RESEARCH ARTICLE

Inferring Atmospheric Particulate Matter Concentrations from Chinese Social Media Data

Zhu Tao[1¤], **Aynne Kokas**[2], **Rui Zhang**[3], **Daniel S. Cohan**[3], **Dan Wallach**[1]*

1 Department of Computer Science, Rice University, Houston, Texas, United States of America,
2 Department of Media Studies, University of Virginia, Charlottesville, Virginia, United States of America,
3 Department of Civil and Environmental Engineering, Rice University, Houston, Texas, United States of America

¤ Current address: Googleplex, Mountain View, California, United States of America
* dwallach@rice.edu

Abstract

Although studies have increasingly linked air pollution to specific health outcomes, less well understood is how public perceptions of air quality respond to changing pollutant levels. The growing availability of air pollution measurements and the proliferation of social media provide an opportunity to gauge public discussion of air quality conditions. In this paper, we consider particulate matter (PM) measurements from four Chinese megacities (Beijing, Shanghai, Guangzhou, and Chengdu) together with 112 million posts on Weibo (a popular Chinese microblogging system) from corresponding days in 2011–2013 to identify terms whose frequency was most correlated with PM levels. These correlations are used to construct an Air Discussion Index (ADI) for estimating daily PM based on the content of Weibo posts. In Beijing, the Chinese city with the most PM as measured by U.S. Embassy monitor stations, we found a strong correlation (R = 0.88) between the ADI and measured PM. In other Chinese cities with lower pollution levels, the correlation was weaker. Nonetheless, our results show that social media may be a useful proxy measurement for pollution, particularly when traditional measurement stations are unavailable, censored or misreported.

Citation: Tao Z, Kokas A, Zhang R, Cohan DS, Wallach D (2016) Inferring Atmospheric Particulate Matter Concentrations from Chinese Social Media Data. PLoS ONE 11(9): e0161389. doi:10.1371/journal.pone.0161389

Editor: Juan A. Añel, Universidade de Vigo, SPAIN

Received: March 7, 2016

Accepted: August 4, 2016

Published: September 20, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

User generated social media data are widely seen as an important source for the observations of crowds to be harnessed for a variety of applications. Perhaps most famously, Google's "Flu Trends" project found a significant correlation between localized search engine queries related to illness symptoms and the subsequent growth in emergency room visits [1–2]. Internet data are used to understand public sentiment for a variety of applications, ranging from consumer marketing to stock trading [3–7]. Furthermore social media like Twitter, Facebook and Weibo provide a wide range of content and relatively precise timestamp, facilitating time series research for acute and chronic conditions [8–20].

One such condition that could be characterized by social media data is air quality. Ambient air pollution is estimated to kill 3.7 million people per year worldwide [21]. Most of these deaths occur due to fine particulate matter (i.e., $PM_{2.5}$, denoting particles with aerodynamic diameter less than 2.5 microns), for which concentrations are especially high in cities of China. Ground-level measurements provide limited information about air quality in China, due to sparsity of monitors and withholding of some data [22–24]. Although capabilities are improving for measuring air quality from space, satellites face limitations for characterizing short-term variability in conditions at the surface [25]. Meanwhile, people frequently comment about the weather, a propensity that extends beyond spoken conversations to the online world. Air quality is thus an area that is well suited for investigating the potential for social media-based indicators to infer ambient conditions.

Here, we show how messages publicly posted to Weibo, the most popular microblogging service in China with over 300 million users [26], can be used to construct an "Air Discussion Index" (ADI) characterizing air quality conditions. Our approach derives the terms whose use correlates most directly with pollution metrics, rather than a *priori* selection of terms. Building upon prior data mining of Weibo messages [27], we show how meaningful inferences about pollutant conditions can be extracted from noisy social network data, despite the difficulty of computer processing for Chinese text.

Material and Methods

Collection of Weibo Posts and Air Quality Data

A statistical technique is used to identify words or phrases most associated with varying air quality conditions. An index of those terms, the ADI, can then be used to characterize the relationship between $PM_{2.5}$ and social media posts.

For Weibo, we utilize public timeline posts collected for July 23, 2011 to May 15, 2013 from our previous research [28–29]. The public timeline posts were queried roughly once every four seconds, for which the Weibo server returned roughly 200 recent posts responding to each request. Public timeline is one of the Application Programing Interface API (http://open.weibo.com/), a set of routines and standards for accessing Weibo database, provided by Weibo to access their posts. The posts returned by public timeline API can be considered as real-time random sampling from total posts population coming to Weibo system. In total 500 Gigabytes of collected data, including 112 million Weibo posts from the four cities with available $PM_{2.5}$ measurements described below, were stored and processed on a four-node cluster using Hadoop [30] and HBase [31].

For $PM_{2.5}$, we utilize air quality reports from monitoring stations located at U.S. embassies or consulates in four cities: Beijing, Shanghai, Guangzhou, and Chengdu. Air quality readings from the U.S. consulate in a fifth city, Shenyang, were excluded because their air quality data overlaps with our Weibo data by only one month. $PM_{2.5}$ data are collected by crawling U.S. Embassy and consulate air quality Twitter accounts, which report $PM_{2.5}$ and ozone readings and corresponding U.S. Environmental Protection Agency air quality index (AQI) values hourly and/or daily. Since the format and content of the posts changed several times over the time period of the sample, we compute average daily (12 a.m.–11 p.m.) $PM_{2.5}$ from the hourly data. The number of days in each city with available $PM_{2.5}$ data, which means we collected $PM_{2.5}$ value from Twitter successfully, and successful collection of Weibo posts, along with the average number of collected posts on those days, is shown in Table 1.

Official Chinese measurement data of $PM_{2.5}$ became available in 2013, when China began to regulate $PM_{2.5}$ as a criteria pollutant with an ambient air quality standard. Given the short overlap period of those data and our Weibo posts, we use the Chinese $PM_{2.5}$ data only for an inter-comparison to assess the representativeness of the U.S.-reported data.

Table 1. Information of Weibo posts and air quality condition in four mega cities of China for this study.

City	Twitter site (twitter.com)	Twitter reports starting date	Number of valid days	Average daily Weibo posts	Average daily $PM_{2.5}$ concentration (μg/m^3)
Beijing	../BeijingAir	Feb 19, 2009	528	89000	100.1
Shanghai	../CGShanghaiAir	May 15, 2012	281	83000	53.4
Guangzhou	../Guangzhou_Air	Jun 15, 2011	462	81000	54.5
Chengdu	../CGChengduAir	Jun 28, 2012	207	21000	93.4

doi:10.1371/journal.pone.0161389.t001

Key Term Extraction in Weibo Post

Even though our Weibo crawler returned only original posts and excludes retweets, roughly 10% of the captured posts were identical. The main sources of these identical posts are the "posting machines" which enable users to generate posts automatically. Two examples are PiPi Timing Machine (weibo.pp.cc/time/) and Weibo Tong (wbto.cn/). These identical posts create noise in terms of forming topical trends, which do not reflect users' real opinion. To alleviate this noise, the MD5 message-digest algorithm [32] is used to remove identical posts.

