Twitter data analysis to understand societal response to air quality

Supraja Gurajala[1,2]
[1]Clarkson University
[2]SUNY Potsdam
gurajas@clarkson.edu

Jeanna N. Matthews[1]
[1]Clarkson University
jnm@clarkson.edu

ABSTRACT

Air quality is recognized to be 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. The ability of people to understand air quality information and take actions to protect their health is not clear. Recent studies have shown that even when public get data on air quality and understand its importance, they do not exhibit a pro-environmental behavior to address the problem. All existing studies on public attitude and response to air quality are based on offline studies, with a limited number of survey participants and over limited number of geographical locations. For a larger survey size and global set of locations, we analyzed Twitter data collected over a period of nearly two years. We identify a limited number of hashtags (3) that can best correlate the frequency of tweets with local air quality (PM2.5) in three major cities around the world: Paris, London, and New Delhi. Using tweets with just these three hashtags, we determined that people's response to air quality in the 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 amongst 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. Our approach will allow for global analysis of public response to air quality and aid public health officials respond appropriately.

CCS CONCEPTS

• **Networks → Online social networks**; • **Computing methodologies → Supervised learning by classification**; • **Applied computing → Earth and atmospheric sciences**;

KEYWORDS

Online social networks, human behavior, Twitter, air quality, PM, data mining, correlations, machine learning

SMSociety '18, July 18–20, 2018, Copenhagen, Denmark

ACM ISBN 978-1-4503-6334-1/18/07…$15.00
https://doi.org/10.1145/3217804.3217900

1 INTRODUCTION

Ambient air pollution is one of the most important risk factors for public health globally [32]. Amongst the different air quality parameters regulated by global environmental agencies, the mass concentration of particulate matter (PM) smaller than 2.5 µm, i.e., $PM_{2.5}$, is one of the most significant from a health perspective [19]. It is estimated that exposure to high PM pollution resulted in ~3.7 million premature deaths worldwide in 2012 [32] 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.

The long-term epidemiological studies might 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 the extent of the recognition of the problem by the public is not entirely clear [3, 26]. 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 using simple color-coded indices [33]. In spite of these measures, people usually fail to minimize their exposure to air pollution on a daily basis [3]or take effective mitigation actions [37], resulting in air pollution exposure becoming a major public health issue. To minimize air population-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 [18]. Traditional survey tools - such as personal interviews [42] - have often been used in this context, to document feelings and sentiments experienced by people exposed to different levels of ambient air pollution. Recently, [6] 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 towards air pollution issues. They determined that information about air pollution events, often obtained from traditional media, was focused on short term, alarmist issues, without a focus on 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 [37] and thus, efforts to disseminate information

about individual responsibility to tackle this problem is important, while also understanding its effectiveness.

For effective information dissemination and to survey public response, social media platforms such as Twitter and Facebook, could be very useful. 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 [11]. Twitter messages have also been analyzed to track public-health issues such as flu epidemics [1, 8, 20, 25], smoking [15, 24, 31], exercise [43], and mental health trends [9, 13] and personal health concerns such as cancer [41]. Recently, Twitter analysis has been extended to not just track events, but to understand human social interactions, perceptions, and sentiments [31, 36, 38]. Twitter, therefore, should provide data relevant for understanding human behavior and response to ambient air quality.

2 RELATED WORK

There have been several studies analyzing the relevance of social media data for ambient air quality. Wang et al. [40] analyzed Sina Weibo messages from 74 cities in China and determined that messages related to air quality were closely related to the annual particle pollution levels. Mei et al[23] used a machine learning algorithm with air pollution related tweets from Sina Weibo to demonstrate the relation between AQI values and tweet frequency. Jiang et al. [16] also used a machine learning method to monitor the dynamics of AQI based on air pollution-related posts on Sina Weibo. Almost all existing studies relating air quality to social media posts have been based on data from Sina Weibo, and only for air quality in China. Also, most of these studies were focused on demonstrating an ability to predict local air quality based on the frequency of the posts.

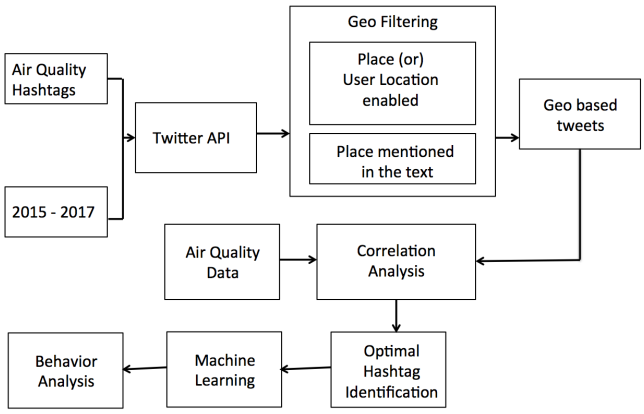

Figure 1: Flow chart illustrating our data collection and analysis procedure.

