LUND UNIVERSITY
School of Economics and Management
Department of Informatics

Using Social Media for air pollution detection

The case of Eastern China Smog

Master thesis 15 HEC, course INFM10 in Information Systems
Presented in June, 2017

Authors: Han Gao
Yu Shi

Supervisor: Zafeiropoulou Styliani

Examiners: Markus Lahtinen
Azadeh Sarkheyli

Using Social Media for air pollution detection: The case of Eastern China Smog

Authors: Han Gao and Yu Shi

Publisher: Dept. of Informatics, Lund University School of Economics and Management.

Document: Master Thesis

Number of pages: [41]

Keywords: air pollution surveillance, social media analytics, text analysis, time series analysis, co-word network analysis

Abstract:

Air pollution has become an urgent issue that affecting public health and people's daily life in China. Social media as potential air quality sensors to surveil air pollution is emphasized recently. In this research, we picked up a case-2013 Eastern China smog and focused on two of the most popular Chinese microblog platforms Sina Weibo and Tencent Weibo. The purpose of this study is to determine whether social media can be capable to be used as 'sensors' to monitor air pollution in China and to provide an innovative model for air pollution detection through social media. Based on that, we propose our research question, how a salient change of air quality expressed on social media discussions to reflect the extent of air pollution. Hence, our research (1) determine the correlation between the volume of air quality-related messages and observed Air quality index (AQI) with the help of time series analysis model; (2) investigate further the impact of a salient change of air quality on the relationship between the people's subjective perceptions regarding to air pollution released on the Weibo and the extent of air pollution through a co-word network analysis model. Our study illustrates that the discussions on social media about air quality reflect the level of air pollution when the air quality changes saliently.

Content

Figures

Tables

1 Introduction

With the rapid increment of private cars, factories and the use of fossil fuels in China recently, outdoor air pollution gradually becomes an urgent issue in most regions, which harms seriously public health. Particulate air pollution not only spreads respiratory disease among people, but also increase the mortality of cardiovascular and respiratory disease among older people (Seaton et. al, 1995). According to world health organization(WHO), China has the highest number of death of outdoor air pollution in all over the world, which is more than 1 million people in 2012. Therefore, it is important to avoid the exacerbation of air pollution and enhance public protected awareness.

Air quality index(AQI) is a scale that precisely indicates the levels of air pollution. It is created based on data coming from sensors in monitoring stations that measure the urban real-time quality and pollution level. Most websites base on AQI to show the air condition. However, Jiang et. al (2015) claim that the monitoring stations as requirements to collect and calculate data for AQI are huge and costly in order to surveil air quality, which means those monitoring stations need a large amount of labour and money to maintain and manage. Some cities and most rural regions still lack monitor stations or even any approaches to achieve the surveillance of air quality(ibid).

At the same time, with social media expansion, people move to online social space. They create huge amounts of content on social media by discussing and commenting their experiences, feelings and even events happening around them. Social media have become one of the main data sources when it comes to "Big Data" era (Watson, 2014). As a consequence, social media as a data source has attracted the interest of many researchers with focuses on real world events, for example, 2012-2013 flu outbreak (David, Michael and Mark, 2013), Queensland floods (Ahmed and Sinnappan, 2013), Haitian earthquake (Yates, D. and Paquette, S., 2011), Canberra fire (Eustace, J. and Alam, S., 2012). Twitter, that is one of the most popular social media website in the world, has great potential to perform disaster surveillance based on its massive and instant public short status messages everyday (Broniatowski, Paul and Dredze, 2013). Likewise, 'Chinese twitter'--Weibo, has been used in previous studies to detect air pollution (Jiang et. al, 2015; Wang, Paul and Dredze, 2015).

To give further an insight for precise 'social sensors' on monitoring air pollution focusing on China, in this study, diverse data sources are taken considerations to investigate the correlation between them in order to find out a qualified alternative of monitoring stations to sniff air pollution. According to Hevner and Chatterjee (2010), the data derived from broader fields can help social science researchers to design effectively, analyse from different perspectives so that can mitigate human bias and better understand the data. Thus we consider the both popular Chinese microblog--Sina Weibo and Tencent Weibo so as to strengthen the social media data analysis of air pollution surveillance.

Given that there are extremely small number of studies in regards to the topic that social media analysis and air pollution in China, it is not enough for a systematic knowledge or theory to demonstrate that social media can be qualified to monitor air pollution timely in China instead of monitor stations. Eustace and Alam (2012) mentioned that it is beneficial to utilize case studies in

times of "available or existing knowledge base is poor, in essence" and "the boundaries between phenomenon and context are not clearly evident". Hence, we select an event-- 2013 Eastern China smog, that is the most serious and widest air pollution in the East China during 2013 as our case study to explore further the correlation between social media data and observed AQI.

1.1 Problem area

In the end of Jiang et. al (2015)'s research, they proposed an assumption for the occurrence of a salient peak of inferred AQI (in April), which is that the temporal changes of daily air quality would lead to the highest correlation coefficient between the frequency of filtered social media messages and observed AQI compared to other months. That is to say that if a change of air quality happened in daily life is unexpected and obvious, namely a 'temporal change', it would attract high attentions on social media and result in high correlation. Furthermore, when the 'temporal change' is salient enough, it could attract excessive attentions thereby lead to a 'fake peak', which may interrupt the reflection of a vertex with the worst air quality. Therefore, the 'temporal change' is a possible error for social media to surveil and reflect the level of outdoor air pollution accurately.

1.2 Research question

Based on aforementioned problem, we think that in a specific period, the temporal change of air quality is more salient, its influence on the reflection of the extent of air pollution is more significant. Hence, a salient change of air quality will be the research target of this study. The hypothesis would be that a salient change of air quality may influence the occurrence of a peak of air pollution afterwards. Therefore, in this research, in order to explore whether the social media can precisely detect the level of air pollution, we clarify herein the research question is that ***How does a salient change of air quality expressed on social media discussions reflect the extent of air pollution?***

1.3 Purpose

To explore whether social media can be a qualified sensor of air pollution when the serious air pollution outbreak, we focus on a specific period --2013 Eastern China smog as our case study. We tend to use correlation analysis to determine the relationship between the volume of air quality-related messages and the level of air pollution referring to observed AQI within our case. Based on that, we concentrate on the text analysis to explore further whether a salient change of air quality can influence people to reflect a peak of air pollution on social media.

Therefore, we aim to (1) determine the correlation between the volume of air quality-related messages from Sina Weibo, Tencent Weibo and observed AQI with the help of time series analysis model; (2) investigate further the content of air quality-related messages for the impact of

a salient change of air quality on the reflection of level of air pollution through co-word network text analysis model.

Collectively, our purpose is to investigate whether social media can accurately monitor the extent of air pollution in China. The contribution of this study would be an extended model that uses multiple social media sources and that it could be used to detect air pollution so that it can improve the public awareness of air pollution and help government to cut the cost of building monitor stations in the future work.

1.4 Delimitation

This research investigates the correlation between the surveillance of air pollution and social media platforms. We focus on the impact of the negative change of air quality on social media platforms to reflect the extent of air pollution. To solve our research question, we pick up a specific case--2013 China Eastern smog with the momentous influence in China to study and access data from two different Chinese social media services that are Sina Weibo and Tencent Weibo via web crawler and API. The data is collected is from 2nd to 14th, December in 2013. The case will pay attention to a part of cities that are influenced the most in this event rather than a country level.

Given that our research bases on Chinese social media services, Chinese text analysis need to be considered predominantly. However, Chinese lexical semantics are considerably complicated. Hence, we do not tend to explore the linguistic analysis such as double meaning of the words when processing the social media text. Instead, obvious keywords in understandable and air quality related messages are paid high attention to. Besides, translation of text into English is also indispensable in this study.

2 Theoretical Background

In this section, we will mention the theoretical background for the research. Social media as a foundation of the whole study is introduced firstly. Then we illustrated social media analyst. Given that the emergency air pollution outbreak was the target in this research, we thereby explained that how social media analyst was applied into emergency managements. To be specific, the two previous researches (Jiang et al., 2015; Wang, Paul and Dredze, 2015) that we focused on mainly were highlighted within the filed of air pollution detection in China. Furthermore, we elaborated the two models that facilitated us to do the social media analyst in this research.

