The Impact of Ambient Air Pollution on Chinese Expressed Happiness through Social Media[1]

Yiren Wang

Duke University yw219@duke.edu

Version: December 2019

Abstract

Faced with severe outdoor air pollution, people's expressed happiness on social media will be decreased. Daily $PM_{2.5}$ concentration reports and visible weather quality can serve as stimulation information to motivate individuals to tweet on Sina Weibo, which is the biggest microblog platform in China like Twitter and has more than 200 million daily active users. In this study, I investigated the association between individual expressed happiness and the instant daily ambient air pollution levels (using Air Quality Index, short for AQI). Approximate ten thousand posted Weibo contents data were analyzed from the three biggest Chinese cities (Beijing, Shanghai, and Guangzhou) in 2019. The result shows that mainly when the AQI gets worse from 0 to 150, the overall expressed happiness in Weibo (public sentiment) would decrease by 5.73%. This study suggests the instant outdoor air quality contributes to the residents' emotions and local governments should pay more attention to the public's mental health and accelerate the acceleration of solving environmental threats.

Keywords: Expressed Happiness; Public Sentiment; Air Quality Index (AQI).

JEL Classifications: Q5, H2

[1] I am grateful for the helpful comments from my professor Dr. Moon Joon Kim, my classmates Fanyu Wang, Tianqi Wu, and Yingjie Chen. Any errors remain the author's responsibility. For correspondence, please contact Yiren Wang, International Master of Environmental Policy, at No.8 Duke Avenue, Kunshan, Jiangsu, China, 215316.

1 Introduction

Air pollution is becoming a prominent problem in China. It is always charged with the rapid industrialization and high population density. The emissions from industrialization, coal burning, and motorization lead to the deterioration of China's environment and urban life quality (Zheng and Kahn 2013, Ebenstein et al. 2015). Since 2008, Chinese citizens have become more vocal about ambient air quality issues and real-world environments (Kay 2015). Especially in large and medium cities, health concerns have become more of a problem (He et al. 2001). On polluted days, more people purchased self-protection products such as particulate-filtering facemasks and reduced their outdoor times (Zhang and Mu 2018). Air pollution incidents can also even drive people to escape. With the worsening of 100 points in the air quality index, there is a 49.60% increase in population outflow in China (Cui et al. 2019).

The air pollutants such as $PM_{2.5}$, PM10, sulfur dioxide, nitrogen dioxide, and carbon emissions do great harm to human's physical health (Franklin et, al. 2007, Darçın 2014). Meanwhile, researchers have examined the effect of the relationship between environmental pollution and subjective well-being (Welsch 2002, Cuñado & De Gracia 2013, Li et al. 2018). In China, Zhang and Wang (2019) found out that the control of $PM_{2.5}$ concentrations can help decrease happiness inequality and increase the individuals' levels of happiness in the long term. Meanwhile, some studies have matched self-reported well-being data from surveys with daily air quality (Zhang et al. 2017, Chen et al. 2017).

Social media is viewed as the gather of pubic opinions. In recent years, social media, such as Twitter and Facebook, have billions of active users worldwide. Sina Weibo is a popular Chinese biggest microblog platform synonymous with Twitter. According to its financial report of the fourth quarter of 2018, Weibo had 462 million monthly active users and 200 million daily active users. Among all the registered users, 35% were at the age of 18 to 22, 40% were at the age of 23 to 30, and 14% were at the age of 31 to 40. The rest 11% were teenagers from 16 to 17 and middle and elder generations above 40 years old. Sina Weibo is not just a major communication channel, at the same time, the content from it is the reflection of real-time updates of social activities. When $PM_{2.5}$-related hazard is serious, air pollution would be the spotlight of public attention (Sha et al. 2014).

The natural language processing (NIP) has provided a plethora of methods to understand the sentiment expressed by the writers. By conducting a comprehensive meta-analysis, Siebert et al (2019) have assessed the accuracy of language models for sentiment analysis. It proves that on average, 91.48% of language model-based sentiment classifications can automatically capture writers' opinions and match human perception. When it comes to the environmental fields, by using the 'Tencent' natural language processing (NLP) platform, Zheng et al (2019) try to find the relationship between air pollution and urbanites' expressed happiness on social media in China.

In this research, I investigated the association between individual expressed happiness on Sina Weibo and the instant daily ambient air pollution levels (AQI) in three Chinese cities, Beijing, Shanghai, and Guangzhou. 9089 valid tweet samples from January to March 2019 were investigated as self-report happiness content.

2 Data

I collected four types of data: Weibo tweets contents, Air quality index (AQI), weather and holiday data. The Weibo tweets contents were obtained on the Weibo platform by web crawlers; AQIs were collected from the Ministry of Ecology and Environment (MEP) of China; weather data came from three regional government websites; holiday data was accordance with the national government's schedule. The dataset was composed of results from January 1st to March 18th in three Chinese cities: Beijing, Shanghai, and Guangzhou. The text sentiment analysis and expressed happiness were accomplished by an open AI platform for Chinese text, the Baidu sentiment analysis.

