SI: Networked Influence

SM+S
social media + society

Social Media + Society
July-September 2019: 1–14

DOI: 10.1177/2056305119867656
journals.sagepub.com/home/sms
SAGE

# Understanding Public Response to Air Quality Using Tweet Analysis

**Supraja Gurajala[1], Suresh Dhaniyala[2], and Jeanna N. Matthews[2]**

**Abstract**

Poor air quality is recognized as a major risk factor for human health globally. Critical to addressing this important public-health issue is the effective dissemination of air quality data, information about adverse health effects, and the necessary mitigation measures. However, recent studies have shown that even when public get data on air quality and understand its importance, people do not necessarily take actions to protect their health or exhibit pro-environmental behaviors to address the problem. Most existing studies on public attitude and response to air quality are based on offline studies, with a limited number of survey participants and over a limited number of geographical locations. For a larger survey size and a wider set of locations, we collected Twitter data for a period of nearly 2 years and analyzed these data for three major cities: Paris, London, and New Delhi. We identify the three hashtags in each city that best correlate the frequency of tweets with local air quality. Using tweets with these hashtags, we determined that people's response to air quality across all three cities was nearly identical when considering relative changes in air pollution. Using machine-learning algorithms, we determined that health concerns dominated public response when air quality degraded, with the strongest increase in concern being in New Delhi, where pollution levels are the highest among the three cities studied. The public call for political solutions when air quality worsens is consistent with similar findings with offline surveys in other cities. We also conducted an unsupervised learning analysis to extract topics from tweets in Delhi and studied their evolution over time and with changing air quality. Our analysis helped extract relevant words or features associated with different air quality–related topics such as air pollution policy and health. Also, the topic modeling analysis revealed niche topics associated with sporadic air quality events, such as fireworks during festivals and the air quality impact on an outdoor sport event. Our approach shows that a tweet-based analysis can enable social scientists to probe and survey public response to events such as air quality in a timely fashion and help policy makers respond appropriately.

**Keywords**

online social networks, Twitter, air quality, PM, data mining, machine learning, text classification, topic modeling

## Introduction

Ambient air pollution is one of the most important risk factors for public health globally (Prüss-Ustün, Wolf, Corvalán, Bos, & Neira, 2016). Among the different air quality parameters regulated by global environmental agencies, the mass concentration of particulate matter (PM) smaller than 2.5 μm, that is, $PM_{2.5}$, is one of the most significant from a health perspective (Kelly & Fussell, 2015). It is estimated that exposure to high PM pollution resulted in ~3.7 million premature deaths worldwide in 2012 (Prüss-Ustün et al., 2016) due to ischemic heart disease and strokes (80%), chronic obstructive pulmonary disease or acute lower respiratory infections (14%), and lung cancer (6%). Many (88%) of these deaths occurred in low and middle-income countries where air quality is poorest and monitoring is often inadequate.

Long-term epidemiological studies have established the severity of the air quality problem and informed scientists about the public-health crisis associated with increasingly poor air quality in the emerging economies, but it is unclear whether the general public recognizes and understands the problem (Bickerstaff & Walker, 2001; Oltra & Sala, 2015). Agencies have increasingly tried to bring air quality information to the public with alerts, monitors in public sites with air quality information, coverage in local newspapers, and

[1]SUNY Potsdam, USA
[2]Clarkson University, USA

**Corresponding Author:**
Supraja Gurajala, Department of Computer Science, SUNY Potsdam, 307 Dunn Hall, 44 Pierrepont Avenue, Potsdam, NY 13676, USA.
Email: gurajas@potsdam.edu

using simple color-coded indices ("Review of the UK Air Quality Index," 2011). In spite of these measures, people usually fail to minimize their exposure to air pollution on a daily basis (Bickerstaff & Walker, 2001) or take effective mitigation actions (Sawitri, Hadiyanto, & Hadi, 2015), resulting in air pollution exposure becoming a major public-health issue. To minimize air pollution–related health impacts, it is critical that public information about air quality be transmitted effectively and the response to this information be measured accurately.

Understanding the extent of public access to air quality information and their response and behavioral characteristics requires an extensive social surveying effort (Kelley, Clark, Brown, & Sitzia, 2003). Traditional survey tools, such as personal interviews (Zeidner & Shechter, 1988), have often been used in this context, to document feelings and sentiments experienced by people exposed to different levels of ambient air pollution. More recently, Carducci et al. (2017) analyzed data from a variety of informational sources in Italy, over a period of several months, to study coverage of air quality events and simultaneously used a traditional questionnaire approach to understand citizen awareness and interest toward air pollution issues. They determined that information about air pollution events, often obtained from traditional media, was focused on short-term events when air quality was high and based on alarmist reporting such as highlighting the limited usefulness of air pollution ordinances, while ignoring the role of individual behaviors. Individuals were seen to place the responsibility of pollution mitigation on political institutions rather than on themselves. A pro-environmental behavioral change by individuals is, however, critical if an effective environmental policy is to be developed to tackle air pollution (Sawitri et al., 2015), and thus, efforts to disseminate information about individual responsibility to tackle this problem is important, while also understanding its effectiveness.

Social media platforms such as Twitter and Facebook could be very useful to survey public response and to provide effective information dissemination. As an example, analysis of tweets during an earthquake event showed that information about the event traveled to the public sooner through Twitter than was possible from official agencies such as US Geological Survey (Earle et al., 2010). Recently, Twitter analysis has been extended not only to track events but also to understand human social interactions, perceptions, and sentiments (Prier, Smith, Giraud-Carrier, & Hanson, 2011; Salathé & Khandelwal, 2011; Signorini, Segre, & Polgreen, 2011).

In this article, we analyze air quality tweets to extract topics and understand how societal response to air quality changes in three global cities. This understanding will allow us to survey people's reaction to air quality changes and how they associate these changes with policies and pollution-related events. We discuss different text analysis techniques to extract topics from tweets and determine the evolution of topics with variation in air quality.

## Related Work

Twitter messages have long been analyzed to track public-health issues such as flu epidemics (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Culotta, 2013; Lee, Agrawal, & Choudhary, 2013; Nagel et al., 2013), smoking (Huang, Kornfield, Szczypka, & Emery, 2014; Myslín, Zhu, Chapman, & Conway, 2013; Prier et al., 2011), exercise (Zhang et al., 2013), and mental health trends (De Choudhury, 2013; Harman & Dredze, 2014) and personal health concerns such as cancer (Xu et al., 2016). Recently, Twitter analysis has been extended to not just track events but to understand human social interactions, perceptions, and sentiments (Prier et al., 2011; Salathé & Khandelwal, 2011; Signorini et al., 2011). Mining Twitter content data (tweets) can help us understand how societal response might change over time and help determine public understanding of the problem.

Air quality analysis from social media has been conducted by several researchers, primarily using Weibo data from China. Jiang, Wang, Tsou, and Fu (2015) used a manual approach to understand the sentiments expressed in posts during both periods of relatively good and poor air quality. They provided a qualitative understanding of the tweet content and manually classified post sentiment (positive or negative). They used the frequency of positive and negative posts as individual features in a machine-learning model and showed that it helped improve their correlation of air quality index (AQI) to number of tweets. They did not, however, identify any topics within the tweets or demonstrate the evolution of sentiments over time. Mei, Li, Fan, Zhu, and Dyer (2014) conducted regression analysis to predict air quality in Chinese cities using a spatio-temporal model built with social media and air quality data from several surrounding cities. From their regression analysis, they were able to identify the most correlated words in Weibo posts containing the Chinese character for air pollution (mai). They identified the that the highest positive weights were associated with poor air quality terms (e.g., haze, pollution, indoor) and the highest negative weights were associated with weather (e.g., sunshine, sunny, cold). They, however, did not classify the topics with the posts or analyze how they varied over time and with changes in air quality.

Sachdeva, McCaffrey, and Locke (2017) used social media data to understand population response to air quality during forest fires in California. They used a structural topic model (STM) to extract topics from tweets during one wild fire event and demonstrated that societal response to a particular event can be accurately captured using topic models. Here, we analyze Twitter data to determine underlying human behavior and changes in response characteristics with change in air quality and then discuss techniques to extract

topics from tweets and determine topic evolution with air quality changes and with time of the year.

## Data Collection

### *Tweets: Air Pollution Related*

For this study, we collected air quality–related tweets using Twitter's stream API from September 2015 to May 2018. An authenticated application connects to a public stream comprising of a sample of the tweets being posted on Twitter. A filter indicating which tweets are to be returned is included in each request. For our research, the tweets were filtered using a list of specific hashtags that we identified based on web search for popular air quality–related hashtags (e.g., RiteTag, 2015) and used in prior publications (e.g., Jiang et al., 2015). The list of search terms that were selected for our analysis is listed in Table 1. The JSON version of each tweet returned was saved to a MongoDb database (Pollution stream).