Here, we extend the earlier efforts by analyzing tweets from Twitter platform, where a world-wide audience exists, to determine if the tweet frequencies are correlated with air quality at a global scale. Analysis of tweets will also allow us to understand the underlying human behavior associated with air quality changes globally.

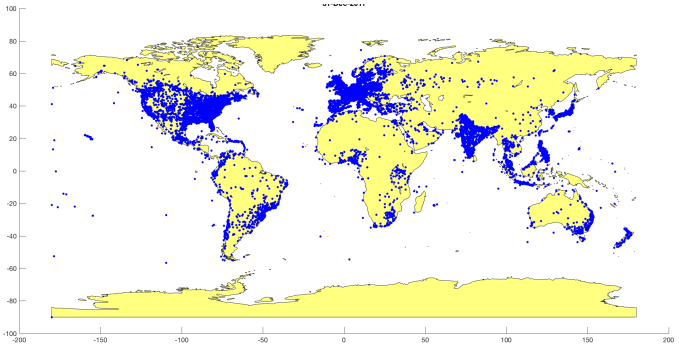

Figure 2: Global distribution of tweets analyzed in this study.

Prior work based on Twitter data has shown that social environmental sensing is possible from personal observations extracted from the platform [34], however, no studies have been conducted for sensing air quality trends and related public response.

The paper will first discuss the details of the Twitter data obtained, the data filtering methods used, sources for PM data, and the correlation analysis conducted to establish the relation between air quality and tweet frequencies for three identified cities. Then, from the correlation analysis, the procedure to obtain the most relevant hashtags for air quality for the different cities will be detailed and the predictions of machine learning algorithms to determine underlying human behavior and response characteristics with change in air quality will be discussed. The overall procedure followed in the paper is illustrated in Figure 1.

3 DATA COLLECTION

The primary method we used to obtain tweets was by accessing Twitter's stream API. In this approach, a single application is authenticated and connected to a public stream comprising of a sample of the tweets being posted on Twitter. Included in the request is a filter indicating which tweets are to be returned. For our research, the tweets were filtered using a list of specific hashtags that were selected based on web-search of popular air-quality related hashtags (e.g.,[35]), use in prior publications (e.g.,[16]), and discussions with air quality scientists (S. Dhaniyala, personal communication, Oct 2015). The list of hashtags that were selected for our analysis (Table 1) that were deemed relevant to air quality. The tweet JSON objects received were then saved to a Mongo database.

For this study, pollution-related tweets, totaling over 20 million, were collected over a period of 2 years (Sep. 2015 to Dec. 2017). In the first 13 months (Sep. 2015 to November 2017), the tweets were collected sporadically. Over the last 13 months (Nov 2016 to Dec 2017), the tweets were collected continuously, except for the month of Jan 2017. The data over the entire time period is generated identically, i.e., with the same hashtags and data collection speed. Also, the data collected in the initial time period is only a small fraction of the total data and has the same geographical spread as the rest of our data set. Thus, the analysis of the entire set does not bias our analysis in any manner.

The global distribution of tweets that we collected with *place* enabled is shown in Figure 2. It is seen that most of our tweets

