Using social media to analyze air pollution in China

Dan Li [1] · **Weiwei Zhou**[1] · **Meng Niu**[1] · **Xianyun Tian**[2]

Abstract

Air pollution in China has increased due to the rapid industrialization of the country, which has caused severe environmental and health problems. As a way of exploring the feasibility of using social media to understand air pollution in China, this study presents a multi-faceted analysis of 630 million air pollution-related postings on Sina Weibo (China's equivalent of Twitter). From the thematic analysis used to identify themes of filtered air-pollution-related postings, it was found that disclosure and dissatisfaction with air pollution were the most common themes. Based on a collection of health-related self-reports, respiratory-related diseases and headache were found to be the most common health effects in air pollution areas. It was found that individual posting frequencies were strongly correlated to the pattern of China's air quality index (AQI): users living in areas of high air pollution were more active on Sina Weibo than users living in areas with negligible air pollution. Furthermore, complaints of air pollution were found to be more serious in southern cities where industries and economies are generally more developed. The study found that social media messages could augment data on perceptions of air pollution, which is traditionally hard to capture, and complement data on air pollution related health effect based on informal self-reports. Social media can also complement air pollution surveillance data, which is currently expensive to retrieve, at a much lower cost.

Keywords: social media; air pollution; public health; haze; Sina Weibo; public perception

1. Introduction

In recent years, social media has served as a communication and information-sharing platform for citizens around the world, including China. The diverse social media messages are taken into consideration to detect and investigate changes in the real world. And users of social media find it easier to express their opinions online and thus make it possible for the platform to provide real-time updates of social and human activities. Sina Weibo, as the most influential Chinese micro-blogging website (similar to Twitter), enables more people to communicate with others and to be actively involved in social events; it has more than 500 million registered users and processes over 100 million postings per day. The content shared on Sina Weibo involves a wide range of topics, including social, economic, political, and environmental issues (Yu et al., 2011). Recent studies have demonstrated that postings from social media are reflection of collective human behaviors and can further monitor real-time updates of social events (Sasahara et al., 2013). Therefore it can be used as a rich data source for many different research goals.

Dan Li
ld@nefu.edu.cn

[1] Information and Computer Engineering College, Northeast Forestry University, Harbin, PR China
[2] School of Management, Harbin Institute of Technology, Harbin, PR China

Air pollution has become a major concern in China, as it is having a great impact on human health and the environment. Specifically, air pollution causes eye irritation, allergies, as well as respiratory and cardiovascular diseases; in extreme cases it can even cause death (Xu et al., 2013). It is estimated that air pollution in China is killing 4000 people every day. In addition, due to the health impacts of air pollution, large amounts of money are being lost in economic productivity, medical expenses, lost wages, and lost leisure time (Gerking and Stanley, 2017).

It is therefore important to enhance public awareness air pollution. One solution might be to measure public perception and response to air pollution (Mollema et al., 2015), such as people’s behaviors to prevent discomfort, and use this knowledge to advise the public on how to protect themselves from air pollution. Social media is a perfect candidate for disseminating this knowledge and providing information for researchers. Since people often discuss air pollution issues on Sina Weibo, it is a great tool for obtaining relevant data.

Recent studies have demonstrated that social media has broad applicability for public health research (Cheng et al., 2010), including disease surveillance, mental health trends, and health perceptions. Wang et al. (2014) comprehensively explored a variety of Chinese public health topics in social media and discovered that air pollution was being discussed, which had not been found earlier with Twitter. This indicates that air pollution is a specific health issue in China. Some studies have shown that levels of air pollution can be estimated based on complaints of poor air quality in social media. Some researchers have identified social media as a potential air quality sensor (Jiang et al., 2015; Jiang et al., 2016; Mei et al., 2014). Different from physical sensors of pollutant levels, social media-based analysis measures responses to public health issues related to air pollution. For example, a case study focusing on the 2013 Harbin haze disaster showed the public's response to air quality, thereby providing a method to guide people correctly (Wang and Bai, 2014). The results showed the feasibility of understanding people’s behavior in a haze disaster through social media. With the knowledge of public awareness, concern, attitudes, health effects, and behavioral response to air pollution, it is possible to direct the public on how to best protect themselves and guide public policy to reduce air pollution.