2.1 Social media

The term 'Social media' has been familiar with blogs, public forums, online chat room, Wikipedia, Instagram, YouTube, LinkedIn, Facebook and Twitter for people under the era of Internet technology with high-speed development (Power and Kibell, 2017). Unlike traditional Internet and communication technologies, social media enables to manage online information via creating, sharing and consuming blogs, tweets, Facebook entries, movies, pictures, and so forth in the network (Yates and Paquette, 2011; Kietzmann et al., 2011). Also it supports an unprecedented platform to assure human beings to connect and interact with each other anywhere and anytime (Zafarani, Abbasi, and Liu, 2014).

To define and make sense social media, Kaplan and Haenlein (2010) stated that it was necessary to trace back to where they came from and what they included. An early social networking probably started from the communities with online diary writing, namely "weblog" about 27 years earlier (Kaplan and Haenlein, 2010). What interesting is that the noun 'weblog' is divided into a sentence 'we blog' due to accessibility and popularity of Internet (ibid). Kietzmann et al. (2011) attributed the origins of social media to growing increment bloggers. As the list of the aforementioned applications that we familiar with today, various types of social media are encompassed. Based on the two essences of social media that are media research (social presence, media richness) and social processes (self-presentation, self-disclosure), Kaplan and Haenlein classified social media into six different types, for example, blogs like twitter, Weibo with low media research but high social process, Facebook with moderate social research and high social processes, and collaborative projects like wikipedia with high both of media research and social processes. Furthermore, Kietzmann et al. (2011) proposed seven functional blocks of social media: identity, conversations, sharing, presence, relationships, reputation, and groups, in order to understand and explain the functional characteristics of diverse social media activities. A general definition is herein given "Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content (Kaplan and Haenlein, 2010)" .

2.2 Social media analytics

Wasserman and Faust (1994) suggested that social network analysis is distinct and is based on "an assumption of the importance of relationships among interacting units" and "the unit of analysis in network analysis is not the individual, but an entity consisting of a collection of individuals and the linkages among them". With the development of the information technologies for public entertainment, social media has become ubiquitous and indispensable in daily life for social network (Asur and Huberman, 2010), which means social media analysis is essential for social network analysis. Hence, social media data as an essential actor (Conway and Connor, 2016; Zafarani, Abbasi, and Liu, 2014).

Given that those social media data is produced when people post their emotions, comment their insights, complains to the life, society or municipality and share others posts on social media platforms Yates and Paquette (2011) thought that the social media analysis can load small knowledge chunks such as text messages, images, short videos, blog posts and web links to help the scientific research to investigate trends, issues and urgent needs for people in different disciplines as a significant data source. Yang, Kiang and Shang (2015) explored adverse drug reactions through analyzing drug use experience-related posts that were derived from web forums and discussion boards. Xiang and Gretzel (2010) performed a search engine mining method to analyze online travel information for travelers based on social media websites and the research found that social media sites a crucial information source influenced travellers substantially.

Regarding the application of social media analytics(SMA), Zafeiropoulou, Sarker and Carlsson (2015) have researched the trend of SMA in research and practiced for IS researchers. They found that the main applications of social media analytics are specific application domains, business operations and technology-related issues. The application field contains crime detection, healthcare, politics, travel and tourism and so on.

Because it can show real-time updates of social activities and affect collective human behaviors (Jiang et al., 2015), social media analysis is often applied to help spread instant information for official departments in the real world events, especially nature disaster. Even though nature disaster brought inevitable loss and devastation, social media played an irreplaceable role in conducting the response and rescue activities in the Queensland floods of Australia, which was in the result of the effective, reliable and timely information that was flowed to related communities and business (Ahmed and Sinnappan, 2013). Our research will discuss the application of social media in air pollution detection, which belongs to the emergency management. Therefore, in the following chapter we will discuss more about social media analytics for emergency management.

2.2.1 Social media analytics for emergency management

Emergency management is a specific application of social media analytics. For example, a case study conducted by Eustace and Alam (2012) about Mitchell factory fire in Canberra, Australia in 2011. They found that the use of twitter allows people to understand the emergent situation timely and help official channels and public entities to disseminate information to the mass by social media analysis and an interview of Emergency Services Agency(ESA). In this case, a tight link between online world and offline events in an emergency context is demonstrated through evaluation framework analytics method, which involves 513 tweets from Twitter with hashtags

#CanberraFire released during 5 days of the event lasting (ibid). Geo-tagged used in social media also enable to conduct a situational analysis to understand the place and the degree affected of the emergency according to the messages posted by e-participant (Eustace and Alam, 2012). Another case investigated that social media enhanced the process of communication among various communities during the event of Queensland floods in Australia during 2010 to 2011 (Ahmed and Sinnappan, 2013). The authors showed that social media can be as mainstream media where dissemination of official, instant and accurate information, especially in the crises and disaster situations by conceptual framework and content analysis(ibid). In their case, the data derived from two online national newspaper websites also suggested the role of traditional communication platforms in emergency, such as twitter and Facebook (ibid).

2.2.2 Social media analytics for air pollution management

As mentioned above, air pollution becomes a serious problem in China in recent years. There are two studies that we mainly focus on in this study showed that the Sina Weibo is able to achieve a real-time surveillance of air pollution with lower cost and higher public attention compared to expensive monitor stations and the result illustrated the high correlation between the volume of Weibo and observed AQI (Jiang et al., 2015; Wang, Paul and Dredze, 2015). Wang, Paul and Dredze(2015) investigated the value of Chinese social media for monitoring air quality trends and related public perceptions and response. They collected 93 million messages from Sina Weibo and covered 74 cities in China. The authors defined many topics which are related with air pollution in social media and found quantitatively that message volume in Sina Weibo is indicative of true particle pollution levels. Besides, they researched qualitatively messages which contain rich details including perceptions, behaviors, and self-reported health effects. Jiang, et al.(2015) focused on only one city Beijing rather than the country level. They divided tweets about air pollution in social media into three categories, retweet messages ,mobile app messages and original individual messages. They investigate the relationship between these three kinds of tweets in Sina Weibo and daily AQI. Moreover, the result of content analysis illustrated a strong public awareness on the air quality-related deterioration, protection measurements, and health issues as well (Jiang, et al., 2015) .

2.3 Time Series Analysis model

Time series analysis is a model to analyze time series data in order to extract meaningful statistics and other characteristics of the data. A time series is a series of data points indexed in time order (Zissis et al.,2016). Time series analysis is a theory and method which is based on the system time series data of observations, by curve fitting and parameter estimation to build mathematical model. There are several types of motivation available for time series analysis such as exploratory analysis, curve fitting, function, approximation prediction and forecasting and so on (Shumway and Stoffer, 2011).

Time series analysis has been applied in social media analytics by many researchers before. Asur and Huberman (2010) have used time series analysis to predict real-world outcomes such as box-office revenues for movies. O'Connor, et al.(2010) have applied time series analysis in several surveys about consumer confidence and political opinion over the 2008 to 2009 period, and have found that they correlate to sentiment word frequencies in contemporaneous Twitter messages. Broniatowski, Paul and Dredze (2013) focused on the full 2012-2013 influenza season and

developed influenza infection detection algorithm that automatically distinguishes relevant tweets from other chatter through time series model.

In this study, we will choose one of the most widely used time series analysis approaches autoregressive integrated moving average(ARIMA) model to fit data. ARIMA model is a generalization of an autoregressive moving average, which is fitted to time series data either to better understand the data or to predict future points in the series. ARIMA can be use less parameters to fit the sequence better. In this model, stationarity and the technique of differencing time series are two important concepts (Wei,1994). A stationary time series is whose properties do not depend on the time at which the series is observed. Differencing in statistics is a transformation applied to time-series data in order to transfer it stationary. We will transfer the figures of frequency and observed AQI into two stationary lines to compare their relationship.