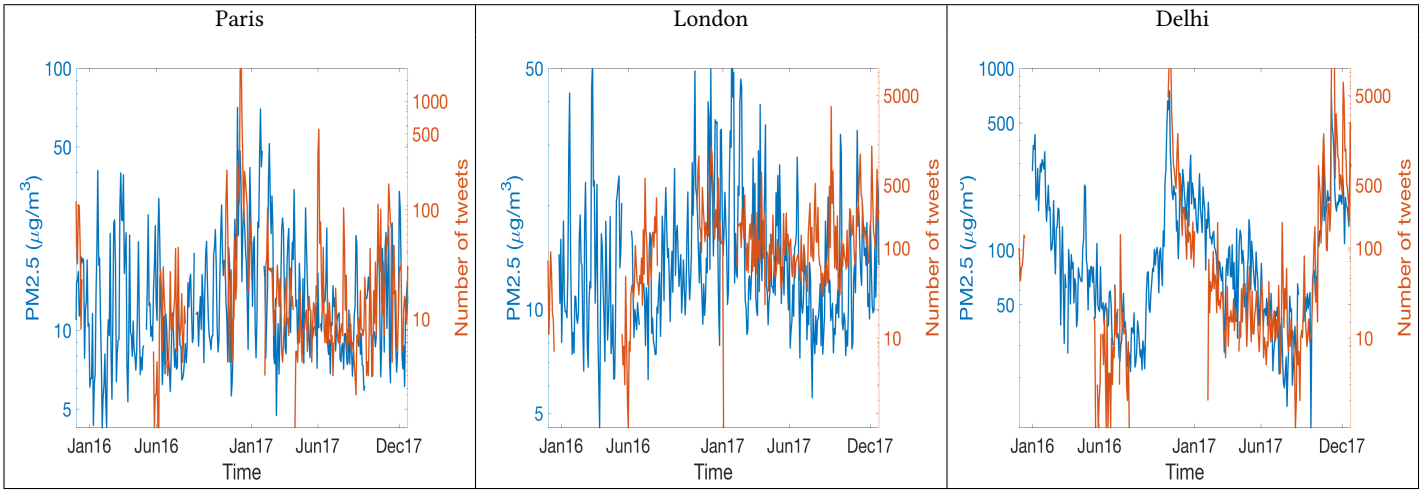

Figure 3: Temporal trends in air quality and the total number of tweets associated with selected air quality hashtags. The PM data is averaged over 48 hours and the tweet data is accumulated over the same time period.

associated with air pollution were collected from US, Europe, and India. As Twitter is largely inaccessible in China, tweets from China constitute only a small fraction (~0.2%) of all our tweets. Amongst the three regions with a large number of tweets, the problem of air pollution is most severe in India and in certain European cities than in the US. Thus, it was decided that further analysis would be limited to these geographical localities. Within these areas, we decided to concentrate on three major cities: New Delhi, Paris and London, because of two reasons: air pollution data is readily available in these cities at an hourly rate (or higher frequency) and air pollution in these sites varies significantly over the course of a year [10], [30], [5].

The tweet collection for the three cities was then compiled from the entire tweet dataset if the city name was indicated either in the attribute "place," or anywhere in the tweet, or given in the user location. The number of tweets analyzed for cities New Delhi, Paris and London are listed in Table 2.

Table 1: Hashtags Used

#AIRPOLLUTION	#OZONE	#POLLUTION
#AIRQUALITY	#HAZE	#SMOG
#CLEANAIR	#EMISSIONS	#PM25
#PARTICLES	#PM2.5	#PM10
#PM1	#PARTICULATES	

Table 2: Number of tweets collected for each city

Place	Number of Tweets
Delhi	1005240
Paris	593097
London	655897

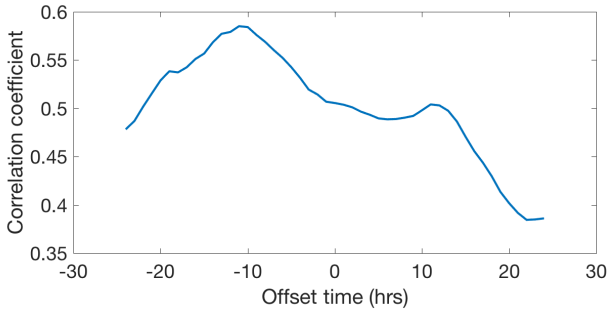

Figure 4: Time for a selected combination of hashtag (smog) and city (New Delhi).

The $PM_{2.5}$ data for the three sites were obtained from their respective monitoring agencies or the US embassy monitoring station (for New Delhi). For Paris, hourly data was obtained form the site *Paris (City Center)* [28]. For London, 15minute data was obtained for two sites, Farringdon St and Sir John Cass school [21], and the data from the two sites was averaged. For New Delhi, hourly data was obtained from the US embassy site [2].

4 DATA ANALYSIS AND RESULTS

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 was 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 hours) is shown in Figure 3. All three cities, show a correlation