An analysis by Kay et al. (2015) who applied a simple keyword-based filter to select messages and perform a text classification experiment, concluded that Sina Weibo postings contain rich details including perceptions, behaviors, and self-reported health effects of air pollution. However, their methodology was not able to identify what the specific themes were and the training set of 170 messages was small. In other words, new algorithms to select the postings with large sample sizes need to be developed. Although the feasibility of using social media to understand public health and public response to air pollution has been validated, the use of social media and some statistical characteristics of air pollution have not been well studied.

The main purpose of this study therefore is to investigate the public health effect and public perception of air pollution, measuring the statistical characteristics and identifying air quality trends. Firstly, we applied a self-designed web crawler to extracted massive air pollution-related postings from Sina Weibo. Then in the data selection phase, a keyword filter method and different machine learning algorithms were used to accurately classify the filtered postings. Secondly, we conducted some experiments to evaluate the ability of social media, including exploring themes and health-related topics of the filtered postings, identifying the relationship with AQI and analyzing statistical characteristics of air pollution.

2. Research Methodology

2.1 Data collection

Sina Weibo does not provide "streaming" API tools, which are commonly used to collect random subsamples from Twitter. Instead, we used free API for Sina Weibo to query all the information from a given user but without publishing any individual information. Starting with a set of 1 million randomly sampled users, we collected all the postings and personal information of every user. During the period October 5 to October 25, 2017, we mined a collection of approximately 635 million (635,319,848) messages from Sina Weibo using the Sina Weibo API and a self-designed web crawler. The detailed information contains several attributes, such as nickname, date, content of original individual messages or reposted messages, numbers of comments, the number of followings, and the number of followers. Since we mainly relied on the Sina Weibo text and counted embedded re-posts as part of the text, multimedia messages such as images and videos were not included in our dataset.

2.2 Text pre-processing and classification

In order to filter only air pollution-related data, a keyword-set including the following keyword: "air pollution" ("空气污染", "大气污染"), "haze"("霾"), "pm2.5", "PM2.5", "smog"("烟雾") was built. Messages that contain any of those keywords were identified and a collection of 440,907 air pollution-related postings were selected. However, not all messages matching our keyword-set are related to air pollution. To obtain a more accurate and smaller dataset, a new text classifier was trained using a machine-learning algorithm to distinguish air pollution-related postings from non-related postings.

A machine learning classifier learns to assign a category or tag to a text by using mathematical models and algorithms. It associates a particular input text to a corresponding label. But at first, a training dataset is necessary for the classifier to learn. Thus we labeled 10,000 postings randomly selected from the filtered dataset by identifying whether it is related to air pollution.

Two annotators independently performed the labeling data process and were required to discuss their findings with each other until they reached agreement. At the same time, a third annotator labeled the same 10,000 posts without any discussion. We took the majority of the three lists of code as the final code of the posts and the agreement between the two primary annotators was measured using Cohen's kappa score. We then experimented with a support vector machine (SVM), which is a supervised machine-learning algorithm, using the coded posts as training data. Ten-fold cross-validation method and accuracy were chosen as two performance estimation methods of our classifier. With a high accuracy of 88.79%, we applied our classifiers to the 440,907 posts and 187,234 posts were selected. The flowchart of text pre-processing and classification is presented in Fig.1.

2.3 Analysis of content

To measure public perception and concern about air pollution, an analysis of topics is necessary. First 1,600 postings were randomly selected from our dataset that contained only the air pollution-related posts as a sample to perform coding. Two annotators of our team were then trained to determine the most common themes in the 1,600 postings. Topics that occur repeatedly were defined as themes (Ryan and Bernard, 2016). The 1,600 postings were coded as follows: (1) disclosure of air pollution; (2) dissatisfaction with the poor air quality; (3) report of behavioral changes (e.g. wearing a mask or staying indoors); (4) concerns for his or her health; (5) evaluation of government policy; (6) banter about the poor air quality; (7) appeal to the government to reduce air pollution.

The two annotators discussed their results and then disagreements were solved. In order to enhance credibility, a third annotator was asked to code 300 postings randomly selected from the 1,600 postings. For each theme, inter-coder reliability was as shown in Table 1 below.