In general, there are four steps to build ARIMA model (Bisgaard and Kulahci, 2011). The first step is to obtain observed time series data. The second step is to plot data map to judge whether the series is stationary, if not it need to be transferred into stationary time series. The third step is autocorrelation and cross-correlation analysis. Then we can obtain the model and test the model at last. The process is shown below in Figure 2.1.

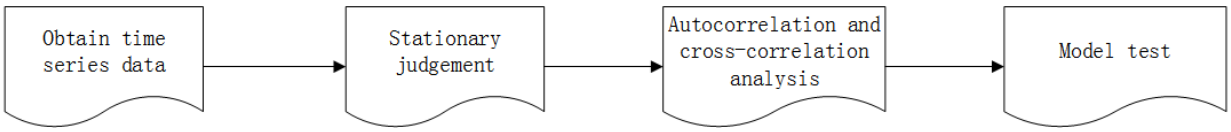

Figure 2.1: Process of building ARIMA model (Bisgaard and Kulahci, 2011)

The purpose to use time series analysis in our research is that we want to research the correlation between the daily frequency of tweets about air pollution and observed AQI of the whole period through curve fitting. Besides, we want to research the salient change of air quality during a specific time. We will compare the figures of observed AQI and social media data when the salient change occurs.

2.4 Co-word network analysis model

In bibliometrics, keywords analysis plays a significant role in retrieving and condensing important information from documents (Zhang et al., 2017). Keywords can not only summary the main topic of the text but also reveal the relationship between the knowledge systems (ibid). Co-word network model, as a content analysis method is based on "the assumption that a scientific field could abstract a set of signal-words to mark literature and reflect its core contents" (Hang et al., 2012) and investigate the relationship between the two topics. That is to say that high frequency words appear simultaneously within the same text can determine a specific topic. More times of the co-occurrence keywords in the same context, more important the themes that they refer to (Hang et al., 2012). In addition, "co-word, like co-citation and co-author analysis, is carried by exploring the co-occurrence and co-absence of keywords" (Zhang et al., 2017). According to Wang, Cheng and Lu (2014), a co-word network analysis process that is shown as Figure 2.2, starts from defining the research questions for collecting related data; it is of the second step to make sure the analysis units, which means define whether the way to analyze is based on the co-

topic or co-keywords; then singling out high frequency keywords or topic-related words to build a co-word matrix; finally getting the result via data analysis.

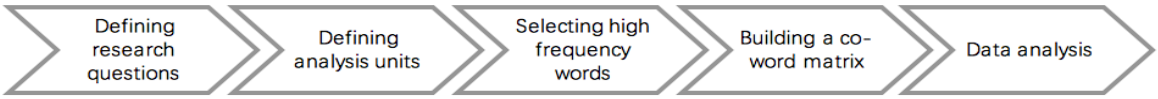

Figure 2.2: Co-word network analysis process (Wang, Cheng and Lu, 2014)

Co-word network has been used in many specific fields. Previous research has shown that using a dynamic co-word network model to map how the knowledge system of China's urbanization research evolves generally (Zhang et. al., 2017). Wang, Cheng and Lu (2014) proposed a community detection algorithm for the evolution analysis based on co-words network model. Co-word network model not only serves for science mapping but also is a prominent approach applied in other knowledge networks. In Hellsten, Dawson and Leydesdorff 's (2010) research, the model was used to compare the development of implicit frames in public debates to determine the changes on artificial sweeteners in New York Times newspaper. To build a word network to uncover latent crisis frames, co-word network model was also performed in the social media text analysis (Van der Meer and Verhoeven, 2013).

In this study, co-work network analysis model is taken as a main method in air pollution-related Weibo messages analysis to further explore the relationship between the evident change of air quality and the level of air pollution. Given that the complication of the text data, we tend to pick up topic-related words from the change day and the worst pollution day with the highest AQI value to consider and compare with each other.

3 Research methods

In this section, we will introduce the research methodology of our research. The first thing we discuss is our research strategy which includes the reason we choose 2013 China Eastern smog as our case and the two Chinese social medias as our data source. Then we will introduce how to collect data and what kind of data we collect in this research. At last, we use correlation analysis and text analysis to investigate the data we collect to answer our research question.

3.1 Research strategy

In order to answer our research question that how a salient change of air quality expressed on social media discussions to reflect the extent of air pollution, we tend to use time series model in the correlation analysis to determine the relationship between the volume of air quality-related messages and the level of air pollution referring to observed AQI within our case. Based on that, we pay attention to the text analysis to explore further whether a salient change of air quality can influence people to reflect a peak of air pollution on social media with a co-word network analysis model. According to our purpose that to investigate whether social media can accurately surveil the level of air pollution based on the social media data, we focus on a serious event--2013 Eastern China smog to analyse. Also, two Chinese microblog platforms are used as data sources to help mitigate human bias and better understand the social media data. Hence, this study contains two parts, correlation analysis and text analysis.

Firstly, our research strategy is to collect the air quality-related messages from the case. And then, we will determine the correlation between the volume of air quality-related messages from Sina Weibo, Tencent Weibo and observed AQI with the help of time series analysis model. The social media data here we use is the daily frequency of tweets about air pollution in the target cities. The first step that verifying the relationship between social media data and observed AQI is the Pearson correlation analysis that is same as Jiang, et al. (2015) and Wang, Paul and Dredze (2015) both used in their researches. Then in order to explore and locate the salient change of air quality we focus on the whole period of the case through time series analysis. In addition, we build a special time series model --ARIMA to fit the figures of social media data and observed AQI to get the inferred value and compare their overall relationships.

According to the time series analysis within a specific period that 2013 China Eastern smog, we pick up the two specific days to analyse the content of Weibo messages, one of which is the salient change of air quality and another one is the peak of air pollution. A text analysis will focus on the two days to build a Co-word network analysis model. Text analysis is the process of deriving high-quality information from text (Miner and Hill, 2012). We will apply this method to focus on Sina and Tencent Weibo so as to investigate further the content of air quality-related messages for the impact of a salient change of air quality on the reflection of level of air pollution through co-word network text analysis model.

3.1.1 Chinese twitter: Sina Weibo and Tencent Weibo for air pollution

Sina Weibo, launched by SINA corporation in August 2009 and Tencent Weibo, launched by Tencent in April 2010, are the two most popular microblog services in China, which contain the most users in Weibo services and play an important role in the social media network (Zheng, 2013) The both Weibo continuously produce huge amounts of data involved daily life of social users, which includes many kinds of attributions like time, users' emotions (Wang et al., 2017). Users can use devices with GPS to edit microblog, such as smart phones. They also can add the location in Weibo that users post.

With the worsening of air quality in China, more and more people pay attention to the problem of air quality than before via Weibo platforms. Meanwhile, the data with location information from Weibo services in some ways can reflect local air quality (Wang et al., 2017). Previous researches have shown that social media can be "sensors" to monitor and detect outdoor air quality within the city preliminarily (Wang et. al, 2017; Jiang et. al, 2015; Wang, Paul and Dredze (2015). For example, Wang, Paul and Dredze (2015) collected and analysed 93 million data for 74 cities in 2013 from Sina Weibo, through keywords filtering and a probabilistic topic model LDA, and demonstrated that there was a high correlation between the volume of air pollution-related messages and particle pollution levels.

Jiang et. al (2015) in their article used Gradient Tree Boosting to analysis social media data that they collected within Beijing during 2013 from Sina Weibo, inferred AQI value by function fitting. Their result also showed that most value of inferred AQI were close to observed AQI from monitoring stations. What's more, they claimed different types of air pollution-related messages from Sina Weibo by social media analytics, which were retweets, mobile app messages, and original individual messages. However, they only focused on one city which is Beijing during 2013, they used the month average data to analyse the correlation, which is too broad and lack of representativeness.