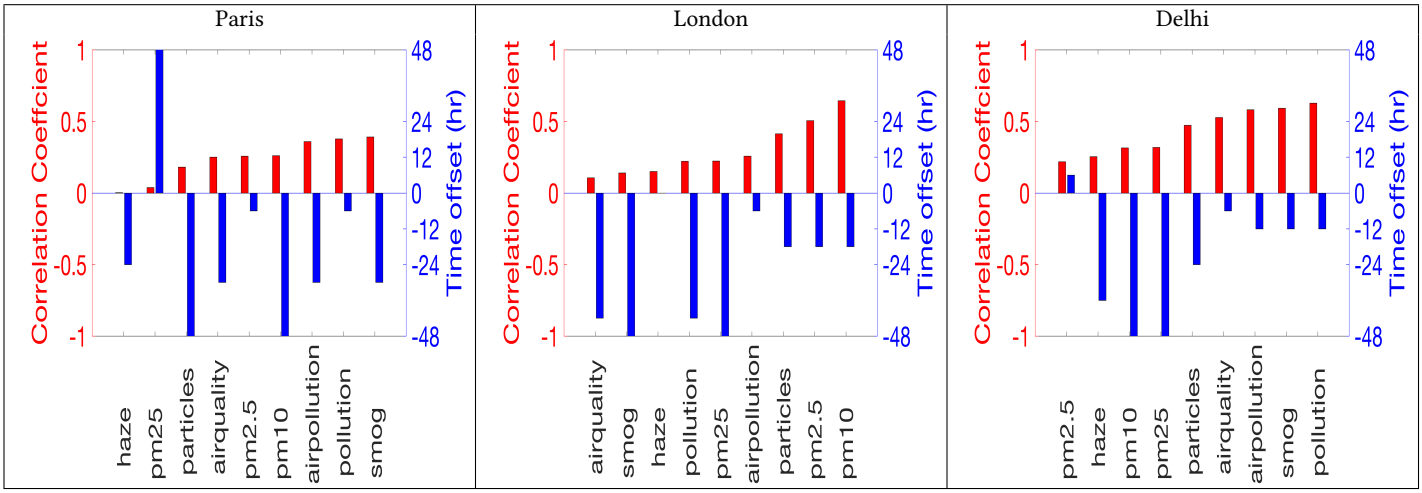

Figure 5: The correlation and time-shifts associated with different hashtags and cities studied.

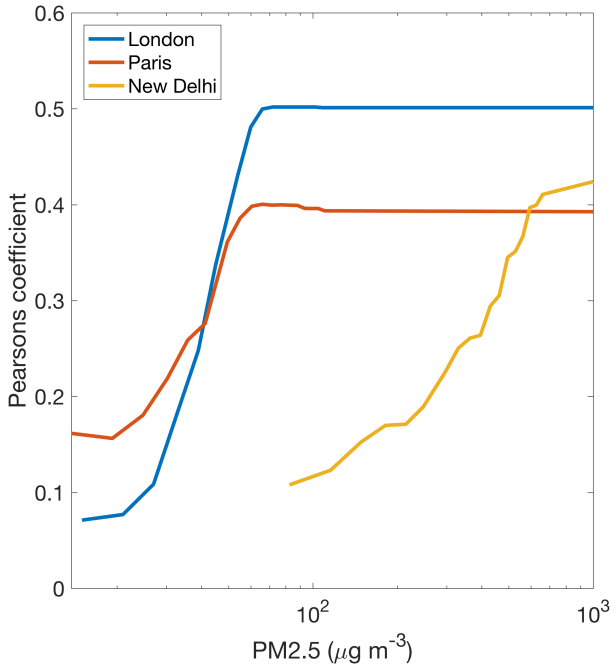

Figure 6: Change in correlation coefficient with $PM_{2.5}$ cutoff values. At any given $PM_{2.5}$ value (x-axis), the correlation coefficient was calculated for all tweets at times when the PM value were greater than the cutoff value.

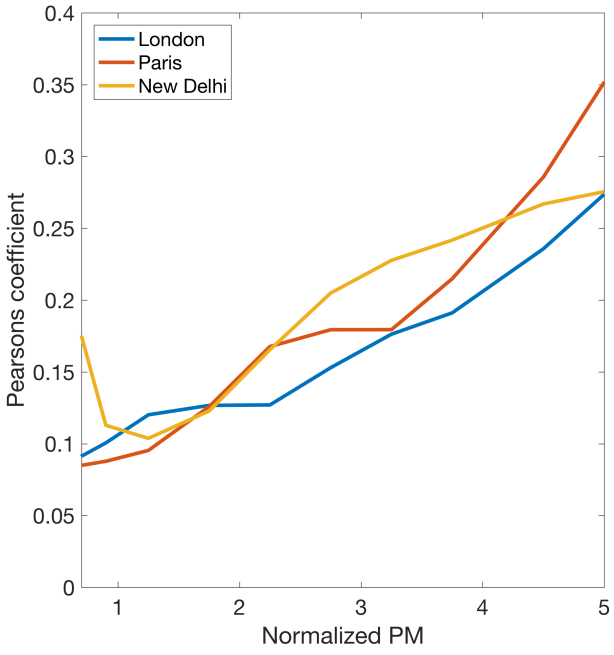

Figure 7: Change in correlation coefficient with normalized $PM_{2.5}$ cutoff values. At a selected $PM_{2.5}$ value (x-axis), the correlation coefficient was calculated for all tweets at times when the PM value were greater than the cutoff value.

in temporal variation with the tweet number. This provides some initial validation that our selection of hashtags is reasonable.

