Preventive Medicine 121 (2019) 86–93

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Feasibility of using social media to monitor outdoor air pollution in London, England

Yulin Hswen[a,b,*], Qiuyuan Qin[b,c], John S. Brownstein[b,d], Jared B. Hawkins[b,d]

[a] Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[b] Informatics Program, Boston Children's Hospital, Boston, MA, USA
[c] Department of Statistics, Boston University, Boston, MA, USA
[d] Department of Pediatrics, Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Keywords:
Air pollution
PM2.5
Surveillance
London
Urban area
Social media
Twitter

ABSTRACT

Air pollution is a serious public health concern. Innovative and scalable methods for detecting harmful air pollutants such as PM2.5 are necessary. This study assessed the feasibility of using social media to monitor outdoor air pollution in an urban area by comparing data from Twitter and validating it against established air monitoring stations. Data were collected from London, England from July 29, 2016 to March 17, 2017. Daily mean PM2.5 data was downloaded from the LondonAir platform consisting of 26 air pollution monitoring sites throughout Greater London. Publicly available tweets geo-located to Greater London containing air pollution terms were captured from the Twitter platform. Tweets with media URL links were excluded to minimize influence of news stories. Sentiment of the tweets was examined as negative, positive, or neutral. Cross-correlation analyses were used to compare the relationship between trends of tweets about air pollution and levels of PM2.5 over time. There were 16,448 tweets without a media URL link, with a mean of 498.42 (SD = 491.08) tweets per week. A significant cross-correlation coefficient of 0.803 was observed between PM2.5 data and the non-media air pollution tweets ($p < 0.001$). The cross-correlation coefficient was highest between PM2.5 data and air pollution tweets with negative sentiment at 0.816 ($p < 0.001$). Discussions about air pollution on Twitter reflect particle PM2.5 pollution levels in Greater London. This study highlights that social media may offer a supplemental source to support the detection and monitoring of air pollution in a densely populated urban area.

1. Introduction

Air pollution in London, England is a serious health concern. In the last week of January 2017 air pollution PM2.5 levels in London were worse than in Beijing, China, a city that is notorious for poor air quality (Broomfield, 2017; Mittal and Fuller, 2017). Readings from air quality monitors were concentrated at 197 micrograms per cubic meter in London compared to 190 in Beijing (Broomfield, 2017; Mittal and Fuller, 2017). PM2.5 refers to fine particles of dust smaller than 2.5 μm in diameter and is the key indicator that the World Health Organization (WHO) uses to measure air pollution because of its association with serious health effects (Cohen et al., 2005; Mittal and Fuller, 2017; Organization WH and UNAIDS, 2006). PM2.5 is produced from combustion of fuels from motor vehicles, wood burning, coal-burning power plants, and other industrial sources (Cohen et al., 2005; Organization WH and UNAIDS, 2006). In London, traffic is a main source of air pollution; however, construction sites, flumes for gas boilers or diesel generators, petrol stations, multi-story car parks and airports are also common sources of air pollution in Greater London (Adams et al., 2001; Charron and Harrison, 2005; Mittal and Fuller, 2017).

Ambient outdoor air pollution poses serious health risks and contributes to increased mortality from stroke, heart disease, lung cancer and chronic respiratory diseases (Apte et al., 2015; Birnbaum et al., 2017; Cohen et al., 2005; Gerlach-Reinholz et al., 2017). Each year, air pollution from outdoor sources contributes to nearly 9500 early deaths in London and over 40,000 deaths in the United Kingdom (Goddard, 2017; Pultarova, 2017). In addition to serious physical health effects, there is mounting evidence that air pollution is associated with greater psychological distress, more frequent emergency department visits for depression, and increased risk of suicide (Ho et al., 2014; Kim et al., 2018; Sass et al., 2017; Szyszkowicz et al., 2009). While disease burden attributed to air pollution is most severe in lower income countries (Landrigan et al., 2017), the European Environment Agency estimates

* Corresponding author at: Harvard T.H. Chan School of Public Health, Department of Behavioral Sciences, 677 Huntington Ave, Boston MA 02115, United States.
E-mail address: yuh958@mail.harvard.edu (Y. Hswen).

https://doi.org/10.1016/j.ypmed.2019.02.005
Received 26 July 2018; Received in revised form 24 December 2018; Accepted 7 February 2019
Available online 08 February 2019
0091-7435/

that over 90% of city dwellers in Europe are regularly exposed to harmful air pollutants (European Environment Agency, 2008). This has prompted outcry among citizens, activism, and increased pressure on local governments to implement measures to improve air quality (Gardiner, 2014; hackAIR, 2018).

A key factor for improving civic response to harmful air pollutants is to develop a better understanding of how air pollution fluctuates over time by geographic area. This is essential in order to identify which populations are most at risk and to support methods to intervene and reduce harmful exposure. Current methods of measuring PM2.5 levels require systematic, long-term assessment of air pollutants through sophisticated air monitoring stations. These monitors are mainly funded by London boroughs as part of their duties under the Local London Air Quality Management (LLAQM) framework (Greater London Authority City Hall, 2018). However, these fixed monitoring stations are only capable of measuring air pollution within a specific radius, and raw data collected from these stations is used to extrapolate the concentration of pollutants throughout an entire region using dispersion models (Devarakonda et al., 2013). This complex process makes it difficult to quickly analyze and identify hazardous levels of air pollution (Devarakonda et al., 2013). Additionally, where air quality monitors are situated is very important, as monitors as little as 5 m apart can return very different measurements because pollution levels can vary largely over time and space. Yet, the size and need for power supply make it difficult to find suitable places to put these monitors. Finally, the equipment and maintenance of these monitoring stations is very expensive, which deters continuous monitoring and further deployment in additional areas of the city (Birnbaum et al., 2017; Blankers et al., 2012; Devarakonda et al., 2013; Jiang et al., 2015). For instance, certain areas may be without air monitoring systems altogether leaving residents in these areas without any advanced warning of harmful pollutants (Jiang et al., 2015).