Table 1 Analysis of content

Topics	Kappa value	Agreement (%)
Disclosure of air pollution	0.881	95
Dissatisfaction with the poor air quality	0.787	89
Report of behavioral changes	0.840	91
Health concern for his or her health	0.901	97
Evaluation of government policy	0.821	89
Banter about the poor air quality	0.782	92
Appeal to the government to reduce air pollution	0.845	95

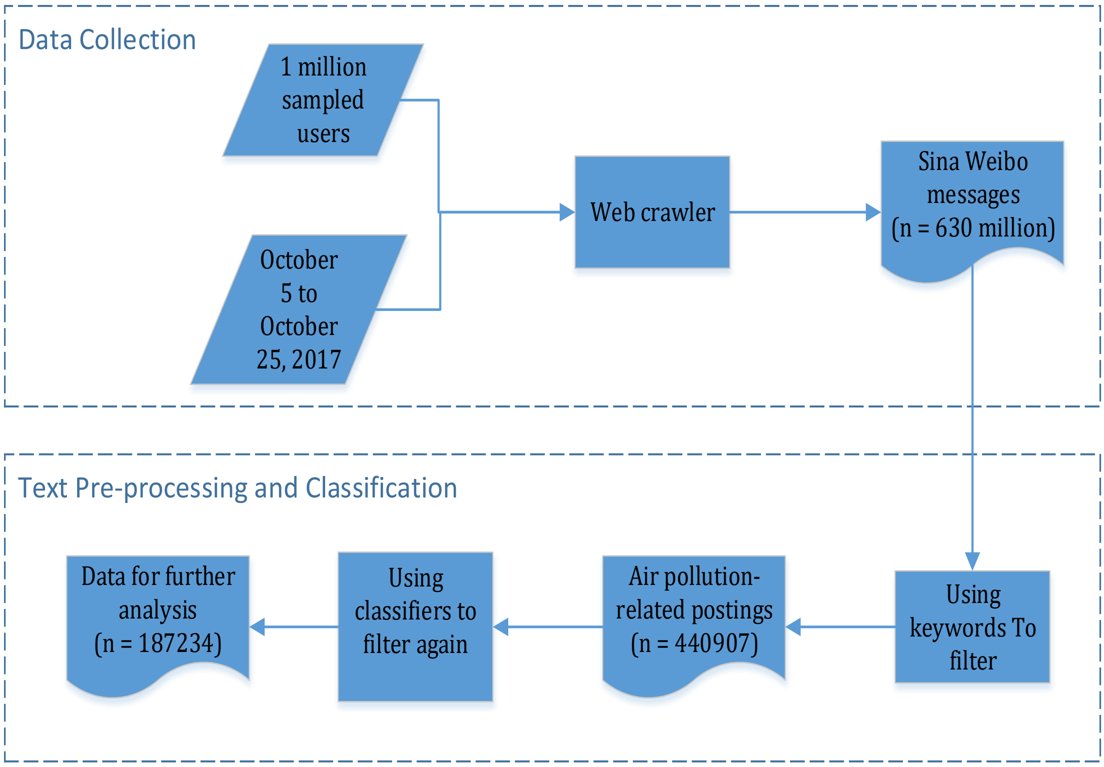


Fig. 1 Flowchart of text pre-processing and classification

2.4 Relationship with AQI (Air quality index)

For analyzing the dynamic relationship between social media messages and air quality in the real world, we performed experiments on Sina Weibo messages and AQI data from eight cities in 2016. Since air pollution occurs in urban areas more frequently than rural areas (D'Amato et al., 2010), eight typical big cities in China were used as target cities and were divided into two groups: (1) first-tier cities, including Beijing, Shanghai, Guangzhou, Shenzhen; and (2) second-tier cities, including Hangzhou, Nanjing, Chongqing, Chengdu. Two sets of data from these eight cities were collected. One was the AQI data obtained from China's Ministry of Environmental Protection. The AQI is an index for reporting daily air quality of cities; the higher the AQI value, the greater the level of air pollution and health concern is. The other set of data was air pollution-related posts of the eight target cities during a one-year period after having been filtered for relevance.

Correlation was investigated on a per-month base because the air quality changes more obviously from month to month. As a result, the average AQI per month of each city and the total messages about air pollution of each city per month in 2016 were used as our dataset.

Two experiments were conducted. First, we computed the Pearson correlation coefficients between the number of postings and the AQI data for each month in 2016. Second, based on the correlation, we drew the trends of AQI and the number of postings, and then analyzed the pattern.

2.5 Health-related topics among users

To identify and characterize health-related topics when people discussed air pollution, a probabilistic topic model, Latent Dirichlet Allocation (LDA), was employed to estimate the prominent health topics (Ou et al., 2015). We used Jieba to tokenize the 187,234 postings, then removed Chinese stop words and infrequently used words from the postings. After preprocessing, the original data became a vocabulary list containing 38,758 words and then the topics describing the symptoms of air pollution were discovered by LDA. With these topics, the keyword-filter method was applied again to obtain a collection of approximately 22100 health-related posts. By calculating the occurrence frequency of each symptom, we were able to evaluate the impacts of air pollution on public health.