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 at a 6-hour frequency. The choice of a 6-hour 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 the data time-shift

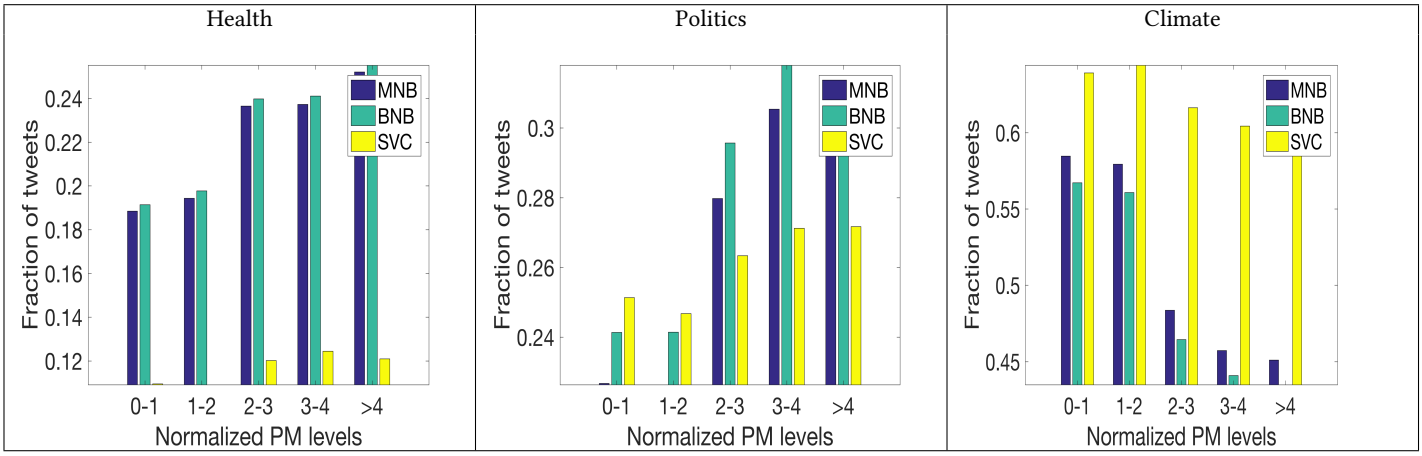

Figure 8: The fraction of tweets associated with the three classes as predicted by the different machine learning algorithms as a function of normalized PM levels.

is shown in Figure 4. 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, maximum correlation coefficient and peak time-shift, were determined considering both the original tweets and retweets associated with each of the hashtags. The results from our analysis (Figure 5) show that, in general, the top three hashtags for each city have a similar strength of correlation with respect to each other. The data was significant with $p < 0.01$ for all hashtags and city combinations. Most of the hashtags have a peak correlation when the tweets are ~6 to 24 hours after the event. Hashtags with a positive time-shift, i.e., 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). These hashtags were then eliminated from our dataset in order to study public response to air quality events. Also, these results suggest that prediction of air quality from tweet data must consider time shifts between events and their associated tweets. 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 hashtags. 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).

The correlation of tweet numbers with $PM_{2.5}$ cut-off 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 cut-off values of $PM_{2.5}$ (Figure 6). 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)[16] for data from Sina Weibo. The public response to air quality (represented by increasing correlation of tweet frequency) 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 7). This result suggests that public response to $PM_{2.5}$ values is driven very much by their chronic exposure. Public response seems to be driven by the relative difference in the $PM_{2.5}$ values that they experience rather than the absolute values. The response seems to saturate only when the normalized PM values exceed 5, an increase that is possibly already perceived as excessive.

5 BEHAVIORAL ANALYSIS

The tweet-collection based on the top three hashtags was then analyzed to determine the evolution of the tweet-content for each city as a function of $PM_{2.5}$ values. The tweet content was analyzed using machine learning algorithms, with the goal of assessing changes in sentiment or behavior of the Twitter users. For this analysis, we first divided the tweets based on the PM values at the tweet time. We considered five normalized $PM_{2.5}$ ranges: 0 to 1, 1 to 2, 2 to 3, 3 to 4, and > 4. For each of these normalized PM values, the tweets were analyzed to determine their content. Considering the dual role of airborne particles in public health and climate change [27], and, therefore, government policy, we categorized the tweets into one of three classes: health, climate, and politics.

5.1 Supervised Learning Algorithms

The text analysis was conducted using three supervised learning algorithms: Bernoulli Naïve Bayes (BNB), Multinomial Naïve Bayes (MNB), and Support Vector Classifier (SVC). Each of these are described below:

5.1.1 Naïve Bayes. This is a simple (naïve) classification method that uses Bayes rule of independence of features or words to categorize tweets. Naïve Bayes (NB) classifiers make the assumption that the order of words in the tweets don't matter, i.e., a 'bag of words' assumption is made.

