB (CPCB 2014) to understand the social media behaviour with respect to change in air quality (Fig. 6).

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.

Temporal variation of tweet responses with PM_{2.5} concentrations

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 was primarily due to lesser number of good air quality days in Delhi and also because good air quality days were 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 volume 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

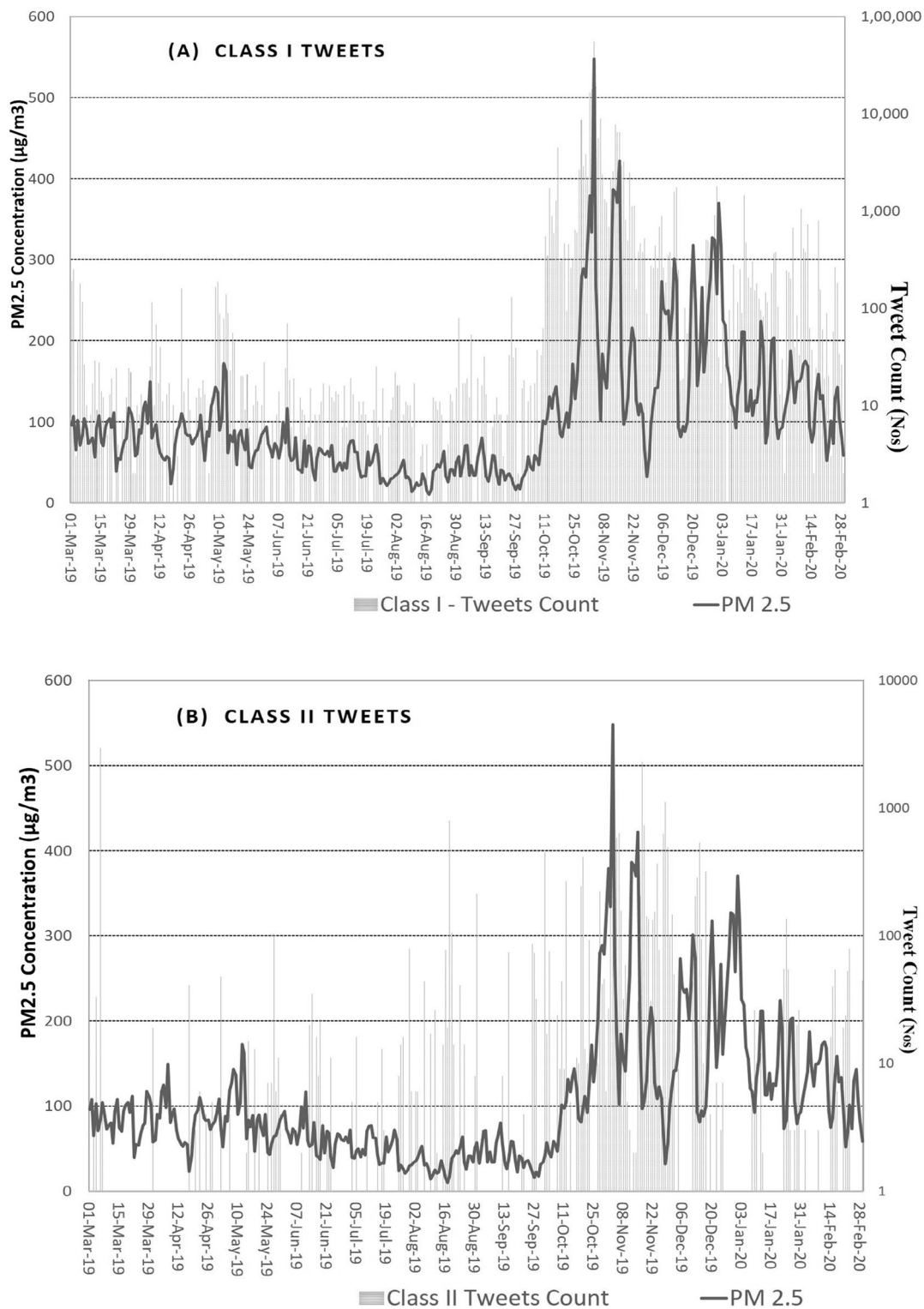


Fig. 3 Time series of PM_{2.5} mass concentration with **a** Class I tweets count and **b** Class II tweets count