2.6 Characteristic analysis of users

In this section, the demographic characteristics of users who posted air pollution-related postings were investigated. The demographic characteristics consist of nickname, gender, the location of an individual, the number of followers, the number of followings, and the number of postings.

For the experimental group, the trained text classifier mentioned above obtained 187,234 air pollution-related postings and the corresponding personal information of each user was crawled by our Sina Weibo crawler using the open API. For the control group, 600 million postings were randomly selected from Sina Weibo.

In order to analyze the geographic distribution of the users who experienced air pollution, we divided the users into provinces (or autonomous region, municipality, special administrative region). The number of users who posted air pollution-related postings for each province were first normalized and then multiplied by 100 to make it more intuitionistic; this was denoted as the UPPI (Users Per Province Index). In order to illustrate air quality of each province, we divided the measure into five categories: Good, Moderate, Unhealthy for Sensitive groups, and Unhealthy. Each category corresponded to a different level of health concern so as to illustrate the air quality and health concerns of each city.

3. Results

3.1 Analysis of content

A codebook consisting of 1,600 randomly selected codes was created, which were air-pollution related. In this process, of the 1,600 postings, 205 postings (12.81%) were found useless (most of them were advertisements), and thus were excluded from our dataset. The themes of the remaining 1395 postings are shown in Table 2.

Table 2 Themes of postings

Theme	N (%)	Example
Disclosure of air pollution	403(28.89)	The haze is back today. 今天雾霾又来了。
Dissatisfaction with the poor air quality	345(24.73)	The haze is everywhere, extremely dirty. 一路雾霾啊,真的是脏到极点了。
Report of behavioral changes	109(7.81)	I dare not run in this severe smog. 雾霾天不敢出去跑步了。
Health concern for his or her health	112(8.03)	The haze is too severe, I have a sore throat. 雾霾太严重啦，喉咙痛呀。
Evaluation of government policy	62(4.44)	Traffic restriction is of no use! 限行也降低不了雾霾!
Banter about the poor air quality	54(3.87)	In the smog, people cannot be seen clearly, so everyone looks more beautiful. 雾霾天，都看不清人，大家都变漂亮了。
Appeal to the government to reduce air pollution	30(2.16)	Why doesn't the government do something about haze! 政府为什么不管管雾霾!

The majority of postings (28.89%) disclosed the existence of air pollution. Nearly a quarter of the postings (24.73%) showed dissatisfaction with the poor air quality. Approximately 8% (8.03%) expressed health concerns, including cough, sore throat, and streaming eyes. Another 8% (7.81%) of the postings disclosed people’s behavioral changes, such as staying indoors, exercising less often, or wearing masks. About 5% (4.44%) were negative evaluation of government policy, while about 2% (2.16%) were people appealing for effective solutions to reduce air pollution. Nearly 4% (3.87%) of the postings was general banter about poor air quality. Overall, the identified themes suggest that poor air quality affects people’s physical and mental health and causes behavioral changes. It is also very apparent from users' who are willing to talk about air pollution-related topics on social media that there is a need for effective methods and relevant regulations to reduce air pollution.

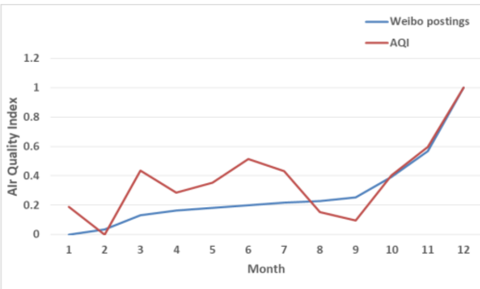
3.2 Relationship with AQI

Based on the Pearson correlation analysis between air-pollution related posting frequency and average monthly AQI, as shown in Table 3, we found that the dynamics of the AQI were consistent with the pattern of individual messages about air pollution for most months. As for Beijing, Shanghai, Hangzhou, Nanjing, Chongqing and Chengdu, the Pearson correlation coefficient of these six cities was above 0.7, and the correlation between posting frequency and the AQI value was significant at the 0.05 level, which shows that the posting frequency was strongly correlated to the AQI value in these six cities. However, for Guangzhou and Shenzhen, the sig. value of both cities was greater than 0.05, meaning little statistical significance, and the Pearson correlation coefficient of these two cities was moderate (0.3–0.5). Hence, posting frequency is not strongly correlated to the AQI value in Guangzhou and Shenzhen, two of the most southern cities in China.