3.1.2 The Case-2013 Eastern China smog

The 2013 Eastern China smog was one of the most serious air pollution in China. It affected almost cities in the east between 2nd and 14th December, 2013, especially some municipalities and provinces such as Tianjin, Hebei, Shanghai, Anhui, Zhejiang and Jiangsu, which was the biggest scale of smog pollution during 2013. This event not only influenced public health, but also disrupted public transportation and daily activities, such as airports, highways, even schools(ibid).6th was the most polluted day in many cities, such as Nanjing, Hangzhou, Shijiazhuang, Shanghai, Tianjin. In Shanghai, the Air Quality Index (AQI) hit 416 on 6th (ibid). Air in Shanghai was reported to have strange taste – astringent and smoky, with an aftertaste of earthy bitterness (Tang,2013). The smoggy weather began to clear up by December 9 (Zhang,2013). Until 14th, the smog gradually cleared and air condition got better.

The reason we want to research this event is that it not only made the government and authority realize that solving the air pollution can not be waited, but also caused the public to emphasize on the discussion about air pollution in social media. During this period, the users of Sina and

Tencent post large quantity of tweets about smog. In addition, the observed AQI of theses six most polluted cities during this period has an evident change.

The six cities are the representative cities in terms of their air pollution level. And also it is easy to collect social media data of these cities. Therefore, we choose Shanghai, Nanjing, Hangzhou, Hefei, Shijiazhuang, Tianjin as our target research cities. So in this research we choose 2013 Eastern China smog as a case study in order to surveil air pollution through social media especially air quality changes saliently. A specific case can help to answer effectively our research question that is whether a salient change of air quality can affect social media to reflect the extent of air pollution. Since the event 2013 Eastern China smog includes a peak of air pollution according to figure 3, which is clear to see the days with salient change of air quality and the worst air quality and compare them. Hence, we tend to focus on data from 2nd December to 14th in order to research what users post and how many messages they released on social media in order to help to answer our research question.

3.2 Data collection

In order to collect the air quality-related messages that users post on social media platforms, we aim to explore the content and frequency of tweets about air pollution during the 2013 Eastern China smog from Sina and Tencent Weibo. Given that our target data, data mining could be a useful way to collect these two kinds of data. Hence, we collected data through a free web crawler named MetaSeeker, which is same as Paul and Dredze (2015) and Jiang et al. (2015). Besides, the Application Program Interface (API) of both Sina and Tencent Weibo is open to researchers, so we can use the free APIs to collect data from these two social media platforms.

The process of air quality-related messages collection includes four steps as Figure 3.1 shows. The first step is to define the data source. As mentioned above, we tend to focus on the event--2013 Eastern China smog, so the time we need is from December second to fourteenth in 2013. And Shanghai is a serious polluted big city in this event, so we can get large quantity of detailed information during this event. All the data is derived from Sina and Tencent Weibo. We put these three limiting conditions into the web crawler to collect specific data. The second step is to set the keywords we will use to search in the Sina and Tencent Weibo, as Jiang, et al. (2015) and Wang, Paul and Dredze (2015) suggested, we used some keywords as our topics to search in social media, such as ‘air pollution’, ‘breathe’, ‘cough’, ‘smog’ and so on. We will search these keywords in both of Sina and Tencent Weibo to see what users post in this event. The third step is to use the web crawler to make a scraping program considering the limiting conditions and keywords to get data from Sina and Tencent Weibo automatically with the help of API. Finally, we need to filter these data which is invalid or not related to our topic.

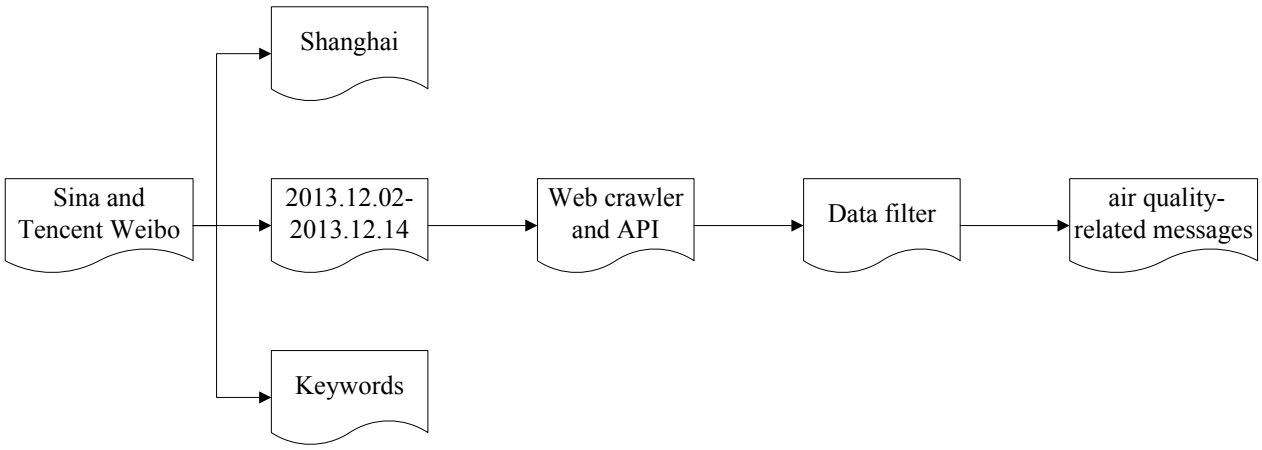

Figure 3.1: Air quality-related messages data collection process

In this research, we collect in total 11000 messages from the both Sina and Tencent Weibo. All the data is related to air pollution and the location is Shanghai from 2nd to 14th December in 2013. Each detailed information contains the date, content, numbers of comment, retweet and praise.

3.3 Data analysis

3.3.1 Pearson correlation and time series analysis

After collecting the data of daily frequency about air pollution in these six cities, we will analyse the correlation between social media data and AQI. The emphasis is to use time series model to fit the daily frequency and observed AQI from 2nd to 14th, December to make nonstationary time series into a stationary time series. And then we will compare these two lines to investigate whether they are related in six target cities. In this part, we will process the data we have collected firstly. Then we will analyse the correlation between daily frequency of tweets and AQI. Next we will test the autocorrelation and cross-correlation before building ARIMA model to fit the figures of frequency and AQI to compare the correlation further and explore the result when there is an evident change of air condition.

a. Data processing

The first step to build time series model is to obtain time series data as mentioned in the theoretical framework. After counting the number of daily message about air pollution in these six cities during the 2013 China eastern smog, we make a statistical table as follows. The frequency of tweets is a time series data.

In addition to the frequency of tweets, as our research strategy has showed, we need to research the correlation between daily number of tweets and real AQI. So we download the AQI of these six cities from the China Meteorological Administration during this period and make a statistical table. Obviously, the AQI of each city is also a time series data. All the data will be illustrated in the Chapter Result

The analytical tool we use is SPSS, which is an easy way to do correlation analysis and time series analysis. So we need import data from Excel into SPSS. Because we have all the data of six cities during this period, we don't need to fill the missing value. So what we need to do to process data is to define time variable. We need to set up time variable before time series analysis. Otherwise, SPSS cannot identify the time data we manually enter. So we define dates by days from 2nd to 14th in Data option bar.

b. Pearson correlation analysis

The second step we need to do is Pearson correlation analysis before building time series model. In order to use social media to surveil air pollution, we need to prove that the tweets on social media can reflect air pollution to some extent. This part will use a quantitative method to analyse the correlation between social media and air pollution. Jiang, et al., (2015) and Wang, Paul and Dredze (2015) have tried to research the correlation between social media and air pollution. As what they did before, we will use daily number of tweets about air pollution to reflect air pollution. And AQI can be used to reflect air condition. But we will go further on this area and focus on the significant event--2013 China eastern smog. We will research the six polluted cities from Sina and Tencent Weibo as mentioned before.

Correlation analysis is used to describe the strength and direction of the linear relationship between two variables(Pallant, 2001). In this part, we will use Pearson correlation analysis to research the relationship between daily number of tweets about air pollution and AQI. This method not only indicates the presence, or absence of correlation between any two variables but also, determines the exact extent, or degree to which they are correlated. Under this method, we can also ascertain the direction of the correlation i.e. whether the correlation between the two variables is positive, or negative. Pearson correlation coefficient is the r value, which describes the strength of variables linear correlation.r=0.10 to 0.29 means small correlation.r=.30 to .49 means medium correlation. r=.50 to 1.0 means large correlation.Sig. (2-tailed) means the significance level. When sig<0.05, it means correlation is significant at the 0.05 level.

c. Autocorrelation and cross-correlation analysis

Because our data is from a specific event, which change over time. the AQI and daily frequency both are time series data. In order to prove their correlation further, we want to research the relationship during this specific period. We will use time series analysis to research these six cities respectively. Here, we will take Shanghai as an example.