Identifying words in Chinese is not an easy task for computer systems. Chinese is different from western languages in that there are no spaces between words in a sentence. Most commonly used Chinese words are composed of two or more characters. Analysis is complicated in social media by the frequent use of neologisms in online discourse [33]. The approach adopted here is to utilize n-grams [29], representing *n* consecutive Chinese characters in a sentence. Some of these sequences of characters form meaningful words, but most of them do not. Trigrams ($n = 3$) were used in a preliminary experiment on a subset of Beijing data, and achieved slightly better correlation than bigrams ($n = 2$). However, considering the vastly larger computational task to analyze trigrams (2 billion in our current Weibo database) instead of bigrams (40 million), we decided to report only the algorithm and performance of bigrams in this paper.

We examined each of the 40 million bigrams in our data database. Rather than using any subjective judgment to select the relevant bigrams, the selection is performed automatically by the algorithms described here. For each term, the number of posts containing is counted daily and aggregated separately by city according to the user's registered city, denoted as *post_count (term, date, city)*. When a term shows up multiple times in a post, it is counted only once. The post count is then divided by the number of all posts in that city on that day in our dataset, *base_post_count (date, city)*, resulting in a fraction of the posts containing that term in a particular city and on a particular day:

$$f(term,\ date,\ city) = \frac{post\_count\ (term, date, city)}{base\_post\_count\ (date, city)} \tag{1}$$

For each city and each term, a linear regression model is built to infer daily $PM_{2.5}$.

$$[P]_{city} = \alpha_0 + \alpha_1 [T]_{city,term} + \varepsilon \tag{2}$$

where $[P]_{city}$ is the vector of the daily $PM_{2.5}$ concentration of a city, $[T]_{city,term}$ is the term's daily fraction vector in which each element is calculated in Eq 1, α_0 and α_1 are the parameters of the linear function, and ε is the error term.

Using the above linear regression with four-fold cross validation [34], we fit the model to four ¾ subsets in each city. Each per-term model is validated by measuring the correlation coefficient between the model's estimates for the ¼ reserved points and U.S. Embassy reported $PM_{2.5}$ for that city at those points. Thus we get four correlation coefficients for each term city pair, and the final score of each candidate term of a city is the mean of those four values. At last, the terms are sorted by this final score for each city to generate a Sorted Term List (STL).

Computation of Air Quality Discussion Index (ADI)

A naive way aggregate the probability in each term set (TS) in the ADI for different day at different city is simply to sum them together as below:

$$ADI_{date,city,TS} = \sum\nolimits_{term \in TS} sign(term, city) \times f(term, date, city) \tag{3}$$

,where *sign*(*term*,*city*) is the sign of α_1 in Eq 2, which is either +1 or -1, depending on whether the term is positively or negatively correlated with $PM_{2.5}$.

However, since the term's base frequencies may differ by orders of magnitude, the effect caused by terms with smaller frequencies will be overwhelmed by those with larger frequencies. Standardization is a method to eliminate the influence of high frequency term. Therefore, instead of using the probability directly, we calculate the standard score or the z-values [35] in each term's daily probability vector (Eq 1) to the mean equal to 0 and the variance equal to 1, which is *norm*(*term*,*date*,*city*). Then we use this normalized vector to compute the ADI by summing the normalized term probability over a period of time and city:

$$ADI_{date,city,TS} = \sum\nolimits_{term \in TS} sign(term, city) \times norm(term, date, city) \tag{4}$$

Selected Weibo terms can be either positively (e.g. "haze") or negatively (e.g. "blue sky") correlate with air pollution condition.

Determination of Term Set and Estimation of $PM_{2.5}$

The algorithms for selecting the terms for the ADI term set (TS) and for evaluating the term set are introduced here as Algorithm 1 and Algorithm 2.

Algorithm 1: Algorithm of ADI term set selection.

```
ADI_TERM_SET_SELECTION(city):
  TS ← ∅; BestPerformance ← 0;
  For each $term_i$ in top N of $STL_{city}$
      Add $term_i$ to TS;
      CurrentPerformance = Evaluate(TS, city)
      If CurrentPerformance > BestPerformance
            BestPerformance ← CurrentPerformance;
      else
            Remove $term_i$ from TS;
Return TS;
```

From the previous step, we get a STL in which the terms are sorted by the mean correlations for each city. To decide how many terms from the STL should be included in the term set (TS), an incremental approach is applied (Algorithm 1). The first term in the STL is added to a candidate term set. The second term in the STL is added to the ADI set, only if this second term increases the ADI infer performance. In this manner, all terms in the STL are scanned in order, with an additional term added to the TS only if this term increases the TS infer performance measured by Algorithm 2.

Algorithm 2: Algorithm of term set evaluation

```
Evaluate(TS, city):
   Compute [ADI] from Eq 4
   Divide data points (both [P] ($PM_{2.5}$ observations) and [ADI]) into four
continues sections;
   [$P_f$] and [$ADI_f$] denote the portion of section f in [P] and [ADI] (f = {1, ..., 4}).
   $\underline{[P_{\bar{f}}]}$ and $\underline{[ADI_{\bar{f}}]}$ denote the complementary vector of [$P_f$] and [$ADI_f$]
```

```
    For each section f
        Learn model parameters $\beta_0$ and $\beta_1$ in Eq 5 with $[P_f]$ and $[ADI_f]$
        $[FitP_f]$ ← Fit the learned model with data $[ADI_f]$
        $R_f$ ← Pearson correlation coefficient between $[P_f]$ and $[FitP_f]$;
    R = average of $R_f$;
Return R;
```

By Algorithm 2, a TS is evaluated by using simple linear regression with four-fold cross validation. In four-fold cross-validation, the original dataset is partitioned into four equal sized sub-datasets. Three sub-datasets are used for training purposes, while the remaining single sub-dataset is kept as the validation data to test the model performance. Then, this four cross-validation process is performed four times, with each of the four sub-datasets used in turn as the validation data. Here, the average value of the four correlations between the model's estimates and reported $PM_{2.5}$ concentrations is assigned as the score for that specified TS. The TS with the highest score becomes the Final Term Set (FTS), which will be used to compute the ADI.

After we decide the term set to compute ADI, $PM_{2.5}$ is inferred from it by applying a linear model analogous to Eq 2, as shown in Eq 5.

$$[P]_{city} = \beta_0 + \beta_1 [ADI]_{city,TS} + \varepsilon \tag{5}$$

Here $[P]_{city}$ is the vector of daily $PM_{2.5}$ as in Eq 2; $[ADI]_{city,TS}$ is a daily aggregated normalized term vector whose elements are calculated as shown in Eq 4; β_0 and β_1 are the parameters of the linear function; and ε is the error term.

Ethics Statement

The protocol of data processing and anonymization was followed by the standards of ethical conduct in Rice University. All Weibo posts used for analysis in this study were from our previous dataset [28], which crawled social media information available to the broad public. No attempt was made to inform Weibo users of the current study. In compliance with the privacy/ethic requirements of Sina Weibo Term of Service, the individual profiles were treated as encrypted sensitivity information and only the aggregate statistic results are reported. The raw data were de-identified before the current analysis began.