The tweets are classified into one of three categories (Health, Politics, Climate) using Bayes Theorem, expressed as:

Figure 9: Word cloud for the three classes for New Delhi. Note that the Words associated with the top three hashtags are removed. The search term is also not shown eg. the term 'health' not shown in the word cloud for class health.

$$C_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{w \in W} P(w \mid c)$$

where C_{NB} is the selected category or class for the tweet, C is one of the 3 categories considered here and $W = (w_1, .., w_n)$ is the feature or word vector associated with a tweet. In the above equation, the Naïve Bayes [39] assumption of conditional independence is made, i.e., the probabilities $P(w \mid c)$ are independent of the category c. NB is often the first-choice algorithm for text classification as it is robust to irrelevant features, low amount of data, and can handle classification even when many features with equal importance exist. Naïve Bayes, however, has some well recognized problems, particularly the assumption of feature-independence ([22]). But in spite of this problem Naïve Bayes has been popular for text classification because of its simplicity, its fast speed, and low storage requirements.

5.1.2 Bernoulli Naïve Bayes (BNB). In the BNB model, the Naïve Bayes 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 Bernoulli Naïve Bayes is based on

$$C_{BNB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{w \in W} P(w \mid c) \prod_{w \notin W} (1 - P(w \mid c))$$

where C_{BNB} is the selected category or class for the tweet. BNB model is best for short documents such as tweets, where occurrence of multiple instances of a word is unlikely or possibly unimportant.

5.1.3 Multinomial Naïve Bayes (MNB). A more appropriate algorithm for text categorization is the multinomial Naïve Bayes (MNB), where a multinomial probability is assumed for the features, accounting for multiple instances of a feature (word) being present in the document (tweet) [12]. In the MNB model, the Naïve Bayes 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.

$$C_{MNB} = \underset{w \in W}{\operatorname{argmax}} P(c) \prod_{w \in W} P(w \mid c)^N$$

where C_{MNB} is the selected category or class for the tweet, 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

5.1.4 Support Vector Classifier (SVC). SVC is a supervised learning method that is particularly effective in high dimensional spaces, i.e. when there is a large feature set. In SVC, learning data is used to determine decision boundaries or hyperplanes to separate tweets into the selected categories [7]. 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) [17] as is the case with tweets. We use a Linear Support Vector Classifier (SVC) as it has been shown to be as accurate as a non-linear model when the feature set is large, as is the case here [14]. 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.

5.2 Results and Discussion

The algorithms were first trained with a set of tweets. The tweets were first preprocessed to remove handles, retweet symbols, urls, emojis, sentences containing single word, and extra spaces. We then extracted features form these preprocessed tweets using a Bag-of-words representation. We used Natural Language Processing Tool Kit (NLTK) [4] for preprocessing and feature extraction. These feature sets where then used to train the model. To build a training set, we used selected search terms (e.g., terms for the class "health": health, sick, disease, lung) and obtained 200 tweets for each class. The obtained training tweets were then manually analyzed to determine if they were relevant for the study and any irrelevant tweets (advertisements, off-topic, etc) were filtered from the training set. The dataset was randomly split into two groups:

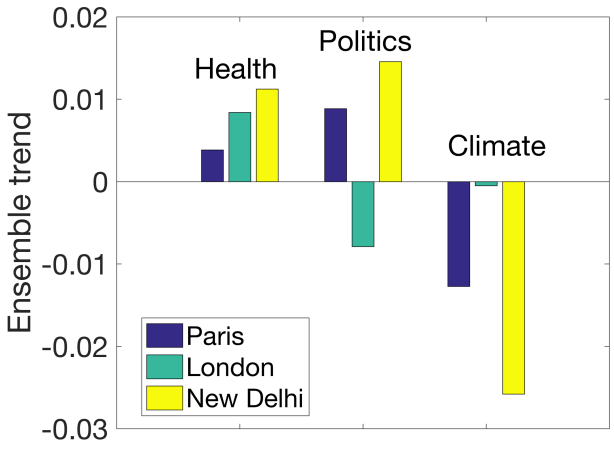

Figure 10: The linear trendline slopes for each class and city combination considering an ensemble of the model predictions.

with 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. The algorithms were implemented using Scikit-learn library in Python [29].

We first analyzed the New Delhi data set that we created considering only the top three correlated hashtags (air pollution, pollution, 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. 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. 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 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 9, 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. Amongst the words in the health collection include: asthma, breathe, health emergency, immunity, etc, all pointing to severe health concerns. In the politics collection, words include: political party names (AAP, BJP, Congress), government policies (Odd-Even, achedin), petitions (my right to breathe) etc, 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 [6] 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 average of the predictions of the three algorithms. The slopes of the linear trendlines for each class and city is shown in Figure 10. 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. This analysis reveals that the public in New Delhi are more concerned with their increase in PM from the base level to 4 times higher than the public in the other cities. The trend for climate is negative, suggesting that at high PM values, the concern for climate takes a back seat. People in New Delhi and Paris seem to associate politics/government with poorer air quality, but not in London.