Weibo Posts: By web crawlers, more than ten thousand original posts from January 1st to March 18th were collected on the Weibo platform. I used two-dimension keywords to filter the posts containing information with "location" and "concern" items at the same time. Three locations were chosen: Beijing (北京), Shanghai (上海), and Guangzhou (广州). They were center cities in three important economic centers in China: Beijing-Tianjin-Hebei (BTH) region, Yangtze river delta, and pearl river delta. Three "concern" terms were selected: haze (雾霾), weather (天气), and emotion (心情). The messages contained with one of these "concern" keywords indicted that the Weibo users have paid attention to the environment and air quality. However, the use of haze showed a direct mention of air pollution; the use of weather told that the user actively may have noticed the nearby environment; the use of emotion may simply mood sharing behaviors. The two-dimension filter (location and concern) made nine combinations of keywords and provides people's responses in different situations.

Beijing, Shanghai, and Guangzhou economic performances were listed the top 10 cities by GDP in 2018 (GDP performance: Shanghai, No1; Beijing, No2; Guangzhou No4), meanwhile, they have different geographical climate and population density (the city population density in 2017: Shanghai, No2; Guangzhou No7; Beijing, No10). Also, due to the large population density, they all have millions of active users on the Weibo platform.

Haze, weather, and emotion, as three keywords in the posts, were typed by the users. They can be viewed as the users' self-reports and different degrees of environmental recognition: the messages should be stimulated by the nearby surroundings to some extent. For "haze", the users' direct mention may come from their notices to the ambient air pollution recently. For "weather", the reasons for sharing could be quite different. They may live under a heavily polluted environment or experience different weather conditions like sunny, rainy and windy. For "emotion", driving forces of posting Weibo may not relate to the air quality or the environment, instead, such messages were more like purely mood-sharing. In a word, from 'haze', 'weather', and 'emotion', the extents of the relationship may get weaker between ambient air quality and content sharing behind the Weibo posts.

However, $PM_{2.5}$ was not chosen as a filter due to the limited related results. As a scientific terminology, it was still not an everyday vocabulary. Instead of individual posts, governments' air quality reports and $PM_{2.5}$ alerts on the Weibo platform were more commonly contained "$PM_{2.5}$" keyword. In addition, air pollution (空气污染), was also not chosen as a filter keyword because large percentages of mentions were the air quality announcements. Other Chinese words and expressions such as air (空气) and environment (环境) hold various meanings in different contexts. Their referential are not unique and are not confined to environmental and pollution fields.

The location-based filter was aimed to tag the instant and true sharing locations of these posts. Though the users' homepages usually contain location information, it may have several kinds of meanings: the birthplaces or permanent resident regions or incorrect information. In addition, users may also share their posts when traveling. To avoid ambiguity, I added direct location information as a filtering factor to increase the accuracy of geospatial data. However, meanwhile, the number of posts containing locations information would largely reduce due to the keyword-based selection criteria.

As the biggest microblog platform, the searching services of Sina Weibo can just retrieve full-text content for the top 50 pages of all tagged tweets with keywords. On one page, 10 relevant tweets were listed from recent to a long time ago. The data collection time was from January to April repeatedly. On account of the 50-page limitation, though multiple sampling collection processes have been tried, the collected Weibo posts in this research were just part of all related tweets on the Weibo platform. The whole dataset can't be traversed. The statistic in this paper was not a quasi-dataset. However, nearly 10000 valid data were applied to find out the relationship between daily air pollution levels and people's expressed happiness on social media.

Expressed happiness data: The Baidu sentiment analysis service has been applied to get the sentiment data. It is an open commercial AI platform for Chinese text, combining the deep learning technology and Baidu big data. The emotional polarity will be returned according to the full-text content data. The results were classified into three types: sentiment polarity, positive probability (hereinafter positive prob) and confidence degree. Sentiment polarity results were categorical: positive (happy), negative (sad) and neutral (in the middle); positive probabilities were continuous from 0 to 1; the range of confidence degree was also continuous from 0 to 1. Different from confident intervals, it showed how many degrees were Baidu sentiment analysis service confident and certain about the return values.

As one of the biggest Internet-related services provided, the Baidu search engine can get exposed to numerous Chinese texts, which means there was no extra targeted training for environmental-related items. The mention of "location" and "concern" information will not influence the sentiment and emotional polarity results. However, the service required the maximum word count: the service provider wouldn't reply to any results when content was more than 150 Chinese characters. Some Weibo posts failed to meet this standard and have been identified as invalid samples. Table 1 shows the valid sample distribution in Beijing, Shanghai, and Guangzhou.

Table 1: Valid sample distribution

Valid samples	Haze	Weather	Emotion	All
Beijing	2806	1439	1525	5770
Shanghai	847	798	979	2624
Guangzhou	220	17	458	695
All	3873	2254	2962	9089

Air quality index (AQI): The daily Air Quality Index (AQI) was obtained from the national government website, Ministry of Ecology and Environment (MEP) of China. For every record, data were collected and summarized from the cities air monitor stations. In China, AQI consists of six criteria pollutants: PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, and O_3.

Weather data: Weather data come from the three regional government websites (Beijing, Shanghai, and Guangzhou). Because of the lack of accurate regional average rainfall data, the daily precipitation data were binary, consisting of 0 and 1. The total rainfalls (rainy and snowy) and the lasting duration were not taken into account. The units of maximum and minimum temperatures were the degree centigrade (°C).