Results and Discussion

Weibo Terms Associated with $PM_{2.5}$

The simple linear regression described in the methodology section is applied to the four cities in Table 1. The last five months of data (January 1, 2013—May 15, 2013) are left out as a validation data set for the overall performance evaluation. A total of 513,537 bigrams were identified for which the frequency of appearance in the total dataset was above a once-per-day threshold. A list with the highest scoring terms, sorted by the mean correlations, was produced for each city (see S1 File for the final term set for each city).

Many of the terms positively correlated with $PM_{2.5}$ appear related to high air pollution levels, including dust, cough and mask. Many of the negatively correlated terms relate to good air quality; examples include the terms sunshine, rain, and blue. By looking at the sign of α_1 in Eq 2, terms are classified into two categories objectively.

The most strongly positively correlated term for Beijing is "雾蒙" (misty foggy) while for Guangzhou it is "灰霾" (dust-haze). The bigram "不健" (part of "不健康", unhealthy) topped the term set for Shanghai. Chengdu's most correlated term "的天" (-like day), has a less obvious connection to air pollution, but it is a common connection phrase in Chinese to most likely describe the condition of air turbidity such as blue sky (negative correlate) or foggy sky/muddy sky (positive correlate). For non-Chinese speaking readers, we provide some descriptions of those terms. The bigrams are not necessarily words, for example "rain", which have some specific meanings. Instead, bigrams are pairs of Chinese characters which appeared next to another in Weibo corpus, for example "s raini". As mentioned earlier, there are no spaces in between the words like English, so in Chinese "s raini" would further become "sraini". Since these bigrams cannot be associated directly with dictionary words, human interpretation is needed to examine the meaning of those bigrams.

For Beijing, we manually examined 500 terms with the highest scores, including both positively and negatively correlated terms (Fig 1; also see S1 Table for the complete list of the top 500 bigrams from training data in Beijing). We subjectively judged about 95% of them to be air quality or weather-related terms. Some obvious positively correlated examples include "haze", "fog", "dusky", and "cannot breathe"; negatively correlated examples include "sunny", "raining", and "blue sky".

High scoring terms in the other three cities were less obviously related to air quality. For Shanghai, the top 20 terms were related to air quality or weather, all showing positive correlations with $PM_{2.5}$. However, many high scoring terms beyond the top 20 had no apparent relationship to air quality or weather. Selected terms in Guangzhou and Chengdu had even less apparent connection to air quality. The less obvious correlation of high scoring terms in Shanghai and Guangzhou probably related to the less severe air pollution level in the two cites comparing with Beijing. The average daily $PM_{2.5}$ concentrations in Shanghai and Guangzhou during the evaluation period is nearly half of the corresponding value in Beijing (53.4 μg/m^3 in Shanghai and 54.5 μg/m^3 in Guangzhou versus 100.1μg/m^3 in Beijing, see Table 1). More importantly, the frequency of extreme polluted day with the measured daily mean $PM_{2.5}$ concentration greater than 300 μg/m^3 in Beijing is much higher during our evaluation period than the other two cities (in total 19 days in Beijing versus 3 days in Shanghai and only 1 day in Guangzhou), which intuitively has the higher chance to trigger people to directly complain the bad air quality at social media platform. The possible reason for the less direct air quality related high scoring term in Chengdu is due the much smaller sample size comparing with other cites (with average daily Weibo posts around 21000 versus 89000 in Beijing, 83000 in Shanghai and 81000 in Guangzhou, see Table 1).

Some of the terms with high scores merit attention. For example the term "U.S. Embassy" in Beijing and "U.S. Consulate" in Shanghai and Chengdu rank within the top 100 terms in their respective cities, with positive correlations to $PM_{2.5}$. This may indicate an influence from the air quality report from the U.S. embassy or consulate. The reasons for positive correlations in Beijing between $PM_{2.5}$ and terms such as "highway closure" and "download this app" are less readily apparent. For example, it is unclear whether bad visibility caused by pollution was associated with any "highway closures," and whether pollution made users more likely to download apps showing the AQI.

Another observation is that season-related terms show up frequently in Guangzhou's list of terms with the highest scores, but are less pronounced in the other cities. Terms related to winter, including winter clothes, tend to have positive correlations with the $PM_{2.5}$. The seasonal nature of word choice in Guangzhou makes it less clear whether word choice such as "鼻塞"(stuffiness) or "喉咙痛" (sore throats) is prompted by higher $PM_{2.5}$ itself, or by seasonal illnesses that tend to peak in the winter.

Estimation of $PM_{2.5}$ from Weibo Data

We apply the incremental approach (Algorithms 1 and 2) to the learning dataset for each city to test which terms should be included in the final ADI used to infer the $PM_{2.5}$ values. This approach selected a different number of terms for the FTS for each city, as shown in Table 2.