In response to the need for more area level air pollution monitoring, in January 2018, the Mayor of London released a comprehensive guide to encourage the public to monitor air quality in London. Citizen-led monitoring can be an important method to gain greater insights about air quality issues in discrete local areas (Greater London Authority City Hall, 2018). However, public monitoring of air pollution still requires individuals to purchase air quality monitors, which are not very accurate and start at nearly $200 USD (approx. £150), something many residents cannot afford (Greater London Authority City Hall, 2018). Hence, a more affordable and scalable real-time system is needed to help detect, respond to, and mitigate health risks of air pollution across Greater London, and globally.

Latest digital technologies such as social media platforms and smartphone applications may afford new opportunities for collecting data about air quality. For example, a study used Facebook advertisements to connect participants with a smartphone application for actively measuring physical and psychological wellbeing during an air pollution crisis in Southeast Asia (Zhang et al., 2014). The potential for social media platforms to facilitate passive monitoring of air quality is also gaining greater attention, especially in urban settings where public response to harmful air pollutants may be especially prominent (Ilieva and McPhearson, 2018). Prior studies have demonstrated that the Twitter platform can act as an effective digital surveillance tool for monitoring a range of public health concerns (Gruebner et al., 2017; Hswen et al., 2017; McIver et al., 2015; Nsoesie et al., 2016a; Nsoesie et al., 2016b). Twitter provides a microblogging medium where users can describe their current status in short and frequent posts, called "tweets" containing 140 characters or less. With over 330 million active users who post > 500 million tweets daily (Statista, 2018), Twitter allows for real-time monitoring of unsolicited content posted by users (Java et al., 2007). Furthermore, Twitter's ability to capture geolocation can facilitate discrete surveillance both locally and nationally (Broniatowski et al., 2013). Since people may use Twitter to express their experiences, behaviors and attitudes about air pollution, Twitter could be a valuable resource to support monitoring of ambient air pollution.

The objective of this study was to determine whether it is feasible to use Twitter for monitoring outdoor air pollution in an urban area by comparing data collected from the Twitter platform and validating it against established air monitoring stations. For this project, the geographic area of London, England was selected because there are well-established air pollution monitoring stations in the city. Additionally, air quality often fluctuates in London and is known to elicit complaints from residents. Therefore, as London residents are highly active on social media, they likely voice their concerns about air quality on Twitter. Specifically, our hypothesis was that discussion about air pollution among London residents captured on Twitter would parallel true ambient air pollution levels measured by official air monitoring stations in London.

2. Methods

2.1. Air pollution data

Air pollution data was acquired from the London Air Quality Network's (LAQN)'s LondonAir platform, which is provided by the Environmental Research Group of King's College of London (Network, 2018). LondonAir measures air pollution in the London and South East England area and is collected from 26 monitoring sites that are owned and funded by local authorities, Business Improvement Districts, Transportation of London and Department of Environment, Food and Rural Affairs (Network, 2018). Daily mean data of PM2.5 was downloaded for each of the 26 air pollution monitoring sites from July 29, 2016 to March 17, 2017, and aggregated by week (Network, 2018). Data from all 26 monitoring sites was aggregated to encompass the entire greater London area, and for matching with the location data of the tweets about air pollution. PM2.5 are particles with a diameter of 2.5 μm or less and are considered a serious threat to health since these "fine particles" can travel more deeply into the lungs and contribute to increased risk of cardiovascular and respiratory diseases (Apte et al., 2015; Cohen et al., 2005; Organization WH and UNAIDS, 2006). As the sources of PM2.5 involve combustion; these particles can soften be "sensed" immediately by the public by way of experiencing immediate difficulty breathing and other health concerns.

2.2. Twitter data

We used the GNIP Historical Powertrack service to collect publicly available tweets from July 29, 2016 to March 17, 2017. Only tweets that were geo-located to the Greater London area (using the GNIP's time zone field for London, UK) and contained terms "air pollution" or #airpollution were included in this analysis. These key words were selected because previous studies identified through probabilistic topic modeling that discussions related to air pollution typically use these key words and these words have enabled effective modeling of air pollution on other social media platforms (Jiang et al., 2015; Wang et al., 2015). Number of tweets about air pollution was aggregated by week and plotted over time for the total 33-week study period.

2.3. Media filtering

Twitter is frequently used for sharing news stories. Therefore we expected that sharing news reports on air pollution could influence users' reactions to air pollution on Twitter. To minimize this potential influence on our findings, we specifically examined discussions about air pollution that did not mention media content. Therefore, we removed any tweets that referenced media sources by removing all tweets containing a URL link from our analysis. Previous studies have shown that tweets containing URL links are typically tweets that users post to share a news article or blog post (Demirsoz and Ozcan, 2017;

Sankaranarayanan et al., 2009).

2.4. Sentiment analysis

As posts on Twitter are unsolicited, users may share their attitudes, thoughts and opinions about topics such as air quality and air pollution. Therefore, tweets sharing user opinions typically express sentiment, such as negative or positive. Since air pollution has negative effects on human health, we hypothesized that users would "sense" the harmful effects of PM2.5 and describe their experience with air pollution negatively. We expected that negative tweets (excluding shared media reports) would correlate more closely with PM2.5 measurements than positive tweets. We used natural language processing (NLP) methods to extract affective information to determine sentiment for each tweet. We used a widely accepted lexicon and rule-based sentiment classifier for microblogs called Valence Aware Dictionary for Sentiment Reasoning (VADER). VADER is based on a pattern library which is trained from human annotated words commonly found in blogs, product reviews and social media posts (Gilbert, 2014). VADAR computes sentiment for each word and generates a compound score for the sentence by summing the sentiment score of each word. Sentiment scores range from −1 (extreme negative) to +1 (extreme positive). Scores of exactly 0.0 are discarded as they indicate that there is not sufficient context. A sentiment score is considered negative if the score is ≤ − 0.5 and positive if the mean compound score is ≥0.5. Mean compound scores between −0.5 and 0.5 are considered neutral. We stratified negative and positive tweets about air pollution and plotted these weekly against PM2.5 data to determine if differences were present between the two sentiment groups.