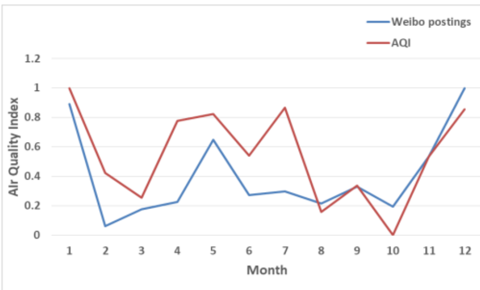
Table 3 Pearson correlation analysis between posting frequency and average monthly AQI

City	Pearson coefficient	Sig. (2-taild)
Beijing	0.843	<0.05
Shanghai	0.674	<0.05
Guangzhou	0.440	0.152
Shenzhen	0.375	0.230
Hangzhou	0.691	<0.05
Nanjing	0.735	<0.05
Chongqing	0.730	<0.05
Chengdu	0.794	<0.05

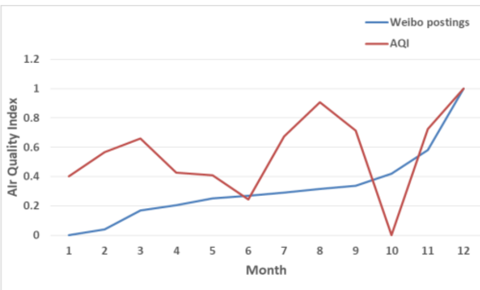
For each city, the trends of postings (blue line) and the AQI (red line) for each month in 2016 is shown in Fig. 2.