Table 2: Average AQIs

Valid samples	Haze	Weather	Emotion	All
Beijing	94.79	53.82	102.13	86.52
Shanghai	94.69	80.95	85.92	87.24
Guangzhou	74.00	97.94	64.47	68.31
All	93.59	63.76	90.95	85.33

Table 3: Average expressed happiness

Positive prob	Haze	Weather	Emotion	All
Beijing	0.64	0.81	0.70	0.70
Shanghai	0.60	0.80	0.71	0.70
Guangzhou	0.68	0.72	0.66	0.67
All	0.64	0.81	0.70	0.70

Holiday data: Holiday and workdays were treated differently. Public holidays like Spring Festivals together with weekend days were coded as 1. The actual workdays were treated as 0. Due to the existence of taking working days off policy around the national holidays, the working days were not simply as same as the five regular working days (Monday, Tuesday, Wednesday, Thursday and Friday). The taking working days off policy and public holidays schedule were obtained from the national government website.

Generally, there were 9087 valid samples in three cities. After removing duplicates, every useful tweet had the following statistics: expressed happiness (sentiment polarity, positive probability, and confidence degree), AQI in the mentioned location and at the tweet posting day, weather data (daily precipitation and temperature), and holiday.

However, the data distributions were not even. As table 1 shows, 63.5% of valid samples were in Beijing, 28.9% were in Shanghai and the rest, only 7.6% were in Guangzhou. The bias may come from the lower engagement of active Weibo users in Guangzhou. The floating population like tourism and short-term workers may also influence the active users in the cities. The frequencies of "concern trigger keywords" mentions seem even. 42.6% of valid tweets were "haze" related; 24.8% posts mentioned "weather" and the rest, 32.6%, had the keyword "emotion".

The daily AQI in table 2 shows the different air quality conditions in these three cities. From January 1st to March 18th, the average air qualities behind collected tweet results were not bad. In Guangzhou, the average was 68.31. In Beijing, the average AQI was 86.52 and in Shanghai, it was 87.24. According to the suggested AQI classification standard (Technical Regulation on Ambient Air Quality Index [on trial]) offered by the national government, days with 0~50 AQI can be viewed as pretty good, 51~100 as in a well good condition, 101~150 as slightly polluted, 151~200 as middle level polluted; 200~300 as high level polluted, and >300 can be identified as extremely serious polluted.

Table 3 shows the average expressed happiness. It can be seen that the average expressed happiness in the three locations are stable (approximately 0.7). Under the keyword "weather", people are happier (0.81); under the keyword “emotion”, people are in the middle, and when directly mentioned “haze”, the average happiness is the lowest (0.64).

3 Methods and results

Firstly, I set the null hypothesis H0: with the increase in outdoor air pollution levels, people’s expressed happiness on social media will decrease. Also, to further the understanding of the two return values of expressed happiness data (sentiment categorical and positive probabilities), two logistic equations were used separately to test the relationship between the relationship of air quality and the public expressed happiness.

3.1 The Baidu sentiment results

3.1.1 Sentiment categorical data

The logistic equation was intended to describe the association. Since the sentiment returns were categorical (positive, negative, and neutral) I recoded them into two binaries: positive was equal to 1, showing the happy moods contained in the tweets; negative or neutral moods were equal to 0.

$$\text{Y sentiment} = \text{Constant} + b(\text{AQI}) \tag{1}$$

3.1.2 Positive probability results

A similar treatment was used to recode the contiguous positive probability values. I recorded the positive probabilities which were more than 0.5 as 1, and 0 was attached to results less than or equal to 0.5. By viewing each data, no results were exactly equal to 0.5.

$$\text{Y positive prob} = \text{Constant} + \text{b}(\text{AQI}) \tag{2}$$

The dependent variable was the binary results of sentiment or positive probability. The independent variable was the Daily AQIs of the locations in each tweet. The results were shown in table 4. Though the parameters of the two equations were different (-0.00205 and -0.00170), the relationships of AQI and expressed happiness were all negative. The potential explanations may relate to the number of tweets identified as neutral: the mechanism of neutral identification was unknown. Though no positive probability results were exactly equal to 0.5, some sentiment behind the posts were identified as neutral. Baidu didn't publicly give the rules of the relationship between positive probability (0~1), sentiment polarities (positive, negative, and neutral), and confidence degree (0~1) on the website or technical manuals.

In the following discussion, I chose "positive probability" as the dependent variable. On one hand, it implied that we may over- or under-estimate the association because we ignore the neutral results identified by the Baidu sentiment service. Maybe the Baidu sentiment service viewed tweet content as neutral though they contained a strong emotional polarity (the absolute values between positive probabilities and 0.5 were more than 0.4). Maybe they were not so confident in their classifications. The confident degree was lower (eg. 0.01) and such tweets were identified as neutral. On the other hand, the reasons for choosing positive probability results instead of sentiment polarity data were that probabilities data had six significant digits.