2.5. Statistical analysis

To compare tweets about air pollution and levels of PM2.5 over time, we used cross-correlation analysis. Cross-correlation analyses the relationship between two trends and calculates the correlation coefficient to determine if there is a lag (displacement) between the two series. The maximum correlation coefficient identifies the time point that the two series (air pollution tweets and PM2.5 levels) correlate most closely and the coefficient reveals how much two series correlate with one another.

The cross-correlation is defined as:

$$r_k(X,Y) = \frac{\sum (X_t - \bar{X})(Y_{t-k} - \bar{Y})}{\sqrt{\sum (X_t - \bar{X})}\sqrt{\sum (Y_t - \bar{Y})}}$$

which calculates the sample cross correlation function (CCF) between a pair of time series at lag k. X_t, Y_t are values at time t of each series, and $\bar{X}$, $\bar{Y}$ are the mean values of each series.

Confidence interval is calculated by $\pm 1.96/\sqrt{n}$. The correlation coefficient value at lag time *k* can be considered statistically significant if it is larger than $1.96/\sqrt{n}$ where n is the sample size (Bracewell, 1986; Nghiem et al., 2016).

2.6. Stationarity through first order differencing

The assumption for cross-correlation function (CCF) is that the data is stationary, where mean and variance are independent of time. To test if the data is stationary we applied the Mann-Kendall test, which is a non-parametric test to detect trends in time series analysis. The null hypothesis of the Mann-Kendall test is that there is no upward or downward trend in the data (Yue et al., 2002). We used the seasonal Mann-Kendall tests in the R package Kendall (McLeod, 2005) to detect whether there were trends in air pollution tweets or PM 2.5 data. The Mann-Kendall test was significant indicating a trend in the data. To perform the cross-correlation function test, we conducted first order differencing on the air pollution tweets and PM2.5 data series to transform the data into "differentiated" data to make it stationary.

3. Results

During the 33-week study period, mean PM2.5 levels were 16.16 μg/m^3 per week (SD = 8.40). PM2.5 levels were lowest (7.44 μg/m^3) during the week of July 29, 2016, and reached a maximum peak of 46.99 μg/m^3 during the week of January 20, 2017. In total, 60,884 geocoded tweets about air pollution were collected from the Greater London area during the study period. The mean number of tweets about air pollution per week was 1844.97 (SD = 1542.45), ranging from 14 tweets during the week of August 12, 2016, to 6062 tweets during the week of January 20, 2017. Of the 60,884 tweets, 73% (n = 44,436) contained a URL link, indicating that these tweets likely shared news stories or blog posts. Among the 16,448 tweets not containing a media URL link, there was a mean of 498.42 (SD = 491.08) tweets per week, ranging from only 1 tweet during the week of August 12, 2016 to a maximum of 2407 tweets during the week of January 20, 2017.

Within the collection of 16,448 tweets without URL links, 3266 tweets were identified as having negative sentiment with a mean of 98.97 (SD = 159.77) negative tweets per week. This ranged from 0 negative tweets during the week of July 29, 2016 to a maximum of 903 negative tweets during the week of January 20, 2017. A total of 21,894 unique users tweeted about air pollution in this dataset. There were 1931 tweets with positive sentiment, and 11,251 tweets with neutral sentiment. Table 1 lists characteristics for these categories of tweets.

3.1. Cross-correlation

Table 2 shows cross-correlation functions between each category of Twitter air pollution tweets and the PM2.5 data series. The maximum correlation for all categories was at lag time 0 ($p < 0.001$), which indicates that there was no lag or lead-time between PM2.5 and tweets about air pollution for any category. A cross-correlation analysis between media and non-media tweets showed a coefficient of 0.835 ($p < 0.001$). There was strong cross-correlation coefficient between PM2.5 data and non-media air pollution tweets at 0.803 ($p < 0.001$), as illustrated in Fig. 1. The cross-correlation coefficient was highest between PM2.5 data and negative air pollution tweets at 0.816 ($p < 0.001$) (Fig. 2). The cross-correlation coefficients between PM2.5 data and other categories of air pollution tweets are listed in Figs. 3-6.

4. Discussion

Our results show that discussions about air pollution on Twitter appear to reflect air particle PM2.5 pollution levels in Greater London. This suggests that the public may act as effective "social sensors" of PM2.5 levels in a densely populated urban area. It was expected that there would be a strong correlation between non-media and media tweets given that both sets of tweets discuss air pollution. However, upon removing media tweets containing URL links to minimize the potential influence of sharing news stories, we found that the maximum correlation coefficient with PM2.5 levels increased from 0.666 to 0.803. As for media tweets alone (excluding non-media tweets), the maximum correlation with PM2.5 levels was reduced to 0.583. Importantly, PM2.5 and non-media tweets had similar peak periods, whereas peak periods for media tweets occurred at different time points and were more frequent than peaks for PM2.5. This shows that news media may emphasize air pollution even at moderate levels whereas the public's response to air pollution on Twitter likely reflects actual concerns and severity of air pollution as demonstrated by PM2.5 monitor readings. This is further reflected by the fact that a higher frequency of tweets was observed during period when PM2.5 levels exceeded 30 μg/m^3 on average, which is the level that the World Health Organization deems harmful to health (Brunekreef and Holgate, 2002). The largest peaks occurred when PM2.5 levels approached 50 μg/m^3, which is also where

Table 1
Description of tweets about air pollution collected for this study and PM2.5 levels over the 33-week study period in Greater London, England.