For this study, pollution-related tweets, totaling over 25 million, were collected over a period of 2 years (September 2015 to May 2018). In the first 13 months (September 2015 to November 2016), the tweets were collected sporadically. Over the last 18 months (November 2016 to May 2018), the tweets were collected continuously, except for the month of January 2017. The data over the entire time period are generated identically, that is, with the same hashtags and data collection speed. Also, the data collected in the initial time period are only a small fraction of the total data and have the same geographical spread as the rest of our data set.

**Table 1.** Hashtags Used to Build Our Database.

| | | |
|---|---|---|
| #AIRPOLLUTION | #OZONE | #POLLUTION |
| #AIRQUALITY | #HAZE | #SMOG |
| #CLEANAIR | #EMISSIONS | #PM25 |
| #PARTICLES | #PM2.5 | #PM10 |
| #PARTICULATES | | |

*Note.* When we search Twitter data for a hashtag, the search results will include tweets in which the hashtag term is used even if the user does not include the "#" symbol.

It is seen that most of our tweets associated with air pollution were collected from the United States, Europe, and India (Figure 1). As Twitter is largely inaccessible in China, tweets from China constitute only a small fraction (0.2%) of all our tweets. Among the three regions with a large number of tweets, the problem of air pollution is most severe in India and more severe in certain European cities than in the United States. For our analysis, we decided to concentrate on three major global cities: New Delhi, Paris, and London. These cities were selected because they have substantial air quality issues, accurate air pollution measurements are readily available at an hourly rate (or higher frequency), and air pollution in these cities varies significantly over the course of a year (Bohnenstengel et al., 2015; Deswal & Verma, 2016; Petit et al., 2015). The number of tweets analyzed for cities New Delhi, Paris, and London are listed in Table 2.

### *Air Quality Data*

For air quality, US Environmental Protection Agency (EPA) regulates the quality of ambient air based on six parameters—four parameters related to concentrations of different gases (CO, NO*x*, SO2, O3) and two related to airborne

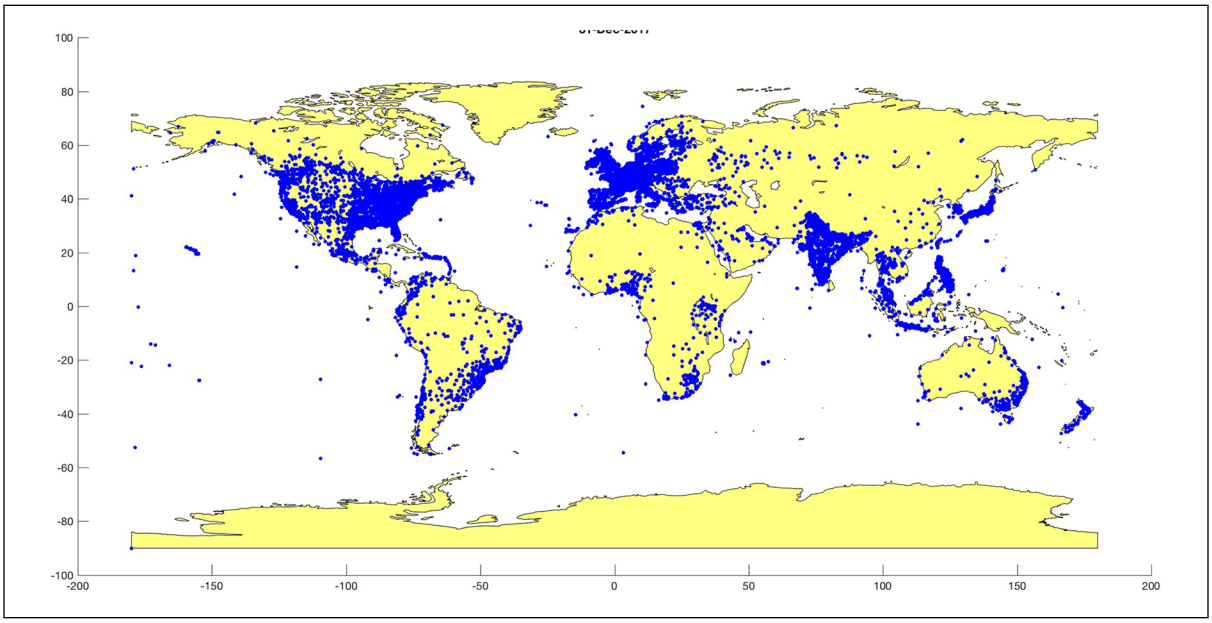


**Figure 1.** Global distribution of tweets analyzed in this study.

**Table 2.** Number of Tweets Collected for Each City.

| Place | Number of tweets |
|---|---|
| Delhi | 1,005,240 |
| Paris | 593,097 |
| London | 655,897 |

particles—mass of PM smaller than 2.5 μm or $PM_{2.5}$ and lead in particles. The combination of these parameters is reported as AQI, but this parameter (i.e., AQI) is calculated differently in different parts of the world. Here, we consider the air quality parameter that is the most important from a human health perspective—$PM_{2.5}$—for all our analysis. $PM_{2.5}$ is the most visible of all parameters—that is, changing $PM_{2.5}$ results in changing ambient light conditions with more polluted places having poorer visibility, and thus, this parameter is what largely drives people's perception of air quality.

The $PM_{2.5}$ data for the different locations were obtained from different data sites, including the US Embassy monitoring station for hourly New Delhi data (AirNow Delhi, 2016); Paris (City Center) site for hourly Paris data (AIRPARIF, 2016), and Farringdon St and Sir John Cass school sites for averaged 15-min London data (London Air, 2016). The air quality data were processed to a 1-hr time resolution for all sites.

## Temporal and Correlation Analysis

The $PM_{2.5}$ and the tweet data for the different sites were first processed to ensure that their sampling frequencies (or time periods) were matched. The $PM_{2.5}$ data were averaged over the selected time period, while the number of tweets was totaled during this time period. Care was taken to ensure that times for PM data (local time) were matched with the tweet times (UTC time). To illustrate the temporal trends in the $PM_{2.5}$ data and the number of tweets for the three sites, a comparison of the two data sets at low resolution (48 hr) is shown in Figure 2. All three cities show a correlation in temporal variation with the number of tweets. This provides some initial validation for our selection of hashtags related to air pollution.

To determine the hashtags most relevant for our study and to quantify the extent of correlation between the number of tweets for a selected hashtag and $PM_{2.5}$, we calculated Pearson's correlation coefficient between the data sets for a 6 hr time resolution. The choice of a 6-hr window was taken so as to smooth out noise in the PM data and improve statistics for the tweet data. As the tweets may either precede or follow an air quality event, the Pearson's correlation coefficient was determined as a function of time-shift between the two sets. For New Delhi, the correlation coefficient calculated for the hashtag "smog" as a function of time-shift is shown in Figure 3. Negative time-shifts represent tweets that temporally follow PM data. The maximum correlation coefficient and the associated time-shift are then noted for each hashtag and city.

For the three cities and all hashtags in Table 1, the two parameters, peak time-shift and the maximum correlation coefficient, were determined considering both the original tweets and retweets associated with each of the hashtags. Our results suggest that prediction of air quality from tweet data must consider time-shifts between events and their associated tweets. Most of the hashtags have a peak correlation when the tweets are ~6 to 24 hr after the event. We find that hashtags with a positive time-shift, that is, their peak correlation is when tweets precede an air pollution event, are either largely unassociated with the event (e.g., hashtags: haze or $PM_{2.5}$) or likely to be associated with public agencies responsible for air quality forecasts (e.g., $PM_{2.5}$ for New Delhi). Thus, we remove these hashtags from our dataset to focus our study on public response to air quality events instead of public agency response.