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 are below $60 \mu\text{g}/\text{m}^3$ as specified in National Ambient Air Quality Standards by CPCB. This was the infamous times when the stubble burning in northern parts of India severely 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 50,000 on the most polluted day of the year when average $PM_{2.5}$ concentration over Delhi was around $550 \mu\text{g}/\text{m}^3$. 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, increasing 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 January 1, 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.

Statistical relationship of tweet response 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. 4a 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 behaviour 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

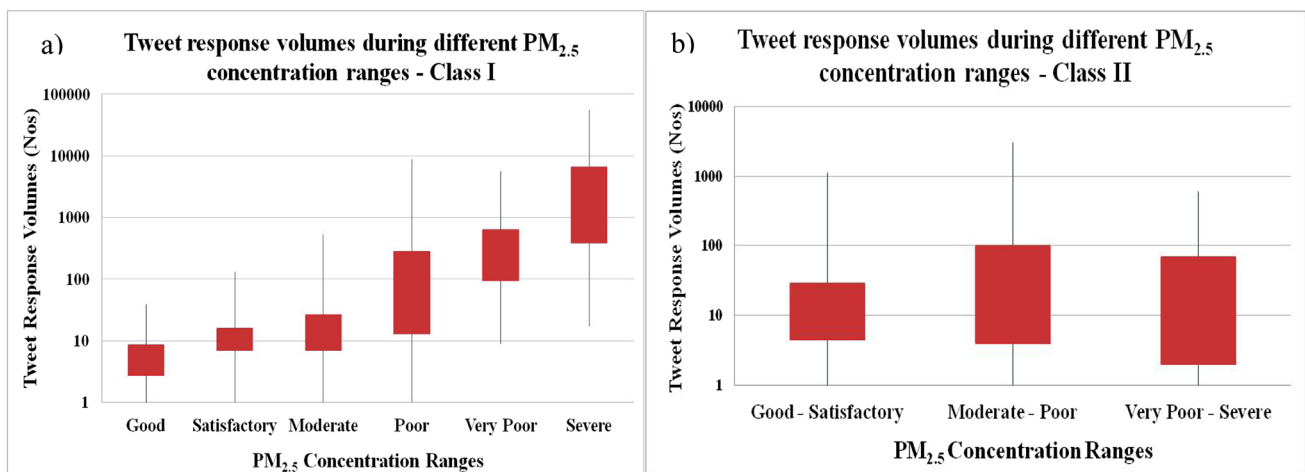


Fig. 4 Statistical variations of the tweet volume responses with $PM_{2.5}$ concentration ranges for a Class I and (b) Class II

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 Δ₃ and therefore had lesser tweet response volumes. During the very poor-severe air pollution days, the tweet 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.

Behavioural analysis from word clouds

Word clouds were generated from the model classified and 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 help in identifying the most frequent words in a season, the nature of emotions expressed through these words and 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 seems 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