Table 4: The effect of using sentiment and positive probability

	(1)	(2)
VARIABLES	Sentiment	Positive prob
AQI	-	-0.00170***
	0.00205***	-
	(0.000499)	(0.000526)
Constant	1.197***	1.370***
	(0.0497)	(0.0523)
Observations	9,087	9,087

*** p < 0.01, ** p < 0.05, * p < 0.1

3.2 Two-dimension keywords

Secondly, I fixed the "location" and "concern" keywords together and separately. AQI was the independent variable. In different locations, various inherent discrepancies exist such as the minimum wages, seasonal climate conditionals, and industrialization levels. Besides, for direct mentions (haze), and non-direct mentions (weather and emotion), the consciousness of air pollution levels in each post varies. The positive probabilities were enlarged by 100 times and the results were shown in table 5.

$$Y_1 \text{ positive prob} = \text{Constant} + b_1(\text{AQI}) + i + t. \tag{3}$$

Table 5: The effect of locations and keywords

	(1)	(2)	(3)	(4)
VARIABLES	positive prob	positive prob	positive prob	positive prob
AQI	-0.0382***	-0.0408***	-0.00149	-0.00263
	(0.00635)	(0.00638)	(0.00641)	(0.00648)
Constant	73.02***	73.50***	63.74***	64.27***
	(0.617)	(0.665)	(0.745)	(0.773)
Observations	9,087	9,087	9,087	9,087
R-squared	0.004	0.006	0.057	0.058
Locations	NO	Yes	NO	YES
Keywords	NO	NO	YES	YES

*** p < 0.01, ** p < 0.05, * p < 0.1

As the results in table 5 shown, the influence of locations was more significant than the effect of “concern” keywords. Generally, the public expressed happiness would decrease by 3.82% when the AQI got worse and increased 100. Considering the effect of locations, the decreased percentage will enlarge to 4.08. Compared with pretty good air quality conditions (when AQI was equal to 0), on high polluted days (when AQI was 250), there was at least a 9.55% happiness possibility decrease. The polluted level would influence people’s daily happiness no matter they noticed the air pollution level and nearby environmental conditions or not.

3.3 Further simulations

3.3.1 Maximum temperature, precipitation and holiday

Thirdly, to better describe the expressed happiness impacting factors, besides AQI, I estimated the effect of other variables like maximum temperature, precipitation and holiday. The data collection period was from January to March, and I chose the maximum temperatures to simulate the sensible temperatures indoor and outdoor in the winter. As mentioned in the data section, the precipitation and holiday variables were binary data; the AQI and maximum temperatures were continuous.

$$Y_2 \text{ positive prob} = \text{Constant} + b_1(\text{AQI}) + b_2(\text{Max temperature}) + b_3(\text{Precipitation}) + b_4(\text{Holiday}) + i + t \tag{4}$$

Table 6: The effect of various factors

	(1)	(2)	(3)	(4)
VARIABLES	positive prob	positive prob	positive prob	positive prob
AQI	-0.0336***	-0.0456***	-0.00168	-0.00974
	(0.00646)	(0.00655)	(0.00660)	(0.00690)
Max temperature	0.663***	0.960***	0.233***	0.402***
	(0.0571)	(0.0647)	(0.0612)	(0.0730)
Precipitation	-0.634	0.826	-0.140	0.884
	(1.989)	(1.997)	(1.955)	(1.970)
Holiday	-4.273***	-2.478***	-2.724***	-2.052***
	(0.770)	(0.788)	(0.762)	(0.778)
Constant	65.10***	63.19***	62.07***	61.55***
	(0.937)	(0.963)	(0.942)	(0.963)
Observations	9,087	9,087	9,087	9,087
R-squared	0.026	0.036	0.060	0.062
Locations	NO	Yes	NO	YES
Keywords	NO	NO	YES	YES

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

The results in table 6 showed the influences of different factors. The location still had a large influence on the results compared with the “concern” keywords. When the “concern” keywords were fixed, the relationship between AQI and expressed happiness was not so robust and significant. For the AQI, when the overall air quality worse, people tended to express more unhappy feelings on the Weibo platform; when it came to the increase of the maximum temperature, people would be easier to feel happier; the effect of precipitation was correlated with the locations; the influence of workdays and holidays was significant but contrary to common sense: people were less happy in holidays. In the discussion part, the effect of expressed happiness in three cities and potential reasons would be reviewed separately.

3.3.2 Potential delay effect

Fourthly, the delay effect of air quality and people's emotions was discussed. Since individual behaviors were unpredictable, their moods may not be just determined by the instant air quality. At the same time, long-term periods of bad weather and high contaminated levels may also ruin their moods. Besides, they may choose to tweet their feelings on social media in the next few days. For instance, the complains posted today were about the events happened yesterday; the anxiety and anger may get accumulated due to the continuous bad weather during the week.

The delay effect in this part accounted for the average AQIs in the past few days. For example, the 1-day delay meant that the averages of two AQIs were calculated: the same day of the Weibo posts and the day before that day. Four situations were created to compare the differences: instant air quality (on the same day), 1-day delay (the average 2-day AQIs), 3-day delay (the average 4-day AQIs), and 7-day delay (the average 8-day AQIs). According to the findings above, the location was fixed to eliminate the potential impact.

$$\begin{aligned} Y_3 \text{ positive prob} &= \text{Constant} + b_1(\text{AQI\_average}) + b_2(\text{Max temperature}) \\ &+ b_3(\text{Precipitation}) + b_4(\text{Holiday}) + i + t \end{aligned} \tag{5}$$

Table 7 provided findings on the delay effect of self-reports and air quality on social media. The delay influences should be attached significances. The parameters of instant, 1-day delay, 3-day delay, and 7-day delay gradually decreased (from -0.0456 to -0.0973). In the short term, the constant bad air quality would expand people's negative emotions.