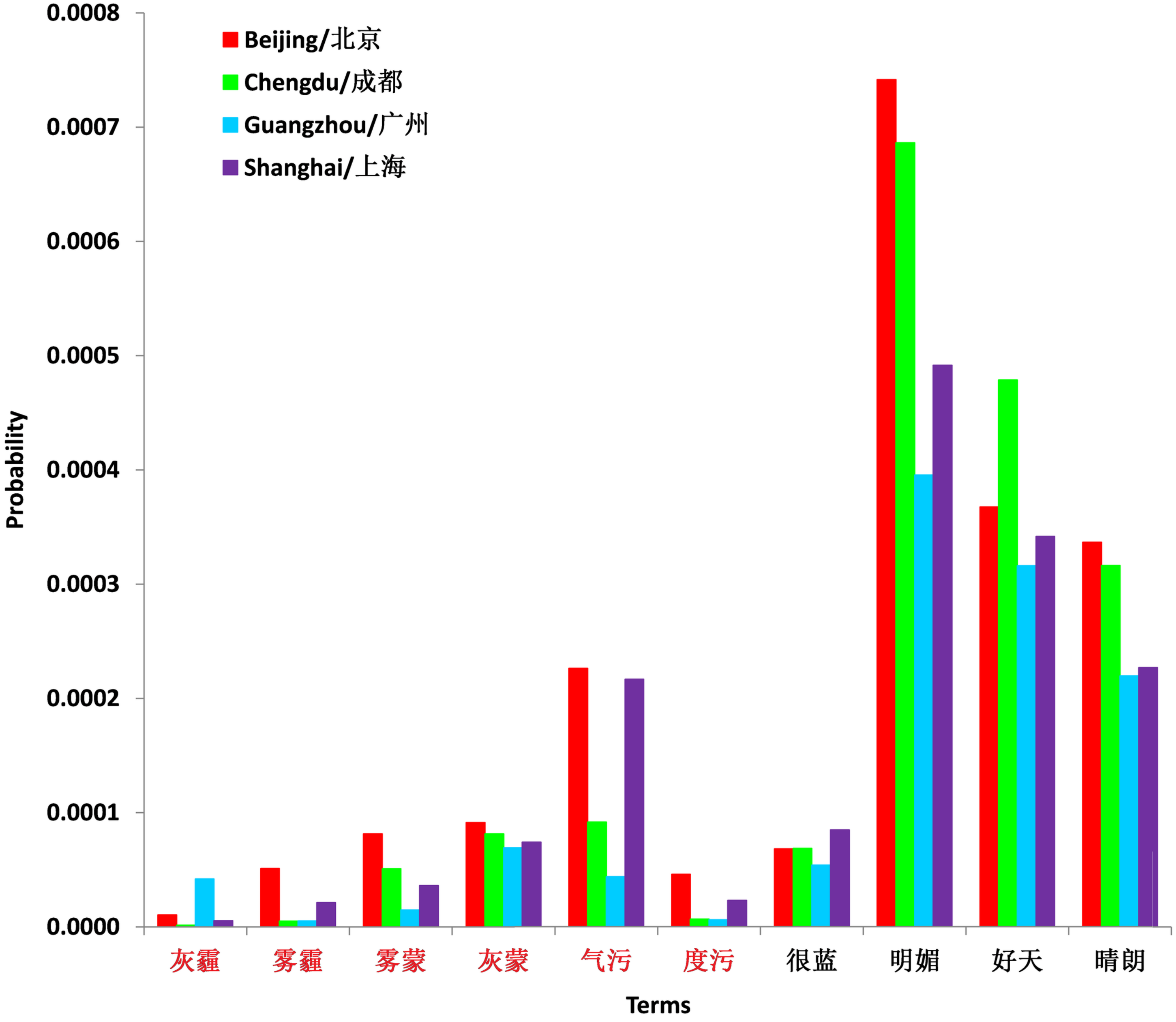

Fig 1. Fraction of Weibo posts containing terms strongly correlated with $PM_{2.5}$. From left to right, positively correlated terms: "dust-haze," "haze," "misty," "dusky," "air pollution," "degree of pollution"; and negatively correlated terms: "very blue," "bright," "good day," and "sunny."

doi:10.1371/journal.pone.0161389.g001

Then we apply the linear model described in Eq 5 with the learned parameters to the reserved test data to see how close the fit is to the observed $PM_{2.5}$ values. Correlations between the ADI-derived $PM_{2.5}$ and observed $PM_{2.5}$ in each city are shown in Table 2.

Among the four cities, the infer ADI from Weibo posts at capital city Beijing achieved the strongest performance in estimating observed $PM_{2.5}$. In Beijing, there were 438 days of learning data points and 90 days of reserved test data set, and the algorithm selected the 20 terms most correlated with $PM_{2.5}$ at the U.S. Embassy for the FTS of the ADI (see S1 File for the complete list). The correlation between the ADI- derived and the observed $PM_{2.5}$ value for the learning data set is 0.81 and for the validation data set the correlation is 0.88 (Table 2 and Fig 2). The

Table 2. Performance of Air Discussion Index in estimating observed $PM_{2.5}$.

City	Number of terms in FTS[a]	Learning Period (valid days)	R[b]: learning period	Validation Period (valid days)	R[b]: validation period
Beijing	20	Jul 23, 2011 –Dec 31, 2012 (438)	0.805	Jan 1, 2013 –May 15, 2013 (438)	0.882
Shanghai	12	May 12, 2012 –Dec 31, 2012 (189)	0.737	Jan 1, 2013 –May 15, 2013 (51)	0.633
Guangzhou	42	Jul 26, 2011 –Jan 31, 2013 (349)	0.649	Feb 1, 2013 –May 15, 2013 (113)	0.425
Chengdu	18	Jul 2, 2012 –Jan 28, 2013 (136)	0.853	Feb 1, 2013 –May 15, 2013 (51)	0.361

[a] FTS: Final Term Set

[b] R: Correlation Coefficient

doi:10.1371/journal.pone.0161389.t002

correlation was especially strong in January 2013 (Fig 3), a month with particularly high levels of pollution, suggesting the ADI may become especially useful as air quality worsens. Since January is included in the validation period, it results in a higher correlation for the validation period than the learning period. Applying the same approach to construct an ADI for the other

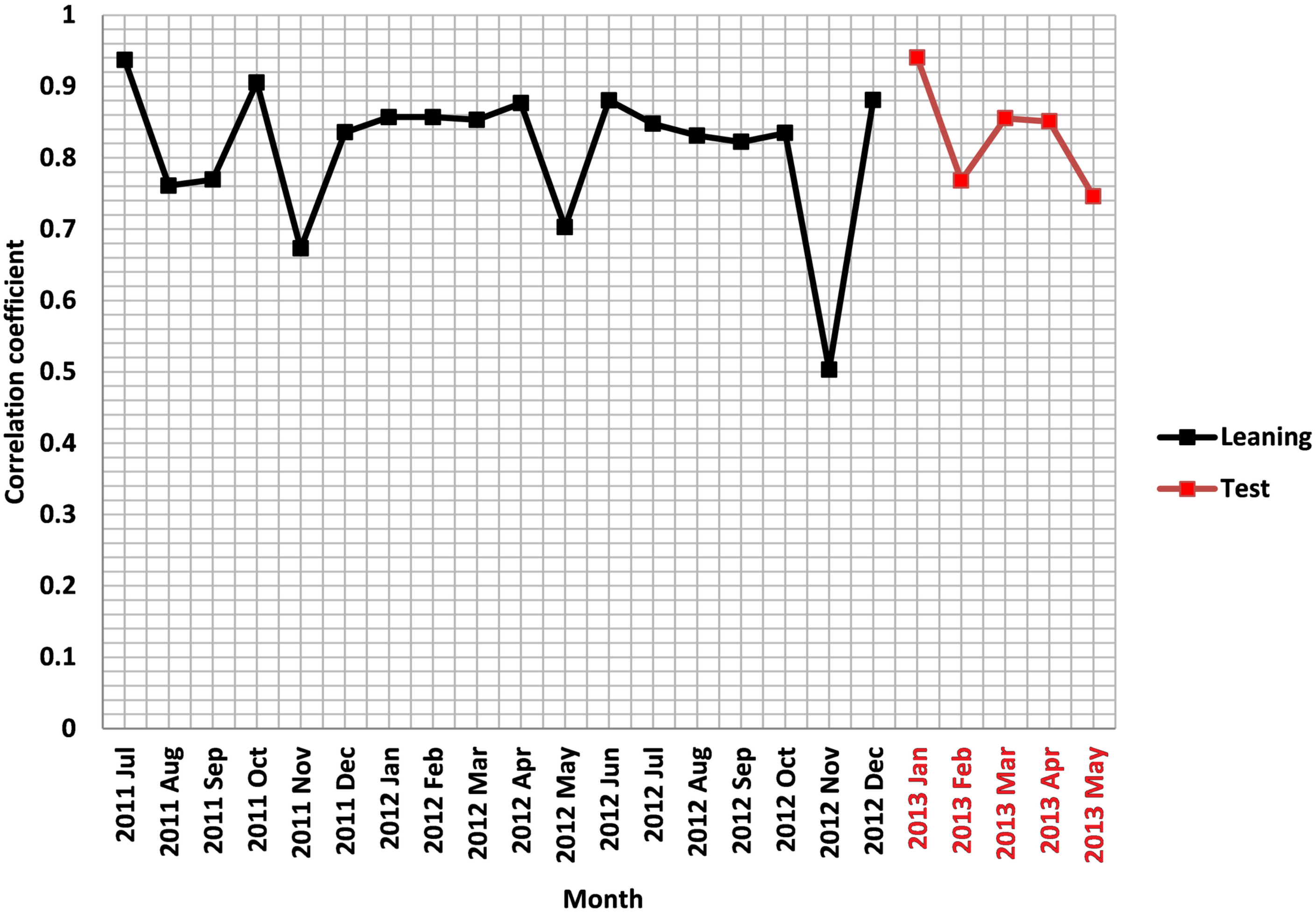

Fig 2. Correlation between observed and estimated $PM_{2.5}$ for Beijing during learning (black) and testing (red) period.