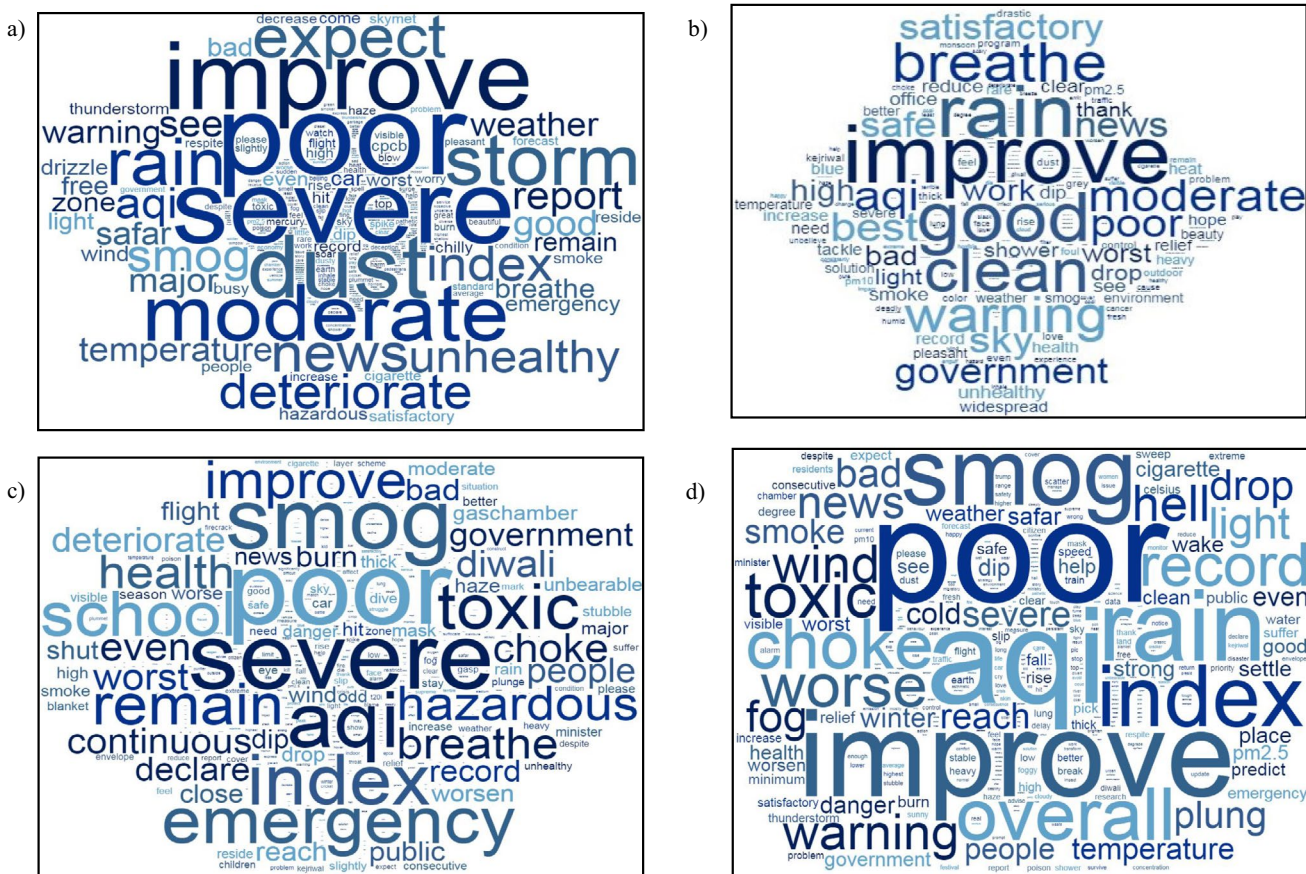


Fig. 5 Word clouds for a) pre monsoon, b) southwest monsoon, c) post monsoon, and d) winter season

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 word cloud 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 manly 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’.

PM_{2.5} estimation results

Deep learning-based model that can estimate PM_{2.5} concentrations from the content of tweets was developed in the present study. Figure 6 shows the percentage accuracy of these estimations for each PM_{2.5} concentration category (CPCB 2014).

Percent 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 easy to estimate with our prediction model. As the PM_{2.5} concentration moved into other categories, mixed tweeting behaviour 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.