To recapitulate, the relationships of AQI and expressed happiness were negative. When the AQI increased 100, the public happiness on social media would decrease by 3.82 %. The polluted levels would influence people's daily happiness no matter they noticed the air pollution level and nearby environmental conditions or not. The location of the cities affected significantly. Invisibility impact of pollution level had towards people's emotion regardless of direct mention environmental-related items. The delay effect matters of air quality and expressed happiness on social media. The lasting of higher contaminated levels would let people feel more anxiety.

Table 7: The effect of a potential delay

	(1)	(2)	(3)	(4)
VARIABLES	instant	1-day delay	3-day delay	7-day delay
AQI	-0.0456***			
	(0.00655)			
1-day delay		-0.0514***		
		(0.00718)		
3-day delay			-0.0696***	
			(0.00869)	
7-day delay				-0.0973***
				(0.0131)
Max temperature	0.960***	0.942***	0.932***	0.993***
	(0.0647)	(0.0643)	(0.0640)	(0.0654)
Precipitation	0.826	-0.0541	-1.035	-1.899
	(1.997)	(1.992)	(1.993)	(2.005)
Holiday	-2.478***	-3.093***	-3.830***	-3.089***
	(0.788)	(0.769)	(0.760)	(0.768)
Constant	63.19***	62.58***	64.31***	65.47***
	(0.963)	(0.927)	(1.002)	(1.110)
Observations	9,087	9,087	9,087	9,087
R-squared	0.036	0.037	0.038	0.037
Locations	YES	Yes	YES	YES

*** p < 0.01, ** p < 0.05, * p < 0.1

4 Discussion

4.1 Two comparisons

The association between the expressed happiness and AQI may be different under different keyword combinations. Therefore, two comparisons (Table 8-1, 8-2) were made between cities and "concern" keywords.

1. In different cities, the effects of ambient air pollution were all negative. As air pollution gets worse, people will be more upset whenever they were. For the AQI parameters, the differences between Beijing, Shanghai, and Guangzhou were not so huge. However, Guangzhou residents were more sensitive to the increase of AQI. One guess was that because the air quality was always stably good in Guangzhou. People can feel small

fluctuations of the worsening trend. The effects of precipitation were different among cities, and not so confident.

2. The influences of "concern" keywords were not confident and robust. But just comparing the parameters, when people mentioned “haze” and “emotion”, the worse of AQI would more obviously influence the individual happiness. Different from the former guess, the sentiment polarity was closer to positive when sharing the “weather”. One potential guess maybe that when air pollution got worse, the direct mention of “weather” can be an eager wish and expectation of future environmental improvement.

Table 8-1: The effect of expressed happiness in three cities

	(1)	(2)	(3)
VARIABLES	Beijing	Shanghai	Guangzhou
AQI	-0.0355***	-0.0337***	-0.0455***
	(0.00647)	(0.00646)	(0.00655)
Max temperature	0.731***	0.662***	0.949***
	(0.0595)	(0.0571)	(0.0642)
Precipitation	0.423	-0.793	0.517
	(2.005)	(2.000)	(1.983)
Holiday	-3.954***	-4.262***	-2.512***
	(0.773)	(0.770)	(0.787)
Constant	62.68***	64.98***	63.05***
	(1.110)	(0.949)	(0.957)
Observations	9,087	9,087	9,087
R-squared	0.028	0.027	0.036
Keywords	NO	NO	NO

*** p < 0.01, ** p < 0.05, * p < 0.1

4.2 Independent variable interpretation

Combing with the actual situation in these three cities, some potential explanations were provided to understand the results of some independent variables.

Precipitation: The attitudes towards may be largely determined by the real situation such as the average hourly rainfall depth, daily total rainfall and the duration time. For instance, if there was only sprinkle for a while, people may not feel strongly unsatisfied with the weather. The city's transportation systems may not be strongly challenged. Less inconvenience was brought into their lives.

Table 8-2: The effect of three different "concern" keywords

	(1)	(2)	(3))
VARIABLES	Haze	Weather	Emotion
AQI	-0.0303***	-0.0125*	-0.0466***
	(0.00661)	(0.00692)	(0.00658)
Max temp	0.654***	0.505***	0.963***
	(0.0687)	(0.0721)	(0.0647)
Precipitation	1.082	0.674	0.868
	(1.980)	(1.977)	(1.997)
Holiday	-2.303***	-2.079***	-2.491***
	(0.781)	(0.780)	(0.788)
Constant	69.30***	62.78***	63.00***
	(1.074)	(0.953)	(0.972)
Observations	9,087	9,087	9,087
R-squared	0.053	0.056	0.037
Locations	YES	YES	YES

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Holiday: As the above graph showed, one of the findings was that people were more likely to express negative feelings during the holidays. One potential clue may be that some companies had different holiday policies rather than the national schedule. In shanghai, some private companies would let their workers enjoy their holiday on Sunday and Monday instead of Saturday and Sunday. However, it was hard to obtain that data in the three cities.