doi:10.1371/journal.pone.0161389.g002

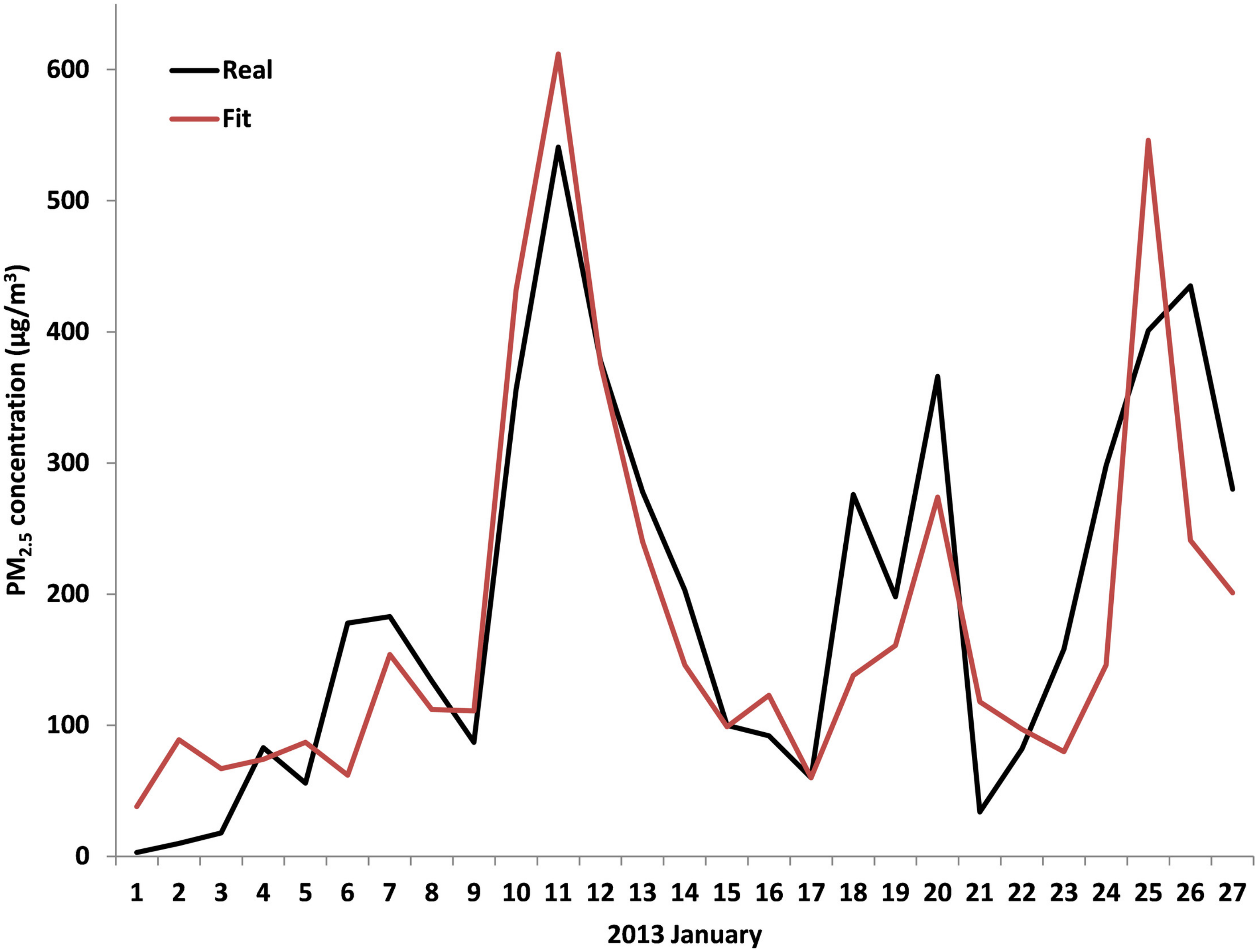

Fig 3. Comparison of ADI model estimates against U.S. Embassy reported $PM_{2.5}$ concentration for Beijing.

doi:10.1371/journal.pone.0161389.g003

three cities yields less accurate fits of their $PM_{2.5}$ than what was achieved in Beijing. The correlation between the fit and the observation for validation periods were 0.63, 0.43, and 0.36 for Shanghai, Guangzhou, and Chengdu, respectively (Table 2).

The performance achieved by the ADI in estimating Beijing's $PM_{2.5}$ concentration is especially impressive given fact that the ADI was constructed based on observations at a single monitor (i.e. hourly twitter data from U.S. Embassy), which may not represent the air quality experienced by all Weibo users inside the geographic boundary of Beijing. A correlation analysis between $PM_{2.5}$ at the U.S. Embassy (North East of downtown; see S1 Fig) and the twelve Beijing Environmental Protection Bureau (BJ-EPB) maintained monitoring sites was given in Fig 4 to better understand the extent to which the embassy data represent air quality conditions throughout Beijing area during different seasons (see S2 Fig for the measured $PM_{2.5}$ concentrations at different sites). For 2013, the correlations between daily mean $PM_{2.5}$ at the U.S. Embassy and the other urban sites, especially the eight sites located inside the 6th Ring Road (BJ1-BJ8, which is considered as the core area in Beijing and has more than 75% of total