Following the theoretical framework, the necessary step before building time series model is autocorrelation and cross-correlation analysis. Cross-correlation diagram is drawn based on the cross-correlation function, which refers stagnant series correlation of two-time series in a specific time. The table and chart below will show the result of cross-correlation analysis. The lag means how many stagnant days between number and AQI. When lag=0, cross correlation is the biggest. It means that there is no lag AQI and daily frequency.

The next step before building time series model is autocorrelation test. The purpose of this test is to make sure that the data can be used to build time series model. Autocorrelation is the correlation of a signal with a delayed copy of itself as a function of delay. Error terms correlated over time are said to be autocorrelated or serially correlated. When error terms are autocorrelated, some issues arise when using ordinary least squares. In theory, the autocorrelation sequences and

time series have the same change cycle. According to the autocorrelation sequences coefficient, we can estimate whether the series has stationarity. Then we can identify the sequence of the model to establishes the corresponding time series model. Based on the standard error values and Box-Ljung Statistic showed in the below table, we can estimate that the series is suitable to build time series model.

d. ARIMA model building

After the data pass autocorrelation and cross-correlation test, we can use ARIMA model build time series model. ARIMA is the abbreviation of autoregressive integrated moving average, which can be applied in stationary series or stationary series through the difference (Tsay, 2005.). ARMA model is the combination of autoregression model and moving average model. ARIMA is the expansion of ARMA model. ARMA model only can be applied in stationary series. However, in most practical applications, the times series is not stationary. So we need to transfer these series into stationary series through different operation. When there is tendency in series, the series can become stationary through specific order difference. When there are a tendency and seasonal cycle, there is a correlation related to the integral multiple of the seasonal cycle in the series. This kind of series need specific order difference and seasonal difference transfer into stationary series (Shumway and Stoffer, 2010).

The general method to build ARIMA model include following four steps as mentioned above. After obtaining the time series data, we de will determine whether the series is stationary based on its scatter diagram, autocorrelation and cross autocorrelation diagram. The next step is to transfer the time series into stationary series until the Sig value of autocorrelation and cross autocorrelation is 0. Then we can build the suitable model according to the identified features. At last, using parametric test to test whether the model make sense according to the mean, variance and correlation function of its subsequence (Wei, 1994).

3.3.2 Text Analysis on Sina Weibo and Tencent Weibo

Aforementioned correlation analysis focuses on the relationship between the volume of information flow, namely air quality-related social media data, and observed AQI. Given that social media radically derives from a subjective public awareness (Jiang et al, 2015; Wang, Paul and Dredze, 2015), we tend to use a text analysis to explore the content of messages from Sina Weibo and Tencent Weibo for better understand. To answer our research question that how a salient change of air quality affects social media to reflect the extent of air pollution during a specific period, a co-word network model is facilitated with the text analysis in this research based on the case. To explore the relationship between the salient change of air quality and the reflection of level of air pollution, we select the peak point of air pollution that has evident differences with the change point to analyse. The difference of the two data is bigger; their relevance should be significant. Hence, in the following, we focus on the comparison with the contents of social media messages that are derived from the moments that an evident change of air quality and a peak of air pollution separately occur through co-word network analysis model.

a. Co-word network analysis model

"The texts of microblog have some special characteristics, such as short and dynamic, which calls for new feature selection methods that are suitable for clustering algorithms to detect the topics from microblog texts", Long et. al (2011) showed that a co-word network feature selection method is efficient for extracting features of microblog texts and performing better microblog topic reflection in terms of conventional text analysis method. This method is applied to investigate the topic-based relationship through analysing a pair of co-occurrence keywords in the same text (Zhang et al., 2017). That is to say, if the both of keywords appear simultaneously to express the topics in the same text, there can be high correlation between the topics that they denote. For instance, there are two keywords that are "air" and "bad", through a co-word network analysis we can verify whether they have inherent relationship to reveal an air pollution-related topic. That will be efficient to explore an outstanding theme from the context in complicated semantic situation, especially Chinese language.

In the text analysis, our objective is to explore and compare the interaction of negative air-quality based on an evident change and an air pollution peak from the perspective of people's. Hence in this case, we tend to build a co-word network for Weibo messages on the day with a salient change and the day with a peak of air pollution respectively, then contrast with each other and see whether the day with a salient change of air quality can lead to a higher public awareness than the real day with the most serious air pollution. Simultaneously, the relationship between the people's subjective perceptions for air pollution and the extent of air pollution can be reached.

b. Defining the analysis units

1) The source of Data

Here we focus on a certain case that is 2013 Eastern China smog with considering one of the most serious polluted city-- Shanghai. Given that our research target is a salient change of air quality, in order to get an evident result for the reflection of the level of air pollution by social media analysis, we will take the highest point of observed AQI to analyze. Therefore, we extract Weibo posts released on the day with a salient change of air quality and the day with a peak of air pollution respectively from Sina Weibo and Tencent Weibo. To be simple, we use **Wc** stands for the whole Weibo messages derived from the day with a salient change of air quality, while **Wp** stands for the whole Weibo messages derived from the day with a peak of air pollution.

2) Data preprocessing

Wang, Paul and Dredze (2015) in their study classified the message in related to air quality firstly, first-hand experience as the second classification criteria to decrease the influence of public awareness. Due to the data that collected initially through first-hand experience and air quality-related posts as one of the advanced filter condition, which means that the 11000 data that derived from Sina and Tencent Weibo directly already are first-hand and air quality-related information, we do not need to think about whether the messages are repeated and topic-relevant.

According to Abdullah et. al, (2017), the messages that are released by government, business, and media, are easy to be re-shared and influence public awareness since people think that that information provided by them is interesting, believable and reliable. And that causes retweeters are subject to the information rather than subjective perceptions of air quality (Abdullah et al.,

2017). Wang, Paul and Dredze (2015) suggested that in the social media analysis of air quality-related text, human behaviours, health concern, description of air quality and requirement about related departments (government and organizations) should be paid attention to. Furthermore, the research also mentioned that if health concern words are in the keywords list, they would be identified key indicators of poor air quality (Wang, Paul and Dredze, 2015).

b. High-frequency topic-related words extraction with TextRank algorithm

Given that our data is derived from Chinese social media platforms and the text made in Chinese is quite complicated, we tend to use a semantic analysis technique to extract topic-related keywords with natural Chinese language processing toolkit SnowNLP in Python programming language. In this research, we aim to compare the reflection of different levels of air pollution from Wc and Wp through topic-related keywords. Wang, Paul and Dredze (2015) suggested that topic-related words can decrease the specificity and improve correlations with other keywords of the whole massages. Hence, topic-related words are taken consideration in this research.

A graph-based ranking algorithm for natural language text analysis---TextRank model is used to extract high-frequency topic-related words in this research. “Graph-based ranking algorithms are essentially a way of deciding the importance of a vertex within a graph, based on global information recursively drawn from the entire graph (Mihalcea and Tarau, 2004)”.