Conclusions

Present study explored the social media behaviour 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

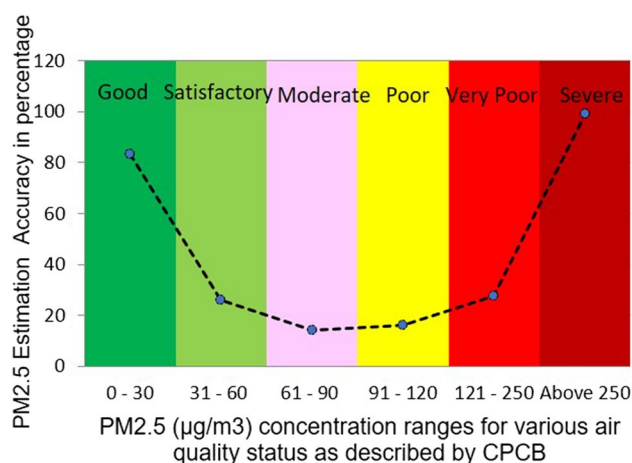


Fig. 6 Variations in accuracy of PM_{2.5} estimation from the tweet content by described methodology

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 is depicted through the words they use in a tweet and 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 estimation by this methodology showed that there was high estimation accuracy (above 80%) for extreme air quality conditions and lower accuracy for moderate air quality conditions. Similar results have been reported by Gurajala and Matthews (2018). For low PM_{2.5} values, the correlation with tweet numbers was poor and it improved with increasing PM. The observation of increasing correlation with increasing PM values is consistent with the findings of Jiang et al. (2015).

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 Indian 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 behaviour 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, and Instagram. Such other social media platforms may also be explored in the 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 estimation 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.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11356-022-22836-w>.

Author contribution Thushara Sudheish Kumbalparambi: data curation, formal analysis, writing - original draft. Ratish Menon: supervision, formal analysis, writing — original draft, writing — review and editing. Vishnu P Radhakrishnan: programming. Vinod P Nair: formal analysis, writing — review and editing.

Data availability PM_{2.5} data is available from CPCB at <https://app.cpcbcr.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.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable

Consent for publication Not applicable

Competing interests The authors declare no competing interests.

References

- Al-Janabi S, Mohammad M, Ali A-S (2020) A new method for prediction of air pollution based on intelligent computation. *Soft computing — a fusion of foundations*. *Methodol Appl* 24(1):661–680. <https://doi.org/10.1007/s00500-019-04495-1>
- Al-Janabi S, Alkaim AF, Al-Janabi E, Aljeboree A, Mustafa M (2021) Intelligent forecaster of concentrations (PM_{2.5}, PM₁₀, NO₂, CO, O₃, SO₂) caused air pollution (IFCsAP). *Neural Comput Appl* 33(1):14199–14229. <https://doi.org/10.1007/s00521-021-06067-7>
- Bahdanau D, Cho K, Bengio Y (2014). Neural machine translations by jointly learning to align and translate. Presented in International Conference on Learning Representations (ICLR)2015, arXiv:1409.0473. Available at <https://doi.org/10.48550/arXiv.1409.0473>
- Castells-Quintana D, Dienesch E, Krause M (2021) Air pollution in an urban world: a global view on density, cities and emissions. *Ecol Econ* 189:107153. <https://doi.org/10.1016/j.ecolecon.2021.107153>
- CCR-CPCB (2020) Central control room or air quality management-CAAQMS dash board, central pollution control board., Delhi. Available at <https://app.cpcbcr.com/ccr/#/caaqm-dashboard-all/caaqm-landing>. Accessed 12 Jan 2021
- Census (2011) Primary census abstracts, registrar general of India, ministry of home affairs, government of India. Available at <https://censusindia.gov.in/census.website/data/census-tables>. Accessed 12 Jan 2021
- Chang Y, Chiao H, Abimannan S, Huang Y, Tsai Y, Lin K (2020) An LSTM-based aggregated model for air pollution forecasting. *Atmospheric Pollution Research* 11(8):1451–1463. <https://doi.org/10.1016/j.apr.2020.05.015>
- CNN (2020) 21 of the world's 30 cities with the worst air pollution are in India. Article published in CNN Health on February 25, 2020. Available at <https://edition.cnn.com/2020/02/25/health/most-polluted-cities-india-pakistan-intl-hnk/index.html>. Accessed 12 Jan 2021
- CPCB (2014) National Air Quality Index. Control of Urban Pollution Series, CUPS/82/2014-15, Published by Central Pollution Control Board, India. Available at https://app.cpcbcr.com/ccr_docs/FINAL-REPORT_AQI_.pdf
- CPCB (2020) Monitoring network under national air monitoring program (NAMP), central pollution control board. Available at <https://cpcb.nic.in/monitoring-network-3/>. Accessed 12 Jan 2021
- Earle PS, Bowden DC, Guy M (2011) Twitter earthquake detection: earthquake monitoring in a social world. *Ann Geophys* 54(6):708–715
- Fan W, Gordon M (2014) The power of social media analytics. *Commun ACM* 57:74–81. <https://doi.org/10.1145/2602574>
- Gholami H, Mohammadifar A, Malakooti H, Esmaeilpour Y, Golzari S, Mohammadi F, Li Y, Song Y, Kaskaoutis DG, Fitzsimmons EK, Collins AL (2021) Integrated modelling for mapping spatial sources of dust in central Asia — an important dust source in the global atmospheric system. *Atmospheric Pollution Research* 12(9):101173. <https://doi.org/10.1016/j.apr.2021.101173>
- Gurajala S and Matthews JN (2018) Twitter data analysis to understand societal response to air quality. *SMSociety '18: Proceedings of the 9th International Conference on Social Media and Society*, July 2018, pp 82–90. <https://doi.org/10.1145/3217804.3217900>
- Hang HT, Rahman A (2018) Characterization of thermal environment over heterogeneous surface of National Capital Region (NCR), India using LANDSAT-8 sensor for regional planning studies. *Urban Clim* 24:1–18. <https://doi.org/10.1016/j.uclim.2018.01.001>