Name	Total number of tweets (*n*)	Mean tweets per week	SD	Min (*n*)	Max (*n*)	Description
Air pollution tweets						
Total tweets	60,884	1844.97	1543.45	15	6062	Tweets about air pollution (contain term air pollution, #airpollution)
Media tweets	44,436	1326.00	1100.91	14	3737	Tweets about air pollution (contain term air pollution, #airpollution) *with* media URL link.
Non-media tweets	16,448	498.42	491.08	1	2407	Tweets about air pollution (contain term air pollution, #airpollution) *without* media URL link.
Sentiment						
Non-media negative tweets	3266	98.97	159.77	0	903	Tweets about air pollution (contain term air pollution, #airpollution) *without* media URL link that are negative.
Non-media positive tweets	1931	58.52.	52.05	0	215	Tweets about air pollution (contain term air pollution, #airpollution) *without* media URL link that are positive
Non-media neutral tweets	44,336	349.94	313.00	0	1403	Tweets about air pollution (contain term air pollution, #airpollution) *without* media URL link that are neutral.
	Mean PM2.5 levels ($\mu g/m^3$)		SD	Min ($\mu g/m^3$)	Max ($\mu g/m^3$)	
PM2.5 levels	16.16		8.40	7.44	46.99	Levels of PM2.5 captured daily from the London Air Quality Network's (LAQN)'s LondonAir platform and aggregated by week.

there was the highest frequency of tweets, thereby potentially indicating that more negative health effects prompted greater response from the public in the form of tweets about air pollution. To investigate this further, we conducted comparative textual analyses of tweets within the 30–32 $\mu g/m^3$ average range and the 6–8 $\mu g/m^3$ average range. We discovered during higher average ranges of PM2.5 levels that users' non-media tweets contained more emotion and negative sentiment such as words like "worry", "suffer" and "shameful". This may explain why the coefficient between the non-media negative tweets and PM2.5 levels was the highest of all of the cross-correlations. In contrast, the media tweets between the PM2.5 average range of 6–8 $\mu g/m^3$ and the PM2.5 average range of 30–32 $\mu g/m^3$ contained fewer emotional terms and consisted of topics such as demanding that the London government "limit", "plan", and "reduce" air pollution.

Consistent with our hypothesis, sentiment analysis of non-media tweets about air pollution showed that negative tweets had the highest maximum cross-correlation of 0.816. Therefore, tweets about air pollution expressing negative sentiment were most predictive of PM2.5 levels. As expected, positive tweets reduced the maximum cross-correlation to 0.418 and were less predictive of PM2.5 levels. Frequency of negative tweets about air pollution fluctuated with comparable highs and lows as PM2.5 levels, suggesting that when PM2.5 levels are high, greater public concern is expressed on Twitter. Whereas during periods with elevated PM2.5 levels, positive tweets had no sharp peaks indicating that Twitter users did not post positive content during these times. Negative non-media tweets about air pollution included tweets such as "terrible air pollution in London this evening. you could taste the buses and the diesel emissions cheating software", "bloody hell it stinks out there. #quaxing + #rushhour + #airpollution." and "dreadful traffic jam on the strand. extremely bad air pollution today!" These tweets suggest that the public is responding to the negative effects of air pollution and voicing their concerns using Twitter, further highlighting the potential for the public to act as "social sensors". This further indicates that the public may be experiencing negative effects of air pollution and are reporting these experiences online as depicted by more negative content compared to positive content.

There are limitations with our study. First, tweets were restricted to English because the sentiment analyzer VADER is programmed for English text. The addition of other languages may increase the correlation between tweets and PM2.5 levels, and will be necessary to generalize these findings to non-English speaking settings. Second, only tweets containing geographic location information (geo-location) by time zone were included. Thus, tweets about air pollution without geographic location details were not captured. Therefore, the data may not fully represent all online communication in Greater London. However, the fact that we found such robust correlations implies that the number of "social sensors" obtained in this study through geographic identification was sufficient to reflect official PM2.5 levels. Additionally, our findings lacked granularity, as we were unable to examine correlations between individual air monitoring sites and communication on Twitter, and instead relied on aggregate data for the entire London area. Future studies should seek to examine how online communication correlates with environmental contaminants such as air pollution at the neighborhood level.

Since we restricted our analyses to Greater London, the third limitation is that our findings may not generalize to other areas. For instance, how London residents react to air pollution and post on Twitter may differ compared to other urban areas, countries, and regions of lower socioeconomic status or rural areas with lower access to the Internet and social media. This is a significant caveat of online big data because populations without Internet access are excluded from analyses. In low-resource settings with less access, there may be insufficient online "social sensors" to detect changes in air pollution. Yet, these lower income areas may be most susceptible to harmful levels of air pollution. In other countries, there may be less of a culture to post thoughts about air pollution on Twitter. However, several studies using

Table 2
Cross correlation coefficient for air pollution tweets and PM2.5 data series collected from Greater London, England.

	Maximum cross-correlation lag time	Maximum cross-correlation coefficient	95% Confidence interval	*P*-value
Total tweets[a]	Lag = 0	0.666	−0.346 − 0.346	< 0.001
Media tweets[b]	Lag = 0	0.583	−0.346 − 0.346	< 0.001
Non-media tweets[c]	Lag = 0	0.803	−0.346 − 0.346	< 0.001
Non-media negative tweets	Lag = 0	0.816	-0.346-0.346	< 0.001
Non-media positive tweets	Lag = 0	0.418	-0.346-0.346	0.018
Non-media neutral tweets	Lag = 0	0.720	-0.346-0.346	< 0.001

[a] Total tweets includes the full sample of air pollution tweets (n = 60,884).
[b] Media tweets refers to only tweets about air pollution that also contain a URL link (n = 44,436) because the link likely indicates sharing of a news story or blog post.
[c] Non-media tweets refers to tweets about air pollution excluding any tweets containing a URL link (n = 16,448).