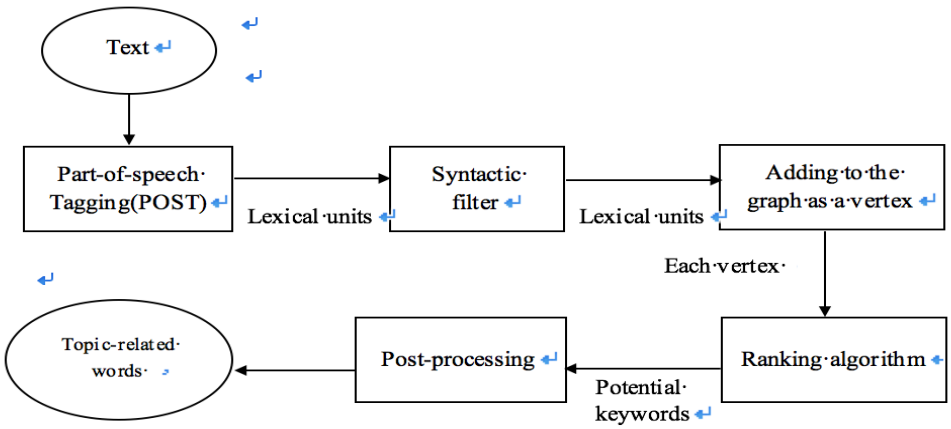

Figure 3.2: TextRank model for high-frequency topic-related words extraction (Mihalcea and Tarau, 2004)

The whole process of extracting topic-related keywords by TextRanking algorithm from the text is completely unsupervised, which is shown as Figure 3.2. According to Mihalcea and Tarau (2004), tokenizing texts from the corpus as a start and then through a semantic analysis method part-of-speech Tagging (POST), the tokenized text is annotated into different lexical units, which aims to enable to pass the syntactic filter. Next, all qualified lexical units are added to the text graph as vertex randomly and an edge links with each other between the co-occurrence lexical units within a window of maximum 2 to 10 words. When the graph is constructed, a ranking algorithm based on Mihalcea and Tarau (2004) works on each vertex continually to ‘vote’ or ‘recommend’ so that can get top T vertices with the high score as potential candidates. That means that when one vertex (Vi) links to another one (Vx), Vi will be casted a vote (the vote number as N_{Vi}) and then Vx will be the next adjacent vertex (as shown Vx=Vx+1), which is

defined simply as Figure 3.3. Thus, “the higher the number of votes that are cast for a vertex, the higher the importance of the vertices (Mihalcea and Tarau, 2004)”, the closer the relationship to the topic of the text as well. Finally, during a post-processing, the T vertices are marked in the text and test whether they fit to an “adjacent rule”, then make up multi-word keywords (ibid). For example, we have the text like “Shanghai smog is quite serious today! (今天上海雾霾相当严重啊！)”. If both Shanghai and smog are selected as potential topic-related keywords by ranking algorithm, they will be collapsed into one single keyword Shanghai smog due to their adjacency.

```
For (i=0; Vi is on the graph; i++)
{NVi =0; Vx=0;
do {NVi= NVi+1;
        if (Vi links to Vx)
        {Vx = Vx+1;}
} while (all Vertex are converged)
}
```

Figure 3.3: A voting process in ranking algorithm (Mihalcea and Tarau, 2004)

d. Building Co-word matrixes

In order to understand further Weibo messages and highlight semantic topics or directions of Wc and Wp to compare clearly, a co-word matrix needs to be built (Yanfang and Nan, 2014). Yanfang and Nan (2014) explained that when two keywords which were essential to express the topic of the specific subject occurred in the same literature, it indicated there must be a particular correlation between the two keywords and semantic topic of the text. Taking an example in our case, if the word ‘breath’ appears in a record, we can not determine that there is relationship between ‘breath’ and air pollution. Instead through counting the frequency of co-appearance of ‘breath’ and ‘air’ or ‘smog’ in the same records, it is convinced that there is specific correlation between health concern and air quality. Higher frequency the both words appeared, higher correlation they have. Then we could infer the extent of air pollution in general.

Hence, based on co-words analysis, we can explore deeply the sentiment regarding to the extent of air pollution from Wc and Wp and will have a clear understanding about them. According to Ding, Chowdhury and Foo (2001), a co-word matrix is created by calculating the co-occurrence frequency of two words with all the other items within the same matrix. Given that the text analysis aims to explore the reflections of the extent of air pollution in the basis of Wc and Wp, it is necessary to set two co-word matrixes to compare with each other. In our case, based on high frequency topic-related words derived by the natural Chinese language processing toolkit SnowNLP in Python programming language, we keep using Python programming language to get co-word matrixes.

e. Visualization of co-word network analysis

To be clear and better understanding the result, we tend to use a technological tool to give a graphical representation of the co-word network. According to Hong, Shin and Kim (2016), NetDraw is "a network visualization tool bundled with UCINET software", which is able to employed to denote the relationship among the keywords and reflect their 'cognitive and attitudinal structure' in a semantic network that indicates what people mainly say. The program NetDraw enable users to save the network and analyse multiple relations ate the same time (Apostolato, I.A., 2013). "The frequency of words and distance among words used for social media messages were examined in the current study," Hong, Shin and Kim (2016) claimed. In the visualization of co-word network, any two co-occurring words as vertices will be linked and the edges stands for potential connections, which also represents the whole relations among syntactic elements from a given text (Hu and Liu, 2004). Also the size of node is related to the edge weights (ibid). Bigger the node is, more important the word is regarding to the topic of the context. Taking account for a contrast between Wc and Wp, a visualized network can centralize the sentiments of Wc and Wp respectively and help to figure out that which day will show higher degree air pollution in order to answer our research question.

3.4 Ethical issues

Fuchs (2017) claimed that social media ethic concerns are not explicit and he illustrated that only 4% social scientists who experienced in analysing social media data discussed about ethical issues in their academic researches. Nevertheless, he mentioned that researchers should keep the research ethics in mind (Fuchs, 2017). "Privacy has been identified as a key ethical concern for population-level social media research (Conway and Connor, 2016)". In this study, we collected air quality-related data from Sina Weibo and Tencent Weibo that are two of the most popular social media platforms. Due to we care about only the frequency and content of the messaged released on social media, so those data attached related information includes date, content and the numbers of comment, retweet and praise without sensitive privacy regarding to personal information such as real name, nick name, age. Hence, those anonymous information will not have harm or benefits crisis on the the social media users and data analysis as well. However, because of the massive data, we did not get acceptance from owners of the social media data before quoting the data (Fuchs, 2017), which could be a privacy issue to be discussed in the following researches.

Jiang et.al (2015) thought that "academic ethical standards" and "public information laws" were not defined well currently. According to the result of the text analysis in our research, it implicates an excessive knowledge presentation on social media. People may over-expressed the extent of air pollution in Wc which was happened in 4th December, 2013. Based on observed AQI shown as Table 1, the degree of air pollution was not such high. An ethical issue that whether or not social media users should take considerations on their public behaviours to avoid to spread rumour or exaggerate the reality needs more discussions.

4. Findings

In this part, we will show the results by the two analysis models on researching the volume and content of Weibo messages respectively.

4.1 Correlation analysis and time series analysis

This part we will discuss the result of the relationship between daily frequency of tweets about air pollution and observed AQI in 2013 Eastern China Smog. We will try to answer the research question how a salient change of air quality expressed on social media discussions to reflect the extent of air pollution according to the quantitative results. The result contains three parts, statistical result, Pearson correlation and time series analysis.

4.1.1 Statistical result

The following table shows the daily frequency of tweets about air pollution of Shanghai, Nanjing, Hangzhou, Hefei, Shijiazhuang, Tianjin during this event from 2nd to 14th, December. Besides, in order to visualize the daily change of the frequency, we draw a line chart to illustrate the trend of each city in the Figure 4.1.