Furthermore, our results (Figure 4) show that, in general, the top three hashtags for each city have a similar strength of correlation with respect to each other. The data

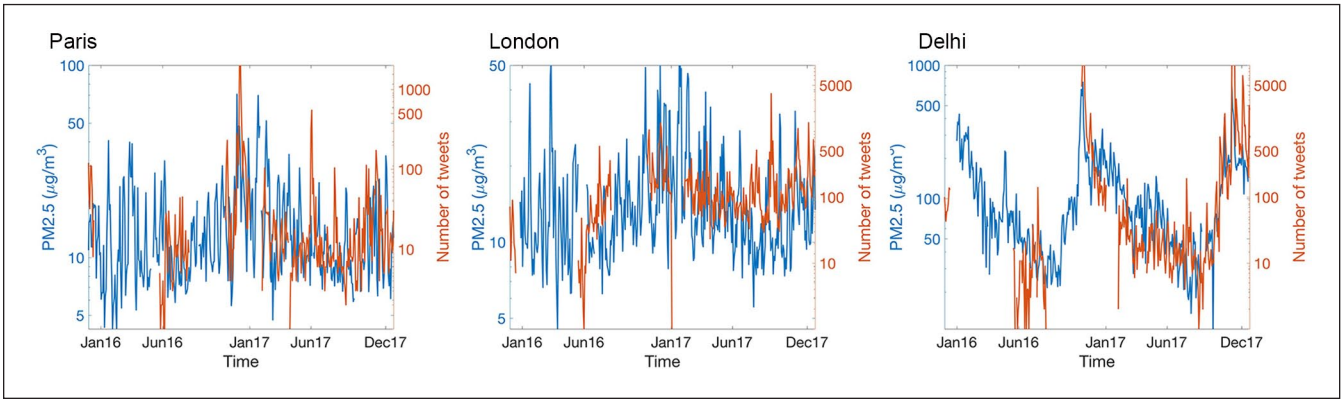


**Figure 2.** Temporal trends in air quality and the total number of tweets associated with selected air quality hashtags.
*Note.* The PM data are averaged over 48 hr and the tweet data are accumulated over the same time period. PM = particulate matter.

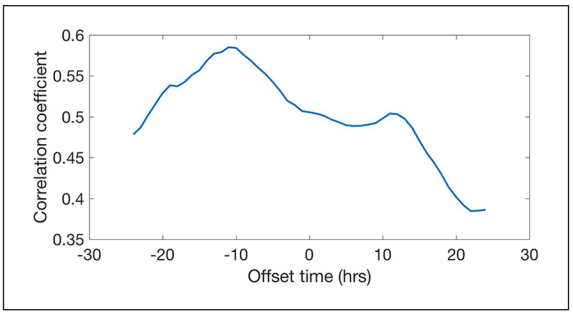


**Figure 3.** Time for a selected combination of hashtag (smog) and city (New Delhi).

were significant with $p < .01$ for all hashtags and city combinations. The top three hashtags were the same for New Delhi and Paris (air pollution, pollution, smog), but different for London (particles, $PM_{2.5}$, PM10). Considering that only a subset of hashtags have a reasonable correlation with $PM_{2.5}$, our further analysis was limited to tweets associated with these top-three hashtags in each city.

The correlation of number of tweets with $PM_{2.5}$ cutoff was studied. For this, we identified times when the $PM_{2.5}$ values were above a selected value and the correlation coefficient was then calculated. For all three cities, the correlation coefficient was seen to increase with increasing cutoff values of $PM_{2.5}$ (Figure 5). For low $PM_{2.5}$ values, the correlation was poor, but the correlations improved with increasing PM. The observation of increasing correlation with increasing PM values is consistent with the findings of Jiang et al. (2015) for data from Sina Weibo.

The public response to air quality (represented by increasing correlation of tweets to PM) occurs at much lower PM values in Paris and London than in New Delhi. When the PM values were normalized for each of the cities with their median values, the correlation coefficients were seen to all lie on the same line for the three cities (Figure 6). This result suggests that public response is driven by the relative difference in the $PM_{2.5}$ values that they experience rather than by the absolute values.

## Tweet Classification and Topic Modeling

Analysis based simply on tweet frequency does not allow us to understand the sentiment of the public response or characterize the evolution of topics over time. To understand societal response to air quality events, the contents of the tweets can be analyzed. In this study, we experiment with both supervised text classification and unsupervised learning classification of tweet contents. We try a variety of models and a number of different model parameters. In the previous section, we had real measurements of air quality with which we could compare our results, but we do not have ground truth data for popular sentiment, and thus, it is harder to pick a "best" model. As a result, we chose to experiment with a variety of models in this study and compare the results obtained.

Prior to analysis, the tweets were first preprocessed to remove handles, retweet symbols, URLs, emojis, sentences containing single word, and extra spaces. We then extracted features from these preprocessed tweets using a bag-of-words (BoW) representation. We used Natural Language Processing Tool Kit (NLTK) (Bird & Loper, 2004) for preprocessing and feature extraction.

### *Supervised Learning*

*Supervised learning algorithms.* Supervised learning text classification algorithms are machine learning tools that can be used for tweet classification based on training data. For supervised learning, we classified the tweets into one of three classes: health, climate, or politics. We picked these topics because airborne particles play an important role in both public health and climate change (Orru, Ebi, & Forsberg, 2017) and government policy is a critical driver of change/action associated with these topics.

To build a training set, we first selected search terms for each topic. For the topic "health," the search terms were health, sick, disease, and lung. For the topic "climate," we just used the term climate. For "politics," we used different terms for different cities. For Delhi, we used the terms politics,

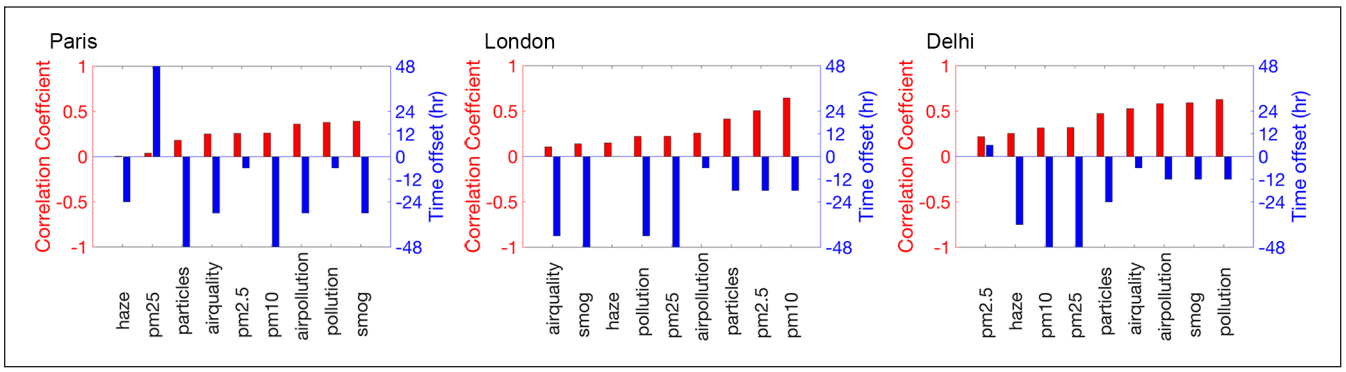


**Figure 4.** The correlation and time-shifts associated with different hashtags and cities studied.

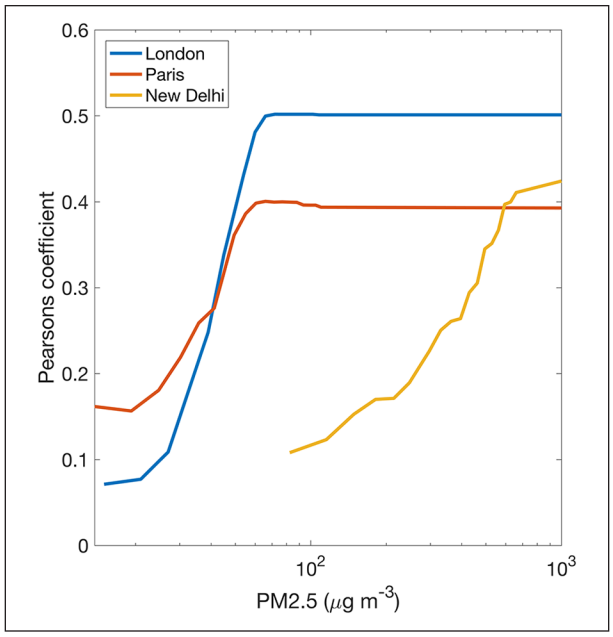


**Figure 5.** Change in correlation coefficient with $PM_{2.5}$ cutoff values.
*Note.* At any given $PM_{2.5}$ value (*x* axis), the correlation coefficient was calculated for all tweets at times when the PM values were greater than the cutoff value. PM = particulate matter.