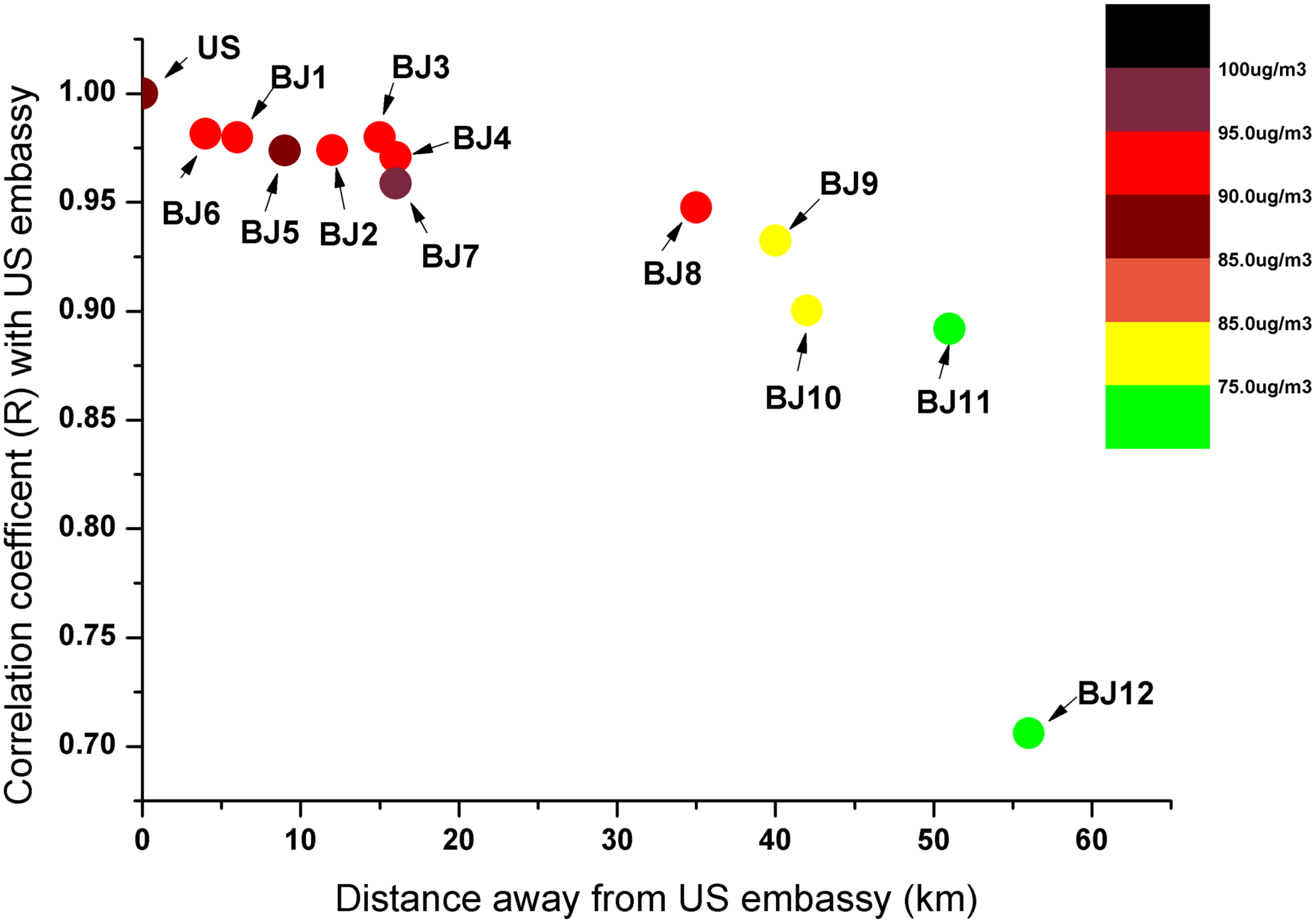

Fig 4. Correlation of daily $PM_{2.5}$ concentrations between U.S. Embassy and 12 BJ-BEP sites. The location of the sites is provided in S1 Fig. The color bar is the annual mean concentration.

doi:10.1371/journal.pone.0161389.g004

residents) are very high (0.95–0.98) and with quite similar pollution levels (90–100 $\mu g/m^3$). Even for the background site BJ12, which is located at the North West upwind direction and around 60 km away from U.S. embassy, the correlation is not low (R = 0.70) and the annual mean PM concentration is also relatively high ($69\mu g/m^3$) Since the air quality in Beijing is dominated by regional sources during episode days [36], the spatial different is not that obvious at those days so that people intend to infer the same bad air at Weibo at different geographic locations in Beijing.

The Beijing ADI also differs from other cities in having both negatively and positively correlated terms ranked among the most influential terms comprising the index. Three reasons might explain the distinctions between Beijing and the other cities. First, Beijing has the highest and most variable $PM_{2.5}$ levels among the four cities (Fig 5). The larger the variance of the response variable, the easier it is to get rid of the noise terms which coincidentally has similar variance as the response variable [29]. Also, the large variance from a high mean may make it more newsworthy for Weibo users to comment on days with relatively clean air quality in Beijing. Secondly, users in Beijing are more interested in the air pollution issue than other cities (Fig 1). As a result, the key terms have obvious fluctuations related to air quality. This property

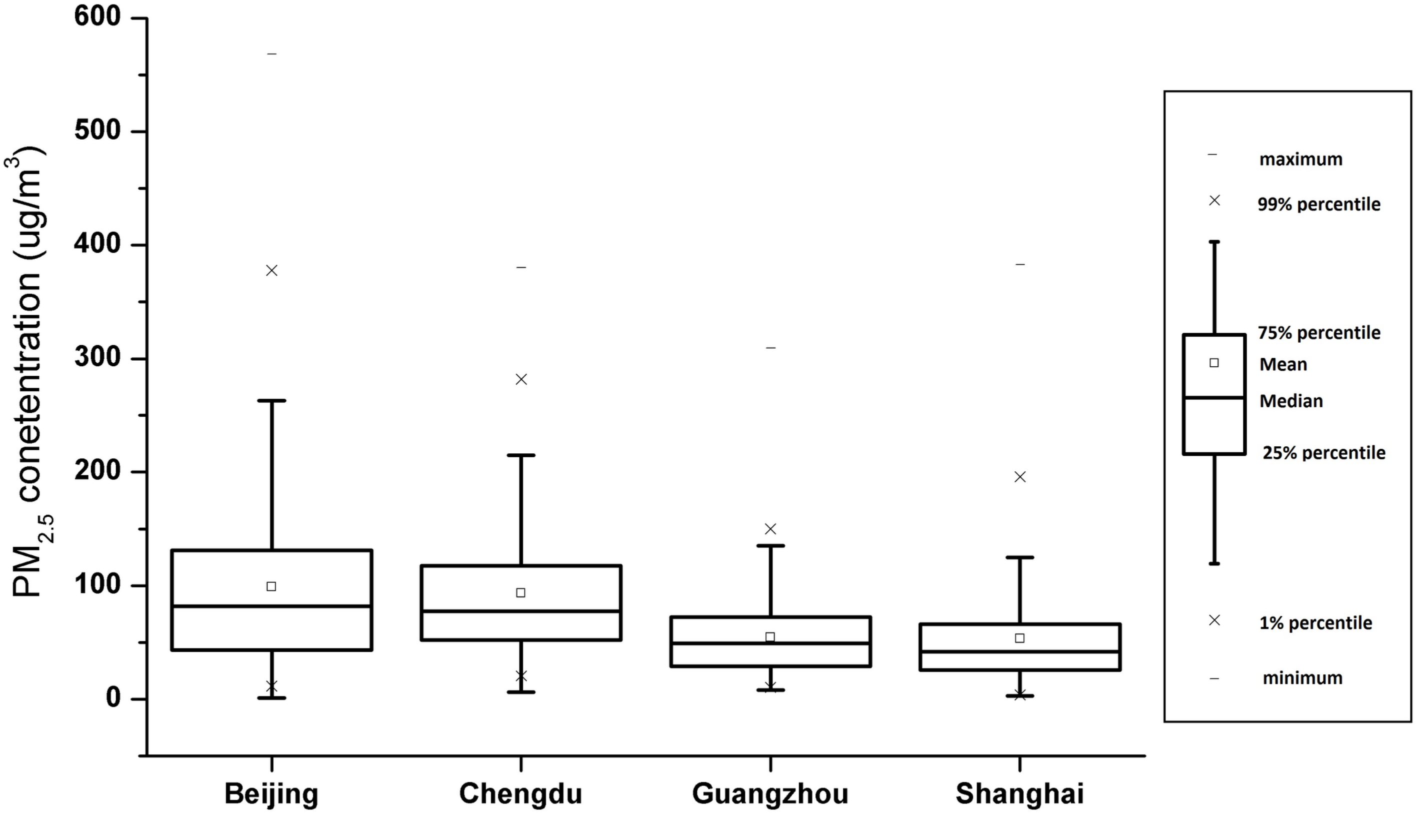

Fig 5. Box plot of daily PM2.5 concentration (μg/*m*^{3}) reported in different U.S. Embassy (Consulate) in China.