Table 4.1: Daily frequency of tweets of six cities

	Shanghai	Nanjing	Hangzhou	Hefei	Shijiazhuang	Tianjin
12/2	863	137	73	27	55	100
12/3	697	213	77	39	42	62
12/4	740	905	676	372	68	119
12/5	884	843	847	647	101	179
12/6	834	846	834	763	133	210
12/7	849	837	865	883	786	626
12/8	861	852	868	830	848	806
12/9	840	845	878	517	432	556
12/10	862	887	850	342	198	330
12/11	797	365	413	138	101	195
12/12	539	212	222	105	56	129
12/13	462	184	205	74	54	85
12/14	270	137	303	67	62	73

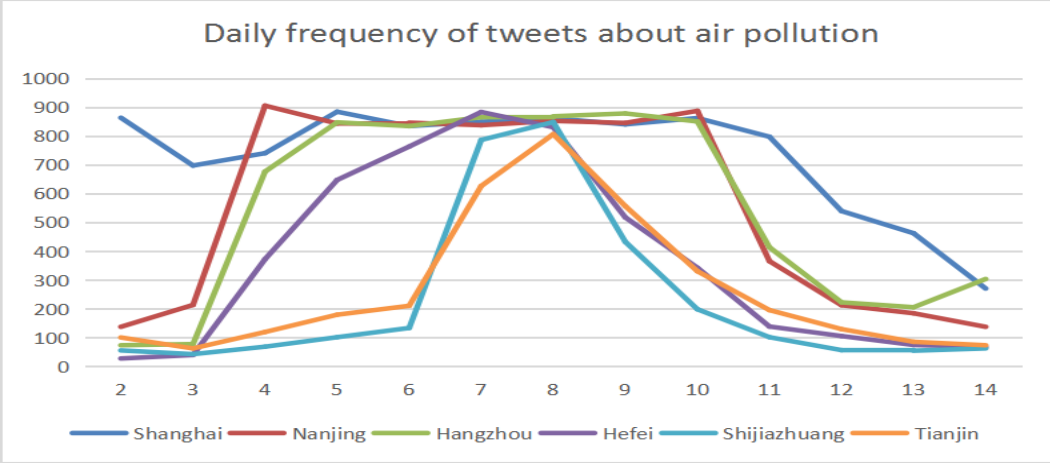

Figure 4.1: Daily frequency of tweets of six cities

The descriptive statistics of data can illustrate some key performance indicators such as average value, maximum value, minimum value and so on. And in order to explore the research question, we will also pay attention to the evident change data. There are all six cities, 13 days from D2 to D14 and two kinds of data, AQI and number. So there are 156 pieces of data all together. Regarding number of tweets, there are total 51882 tweets of six cities during this period. Shanghai ranks NO.1 with the average 730 tweets about air pollution. People in Shanghai post most tweets on D5, which is 883. The number decreased from D10.Nanjing has the similar trend with Shanghai and peak at 905 on D4. Hangzhou's peak is on D9 with 878 tweets. Hefei, Shijiazhuang and Tianjin reach the peak on D7, D8 and D8 respectively with 883,848 and 806. There are large number of tweets about air pollution from D4 to D10 in all cities. Shijiangzhuang has the smallest number of tweets about air pollution in this event. From the figure, we can conclude the frequency of all the cities rise at the beginning of 2013 Eastern China Smog on D3. And from D4 to D9, all the cities reached their peak respectively. Then there is a significant decrease we can see from the picture.

Considering the observed AQI, the following Table 4.2 which is from official statistics--ChinaPM2.5AQI, shows the daily average AQI of Shanghai, Nanjing, Hangzhou, Hefei, Shijiazhuang, Tianjin. And the line chart (see Figure 4.2) visually show the trend of observed AQI in these cities.

Table 4.2:Daily average observed AQI of six cities

	Shanghai	Nanjing	Hangzhou	Hefei	Shijiazhuang	Tianjin
12/2	275	209	147	188	356	284
12/3	181	231	152	211	270	221
12/4	148	377	257	374	290	288
12/5	264	327	278	342	183	125

12/6	416	325	305	363	289	115
12/7	193	365	402	328	400	316
12/8	154	318	393	263	419	294
12/9	216	209	233	242	75	38
12/10	99	100	90	148	134	126
12/11	137	128	126	145	84	69
12/12	119	151	158	175	121	83
12/13	145	163	193	242	135	78
12/14	70	150	183	244	237	118

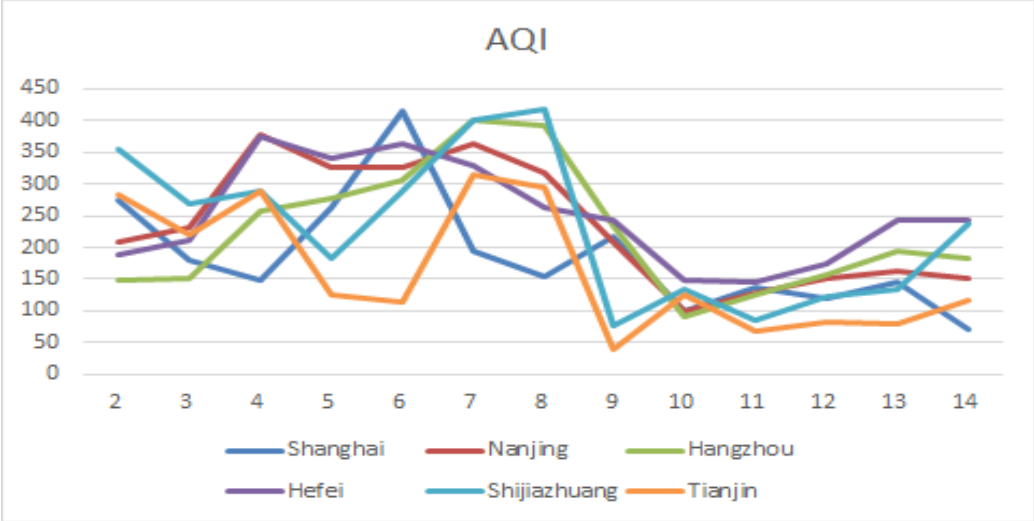

Figure 4.2. A line chart of observed AQI of six cities

Based on the Table 4.2 , Hefei is the most polluted city with the average 251,which means serious contamination according to the AAQS(Ambient Air Quality Standard). Tianjin is the least polluted city the average 166.The highest AQI of Shanghai, Nanjing, Hangzhou, Hefei, Shijiazhuang, Tianjin are 416,377,402,374, 419 and 316 respectively on D6, D4, D7, D4, D8 and D7. So in theses cities, the most serious polluted time is from D4 to D8.

Compared with the Figure 4.1 and Figure 4.2, the observed AQI is more fluctuant. However, the date of peaks of daily frequency and AQI is very closed. In addition, the two figures also have similar trend in terms of the evident decrease after the peak. In spite of the fluctuation of AQI at first, the two figures are relevant to some extent.

4.1.2 Pearson correlation

In terms of correlation analysis, we learn from Jiang, et al. (2015) and Wang, Paul and Dredze(2015) to use Pearson correlation analysis. Based on Table 4.3, the result we find is that during the 2013 China eastern smog event, there this strong correlation between daily frequency of tweets about air pollution and real AQI in Shanghai, Nanjing, Hangzhou and Hefei. And the correlation of Shijiazhuang is medium. So in these five cities, we can conclude that there is a positive correlation between daily number and AQI. When the air quality gets worse, more people will post tweets to comment the air condition or express their opinions about air pollution. But for Tianjin, the correlation is not significant.

Table 4.3: Pearson correlation

	Shanghai	Nanjing	Hangzhou	Hefei	Shijiazhuang	Tianjin
Pearson coefficient	0.525	0.594	0.613	0.650	0.494	0.292
Sig. (2-tailed)	0.046	0.032	0.026	0.016	0.048	0.334

What can we get from these correlation tables is that the sig value of Shanghai, Nanjing, Hangzhou, Hafei, Shijiazhuang are all smaller than 0.05, so the correlation between number and AQI is significant at the 0.05 level. The Pearson correlation coefficients of Shanghai, Nanjing, Hangzhou, Hafei are bigger than 0.5, which means that the strength of correlation of these cities is large. The Pearson correlation coefficient of Shijiazhuang is 0.494. So the correlation is medium. But for Tianjin, the sig value is 0.334 and Pearson correlation coefficient is 0.292. The conclusion is that there is a significant correlation between number and AQI for these five cities except Tianjin. So we can use the result of Pearson correlation analysis to answer part of the research question that social media can reflect the extent of air pollution during a significant polluted time in most areas.