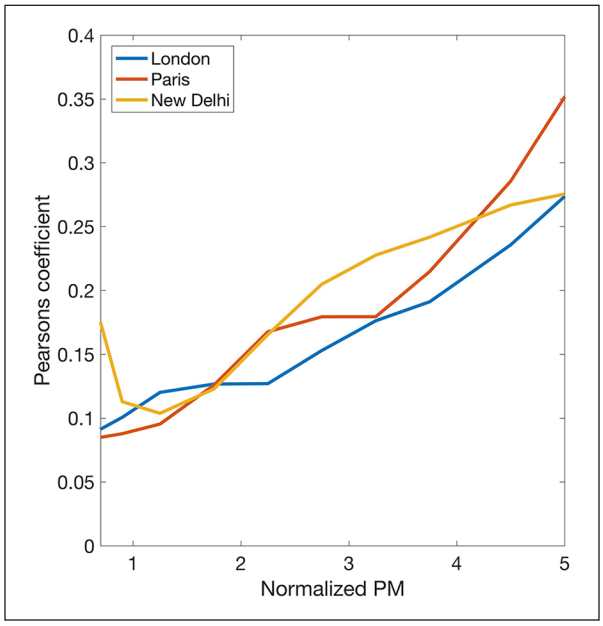


**Figure 6.** Change in correlation coefficient with normalized $PM_{2.5}$ cutoff values.
*Note.* At a selected $PM_{2.5}$ value (*x* axis), the correlation coefficient was calculated for all tweets at times when the PM values were greater than the cutoff value. PM = particulate matter.

government, policy, Modi, and Kejriwal, where the last two search terms are the names of the leaders of India's central government and Delhi's local government. For London, we used the terms politics, government, policy, and Brexit (as this was a popular political topic at that time), and for Paris, we just used politics, government, and policy. For a training data set, we obtained 200 random tweets for each class using these search terms. The training set tweets produce around 1,500 features per topic per city, with sample features shown as word clouds in Figure 7.

We then determined the number of tweets in each of the classes for normalized $PM_{2.5}$ value ranges of 0 to 1, 1 to 2, 2 to 3, 3 to 4, and >4 according to a variety of supervised learning algorithms including Bernoulli Naïve Bayes (BNB), Multinomial Naïve Bayes (MNB), and Support Vector Classifier (SVC).

We describe each of these three models below, but first, we begin with a description of Naïve Bayes (NB) which lays the foundation for understanding some of the others.

*Naïve Bayes*. This is a simple (naïve) classification method that uses Bayes' rule of independence of features or words to categorize tweets. NB classifiers make the assumption that the order of words in the tweets do not matter, that is, a "bag of words" assumption is made.

The tweets are classified into one of three categories (health, politics, and climate) using Bayes' Theorem, expressed as

$$C_{NB} = \arg\max P(c) \prod_{w \in W} P\left(\frac{w}{c}\right)$$

where $C_{NB}$ is the selected category or class for the tweet, $C$ is one of the three categories considered here, and $W=(w1, \ldots, wn)$ is the feature or word vector associated with a tweet. In the above equation, the NB (Spiegelhalter & Knill-Jones, 1984) assumption of conditional independence is made, that is, the probabilities $P(w|c)$ are independent of the category $c$. NB is often the first-choice algorithm for text classification as it is robust to irrelevant features, has low amount of data, and can handle classification even when many features with equal importance exist. NB, however, has some well-recognized problems, particularly the assumption of feature-independence (McCallum & Nigam, 1998). But in spite of this problem, NB has been popular for text classification because of its simplicity, its fast speed, and low storage requirements.

*Bernoulli Naïve Bayes*. In the BNB model, the NB algorithm is used with a multivariate Bernoulli distribution for the feature set. In this model, the features are all assumed to be binary-valued variables. Thus, multiple occurrence of a word in a tweet is no different from a single occurrence. The decision rule for BNB is based on

$$C_{BNB} = \arg\max_{c \in C} P(c) \prod_{w \in W} P\left(\frac{w}{c}\right) \prod_{w \in W} \left(1 - P\left(\frac{w}{c}\right)\right)$$

where $C_{BNB}$ is the selected category or class for the tweet. BNB model is best for short documents such as tweets,

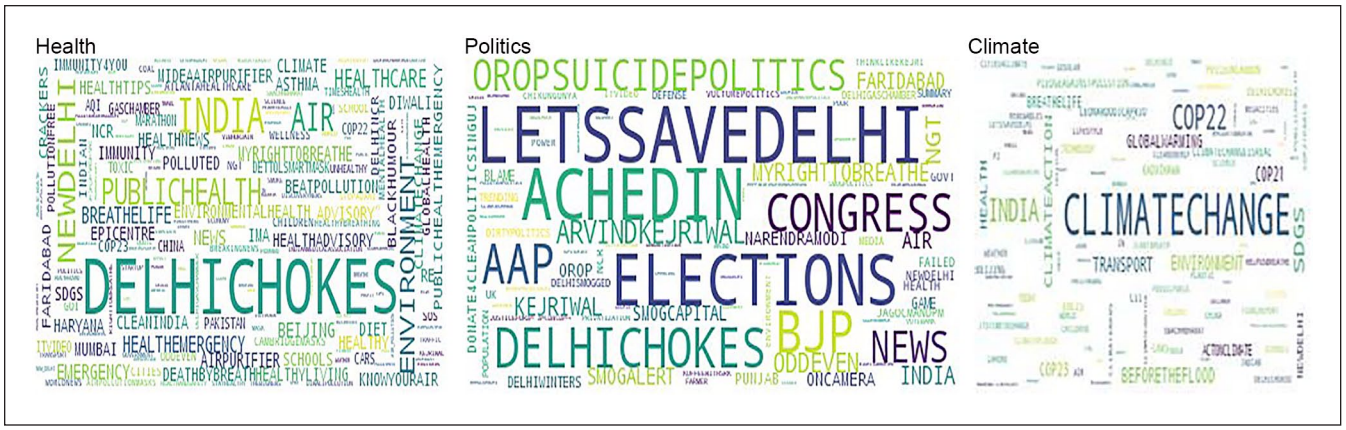


**Figure 7.** Word cloud for the three classes for New Delhi.
*Note.* The words associated with the top three hashtags are removed. The search term is also not shown, for example, the term "health" is not shown in the word cloud for class "health."

where occurrence of multiple instances of a word is unlikely or possibly unimportant.

*Multinomial Naïve Bayes.* A more appropriate algorithm for text categorization is MNB, where a multinomial probability is assumed for the features, accounting for multiple instances of a feature (word) being present in the document (tweet) (Eyheramendy, Lewis, & Madigan, 2003). In the MNB model, the NB algorithm is used with a multinomial distribution for the feature set. In this model, the tweets are represented by a feature vector of integer elements that are the frequency of a word in the tweet

$$C_{MNB} = \arg\max_{w \in W} P(c) \prod_{w \in W} P\left(\frac{w}{c}\right)^{N}$$

where $C_{MNB}$ is the selected category or class for the tweet and $N$ is the number of times that $w$ appears in a tweet. In MNB, the word positions in a tweet are recorded and the frequency of the words is used. To avoid the problem of zero probability when a word does not occur in a tweet, Laplace smoothing is used. The MNB model generally performs better with longer documents.

*Support Vector Classifier.* SVC is a supervised learning method that is particularly effective in high dimensional spaces, that is, when there is a large feature set. In SVC, learning data are used to determine decision boundaries or hyperplanes to separate tweets into the selected categories (Cortes & Vapnik, 1995). SVC can classify documents even with very low ranked features (i.e., a dense sample) and a small set of support vectors (i.e., sparse data) (Joachims, 1998) as is the case for the air quality tweet-based analysis conducted here. We use a Linear SVC as it has been shown to be as accurate as a non-linear model when the feature set is large, as is the current case (Hsu, Chang, & Lin, 2003). SVC does not assume that the features are independent of each other and is optimal for use in cases where the features have some interaction between themselves.

*Results.* For all three models (BNB, MNB, and SVC), the dataset was randomly split into two groups: 90% of the data used for training and 10% used for evaluation. We trained the model 15 times and calculated the accuracy to be greater than 80% for all of the models, comparable to the recall efficiency of other classification studies (e.g., Middleton, Middleton, & Modafferi, 2014). The algorithms were implemented using Scikit-learn library in Python (Pedregosa et al., 2011).

We first analyzed the New Delhi data set that we created considering only the top three correlated hashtags (air pollution, pollution, and smog). For each of the discrete normalized PM levels considered (0-1, 1-2, 2-3, 3-4, >4), we collected a maximum of 10,000 tweets, for computational simplicity. Because the time periods associated with the different PM levels were not the same, we did not always get 10,000 tweets for all levels but the tweets were no less than 6,500. In particular, the highest air quality levels were only for a relatively limited time period, and this limited the number of tweets associated with this level.

The tweets were then classified using the three algorithms into one of the classes: health, climate, politics, or other. The fraction of tweets in each class (normalized by the total of the three classes: heath, climate, and politics) is shown in Figure 8. Interestingly, with increasing PM levels, the fraction of tweets related to health and politics increases, while the fraction for climate decreases.