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. 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.

6 LIMITATIONS OF THIS STUDY

Our tweet collection consisted of both geo-enabled tweets and reference to city names in the tweets. This introduced some noise in the dataset, as some of the tweets were from locations outside our analysis city (based on a visual check). As we collect more data we should be able to conduct our analysis with just the geo-enabled tweets. For the supervised learning study, the assumption of independence of the features inherent in the Naïve Bayes models could be problematic and we would like to explore other classification algorithms including non-linear SVC models. We would also like to explore any dependence of our predictions on the size of the training set.

7 CONCLUSIONS

The impacts of air quality are reasonably well-known scientifically but public attitude to air quality information is less well known. Here, we analyze Twitter data in three global cities to determine similarities and differences in public response to air quality information. We filtered our Twitter collection over two years to obtain

tweets related to air quality in three major global cities: Paris, London, and New Delhi. The number of tweets with just three hashtags was shown to be highly and significantly correlated to PM values in the three cities. Using the optimal hashtags, and normalized PM data, the correlation coefficients suggested that the public in the three cities responded similarly with relative changes in air quality rather than absolute changes. This information suggests that Twitter data from cities without local air quality information can be analyzed to understand relative changes in their air quality.

For further analysis of public response to air quality information, the tweets were analyzed and categorized into three classes related to air quality: health, climate, and politics. Using three text classification algorithms, it was seen that there was a consistency in the trends predicted by the models. For New Delhi, all models agree that the people's tweets on health and politics increased with normalized PM and tweets related to climate decreased. These trends provide a critical clue suggesting that the population (or at-least the Twitter population) recognize the impact of air quality on health and believe that the government should do something about it. The trends in the data for the other cities point a similar increase in health-related tweets with degrading air quality, and a decrease in climate related tweets. The relation of air quality to politics is local - people in New Delhi and Paris seem to associate politics/government with air quality, but those in London do not. Previous studies [6] have argued that when the public blames an environmental problem on institutions rather on the actions of individuals, the effectiveness of policies could be compromised [37]. Therefore, one preliminary conclusion from our results is that individuals are recognizing changes in air quality fairly quickly in time and also its associated health issues. However, they may not be effectively receiving information about individual pro-environmental actions that they should take to address this growing environmental problem.

REFERENCES

[1] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. IEEE, 702–707.

[2] AirNow Delhi accessed since 2016. Environmental Protection Agency. (accessed since 2016). https://www.airnow.gov/index.cfm?action=airnow.global_summary.

[3] Karen Bickerstaff and Gordon Walker. 2001. Public understandings of air pollution: the âĂŸlocalisationâĂŹof environmental risk. *Global Environmental Change* 11, 2 (2001), 133–145.

[4] Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 31.

[5] SI Bohnenstengel, SE Belcher, A Aiken, JD Allan, G Allen, A Bacak, TJ Bannan, JF Barlow, DCS Beddows, WJ Bloss, et al. 2015. Meteorology, air quality, and health in London: The ClearfLo project. *Bulletin of the American Meteorological Society* 96, 5 (2015), 779–804.

[6] Annalaura Carducci, Gabriele Donzelli, Lorenzo Cioni, Giacomo Palomba, Marco Verani, Giulia Mascagni, Giuseppe Anastasi, Luca Pardini, Elisabetta Ceretti, Tiziana Grassi, et al. 2017. Air pollution: A study of citizenâ s attitudes and behaviors using different information sources. *EPIDEMIOLOGY BIOSTATISTICS AND PUBLIC HEALTH* 14, 2 (2017), 1–9.

[7] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[8] Aron Culotta. 2013. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language resources and evaluation* 47, 1 (2013), 217–238.

[9] Munmun De Choudhury. 2013. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-aware multimedia*. ACM, 49–52.

[10] Surinder Deswal and Vineet Verma. 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, 10 (2016), 1000–1005.

[11] Paul Earle, Michelle Guy, Richard Buckmaster, Chris Ostrum, Scott Horvath, and Amy Vaughan. 2010. OMG earthquake! Can Twitter improve earthquake response? *Seismological Research Letters* 81, 2 (2010), 246–251.

[12] Susana Eyheramendy, David D Lewis, and David Madigan. 2003. On the naive bayes model for text categorization. (2003).

[13] GACCT Harman and Mark H Dredze. 2014. Measuring post traumatic stress disorder in Twitter. *In ICWSM* (2014).

[14] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. 2003. A practical guide to support vector classification. (2003).

[15] Jidong Huang, Rachel Kornfield, Glen Szczypka, and Sherry L Emery. 2014. A cross-sectional examination of marketing of electronic cigarettes on Twitter. *Tobacco control* 23, suppl 3 (2014), iii26–iii30.