- Hassani A, Mosconi E (2022) Social media analytics, competitive intelligence, and dynamic capabilities in manufacturing SMEs. *Technol Forecast Soc Chang* 175:121416. <https://doi.org/10.1016/j.techfore.2021.121416>
- IQAir (2019) World air quality report- Region & city PM2.5 ranking. Published by IQAir. Available at <https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2019-en.pdf>. Accessed 12 Jan 2021
- Jackoway A, Samet H and Sankaranarayanan J (2011) Identification of live news events using Twitter. LBSN '11: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, November 2011, pp 25–32. <https://doi.org/10.1145/2063212.2063224>
- Jiang W, Wang Y, Tsou M-H, Fu X (2015) Using social media to detect outdoor air pollution and monitor air quality index (AQI): a geotargeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter). *PLoS One* 10(10):e0141185
- 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, 2627–2632. Available at <https://aclanthology.org/P19-1251/>
- Kent JD, Capello HT Jr (2013) Spatial patterns and demographic indicators of effective social media content during the Horseshoe Canyon fire of 2012. *Cartogr Geogr Inf Sci* 40(2):78–89. <https://doi.org/10.1080/15230406.2013.776727>
- Leffel B, Tavasoli N, Liddle B, Henderson K, Kiernan S (2022) Metropolitan air pollution abatement and industrial growth: global urban panel analysis of PM10, PM2.5, NO2 and SO2. *Environ Sociol* 8(1):94–107. <https://doi.org/10.1080/23251042.2021.1975349>
- Li R, Lei KH, Khadiwala R and Chang KC (2012) TEDAS: A twitter-based event detection and analysis system, IEEE 28th international conference on data Engineering, 1273–1276. <https://doi.org/10.1109/ICDE.2012.125>
- Lindsay BR (2011) Social media and disasters: current uses, future options, and policy considerations. Congressional Research Service, Washington, pp 7–5700
- Middleton SE, Middleton L, Modafferi S (2013) Real-time crisis mapping of natural disasters using social media. *IEEE Intell Syst* 29(2):9–17. <https://doi.org/10.1109/MIS.2013.126>
- Mohammadifar A, Gholami H, Golzari S, Collins AL (2021) Spatial modelling of soil salinity: deep or shallow learning models? *Environ Sci Pollut Res* 28:39432–39450. <https://doi.org/10.1007/s11356-021-13503-7>
- Pant P, Lal RM, Guttikunda SK, Russell AG, Nagpure AS, Ramaswami A, Peltier RE (2019) Monitoring particulate matter in India: recent trends and future outlook. *Air Qual Atmos Health* 12:45–58. <https://doi.org/10.1007/s11869-018-0629-6>
- 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. <https://doi.org/10.3115/v1/D14-1162>
- Poell T, Rajagopalan S (2015) Connecting activists and journalists: Twitter communication in the aftermath of the 2012 Delhi rape. *J Stud* 16(5):719–733. <https://doi.org/10.1080/1461670X.2015.1054182>