While the three models are different in their predictions of the fractions of tweets in each of the categories, they predict the same trends. The people in New Delhi seem to tweet more about health as PM levels go above the median value, suggesting some recognition or concern of health effects of air pollution. The simultaneous increase in tweets related to politics suggests that the people want the government to take

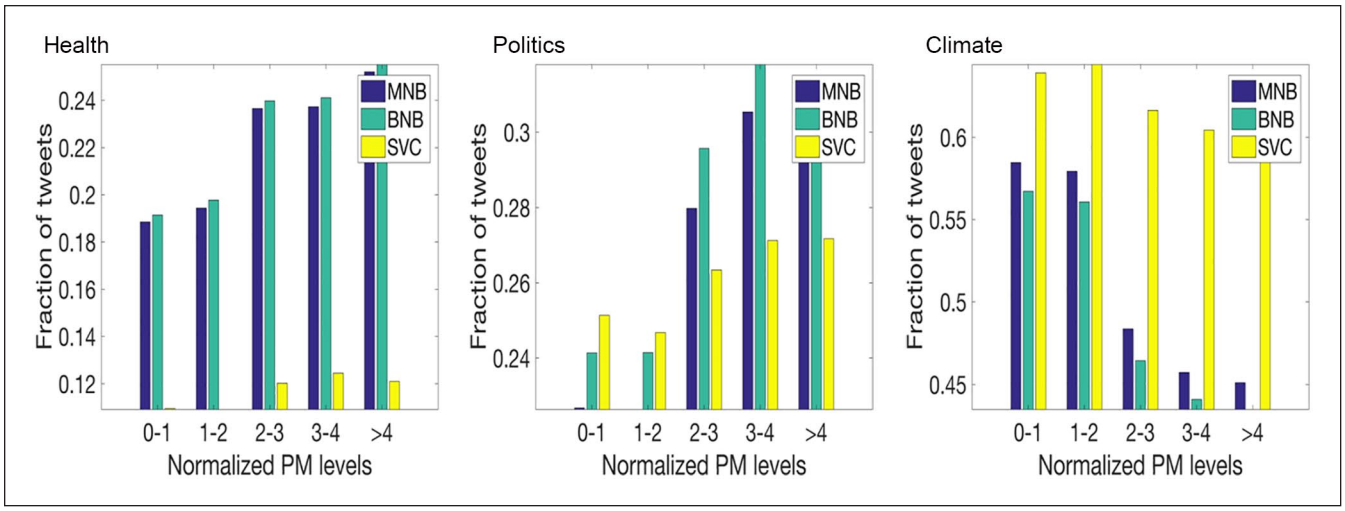


**Figure 8.** The fraction of tweets associated with the three classes as predicted by the different machine learning algorithms as a function of normalized PM levels.
*Note.* PM = particulate matter.

action or they are blaming the politicians for the inaction. The tweets related to climate decrease with increasing PM level, suggesting that when air pollution is high, the primary concern is the acute problem of health, rather than the long-term problem of climate change.

The primary public concerns at high $PM_{2.5}$ levels (normalized values > 4) can be visualized in the word clouds shown in Figure 7, where the words are sized by their frequency. In these word clouds, terms related to the hashtags (air pollution, pollution, and smog) and the classes (e.g., the term "health" for the class "health") are removed. Among the words in the "health" collection include asthma, breathe, health emergency, immunity, and so on, all pointing to severe health concerns. In the "politics" collection, words include political party names (AAP—the political party leading the Delhi state government, BJP—the political party leading the Indian national government, Congress—the main opposition party at the Center, or Federal level), government policies (Odd-Even, achedin—roughly translates to "Good days," was the slogan of the central government ruling party in the last election cycle), petitions (e.g., my right to breathe), and so on, possibly suggesting that the public believe that the pollution should be tackled politically and with policies. The "climate" word collection also has a mix of climate-related terms (COP21) and some air quality–related issues (transport, environment). The trends in the classes of health and politics with increasing air quality suggest that the public focus is on the acute problem (health) and the burden of mitigation is placed on the institution, similar to the findings of Carducci et al. (2017) in Italy.

For the other cities, we followed the same procedure as for New Delhi and calculated the fractions of tweets for the different classes. We then calculated the trends in the three classes with respect to PM levels based on an ensemble

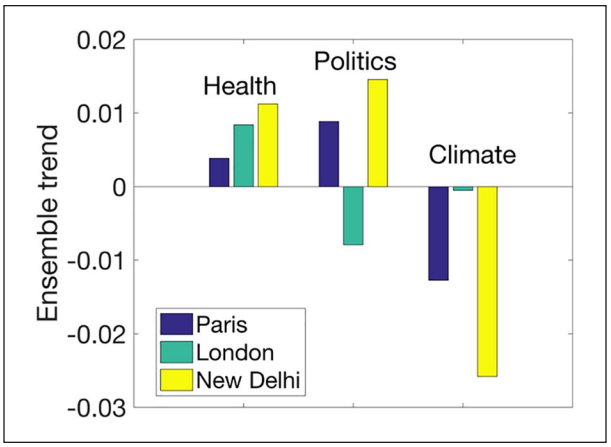


**Figure 9.** The linear trendline slopes for each class and city combination considering an ensemble of the model predictions.

average of the predictions of the three algorithms. The slopes of the linear trend lines for each class and city is shown in Figure 9. Positive correlations represent an increasing recognition of the topic with increasing PM, while negative correlations would represent a decreased interest in the topic with increasing PM. All three cities show a positive trend for health, suggesting that the recognition of the correlation of PM to health is universal, with the strongest correlation being for New Delhi. The people in New Delhi are more concerned with their increase in PM, from a base level to 4 times higher, than the public in the other cities. The trend for climate is negative for all three cities. This is possibly because some climate terms are also related to weather and it has been shown previously (Mei et al., 2014) and also from our unsupervised learning results in the next section, that improved air quality is related to certain weather conditions

(e.g., rain, wind, etc. are associated with reduced air pollution). People in New Delhi and Paris seem to associate politics/government with poorer air quality, while this is not the case in London. We do want to note that we had greater familiarity with politics in Delhi and subsequently picked more relevant search terms for this city (e.g., Kejriwal, Modi) than for others and this could have resulted in a "richer" data set for Delhi compared to the other cities.

The current study suggests that there is some commonality in the three global cities in the public response to air quality as indicated by their similar increase in tweet frequency with normalized PM levels. It is important to note again that the similarity in tweet frequency response is when the air quality values were normalized.

There are also some differences in the global response, with people in New Delhi having the greatest health concern when their PM values increase above the median or typical values. The tweet analysis also seems to indicate that the public associates poor local air quality to local politics in New Delhi, but this is not universally observed. The public response to increasing PM values suggests that there is significant awareness of the air quality problem when the values are high, though it is not clear if people are taking mitigation measures to avoid exposure.

## Unsupervised Learning

With supervised learning, we captured responses to selected topics that were pre-identified and pre-associated with selected search terms. Another way to analyze tweets is to determine topics based on patterns in words within documents, and the relation between different words and the probability of them occurring together in our document. This unsupervised topic modeling approach is used here to determine the wide range of topics associated with air quality in Delhi and the evolution of these topics over time and with air quality.

*Latent Dirichlet Allocation (LDA).* Here, we consider the popular unsupervised learning approach of LDA (Blei, Ng, & Jordan, 2003) model. The LDA model is a hierarchical document topic model that is designed to find topics within documents and word probabilities within the topics. This model assumes a standard BoW representation with the collection of tweets forming a document that is represented as a vector of word counts. The LDA model is one of the most popular models to discover topics from documents (e.g., Doshi Velez, Wallace, & Adams, 2015; Lim, Chen, & Buntine, 2016; Pavlinek & Podgorelec, 2017). LDA is a generative model that identifies topics by recognizing formal statistics-based relationships between the words in the tweets and specifies a probabilistic procedure to generate the topics.

One challenge with the LDA model is that the number of topics identified by LDA is a user input. Specifying a large number of topics results in generating some irrelevant topics, while specifying a small number of topics will result in creating only broad topics. Optimizing the number of topics is important, though not easy. Here, we tried topic numbers ranging from 3 to 20, and when we used a small number of topics (3), we only extracted broad topics, and with a large number of topics (20), it was difficult to distinguish between topics (i.e., the features making up different topics seemed to be similar). We finally selected 10 topics, after manually determining that this selection allowed us to extract seasonal topics associated with air quality (without overwhelming us with a large number of near identical topics).

For our topic modeling, we chose New Delhi because air quality–related tweets were highest in number from this location and also because New Delhi has the worst air quality of the places studied. We classified the Delhi tweets based on the place location information provided by the users. We classified the tweets into one of three categories: if the users provide their place coordinates at the time of the tweet as Delhi, these were classified into a set labeled as $S^1$; if the users provided a manual location information in their profile as Delhi, these tweets were classified into a set labeled as $S^2$; and if they only referred to Delhi in the tweet text, but did not provide any locational information in their profile or as their coordinates, then the set was labeled as $S^3$.