4.3 Limitation

This research has several limitations. Firstly, in the raw data collection period, the searching mechanism set the limitation: they would only provide 50 pages of results. I tried to overcome the barrier by collecting recent data every day from January to April. However, such trials can't guarantee the quasi-data have been obtained. In addition, Sina Weibo didn't show the rule of searching boxes and the priority for rankings 50 pages-results were unknown.

Secondly, the doubts of sentiment analysis results may exist. In the real world, it was hard to quantitatively estimate the positive possibility of people's expression. On social media, pictures may also serve as part of the mood sharing. However, from now on, the Baidu sentiment analysis can't take the texts and pictures into consideration at the same time.

Thirdly, the data only consisted of three months. In April, the Sina Weibo platform announced a new function: users can use the unified-management option to organize the visible status of the tweets. In the self-privacy setting, they can close all the access permissions for half-a- year-ago tweets. Therefore, after that, fewer results can be assessed.

Fourthly, in this research, fewer data was collected to figure out the relationship under the air heavily polluted conditionals. From January to March, most tweets were posted with 0~150 AQI. Among the 9089 samples, posts with 151~235 AQI were only composed of 7.51%. According to the current China AQI classification standard, there were pretty good (0~50), well good (51~100), slightly polluted (101~150), middle level polluted (151~200). For the relationship between air quality and expressed happiness on slightly polluted (101~150), highly polluted (200~300), extremely serious polluted days (>300), this research can provide limited support.

4.4 Policy recommendations

A greener future should be guaranteed by governments. Though air quality is not the only factor influencing the happiness, the effect of serious environmental damages is permanent. Sometimes, the long-term bad environmental quality will gradually become a stereotype of a city. Among the 9089 tweets, some of them inevitably compared the air quality city by city. In Beijing, the averages of air quality from January to March were not too bad: among 80 to 100, ranking as in a well good condition. However, in the past, Beijing's air quality was always linked with sandstorm on the news. Some users said that they really wanted to flee away from Beijing and settle down in places with good air quality in the future. The public environmental health is not only composed of environmental-related diseases but also mental health issues as expressed happiness.

5 Conclusion

To conclude, the findings indicated that in China, the expressed happiness in Weibo posts would decrease together with the worsening of air quality. By correlating the social media content with surveillance AQI data, the relevance of instant outdoor environmental quality and residents' sentiment has been validated. Based on the 9087 Weibo posts from January 1st to March 18th, when the AQI range varied from 0 (air quality pretty good) to 150 (middle level polluted), the average expressed happiness decreased by 5.73%. The short-term high contamination levels may make things worse: the expressed happiness would largely decrease due to the lasting non-good air quality. Besides, residents in different regions held different emotional sensitivity on the impact of air pollution. The results suggest that public psychological health is non-ignorable, and environmental health should be delivered with more insights and attention.

6 References

Sha, Y., Yan, J., & Cai, G. (2014, January). Detecting public sentiment over PM2. 5 pollution hazards through analysis of Chinese microblog. In *ISCRAM*.

Cui, C., Wang, Z., He, P., Yuan, S., Niu, B., Kang, P., & Kang, C. (2019). Escaping from pollution: the effect of air quality on inter-city population mobility in China. *Environmental Research Letters*, *14*(12), 124025.

He, K., Yang, F., Ma, Y., Zhang, Q., Yao, X., Chan, C. K., Cadle, S., Chan, T. & Mulawa, P. (2001). The characteristics of PM2. 5 in Beijing, China. *Atmospheric Environment*, *35*(29), 4959-4970.

Ebenstein, A., Fan, M., Greenstone, M., He, G., Yin, P., & Zhou, M. (2015). Growth, pollution, and life expectancy: China from 1991-2012. *American Economic Review*, *105*(5), 226-31.

Zheng, S., & Kahn, M. E. (2013). Understanding China's urban pollution dynamics. *Journal of Economic Literature*, *51*(3), 731-72.

Zhang, X., Zhang, X., & Chen, X. (2017). Valuing air quality using happiness data: the case of China. *Ecological economics*, *137*, 29-36.

Zhang, X., Zhang, X., & Chen, X. (2017). Happiness in the air: how does a dirty sky affect mental health and subjective well-being?. *Journal of environmental economics and management*, *85*, 81-94.

Franklin, M., Zeka, A., & Schwartz, J. (2007). Association between $PM_{2.5}$ and all-cause and specific-cause mortality in 27 US communities. *Journal of Exposure Science and Environmental Epidemiology*, *17*(3), 279.

Darçın, M. (2014). Association between air quality and quality of life. *Environmental Science and Pollution Research*, *21*(3), 1954-1959.

Cuñado, J., & De Gracia, F. P. (2013). Environment and happiness: New evidence for Spain. *Social Indicators Research*, *112*(3), 549-567.

Li, Y., Guan, D., Tao, S., Wang, X., & He, K. (2018). A review of air pollution impact on subjective well-being: Survey versus visual psychophysics. *Journal of cleaner production*, *184*, 959-968.