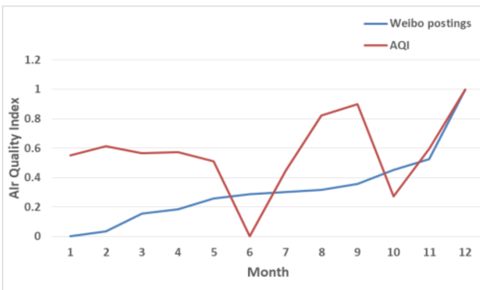
(a) Beijing



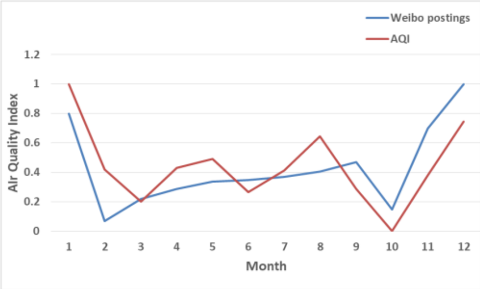
(b) Shanghai



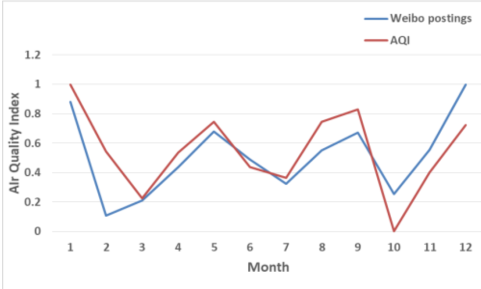
(c) Guangzhou



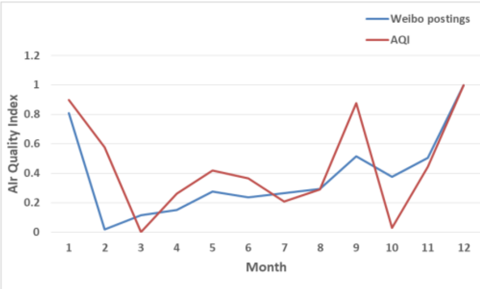
(d) Shenzhen



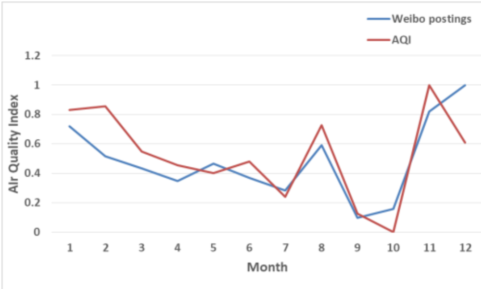
(e) Hangzhou



(f) Nanjing



(g) Chongqing



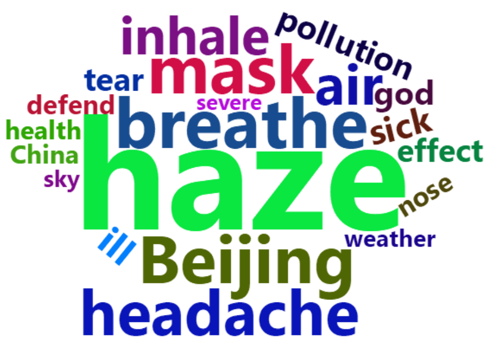
(h) Chengdu

Fig. 2 Trends of postings and AQI for 8 cities from January to December 2016

3.3 Health-related topics among users

The LDA parameters were obtained after iterating the LDA model 1,000 times and classifying 50 topics. Of the 50 topics, 2 topics highly related to air quality and health effects are shown in Fig. 3 as word clouds.

(a) Topic 1



(b) Topic 2

Fig. 3 Topic word clouds

Each topic contains 20 highest-probability words. Larger words are more probably related to the topic. The words were translated from the original Chinese.

Both topic clouds reveal that the word ‘haze’ has a strong correlation with the word ‘Beijing’ which is the only city name revealed in both topics clouds, indicating that air pollution in Beijing is the most serious. The word ‘mask’ is also a frequent word in the word clouds, suggesting that wearing masks is a common behavioral change on haze days.

The two topic clouds demonstrate that respiratory diseases and allergy symptoms are the most common health effects on haze days. The topic word cloud on the left of the figure contains mainly respiratory-related words including the words ‘cough’, ‘respiratory tract’, and ‘throat’. The topic word cloud on the right indicates allergy symptoms including 'headache', 'ill', and 'sick'.

3.4 Characteristic analysis of users

Since expression of dissatisfaction with haze and poor air quality is a common theme on Sina Weibo, we analyzed demographic characteristics of two user groups: the experimental group (users who posted haze-related postings) and the control group. Both are shown in Table 4 below.

Active users tend to have more followings and followers than less active users on social media (Bakshy et al., 2011). The results indicate that users who tend to express concern about poor air quality are more active than typical users. The probable explanation for this is that due to suffering from poor air quality and a lack of related functional departments to complain to, people use Sina Weibo as an open and influential social media to express their dissatisfaction and call upon the government to provide effective solutions.

Table 4 Demographic characteristics of users

Demographic Variables		Experimental Group	Control Group
Followings			
	Quartile 1	204	82
	Median	382	170
	Quartile 3	1168	317
Followers			
	Quartile 1	178	97
	Median	355	188
	Quartile 3	845	338
Postings			
	Quartile 1	988	93
	Median	2515	314
	Quartile 3	7581	821

Users who posted air pollution-related postings were allocated to a province, which was denoted as UPPI (Users Per Province Number) value. Five levels of health concern were divided according to UPPI value to illustrate the air quality of 34 provinces (or autonomous region, municipality, or special administrative region). Fig. 4 and Table 5 show that the air quality of most of them was relatively good.

However, the air quality of economically prosperous provinces, like Beijing, Guangdong, Shanghai was very unhealthy, as was the air quality of industrial provinces like Jiangsu, Shandong, Jilin. This phenomenon suggests that fast-paced urbanization and excessive industrialization may be the cause of air pollution.

4. Discussion

This study explored air pollution-related postings based on social media data in order to get a deeper understanding of air pollution through a multi-faceted analysis. A dataset of 187,234 air pollution-related postings were filtered by a classifier from 600 million general postings on Sina Weibo. The data were analyzed from the following four aspects.

Table 5 Air quality of 34 provinces

UPPI Value	Levels of Health Concern	The Number of cities
0 to 10	Good	19
10 to 20	Moderate	8
20 to 30	Unhealthy for sensitive groups	3 (Jiangsu, Shandong, Jilin)
30 to 50	Unhealthy	2 (Shanghai, Guangdong)
50 to 100	Very unhealthy	1 (Beijing)