Our findings suggest that there is value in looking at all three data sets. The $S^1$ data set is the smallest data set but most closely related to the city of interest. This dataset provides the views of local people and is important when probing local issues such as those related to local governance. The $S^2$ data set represents people associated with the city, but not necessarily living there. This data set is often neglected in analysis requiring tweet location, but because of its large size, this data set can provide more topics than $S^1$ (Gurajala, 2018) and hence is valuable for classification studies. The $S^3$ data set provides no information about user location and will likely contain a significant number of tweeters from outside the location of interest. Topics from this data set about a selected location will likely be relevant or of importance to a global audience.

*Results.* For each of the three datasets, we first generated a BoW model from the documents and then ran the LDA model on each of the BoW data sets. The LDA data set is visualized using a combination of word cloud images and time series plots. We analyzed Delhi data, and selected 10 topics to identify for each of the data sets associated with the different geo-location information data sets—that is, tweets with gps-coordinates ($S^1$), manually entered user locations ($S^2$), and place name in tweets ($S^3$).

For the $S^1$ data set (i.e., tweets with gps-coordinates), the list of words and time series variation associated with two topics are shown in Figure 10. The word cloud in Figure 10a suggests a topic of government/politics while the topic associated with Figure 10c seems to be weather related. The time series plot of the topics is compared with PM values.

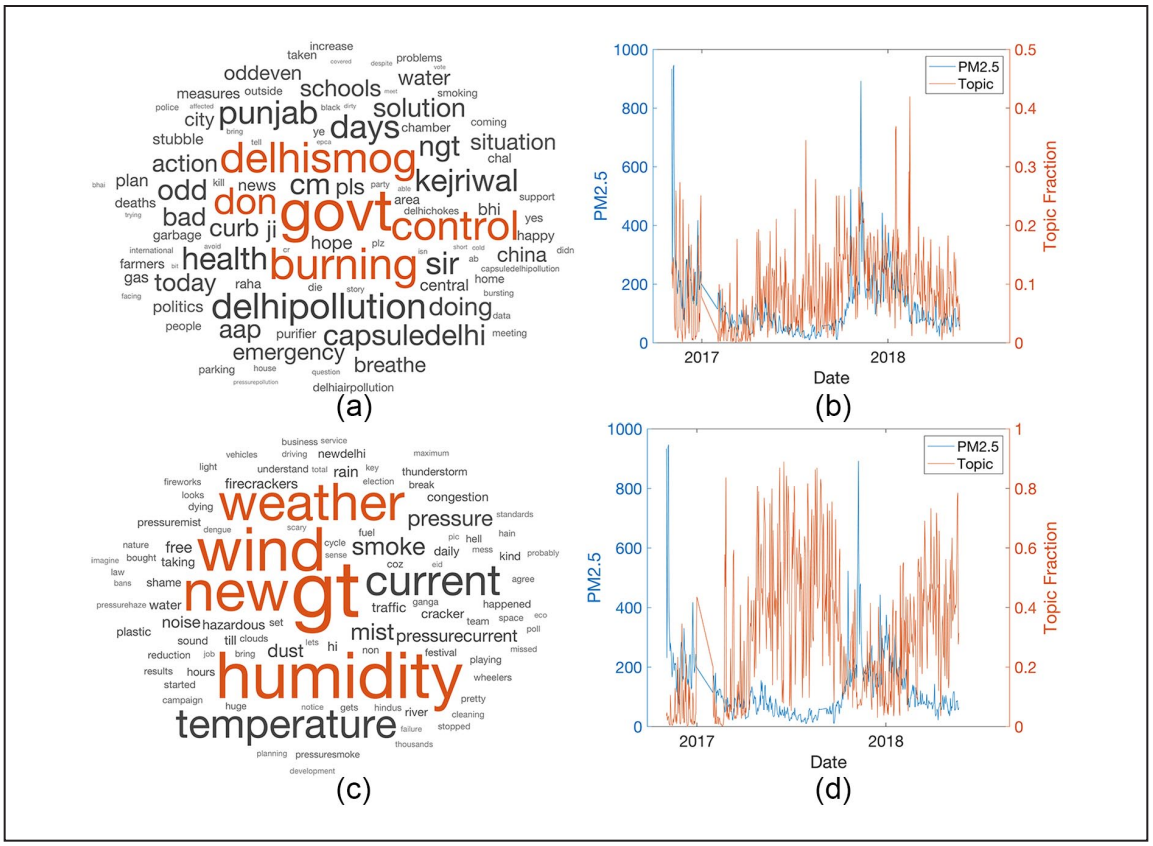


**Figure 10.** Topics extracted from Delhi $S^1$ data (place coordinates) resulted in a "richer" data set for Delhi compared to the other cities: (a) word cloud: government; (b) the time series plot of the government topic frequency and PM data showing largely similar baseline behavior of the topic with air quality; (c) word cloud: weather; and (d) associated time series plot of the weather topic frequency and PM data shows a near inverse relationship that has been identified in earlier air quality studies.
*Note.* PM = particulate matter.

When the PM levels go up, the discussions related to government/politics (Figure 10b) seem to spike. One observation that could be made is that the government/politics spike is greater during the first burst of the event (November 2017) than during a similar burst later (January 2018). This could suggest that the initial intense response to a bad air quality event fades over time. The time series plot of the weather topic (Figure 10d) shows a negative correlation of the topic with PM. A decrease in PM levels is often associated with good weather, as has been observed by other authors (Mei et al., 2014).

For the tweets with manually entered user locations (i.e., $S^2$), the list of words and time series variation associated with two topics are shown in Figure 11. We identify the topics as policy (Figure 11a) and festivals (Figure 11c). Particularly interesting is the topic of festivals, which shows a spike in the fraction of tweets related to a popular Indian festival, Diwali, when air quality related to fireworks is a major discussion issue in India. The LDA model is able to pick up this topic accurately. Diwali is an example of a topic that we would not have thought to pick for our study with the supervised models, but is automatically identified by the unsupervised model. Thus, unsupervised model provides insight into the data and identifies topics that we might not recognize to begin with. On the contrary, analysis of the words within the topic using supervised models can allow policy makers (and others) to further probe a specific desired topic. Unsupervised helps you find information you did not know to look for and supervised helps you answer pointed questions that you had before coming to the data (e.g., Are users more concerned about the impact on their health or climate change?).

From the data set with the place name in the tweet ($S^3$), the list of words and time series variation associated with two topics are shown in Figure 12. We identify the topics as cricket (Figure 12a) and health (Figure 12c). In winter of 2017-2018, there was an international cricket match in Delhi during a bad air quality event when several players were taken sick, and our analysis of the tweets from the tweets with place name in the text (i.e., $S^3$ data set) picked out this topic accurately. In Figure 12d, the time series variation of the topic shows that when there are short PM spikes (e.g Nov 2017), the health-based discussion peaks. When there is a sustained high PM (e.g., January-February 2018), however,