Welsch, H. (2002). Preferences over prosperity and pollution: environmental valuation based on happiness surveys. *Kyklos*, *55*(4), 473-494.

Zhang, P., & Wang, Z. (2019). PM2. 5 Concentrations and Subjective Well-Being: Longitudinal Evidence from Aggregated Panel Data from Chinese Provinces. *International journal of environmental research and public health*, *16*(7), 1129.

Davis, J. J., & O'Flaherty, S. (2012). Assessing the accuracy of automated Twitter sentiment coding. *Academy of marketing studies journal*, *16*, 35.

Zheng, S., Wang, J., Sun, C., Zhang, X., & Kahn, M. E. (2019). Air pollution lowers Chinese urbanites' expressed happiness on social media. *Nature Human Behaviour*, *3*(3), 237.

He, K., Yang, F., Ma, Y., Zhang, Q., Yao, X., Chan, C. K., Chan, T. & Mulawa, P. (2001). The characteristics of PM2. 5 in Beijing, China. *Atmospheric Environment*, *35*(29), 4959-4970.

Ebenstein, A., Fan, M., Greenstone, M., He, G., Yin, P., & Zhou, M. (2015). Growth, pollution, and life expectancy: China from 1991-2012. *American Economic Review*, *105*(5), 226-31.

Zheng, S., & Kahn, M. E. (2013). Understanding China's urban pollution dynamics. *Journal of Economic Literature*, *51*(3), 731-72.