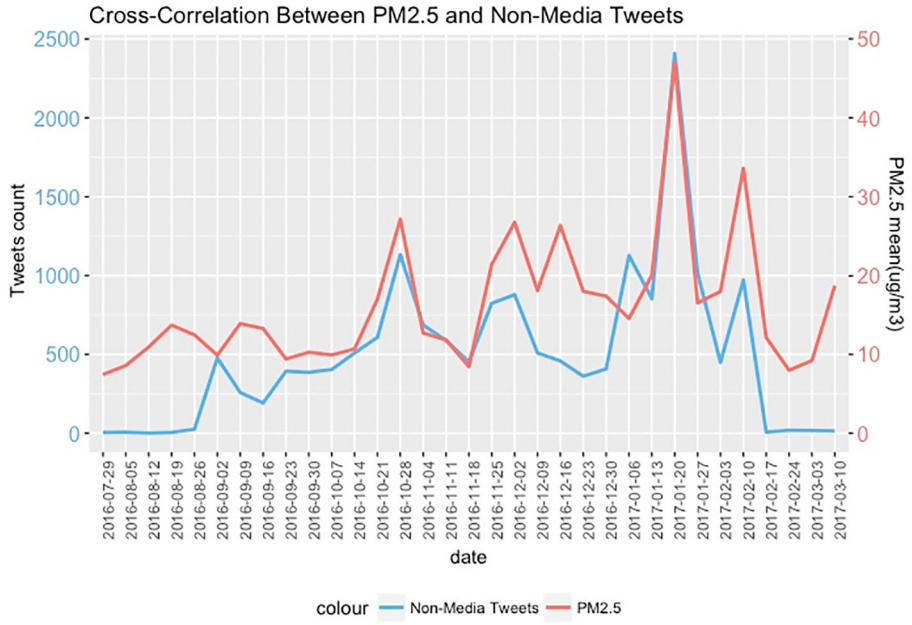

Fig. 1. Cross-Correlation between PM2.5 levels and non-media tweets about air pollution from Greater London, England.

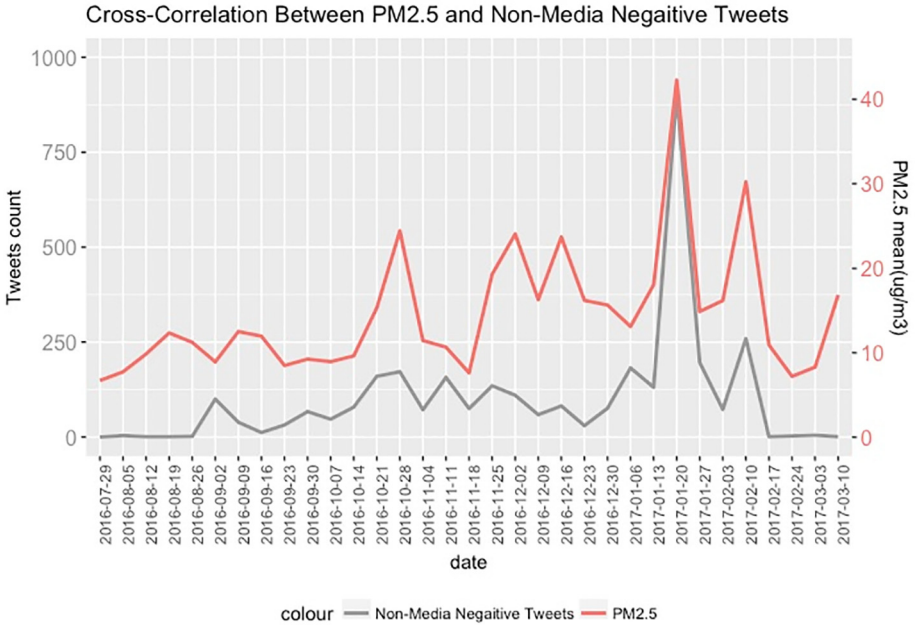

Fig. 2. Cross-Correlation between PM2.5 levels and non-media tweets about air pollution with negative sentiment from Greater London, England.

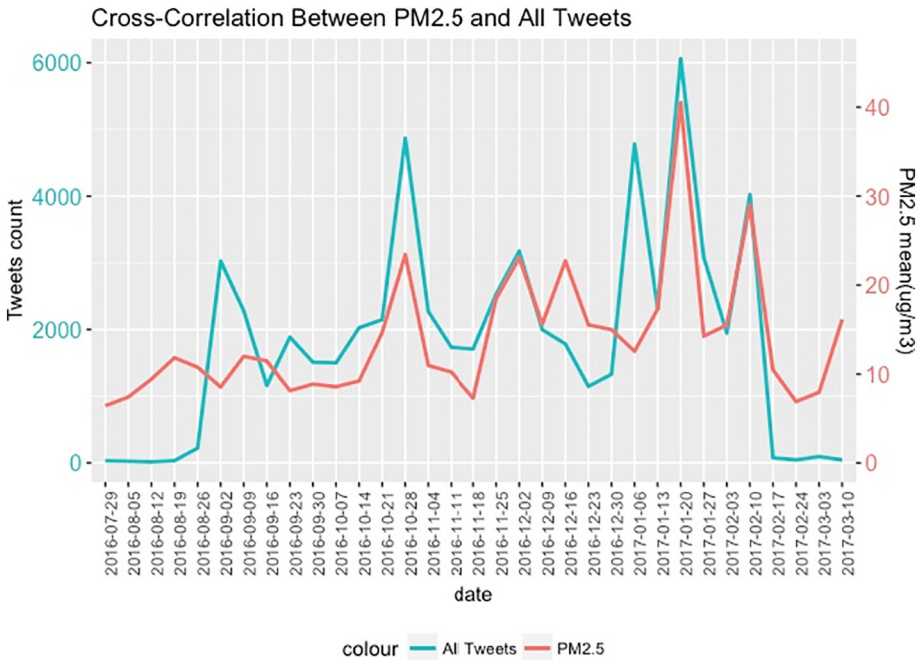

Fig. 3. Cross-Correlation between PM2.5 levels and total tweets about air pollution (tweets containing media URL link and non-media tweets) from Greater London, England.