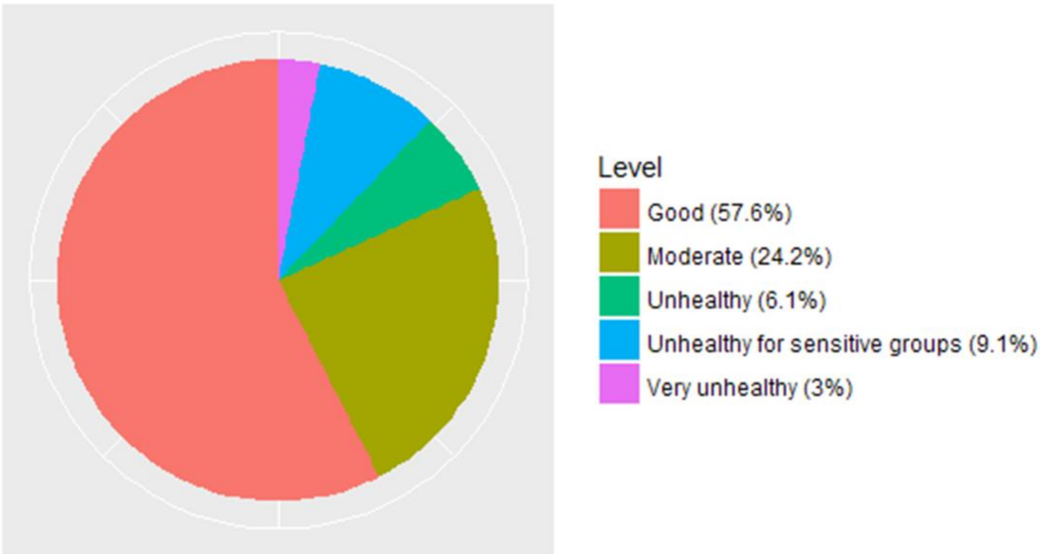


Fig. 4 Pie chart of air quality level of 34 provinces

4.1 What aspects of air pollution were discussed

An analysis of content suggests an alternative way to understand public perceptions, health responses, and behavioral changes. The different proportion of themes suggests that users who posted air pollution-related postings tended to not only disclose the existence of air pollution but also to appeal for governments' help, which may be due to the lack of a direct way to find help from relevant functional departments. Previous studies have suggested that public complaints can not only enhance supervision but also can promote further investigations by officials (Dasgupta and Wheeler, 1996; Noesselt. 2014). Hence, we think our study will arouse public awareness of environmental health issues and help promote legislation on air pollution. What's more, we found that when people felt uncomfortable, they simply expressed their concern for health and tended to go out wearing masks or stay indoors without seeking medical help. It is probable that people just attribute their short period symptoms to an acute allergic reaction and do not consider it a serious potential issue (D'Amato et al., 2010). Further study is required to explore to what degree people get exposed to air pollution before short period uncomfortable reaction will turn into air pollution-related diseases. We believe that our data can arouse public awareness and promote regulations on air pollution, and our methodology can overcome the difficulty of collecting public perception data.

4.2 Feasibility of using social media to monitor air quality

Our study also suggests that social media messages can monitor air quality in the real world. Although previous studies also reach the same conclusion, they performed the experiment by using either a short period of time, a week for example, or merely one city (Jiang et al., 2015; Jiang et al., 2016). Instead, we analyzed the correlation between posting frequency and air quality index of 8 different cities over a period of 12 months, thereby strengthening the correlation. Similarly, a weak correlation between posting frequency and air quality index were found in good air quality areas. As a result, since traditional real-time air pollution measurement and monitoring is costly, we think our study can help monitor the dynamics of air pollution at a much lower cost.

4.3 People’s feelings and responses to air pollution

Our study found that people were more likely to get respiratory-related diseases or feel headache in air pollution areas. Researchers would find it difficult to such obtain such data through current methods, such as interviews, medical records, etc., which are costly and can lead to an incomprehensive conclusion (YI et al., 2007). Postings based on Sina Weibo offers passive and comprehensive self-reporting on a much larger scale in a more cost-effective way. Furthermore, research on the health effects of pollution often focus on more serious outcomes such as disease, while self-reports on social media contain evidence of milder but still important effects (Shah et al., 2015), such as irritability and discomfort (Wang et al., 2015). Reports of symptoms in our study were not sufficient enough to accurately identify the effects of air pollution on people’s health, yet the study has revealed the potential for the data to help quantify health effects of air pollution.