Appendix

Figure 1: Valid sample distribution

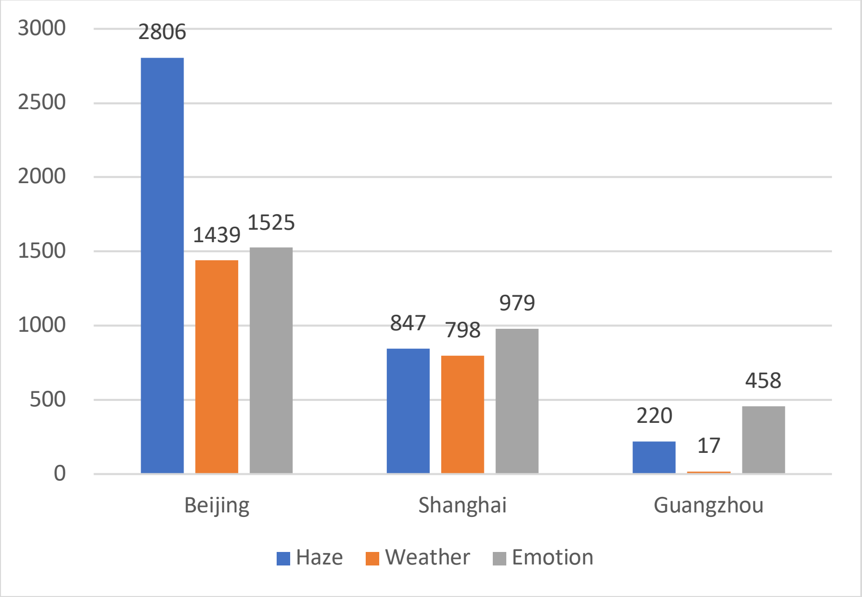


Figure 2: Sentiment polarity distribution

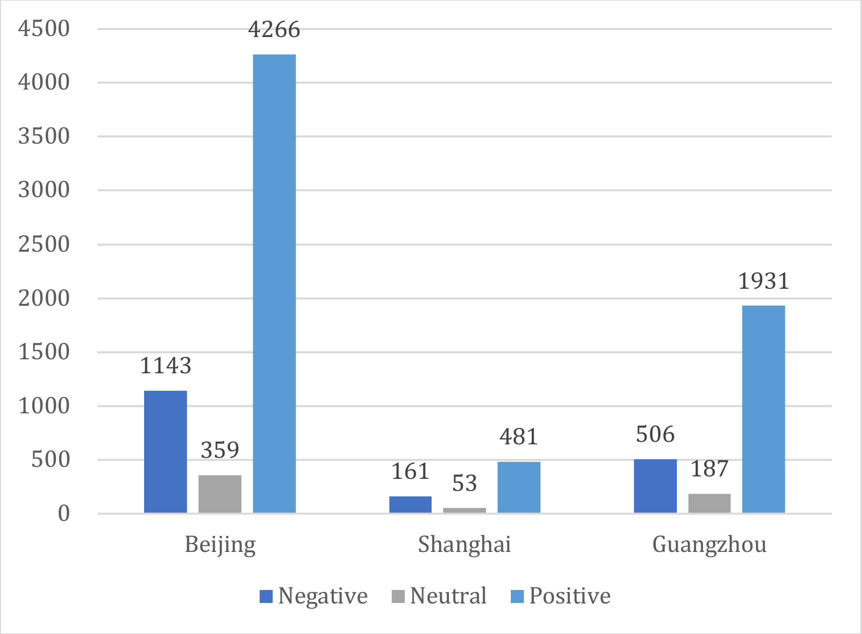


Figure 3: The average positive probability

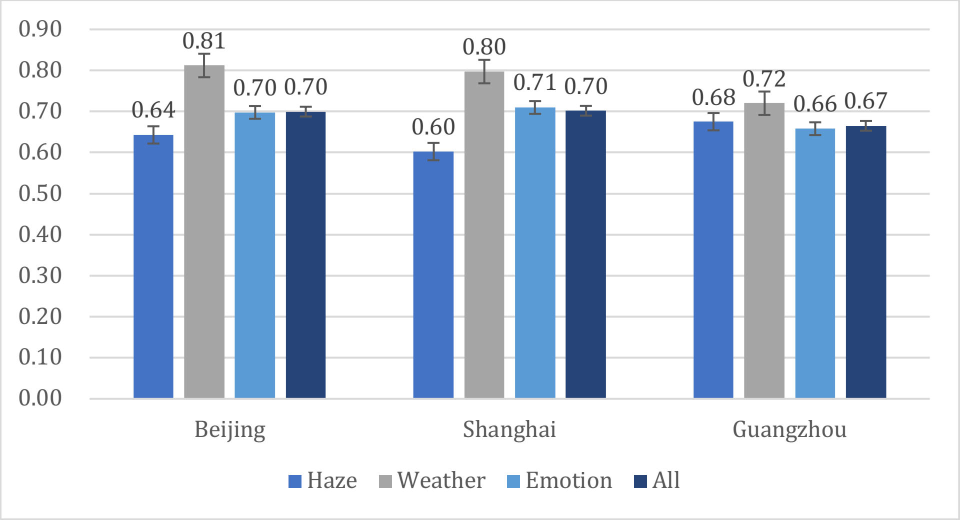


Figure 4: Average AQI distribution

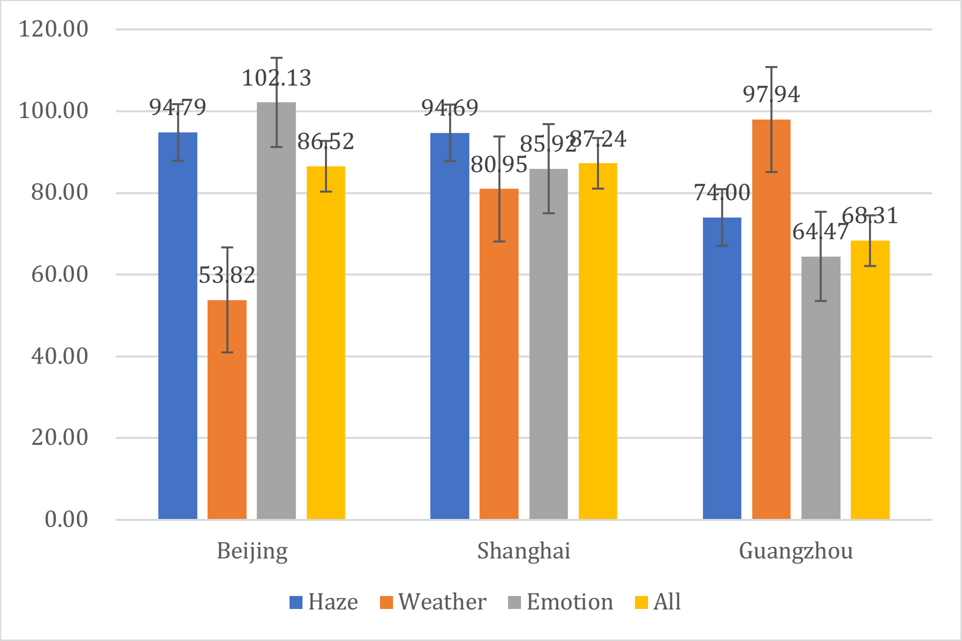
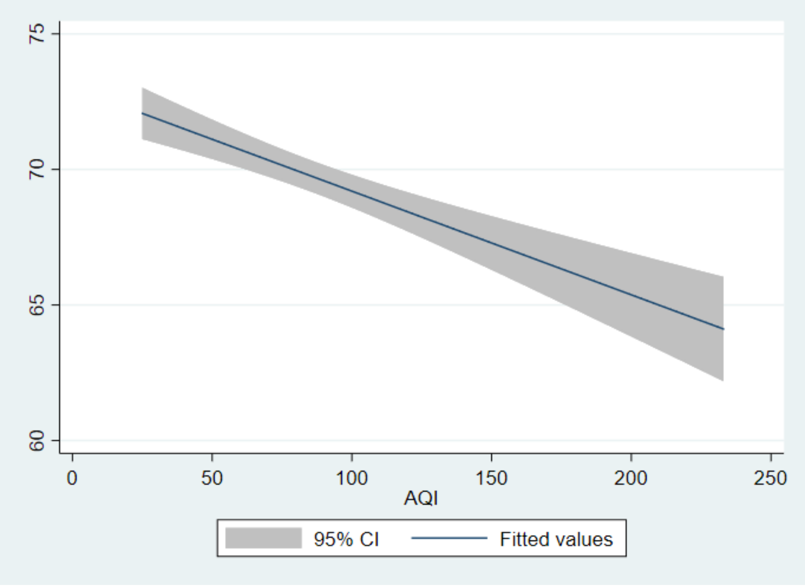


Figure 5: Relationship between AQI and positive probabilities


AQI
95% CI
Fitted values