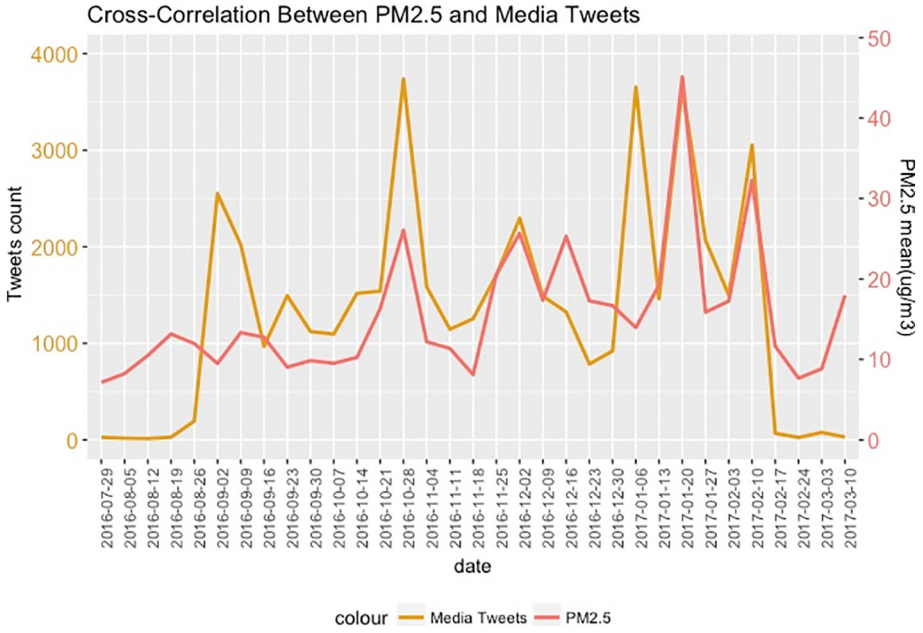

Fig. 4. Cross-Correlation between PM2.5 levels and tweets about air pollution containing media URL links from Greater London, England.

data from Sina Weibo (a popular social media platform in China) have demonstrated feasibility for tracking air pollution in China (Jiang et al., 2015; Wang et al., 2015), where volume of pollution-related messages was highly correlated with particle pollution levels (Wang et al., 2015). Furthermore, one study found that Twitter data appears promising for detecting smoke pollution from wildfires in California, USA (Sachdeva and McCaffrey, 2018), while another study demonstrated that capturing photos posted on social media can supplement existing meteorological data and satellite imagery for monitoring haze events in Indonesia (Khaefi et al., 2018). More recently a systematic review showcased that data from Twitter can provide new opportunities to study health, especially among underrepresented geographic areas and at-risk patient groups (Sinnenberg et al., 2017). Finally, we based our key search terms about air pollution on terms that have been previously used in the literature (Jiang et al., 2015; Wang et al., 2015), and therefore we may not have included a broader range of terms that are potentially reflective of air pollutants. While the search terms that we employed offer greater certainty that we were capturing true online discussions about air pollution, a more comprehensive list of search terms related to health impacts of air pollution may have improved our cross-correlation coefficients. Future research is needed to determine whether the inclusion of a diverse range of search terms, and in particular search terms related to health effects, offer the potential to improve the ability to monitor harmful air pollutants using social media data in urban areas.

Despite these limitations, our results demonstrate the potential to

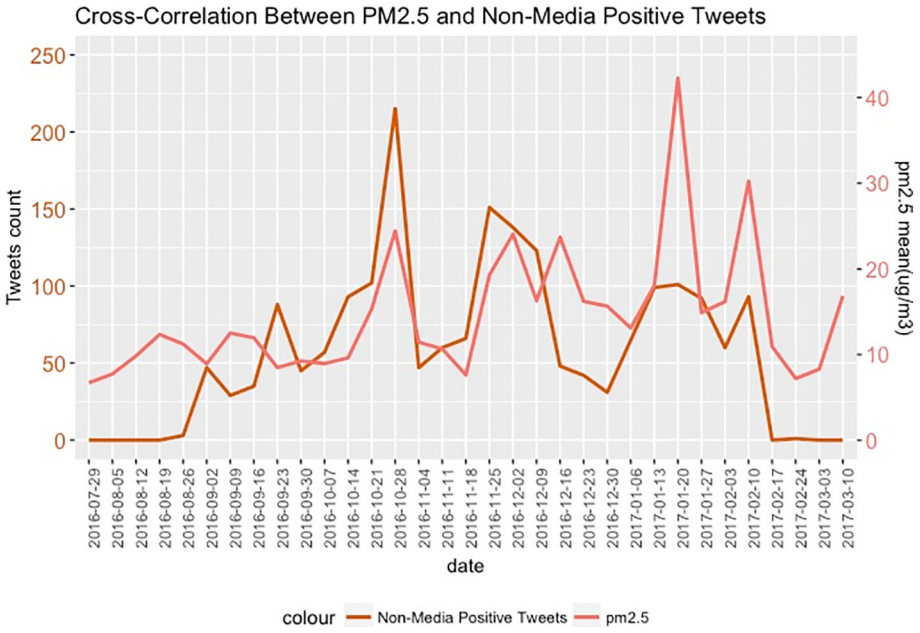

Fig. 5. Cross-Correlation between PM2.5 levels and non-media tweets about air pollution with positive sentiment from Greater London, England.