- Robinson EM, and Fialkowski WE (2010) Air Twitter: using social media and scientific data to sense air quality events. 2010 IEEE International geoscience and remote sensing Symposium, Honolulu. Available at https://wiki.esipfed.org/w/images/b/be/IGARSS_Abstract_AirTwitter.pdf. Accessed 12 Jan 2021
- 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, pp 851–860. <https://doi.org/10.1145/1772690.1772777>
- Shanavas AK, Zhou C, Menon R, Hopke PK (2020) PM10 source identification using the trajectory based potential source apportionment (TraPSA) toolkit at Kochi, India. *Atmos Pollut Res* 11(9):1535–1542. <https://doi.org/10.1016/j.apr.2020.06.019>
- Singh JP, Dwivedi YK, Rana NP, Kumar A, Kapoor KK (2019) Event classification and location prediction from tweets during disasters. *Ann Oper Res* 283:737–757. <https://doi.org/10.1007/s10479-017-2522-3>
- Statista (2020a) Number of active twitter users in selected countries. Available at <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>. Accessed 12 Jan 2021
- Statista (2020b) Number of internet users in India. Available at <https://www.statista.com/statistics/255146/number-of-internet-users-in-india/>. Accessed 12 Jan 2021
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L and Polosukhin I (2017) Attention is all you need. NIPS'17: Proceedings of the 31st international conference on neural information processing systems, December 2017, pp 6000–6010. <https://dl.acm.org/doi/10.5555/3295222.3295349>
- WHO (2020) WHO health topics- air pollution. Available at https://www.who.int/health-topics/air-pollution#tab=tab_1. Accessed 12 Jan 2021
- 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). <https://doi.org/10.48550/arXiv.1811.02906>. Available at <https://arxiv.org/pdf/1811.02906.pdf>
- Wu G, Tang G, Wang Z, Zhang Z, Wang Z (2019) An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. Special Section On Data-Enabled Intelligence For Digital Health. *IEEE Access* 7:113942–113949. <https://doi.org/10.1109/ACCESS.2019.2935223>
- Xu Y, Ho HC, Wong MS, Deng C, Shi Y, Chan TC, Knudby A (2018) Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. *Environ Pollut* 242:1417–1426. <https://doi.org/10.1016/j.envpol.2018.08.029>
- Xu C, Tong T, Zhang W, Meng M (2020) Fine-grained prediction of PM2.5 concentration based on multisource data and deep learning. *Atmos Pollut Res* 11(10):1728–1737. <https://doi.org/10.1016/j.apr.2020.06.032>
- Yigitcanlar T, Regona M, Kankanamge N, Mehmood R, D'Costa J, Lindsay S, Nelson S, Brhane A (2022) Detecting natural hazard-related disaster impacts with social media analytics: the case of Australian states and territories. *Sustainability* 14:810. <https://doi.org/10.3390/su14020810>
- 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. <https://doi.org/10.18653/v1/S18-1040>
- Zhang L, Liu P, Zhao L, Wang G, Zhang W, Liu J (2021) Air quality predictions with a semi-supervised bidirectional LSTM neural network. *Atmos Pollut Res* 12(1):328–339. <https://doi.org/10.1016/j.apr.2020.09.003>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.