4.4 Demographic characteristics

Our demographic analysis suggests that people who post air pollution-related messages tend to be more active on Sina Weibo and that air pollution is more serious in the more developed southern cities of China. Although we cannot determine the exact factors for the latter phenomenon, we surmise that the distinct industrial structure of southern urban areas may account for it. Previous studies tended to explore the effects of air pollution in a large variety of cities across the whole of China (e.g. Jiang et al., 2015). However, our study suggests that it is more effective to use central and southern cities in China as samples to study the correlation between cities and air pollution, and thus help the government to find possible solutions to the air pollution problem.

5. Conclusions

This study gains insight into the role of social media in the public's perception of air pollution and the public health response to it. Sina Weibo messages provide first-hand information about the public's perception of air pollution, which is difficult to capture by traditional methods such as interviews and questionnaires. It was found that people tend to disclose their feelings and complain about air pollution on social media, and that there is a significant correlation between individuals' posting frequency and China's AQI. This may indicate the feasibility of using social media as an alternative tool to monitor air pollution at a much lower cost. The study also provides a method to comprehensively analyze the health effects of air pollution by identifying health-related topics among users along with their individual characteristics, which can complement current medical data of air pollution-related diseases. It is hoped that our findings will help related professionals better understand the public health effects and public perception of air pollution and provide a basis for advice on the best way to protect against air pollution. More importantly, the findings from this study should help with finding efficient and effective measures to reduce and ultimately prevent air pollution.

Reference

Yu L, Asur S, Huberman BA (2011) What trends in Chinese social media arXiv preprint arXiv:11073522

Sasahara K, Hirata Y, Toyoda M, Kitsuregawa M, Aihara K (2013) Quantifying collective attention from tweet stream PloS one 8:e61823

Xu P, Chen Y, Ye X (2013) Haze, air pollution, and health in China The Lancet 382 doi:10.1016/s0140-6736(13)62693-8

Gerking S, Stanley LR (2017) An economic analysis of air pollution and health: the case The Earthscan Reader in Environmental Economics:194

Mollema L et al. (2015) Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013 J Med Internet Res 17:e128 doi:10.2196/jmir.3863

Cheng Z, Caverlee J, Lee K You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on Information and knowledge management, 2010. ACM, pp 759-768

Wang S, Paul MJ, Dredze M Exploring health topics in Chinese social media: An analysis of Sina Weibo. In: AAAI Workshop on the World Wide Web and Public Health Intelligence, 2014. p 59

Jiang W, Wang Y, Tsou MH, 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:e0141185 doi:10.1371/journal.pone.0141185

Jiang W, Wang Y, Tsou MH, Fu X (2016) Using geo-targeted social media data to detect outdoor air pollution ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLI-B2:553-554 doi:10.5194/isprsarchives-XLI-B2-553-2016

Mei S, Li H, Fan J, Zhu X, Dyer CR Inferring air pollution by sniffing social media. In: Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014. IEEE, pp 534-539

Wang S-y, Bai H A case study on micro-blog communication characteristics related to haze during the Harbin haze disaster. In: Management Science & Engineering (ICMSE), 2014 International Conference on, 2014. IEEE, pp 120-125

Kay S, Zhao B, Sui D (2015) Can social media clear the air? A case study of the air pollution problem in Chinese cities The Professional Geographer 67:351-363

Ryan GW, Bernard HR (2016) Techniques to Identify Themes Field Methods 15:85-109 doi:10.1177/1525822x02239569

D'Amato G, Cecchi L, D'amato M, Liccardi G (2010) Urban air pollution and climate change as environmental risk factors of respiratory allergy: an update Journal of Investigational

Allergology and Clinical Immunology 20:95-102

Ou W, Xie Z, Jia X, Xie B Detection of topic communities in social networks based on Tri-LDA model. In: Proceedings of the 4th International Conference on Computer Engineering and Networks, 2015. Springer, pp 1245-1253

Bakshy E, Hofman JM, Mason WA, Watts DJ Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining, 2011. ACM, pp 65-74

Dasgupta S, Wheeler D (1996) Citizen Complaints as Environmental Indicators

Noesselt N (2014) Microblogs and the Adaptation of the Chinese Party - State's Governance Strategy Governance 27:449-468

YI C-I, MA M, CHEN Y (2007) Study on the Present Condition and Development of the Payment of Specific Diseases in China Chinese Hospital Management 7:011

Shah AS et al. (2015) Short term exposure to air pollution and stroke: systematic review and meta-analysis BMJ 350:h1295 doi:10.1136/bmj.h1295

Wang S, Paul MJ, Dredze M (2015) Social media as a sensor of air quality and public response in China Journal of medical Internet research 17