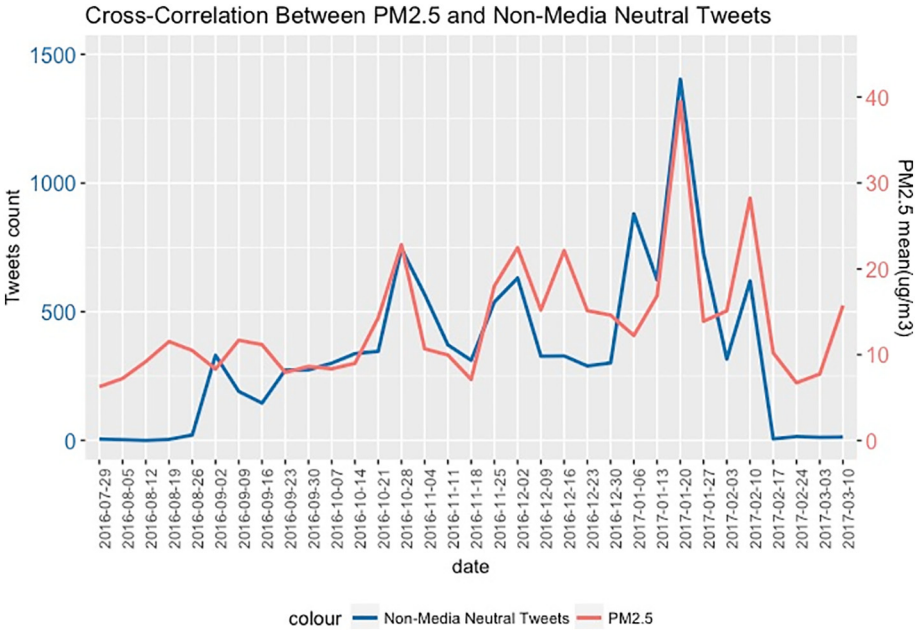

Fig. 6. Cross-Correlation between PM2.5 levels and non-media tweets about air pollution with neutral sentiment from Greater London, England.

use Twitter as a novel surveillance approach for detecting temporal changes in air pollution within a densely populated urban area. An important strength is that we removed media tweets from our analyses and identified that users' unsolicited online posts about air pollution could predict PM2.5 levels without the influence of news reports. Sentiment analysis of tweets has been used to improve prediction of movie box office revenue (Asur and Huberman, 2010), and we determined that sentiment could also improve air pollution prediction using Twitter.

The Mayor of London has committed to reducing concentration levels of PM2.5 according to WHO guidelines by 2030 (Greater London Authority City Hall, 2018). As part of this goal he has urged the public to take action in monitoring air pollution (Greater London Authority City Hall, 2018; Mittal and Fuller, 2017). However, costly and technical air pollution monitoring systems do not afford the opportunity for each person to take part in reporting air pollution hazards. Therefore, it is essential to find alternative methods to supplement existing efforts for detecting air pollution. Our study provides an alternative method to expensive air quality sensors by leveraging online discussion about air pollution on Twitter to track fluctuations in PM2.5 levels consistent with established air monitoring systems.

In areas without air pollution monitoring systems, the approach described here may afford new opportunities to detect air pollution hazards and prevent health risks to the public. Since air pollution monitors only cover certain areas and extrapolation is used to estimate air pollution outside monitoring areas, online big data can potentially yield fine-grained details necessary to fill in gaps in data from existing reporting systems. This type of citizen-led monitoring can be used to better understand the public's interaction with air quality issues and use these discussions as a resource to reduce their exposure or engage the

public in making changes in their behavior that could lead to improvements in air quality. As such, our study highlights an important opportunity for social media to provide a supplemental source to assist in the detection and monitoring of air pollution, and to help mitigate population health risks.

Acknowledgments

This study was funded by: the Canadian Institutes of Health Research (to YH); the Weatherhead Center for International Affairs, Harvard University (YH); the Robert Wood Johnson Foundation Grant 73495 (to YH, JBH); and the NIH/National Human Genome Research Institute Grant 5U54HG007963-04 (to JSB, JBH). The funders played no role in the study design; collection, analysis, or interpretation of data; writing of the manuscript; or decision to submit the manuscript for publication.

Conflict of interest statement

No financial disclosures were reported by any of the authors of this paper.

References

Adams, H., Nieuwenhuijsen, M., Colvile, R., McMullen, M., Khandelwal, P., 2001. Fine particle (PM2. 5) personal exposure levels in transport microenvironments, London, UK. Sci. Total Environ. 279 (1–3), 29–44.

Apte, J.S., Marshall, J.D., Cohen, A.J., Brauer, M., 2015. Addressing global mortality from ambient PM2. 5. Environ. Sci. Technol. 49 (13), 8057–8066.

Asur, S., Huberman, B.A., 2010. Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. vol. 01. IEEE Computer Society, pp. 492–499.

Birnbaum, M.L., et al., 2017. Using digital media advertising in early psychosis intervention. Psychiatr. Serv. 68, 1144–1149. https://doi.org/10.1176/appi.ps.201600571.

Blankers, M., Nabitz, U., Smit, F., Koeter, M.W.J., Schippers, G.M., 2012. Economic evaluation of internet-based interventions for harmful alcohol use alongside a pragmatic randomized controlled trial. J. Med. Internet Res. 14, e134. https://doi.org/10.2196/jmir.2052.

Bracewell, R.N., 1986. The Fourier Transform and Its Applications. McGraw-Hill, New York.

Broniatowski, D.A., Paul, M.J., Dredze, M., 2013. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 8 (12), e83672.

Broomfield, M., 2017. London's air Pollution Worse than Beijing's as Smog Chokes UK Capital. The Indepedent.

Brunekreef, B., Holgate, S.T., 2002. Air pollution and health. Lancet 360 (9341), 1233–1242.

Charron, A., Harrison, R.M., 2005. Fine (PM2.5) and coarse (PM2.5-10) particulate matter on a heavily trafficked London highway: sources and processes. Environ. Sci. Technol. 39 (20), 7768–7776.

Cohen, A.J., et al., 2005. The global burden of disease due to outdoor air pollution. J. Toxic. Environ. Health A 68 (13–14), 1301–1307.

Demirsoz, O., Ozcan, R., 2017. Classification of news-related tweets. J. Inf. Sci. 43 (4), 509–524.

Devarakonda, S., Sevusu, P., Liu, H., Liu, R., Iftode, L., Nath, B., 2013. Real-time air quality monitoring through mobile sensing in metropolitan areas. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. ACM, pp. 15.