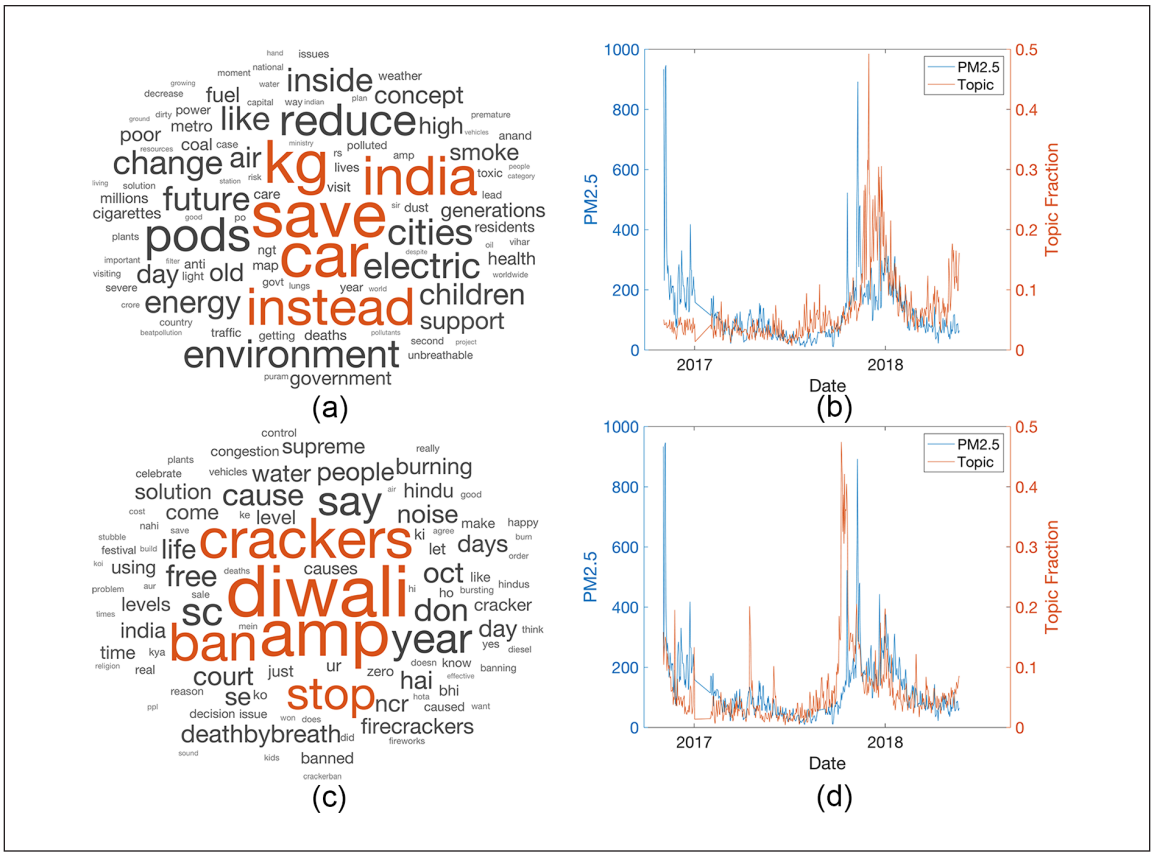


**Figure 11.** Topics extracted from Delhi $S^3$ data (manually entered user locations): (a) word cloud: policy; (b) associated time series plot of the Policy topic frequency and PM data shows that this topic trends with air quality, with large spikes during the poor air quality periods around January 2018; (c) word cloud: festivals; and (d) associated time series plot of the Festivals topic frequency and PM data shows that this topic trends with air quality changes during a festival when fireworks-related air quality issues arise.
*Note.* PM = particulate matter.

there is decreased interest in the health topic. Further analysis of words within the topics can provide social scientists with an understanding of public sentiment associated with these topics.

## Conclusion

In this article, we establish the possibilities that social media data mining provides to understand societal response to air quality. We analyzed 2 years of Twitter data in three diverse cities (Paris, London, and New Delhi) to determine similarities and differences in public response to air quality information. The number of tweets with just three hashtags (the top three for each city) was shown to be highly and significantly correlated to PM values. Using these best performing hashtags, and normalized PM data, the correlation coefficients suggested that the public in the three cities responded similarly to relative changes in air quality rather than absolute levels.

We further analyzed the societal response to air quality using text classification and topic modeling techniques. Using a text classification, supervised learning approach, we classified the tweets into one of four topics—health, climate, politics, or other. We used three different models to classify all the tweets into one of the selected topics and determined that people in different cities responded differently to air quality in terms of the three topics. In all cities, with increasingly poor air quality, there was an increasing concern about health, but only in Delhi and Paris was there a call for more government action as air quality worsened. In all cities, with increasingly poor air quality, the topic related to weather/climate became less important. Thus, supervised learning provides a means to ascertain changes in importance of a topic of interest to us. Unlike comparing our tweet frequency analysis to real air quality sensor data, ground truth data on public sentiment or topics would be challenging to obtain. In this study, we experimented with a variety of supervised learning models and compared their results. We also used an ensemble of the average prediction from the combination of the models.

Using an unsupervised, LDA topic generation model with the Delhi data set, we extracted topics just from the analysis of the BoW representation of the data set. We compared the evolution of the topics against time and air quality data. This approach allowed us to identify niche topics such

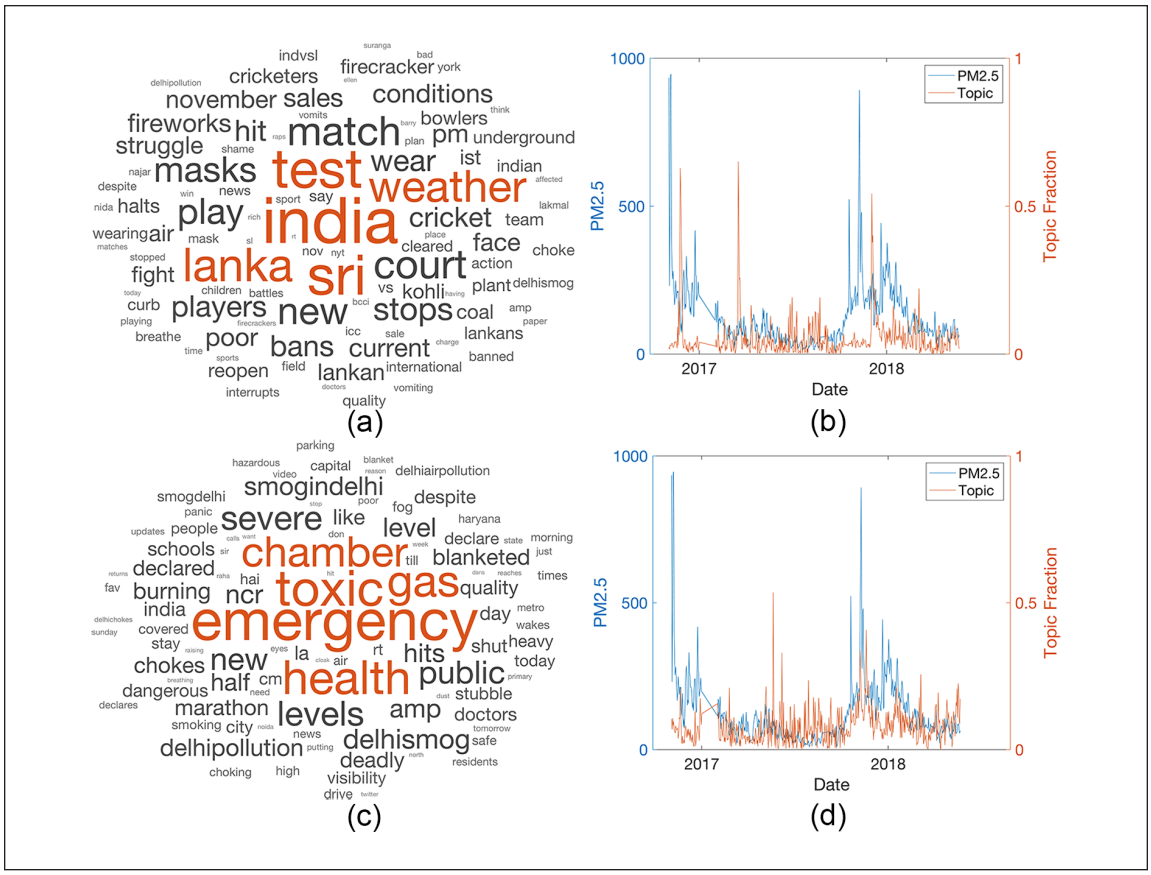


**Figure 12.** Topics extracted from Delhi $S^3$ data (manually entered user locations): (a) word cloud: cricket; (b) associated time series plot of the Cricket topic frequency and PM data shows that this topic trends with an air quality event associated with a major sporting event affected by a severe air quality episode; (c) word cloud: health; and (d) associated time series plot of the Health topic frequency and PM data shows that this topic spikes with poor air quality.
*Note.* PM = particulate matter.

as fireworks and cricket-related, that we would not have picked for analysis with supervised models. A combination of supervised and unsupervised learning might provide some interesting insight into the sentiments associated with the topics. For example, while we see that policy/government-related tweets increase as air quality worsens, it is not clear if the people are appreciative of the measures being taken or are demanding additional action, or placing the blame on the local state government or the central (federal) government. For such insight, we would have to combine supervised learning with unsupervised topic identification.

## Future Work

As future work, we would like to conduct additional investigation with supervised learning to determine if the choice of words used in building a topic changes the degree to which PM level and the number of tweets are correlated. As part of this investigation, we would like to compare the correlation of the frequency of supervised and unsupervised health topic tweets with respect to PM values. Also, we would like to explore advanced topic modeling algorithms, such as those based on a word distance formulation, that can automatically generate an ideal number of topics and also label topics appropriately using training data. For example, currently, when we increase the number of topics (to 20 or above) in the LDA model, several topics with a similar focus of government and policy were obtained, which could be automatically combined using advanced topic model formulations. Finally, further analysis of the unsupervised results will allow us to determine user sentiment associated with different topics and help us better understand public response to air quality events and policies.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Supraja Gurajala https://orcid.org/0000-0001-9770-989X


## References


Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011). Predicting flu trends using Twitter data. In *2011 IEEE conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 702–707). New York, NY: IEEE.

AirNow Delhi. (2016). *Environmental protection agency*. Retrieved from https://www.airnow.gov/index.cfm?action=airnow.global_summary

AIRPARIF. (2016). Paris air quality monitoring network. Retrieved from https://www.airparif.asso.fr/en

Bickerstaff, K., & Walker, G. (2001). Public understandings of air pollution: The "localisation" of environmental risk. *Global Environmental Change*, *11*, 133–145.