[16] Wei Jiang, Yandong Wang, Ming-Hsiang Tsou, and Xiaokang Fu. 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 (2015), e0141185.

[17] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer, 137–142.

[18] Kate Kelley, Belinda Clark, Vivienne Brown, and John Sitzia. 2003. Good practice in the conduct and reporting of survey research. *International Journal for quality in health care* 15, 3 (2003), 261–266.

[19] Frank J Kelly and Julia C Fussell. 2015. Air pollution and public health: emerging hazards and improved understanding of risk. *Environmental geochemistry and health* 37, 4 (2015), 631–649.

[20] Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time digital flu surveillance using twitter data. In *The 2nd Workshop on Data Mining for Medicine and Healthcare*.

[21] London Air accessed since 2016. London Air. (accessed since 2016). https://www.londonair.org.uk/LondonAir/Default.aspx.

[22] Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, Vol. 752. Citeseer, 41–48.

[23] Shike Mei, Han Li, Jing Fan, Xiaojin Zhu, and Charles R Dyer. 2014. Inferring air pollution by sniffing social media. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 534–539.

[24] Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. 2013. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research* 15, 8 (2013).

[25] Anna C Nagel, Ming-Hsiang Tsou, Brian H Spitzberg, Li An, J Mark Gawron, Dipak K Gupta, Jiue-An Yang, Su Han, K Michael Peddecord, Suzanne Lindsay, et al. 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, 10 (2013).

[26] Christian Oltra and Roser Sala. 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, 4 (2015), 361–375.

[27] Hans Orru, KL Ebi, and B Forsberg. 2017. The interplay of climate change and air pollution on health. *Current environmental health reports* 4, 4 (2017), 504–513.

[28] Paris Air Quality Monitoring Network accessed since 2016. AIRPARIF, Paris Air Quality Monitoring Network. (accessed since 2016). https://www.airparif.asso.fr/en.

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[30] J.-E. Petit, O. Favez, J. Sciare, V. Crenn, R. Sarda-Estève, N. Bonnaire, G. Močnik, J.-C. Dupont, M. Haeffelin, and E. Leoz-Garziandia. 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, 6 (2015), 2985–3005. https://doi.org/10.5194/acp-15-2985-2015

[31] Kyle W Prier, Matthew S Smith, Christophe Giraud-Carrier, and Carl L Hanson. 2011. Identifying health-related topics on twitter. In *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, 18–25.

[32] Annette Prüss-Üstün. 2016. *Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks*. World Health Organization.

[33] Review of the UK Air Quality Index. 18 Mar 2017. Review of the UK Air Quality Index. 2011. (18 Mar 2017). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/304633/COMEAP_review_of_the_uk_air_quality_index.pdf.

[34] Marina Riga and Kostas Karatzas. 2014. Investigating the relationship between social media content and real-time observations for urban air quality and public

health. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*. ACM, 59.

[35] RiteTag 2015 (accessed sept 2015). Popular hashtags for pollution on Twitter and Instagram. (2015 (accessed sept 2015)). https://ritetag.com/best-hashtags-for/pollution

[36] Marcel Salathé and Shashank Khandelwal. 2011. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology* 7, 10 (2011), e1002199.

[37] Dian R Sawitri, H Hadiyanto, and Sudharto P Hadi. 2015. Pro-environmental Behavior from a SocialCognitive Theory Perspective. *Procedia Environmental Sciences* 23 (2015), 27–33.

[38] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 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 (2011), e19467.

[39] David J Spiegelhalter and Robin P Knill-Jones. 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)* (1984), 35–77.

[40] Shiliang Wang, Michael J Paul, and Mark Dredze. 2015. Social media as a sensor of air quality and public response in China. *Journal of medical Internet research* 17, 3 (2015).

[41] Songhua Xu, Christopher Markson, Kaitlin L Costello, Cathleen Y Xing, Kitaw Demissie, and Adana AM Llanos. 2016. Leveraging social media to promote public health knowledge: example of cancer awareness via Twitter. *JMIR public health and surveillance* 2, 1 (2016).

[42] Moshe Zeidner and Mordechai Shechter. 1988. Psychological responses to air pollution: Some personality and demographic correlates. *Journal of Environmental Psychology* 8, 3 (1988), 191 – 208. https://doi.org/10.1016/S0272-4944(88)80009-4

[43] Ni Zhang, Shelly Campo, Kathleen F Janz, Petya Eckler, Jingzhen Yang, Linda G Snetselaar, and Alessio Signorini. 2013. Electronic word of mouth on twitter about physical activity in the United States: exploratory infodemiology study. *Journal of medical Internet research* 15, 11 (2013).