European Environment Agency, 2008. Air Pollution: Air Pollution Harms Human Health and the Environment. Copenhagen, Denmark.

Gardiner, B., 2014. Air of Revolution: How Activists and Social Media Scrutinise City Pollution. The Guardian. Guardian News and Media Limited, London, England.

Gerlach-Reinholz, W., Drop, L., Basic, E., Rauchhaus, M., Fritze, J., 2017. Telefoncoaching bei Depression Telephone coaching for depression. Nervenarzt 88, 811–818. https://doi.org/10.1007/s00115-017-0316-0.

Gilbert, C.H.E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International Conference on Weblogs and Social Media (ICWSM-14), . Available at (20/04/16) http://compsocialgatechedu/papers/icwsm14vaderhuttopdf.

Goddard, C., 2017. All about… air pollution. Nurs. World 2017 (12), 21–24.

Greater London Authority City Hall, 2018. Guide for Air Quality in London. The Queen's Walk, Greater London Authority City Hall.

Gruebner, O., Lowe, S.R., Sykora, M., Shankardass, K., Subramanian, S., Galea, S., 2017. A novel surveillance approach for disaster mental health. PLoS One 12 (7), e0181233.

hackAIR, 2018. Citizen activism: tackling air pollution across Europe. In: hackAIR Collective Awareness for Air Quality, . http://www.hackair.eu/citizen-activism-tackling-air-pollution-across-europe, Accessed date: 29 June 2018.

Ho, R.C., Zhang, M.W., Ho, C.S., Pan, F., Lu, Y., Sharma, V.K., 2014. Impact of 2013 south Asian haze crisis: study of physical and psychological symptoms and perceived dangerousness of pollution level. BMC Psychiatry 14 (1), 81.

Hswen, Y., Naslund, J.A., Chandrashekar, P., Siegel, R., Brownstein, J.S., Hawkins, J.B., 2017. Exploring online communication about cigarette smoking among Twitter users who self-identify as having schizophrenia. Psychiatry Res. 257, 479–484.

Ilieva, R.T., McPhearson, T., 2018. Social-media data for urban sustainability. Nature Sustain. 1 (10), 553.

Java, A., Song, X., Finin, T., Tseng, B., 2007. Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM, pp. 56–65.

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.

Khaefi, M.R., Pramestri, Z., Amin, I., Lee, J.G., 2018. Nowcasting air quality by fusing insights from meteorological data, satellite imagery and social media images using deep learning. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, pp. 393–396.

Kim, Y., et al., 2018. Air pollution and suicide in 10 cities in Northeast Asia: a time-stratified case-crossover analysis. Environ. Health Perspect. 126 (3), 037002.

Landrigan, P.J., et al., 2017. The Lancet commission on pollution and health. Lancet 391 (10119), 462–512.

McIver, D.J., et al., 2015. Characterizing sleep issues using Twitter. J. Med. Internet Res. 17 (6), e140.

McLeod, A.I., 2005. Kendall Rank Correlation and Mann-Kendall Trend Test. (R Package Kendall).

Mittal, L., Fuller, G., 2017. London Air Quality Network: Summary Report. King's College, London, pp. 2016 (Technical report).

Network, L.A.Q., 2018. PM2.5 Data. In: London Air Quality Network. King's College of London. https://www.londonair.org.uk/LondonAir/General/about.aspx.

Nghiem, L.T., Papworth, S.K., Lim, F.K., Carrasco, L.R., 2016. Analysis of the capacity of Google Trends to measure interest in conservation topics and the role of online news. PLoS One 11 (3), e0152802.

Nsoesie, E., Hawkins, J., Tuli, G., Kluberg, S., Brownstein, J., 2016a. The use of social media and business reviews for foodborne illness surveillance. Int. J. Infect. Dis. 53 (70).

Nsoesie, E.O., et al., 2016b. Social media as a sentinel for disease surveillance: what does sociodemographic status have to do with it? PLoS Curr. 8.

Organization WH, UNAIDS, 2006. Air Quality Guidelines: Global Update 2005. World Health Organization.

Pultarova, T., 2017. Better hold your breath: London's killer air. Eng. Technol. 12 (4), 42–45.

Sachdeva, S., McCaffrey, S., 2018. Using social media to predict air pollution during California wildfires. In: Proceedings of the 9th International Conference on Social Media and Society. ACM, pp. 365–369.

Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J., 2009. Twitterstand: news in tweets. In: Proceedings of the 17th ACM Sigspatial International Conference on Advances In Geographic Information Systems. ACM, pp. 42–51.

Sass, V., Kravitz-Wirtz, N., Karceski, S.M., Hajat, A., Crowder, K., Takeuchi, D., 2017. The effects of air pollution on individual psychological distress. Health Place 48, 72–79.

Sinnenberg, L., Buttenheim, A.M., Padrez, K., Mancheno, C., Ungar, L., Merchant, R.M., 2017. Twitter as a tool for health research: a systematic review. Am. J. Public Health 107 (1), e1–e8.

Statista, 2018. Number of Monthly Active Twitter Users Worldwide From 1st Quarter 2010 to 1st Quarter 2018 (In Millions). Statista. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users, Accessed date: 30 June 2018.

Szyszkowicz, M., Rowe, B., Colman, I., 2009. Air pollution and daily emergency department visits for depression. Int. J. Occup. Med. Environ. Health 22 (4), 355–362.

Wang, S., Paul, M.J., Dredze, M., 2015. Social media as a sensor of air quality and public response in China. J. Med. Internet Res. 17 (3), e22.

Yue, S., Pilon, P., Cavadias, G., 2002. Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. J. Hydrol. 259 (1–4), 254–271.

Zhang, M.W., Ho, C.S., Fang, P., Lu, Y., Ho, R.C., 2014. Usage of social media and smartphone application in assessment of physical and psychological well-being of individuals in times of a major air pollution crisis. JMIR mHealth uHealth 2 (1), e16.