doi:10.1371/journal.pone.0161389.g005

fits the motivation of our linear regression method. Finally, we have a longer data record, or learning period, for Beijing than for the other three cities (Table 2).

Besides $PM_{2.5}$, we also attempted to apply this method to PM_{10} (particles smaller than 10 microns in diameter) data in Beijing. However, the resulting ADI achieved a correlation of only 0.56 for PM_{10}. $PM_{2.5}$ particles are far more potent than larger particles (which together comprise PM_{10}) in attenuating visible sunlight (thereby forming a visible haze and reducing visibility) and impacting human health, and $PM_{2.5}$ is the main basis for air quality alerts. These factors may contribute to $PM_{2.5}$ being more strongly associated than PM_{10} with the content of Weibo posts.

Comparison with Similar Studies

Our work builds on a corpus of research analyzing both Weibo data and air quality data to measure social and environmental phenomena. We used the pointillism approach for natural language processing of social media to track trends and discover memes. This method, unlike the conventional approach starting from the pre-selected terms, discovers the key terms related to air quality based on mean correlations. In one similar study to utilize Weibo data and device web data to predict pollution related health hazards in China [37], the researchers chose a subset of 38 predefined terms, which have high correlation with API, and used them to monitor the health situation of Weibo users. In another recent Weibo data mining study, the researchers collected social media messages about "outdoor air pollution" in Beijing in 2012 by using the "advanced search" tool in Sina API based on keywords "Beijing" and "air pollution" to do geo-targeted spatiotemporal analysis and infer the AQI in Beijing [38]. Predefined terms are biased by researchers' opinions and are less successful at discovering the new syndromes or capturing new events. On the other hand, the terms we use here are collected objectively from 40 million terms (bigrams) in our database. Those terms can be updated periodically.

The choice of Chinese words segmentation matters for Weibo posts analysis. Whether to use monograms, bigrams, trigrams, quadgrams and so forth for key term extraction is a trade-off between efficiency and accuracy. Multi-grams extraction tends to contain more information in each search but requires more computation resources, while the monograms search is efficient but cannot give us order information of the phrase. Unlike the similar studies [28–29] to use trigrams to extract topics from Weibo posts, we used bigrams in this study due to the more manageable computational intensity. Furthermore, nearly the same strong correlations were achieved with $PM_{2.5}$ data by using bigram extraction instead of longer grams.

Several factors lead this study to achieve somewhat lower correlations between a Weibo-based ADI and $PM_{2.5}$ than the Google flu trends project [1–2] achieved between search queries and flu incidence. First, the Google queries are typically longer and more specific than a bigram. Secondly, the flu trends research used the query's IP address while we use the user's registered city to determine location. It was estimated that about 15% of the time, the city claimed by Weibo users differs from the actual location of sending out posts [37–38]. Thus, our data tends to be more noisy in terms of the mislabeling the geographical information of posts. Lastly, our study used only one monitoring site's data per city, which did not fully represent conditions throughout the city. Even with these data limitations, our approach achieved a mean correlation of 0.81 (learn data) and 0.88 (validation data) for Beijing, compared to 0.90 and 0.97 respectively for the Google flu trends study.

Conclusion

In this paper, we show how microblogs posted on China's popular social media "Weibo" can be used as pollution metric to characterizing the local air quality conditions. We considered $PM_{2.5}$ measurements from four Chinese megacities (Beijing, Shanghai, Guangzhou, and Chengdu) together with 112 million posts on Weibo from days in 2011–2013, which can be break into 40 million bigrams to identify the key terms whose frequency was most correlated with $PM_{2.5}$ levels. These correlations are used to construct an "Air Discussion Index" (ADI) for inferring the daily PM levels based on the content of Weibo posts. In Beijing, the capital city of China with the most frequent and long term ambient $PM_{2.5}$ records from U.S. Embassy monitoring station and most abundant Weibo posts related with air quality, we found a strong correlation ($R = 0.88$) between the ADI and measured PM. In other Chinese cities with lower pollution levels and fewer related Weibo posts, the correlation was weaker. Nonetheless, our results show that social media may be a useful proxy measurement for pollution, particularly when traditional measurement stations are unavailable, censored or misreported.

Our approach derives the terms whose use correlates most directly with pollution metrics, rather than a priori selection of terms. Building upon prior data mining of Weibo messages, we show how meaningful inferences about pollutant conditions can be extracted from noisy social network data, despite the difficulty of computer processing for Chinese text.

To assure accountability in China's growing pollution problem, additional public information on environmental issues is urgently needed [39]. The ADI marks a foundation for a future of publicly generated air quality metrics drawn from social media data. The methodology presented in this paper offers an important foundation for similar work in other densely populated areas like South and South East Asia with large social media user bases and limitations to air quality monitoring infrastructure.

Currently we extract ADI per city. As future work, we want to extract common ADI that have common correlation among all Chinese speaking cities in pollution related topics. This can help us estimate the air quality even though we don't have enough posts corpus.

Supporting Information

S1 File. The final term set for each city.
(PDF)

S1 Fig. Locations of the US Embassy and the Beijing Environmental Bureau monitoring sites in Beijing, China.
(PDF)

S2 Fig. Box plot of measured $PM_{2.5}$ daily average concentrations ($\mu g/m^3$) from U.S. Embassy site and 12 BJ-EPB sites in Beijing in 2013.
(PDF)

S1 Table. Top 500 bigrams from training data in Beijing. Shown as the correlation coefficient R value and its corresponding bigram.
(PDF)

Acknowledgments

The authors thank Prof. Yang Liu from Emory University for providing the Chinese air quality data for this study. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Author Contributions

Conceptualization: DW AK DSC.

Formal analysis: ZT RZ.

Methodology: DW ZT DSC.

Project administration: DW.

Software: ZT.

Visualization: ZT RZ.

Writing – original draft: ZT AK RZ.

Writing – review & editing: DSC DW AK.