Bird, S., & Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions* (p. 31). Stroudsburg, PA: Association for Computational Linguistics.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bohnenstengel, S., Belcher, S., Aiken, A., Allan, J., Allen, G., Bacak, A., . . .Zotter, P. (2015). Meteorology, air quality, and health in London: The ClearfLo project. *Bulletin of the American Meteorological Society*, *96*, 779–804.

Carducci, A., Donzelli, G., Cioni, L., Palomba, G., Verani, M., Mascagni, G., . . .Gelatti, U. (2017). Air pollution: A study of citizen's attitudes and behaviors using different information sources. *Epidemiology Biostatistics and Public Health*, *14*(2), 1–9.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Culotta, A. (2013). Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation*, *47*, 217–238.

De Choudhury, M. (2013). Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-Aware Multimedia* (pp. 49–52). New York, NY: ACM.

Deswal, S., & Verma, V. (2016). Annual and seasonal variations in air quality index of the national capital region, India. *World Academy of Science, Engineering and Technology, International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering*, *10*, 1000–1005.

Doshi Velez, F., Wallace, B. C., & Adams, R. (2015). Graph-sparse LDA: A topic model with structured sparsity. In *Proceedings of the twenty-ninth AAAI conference on Artificial Intelligence* (pp. 2575–2581). Palo Alto, CA: AAAI.

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., & Vaughan, A. (2010). Omg earthquake! Can Twitter improve earthquake response? *Seismological Research Letters*, *81*, 246–251.

Eyheramendy, S., Lewis, D. D., & Madigan, D. (2003). *On the naive Bayes model for text categorization*. Retrieved from https://pdfs.semanticscholar.org/d9b4/fe6d4450257414ef1efa2b2aa80f1183edc9.pdf?_ga=2.208197394.1435775650.1563510602-1737197778.1561626250

Gurajala, S. (2018). *Social media sensing: Towards accurate prediction and analysis of events* (Doctoral dissertation). Clarkson University, Potsdam, NY.

Harman, G., & Dredze, M. H. (2014). Measuring post traumatic stress disorder in Twitter. In *Proceedings of the eighth international AAAI conference on Weblogs and Social Media*. Retrieved from https://pdfs.semanticscholar.org/51d4/4aaef206fd165e1f16f7047b09cf39a5a562.pdf?_ga=2.44824740.1435775650.1563510602-1737197778.1561626250

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). *A practical guide to support vector classification*. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Huang, J., Kornfield, R., Szczypka, G., & Emery, S. L. (2014). A cross-sectional examination of marketing of electronic cigarettes on Twitter. *Tobacco Control*, *23*(Suppl. 3), iii26–iii30.

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 geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter). *PLoS ONE*, *10*(10), e0141185.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European conference on Machine Learning (pp. 137–142). Berlin, Germany: Springer.

Kelley, K., Clark, B., Brown, V., & Sitzia, J. (2003). Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care*, *15*, 261–266.

Kelly, F. J., & Fussell, J. C. (2015). Air pollution and public health: Emerging hazards and improved understanding of risk. *Environmental Geochemistry and Health*, *37*, 631–649.

Lee, K., Agrawal, A., & Choudhary, A. (2013). Real-time digital flu surveillance using Twitter data. In The 2nd workshop on Data Mining for Medicine and Healthcare. Retrieved from http://users.eecs.northwestern.edu/~kml649/publication/LeeAgrCho13-SDM-DMMH13.pdf

Lim, K. W., Chen, C., & & Buntine, W. (2016). Twitter-network topic model: A full Bayesian treatment for social network and text modeling. NIPS 2013 Topic Models: Computation, Application and Evaluation, pp. 1–5. arXiv:1609.06791.

London Air. (2016). Retrieved from https://www.londonair.org.uk/LondonAir/Default.aspx

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI-98 workshop on Learning for Text Categorization* (Vol. 752, pp. 41–48). Palo Alto, CA: AAAI.

Mei, S., Li, H., Fan, J., Zhu, X., & Dyer, C. R. (2014). Inferring air pollution by sniffing social media. In *2014 IEEE/ACM international conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 534–539). New York, NY: IEEE.

Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29*(2), 9–17.

Myslín, M., Zhu, S. H., Chapman, W., & Conway, M. (2013). Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of Medical Internet Research*, *15*, e174.

Nagel, A. C., Tsou, M. H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., . . .Sawyer, M. H. (2013). The complex relationship of realspace events and messages in cyberspace: Case study of influenza and pertussis using tweets. *Journal of Medical Internet Research*, *15*, e237.

Oltra, C., & Sala, R. (2015). Communicating the risks of urban air pollution to the public. A study of urban air pollution information services. *Revista Internacional de Contaminación Ambiental*, *31*, 361–375.

Orru, H., Ebi, K., & Forsberg, B. (2017). The interplay of climate change and air pollution on health. *Current Environmental Health Reports*, *4*, 504–513.

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications*, *80*, 83–93.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . .Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Petit, J. E., Favez, O., Sciare, J., Crenn, V., Sarda-Estève, R., Bonnaire, N., . . .Leoz-Garziandia, E. (2015). Two years of near real-time chemical composition of submicron aerosols in the region of Paris using an aerosol chemical speciation monitor (ACSM) and a multi-wavelength aethalometer. *Atmospheric Chemistry and Physics*, *15*, 2985–3005. doi:10.5194/acp-15-2985-2015

Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011). Identifying health-related topics on Twitter. In International conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 18–25). Berlin, Germany: Springer.

Prüss-Ustün, A., Wolf, J., Corvalán, C., Bos, R., & Neira, M. (2016). *Preventing disease through healthy environments: A global assessment of the burden of disease from environmental risks*. Geneva, Switzerland: World Health Organization.

Review of the UK air quality index. (2011). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/304633/COMEAP_review_of_the_uk_air_quality_index.pdf

RiteTag. (2015, September). *Popular hashtags for pollution on Twitter and Instagram*. Retrieved from https://ritetag.com/best-hashtags-for/pollution

Sachdeva, S., McCaffrey, S., & Locke, D. (2017). Social media approaches to modeling wildfire smoke dispersion: Spatiotemporal and social scientific investigations. *Information, Communication & Society*, *20*, 1146–1161. doi:10.1080/1369118X.2016.1218528

Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology*, *7*(10), e1002199.

Sawitri, D. R., Hadiyanto, H., & Hadi, S. P. (2015). Pro-environmental behavior from a social cognitive theory perspective. *Procedia Environmental Sciences*, *23*, 27–33.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza a H1N1 pandemic. *PLoS ONE*, *6*(5), e19467.

Spiegelhalter, D. J., & Knill-Jones, R. P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society: Series A (General)*, *147*, 35–77.

Xu, S., Markson, C., Costello, K. L., Xing, C. Y., Demissie, K., & Llanos, A. A. (2016). Leveraging social media to promote public health knowledge: Example of cancer awareness via Twitter. *JMIR Public Health and Surveillance*, *2*, e17.

Zeidner, M., & Shechter, M. (1988). Psychological responses to air pollution: Some personality and demographic correlates. *Journal of Environmental Psychology*, *8*, 191–208. doi:10.1016/S0272-4944(88)80009-4

Zhang, N., Campo, S., Janz, K. F., Eckler, P., Yang, J., Snetselaar, L. G., & Signorini, A. (2013). Electronic word of mouth on Twitter about physical activity in the United States: Exploratory infodemiology study. *Journal of Medical Internet Research*, *15*, e261.

## Author Biographies

Supraja Gurajala is an assistant professor in the Computer Science Department in SUNY Potsdam NY. She has her PhD in computer science from Clarkson University. Her research interests are in the fields of social media data analytics, databases, computer networks, and security.

Suresh Dhaniyala (PhD, University of Minnesota, Minneapolis) is the Bayard D. Clarkson distinguished professor in the Mechanical and Aeronautical Engineering Department at Clarkson University. He is the co-director of the Center for Air Resources Engineering and Sciences (CARES) at Clarkson University, on the editorial board of Aerosol Science and Technology, and an active member of the American Association for Aerosol Research (AAAR). His current interests are in the fields of air quality monitoring, aerosol research, air sensors, and data analytics.

Jeanna N. Matthews (PhD, University of California Berkeley) is an associate professor at Clarkson University in Potsdam, New York, USA. She has published work in a broad range of systems topics from virtualization and cloud computing to social media security and distributed file systems. She is an affiliate at the Data and Society Research Institute in Manhattan and a founding co-chair of the ACM Technology Policy Subcommittee on Artificial Intelligence and Algorithm Accountability. Her current work focuses on securing societal decision-making processes and supporting the rights of individuals in a world of automation.