References

1. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS One. 2011, 6, (8), e23610. doi: 10.1371/journal.pone.0023610 PMID: 21886802
2. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2008, 457, (7232), 1012–1014. doi: 10.1038/nature07634
3. Forman C, Ghose A, Wiesenfeld B. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. Info Sys Res. 2008, 19, (3), 291–313. doi: 10.1287/isre.1080.0193
4. Antweiler W, Frank MZ. Is all that talk just noise? The information content of internet stock message boards. J Finance. 2004, 59, (3), 1259–1294. doi: 10.1111/j.1540-6261.2004.00662.x

5. Mishne G, Glance NS. Predicting movie sales from blogger sentiment. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006, pp. 155–158.

6. Spann M, Skiera B. Internet-based virtual stock markets for business forecasting. Manage Sci. 2003, 49, (10), 1310–1326. doi: 10.1287/mnsc.49.10.1310.17314

7. Dhar V, Chang EA. Does chatter matter? The impact of user-generated content on music sales. J Interact Mark. 2009, 23, (4), 300–307. doi: 10.1016/j.intmar.2009.07.004

8. Ao J, Zhang P, Cao Y. Estimating the locations of emergency events from Twitter streams. Procedia Comput Sci. 2014, 31, 731–739. doi: 10.1016/j.procs.2014.05.321

9. Power R, Robinson B, Ratcliffe D. Finding fires with Twitter. In: Australasian Language Technology Association Workshop, Brisbane, Queensland, 2013, pp. 80.

10. Lampos V, Cristianini N. Nowcasting events from the social web with statistical learning. ACM Transactions on Intelligent Systems and Technology (TIST). 2012, 3, (4), 72. doi: 10.1145/2337542.2337557

11. Sakaki T, Okazaki M, Matsuo Y. In earthquake shakes Twitter users: Real-time event detection by social sensors. 2010, ACM: 2010; pp 851–860.

12. Crooks A, Croitoru A, Stefanidis A, Radzikowski J. #Earthquake: Twitter as a distributed sensor system. Trans GIS. 2013, 17 (1), 124–147.

13. Earle PS, Bowden DC, Guy M. Twitter earthquake detection: earthquake monitoring in a social world. Ann Geophys. 2012, 54 (6).

14. Miyabe M, Miura A, Aramaki E. In use trend analysis of Twitter after the great East Japan earthquake. 2012, ACM: 2012; pp 175–178. doi: 10.1145/2141512.2141571

15. Muralidharan S, Rasmussen L, Patterson D, Shin JH. Hope for Haiti: An analysis of Facebook and Twitter usage during the earthquake relief efforts. Public Relat Rev. 2011, 37 (2), 175–177.

16. Nguyen TM, Koshikawa K, Kawamura T, Tahara Y, Ohsuga A. In Building earthquake semantic network by mining human activity from Twitter, 2011 IEEE International Conference on Granular Computing (GrC), 2011; pp 496–501.

17. Mendoza M, Poblete B, Castillo C. In Twitter Under Crisis: Can We Trust What We RT? In: ACM: 2010. 2010, pp 71–79.

18. He H, Yu Z, Guo B, Lu X, Tian J. Tree-based mining for discovering patterns of reposting behavior in microblog. In: Advanced Data Mining and Applications, Springer, 2013; pp 372–384.

19. Lu X, Yu Z, Guo B, Zhou X. Modeling and predicting the re-post behavior in Sina Weibo. In Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, 2013, pp. 962–969.

20. Zhang HP, Zhang RQ, Zhao YP, Ma BJ. Big data modeling and analysis of microblog ecosystem. Int J Autom Comput. 2014, 11, (2), 119–127. doi: 10.1007/s11633-014-0774-9

21. World Health Organization. Burden of disease for household air pollution for 2012. World Health Organization, 2014. Available: http://www.who.int/phe/health_topics/outdoorair/databases/HAP_BoD_results_March2014.pdf. Accessed 23 July 2016.

22. Hsu A, de Sherbinin A, Shi H. Seeking truth from facts: The challenge of environmental indicator development in China. Environ Dev. 2012, 3, 39–51.

23. Liu L, Zhang B, Bi J. Reforming China's multi-level environmental governance: Lessons from the 11th Five-Year Plan. Environ Sci Policy. 2012, 21, 106–111.

24. Wang J, Hu M, Xu C, Christakos G, Zhao Y. Estimation of citywide air pollution in Beijing. PLoS One. 2013, 8(1):e53400, doi: 10.1371/journal.pone.0053400 PMID: 23320082

25. Martin RV. Satellite remote sensing of surface air quality. Atmos Environ. 2008, 42, (34), 7823–7843. doi: 10.1016/j.atmosenv.2008.07.018

26. Cao B. Sina's Weibo outlook buoys Internet stock gains: China overnight. Bloomberg. 28 February 2012. Available: http://www.bloomberg.com/news/2012-02-28/sina-s-weibo-outlook-buoys-internet-stock-gains-in-n-y-china-overnight.html. Accessed 23 July, 2016.

27. Zhuang X, Wang Y, He H, Liu J, Wang X, Zhu T et al. Haze insights and mitigation in China: An overview. J Environ Sci. 2014, 26 (1), 2–12.

28. Zhu T, Phipps D, Pridgen A, Crandall JR, Wallach DS. The velocity of censorship: high-fidelity detection of microblog post deletions. In: Proceedings of the 22nd USENIX conference on Security, Washington, D.C., 2013, pp. 227–240.

29. Song P, Shu A, Zhou A, Wallach DS, Candall JR. A pointillism approach for natural language processing of social media. In: Proceedings of the 2012 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'12), Hefei, China, 2012.

30. Shvachko K, Kuang H, Radia S, Chansler R. In The Hadoop Distributed File System, 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010. pp 1–10.

31. HBase Development Team. HBase: A Distributed Database for Large201 Datasets. The Apache Software Foundation, Los Angeles, CA. 2012. Available: http://hbase.apache.org. Accessed: 23 July, 2016.

32. Rivest R. The MD5 message-digest algorithm. 1992. Available: https://www.ietf.org/rfc/rfc1321.txt. Accessed: 23 July, 2016.

33. Lin MY, Chiang TH, Su KY. A preliminary study on unknown word problem in Chinese word segmentation. In: Proceedings of Rocling VI, 1993; pp 119–137.

34. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th international joint conference on Artificial intelligence, San Francisco, CA, 1995, pp. 1137–1143.

35. Kreyszig E, Norminton EJ. Advanced Engineering Mathematics, Vol. 1022. New York, Wiley, 1993.

36. Zhang R, Jing J, Tao J, Hsu SG, Wang G, Cao J, Lee CSL, Zhu L, Chen Z, Zhao Y, Shen Z. Chemical characterization and source apportionment of $PM_{2.5}$ in Beijing: seasonal perspective. Atmos Chem Phys. 2013, 14(13), 7053–7074.

37. Chen J, Chen H, Zheng G. Pan JZ, Wu H, Zhang N. Big smog meets web science: smog disaster analysis based on social media and device data on the web. In: Proceedings of the companion publication of the 23rd international conference on World wide web companion, 2014, pp. 505–510.

38. Jiang W, Wang Y, Tsou MH, Fu X. 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. 2015, 10, (10), e0141185. doi: 10.1371/journal.pone.0141185 PMID: 26505756

39. Economy E, Levi M. By All Means Necessary: How China's Resource Quest is Changing the World. Oxford University Press, 2014.