4.1.3Time series analysis

Table 4.4:Cross correlation

Lag	Cross Correlation	Std. Error[a]
-7	-,351	,000
-6	-,096	,000
-5	,118	,000
-4	,387	,000
-3	,396	,000

-2	,452	,000
-1	,326	,000
0	,525	,000
1	,268	,000
2	-,048	,000
3	-,173	,000
4	-,041	,000
5	-,201	,000
6	-,193	,000
7	-,104	,000

Table 4.5:Autocorrelation

Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic
			Sig.[b]
1	,547	,248	,027
2	,254	,238	,050
3	-,030	,226	,010
4	-,150	,215	,004
5	-,157	,203	,013
6	-,198	,189	,024
7	-,230	,175	,003
8	-,172	,160	,007
9	-,132	,143	,017
10	-,047	,124	,000
11	-,046	,101	,000

In addition to Pearson correlation analysis, we use ARIMA to build time series model to fit the figures of frequency and AQI to compare the correlation further and explore the result when there is an evident change of air condition. Before building the ARIMA model, we need to test the autocorrelation and cross-correlation whether the time series data has stationarity after difference. From the result of Table Autocorrelation and Cross-correlation, the Sig of Box-Ljung Statistic is <0.05 during the whole period. The concomitant probability of Box-ljung Statistic is approximate from chi square distribution. As a result, the time series is auto-correlative stationarity time series, which can be built ARIMA model.

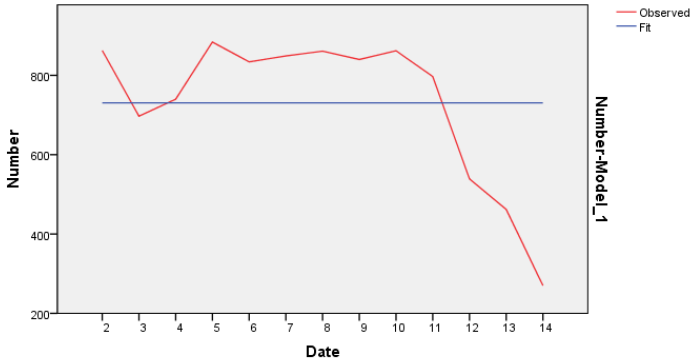

Figure 4.3: ARIMA of frequency in Shanghai

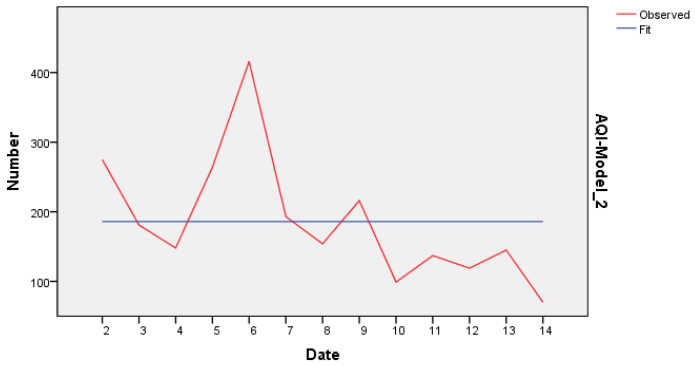

Figure 4.4:ARIMA of AQI in Shanghai

After the autocorrelation and cross-correlation test, we will use SPSS to build ARIMA model for daily frequency and AQI of theses six cities. During the process of creating models in SPSS, we keep the stationary R square, normalized BIC, goodness of fit, parameter estimates in the statistics option, which can be used to judge the goodness of model. In the save option, we keep predicted values and noise residuals as our output result. The two figures above are the output figures of tweets frequency and AQI in Shanghai. The red line is the observed data we collect, while the blue line is the stationary fit line according to the result of ARIMA model. We also get figures of tweets frequency and AQI in other cities.

We try to compare these two fit lines of each city to find their time-dependent relationship. The lines of frequency and AQI have similar shape and trend. But the figure of frequency will rise dramatically compared with observed AQI. Besides, the rise and decrease of frequency will be one day or two days later. And the degree of number decrease is more significant than AQI especially when AQI returns to normal level. The overall fit stationary lines of frequency and observed AQI have proportional relation. After ARIMA model fit, the inferred frequency of tweets about air pollution is about three to four times as inferred AQI in these cities. So there is a correlation between frequency and AQI during this period. As for the evident change of air condition, we focus on the time point when the AQI and frequency change significantly. Regarding the change of AQI, the figure of Shanghai rises significantly from D4 to D6, while other cities increase with fluctuation in the first few cities. As for the frequency of tweets, most cities increase rapidly from D3 except Shanghai. Considering the declining stage, the drop of frequency occurs one day before or after the decline of AQI. In conclusion, through the analysis of the frequency in 2013 China Eastern smog, we can find that as air pollution becomes more serious, the relevance between social media data and AQI is more evident.

The time series analysis focus on the correlation of the whole period during this event. And based on the result, we can draw the conclusion that during the case, the correlation is significant in most cities in 2013 China Eastern smog. Then we research further the relationship between a salient change of air condition and the level of air pollution based on text analysis.

4.2 Co-word network analysis

According to the line chart shown as Figure 4.2, it is clear to determine the day with a salient change of air quality (the lowest point) and a peak of air pollution (the highest point), which are 4th and 6th December, 2013. Thus Wc is the Weibo posts from 4th December and Wp comes from 6th December.

Based on 3.3.2, we randomly selected 5000 messages which is same as Wang, Paul and Dredze (2015) and manually filter out messages that likely coming from government, business, and media. then we got 1330 messages including Wc with 692 and Wp with 638 as data to analyze in the co-word network model.

4.2.1 High-frequency topic-related words extraction

Due to limited time, we deleted low frequency topic-related words and keep 10 top keywords to analyze, which are crucial to express the sentiment of the content. The keywords that derived from Wc and Wp by a TextRank algorithm have been translated from Chinese into English shown as Table 4.5. From the table 4.5, the words ‘雾霾’(smog)occurs over 310 times from the both Weibo message released in the salient change day (Wc) and peak day (Wp). However, the frequency of the topic ‘严重’ (serious) talked about in Wc is approximately three times than Wp, while Wc talks less about the topic ‘爆表’(over-peak) than Wp. What interesting is that the both words can implicate generally the extent of air pollution but the result seems be contradictory. Likewise, ‘污染’ (pollution) occurs in Wc at 25 times but in Wp at 12 times. Health concerned

words are also extracted, such as '呼吸' (breath). In addition, due to fully unsupervised machine learning approach (Mihalcea and Tarau, 2004), the Chinese words '将要' and '趋势' that both express the same meaning were extracted in the list. We thereby summed the frequency of the both words and translated them into 'tendency'. In the day with a salient change of air quality people talk about the topics around 'tendency' 58 time while 74 times in the day with a air pollution peak.

Table 4.5: the frequency of topic-related words occurred in Wc and Wp

	smog	serious	shanghai	people	tendency	air	mask	pollution	breath	overpeak
Wc	310	116	106	64	58	29	26	25	16	4
Wp	311	42	101	95	74	41	49	12	20	26

4.2.2 Co-word matrix and visualized network

Based on the top 10 keywords shown as Table 4.5, we built 10*10 co-word matrixes for Wc and Wp separately shown as Table 4.6 and Table 4.7. The numbers in matrix stand for the frequency of co-occurrence of both corresponding words. For example, in Table 4.6, the words 'shanghai' and 'smog' appear in the same records at 103 times, which means the there are at least 103 records emphasizing smog in Shanghai. In the opposite, zero means the two words are not mentioned directly at the same time in people's posts. That is to say, there could not be relationship between them. Comparing with Table 4.6 that shows a co-word matrix of Wc and Table 4.7 that shows the matrix of Wp, we can see there is a quite high awareness about the topic serious smog in Wc with 112 while 41 in Wp, which implicates that a significant change of air quality enables to make people much sensitive and overstressed. However, in spite of the air quality-related transition increased public awareness significantly, people did not take much more considerations on a tendency of smog until the peak of air pollution happened. Table 4.6 shows that the frequency of the co-occurrence of 'smog' and 'tendency' is 55 while Table 4.7 shows 72 times that is highlighted in Wp. Likewise, people in Wp expressed more concerns about the a trend of air quality, which are 22 and 17 respectively. The result seems to deviate from an expectation that people would take account for how the smog goes and when the air pollution goes better when they get a shock by the salient change of air quality.

In the other hand, the theme about 'over-peak smog' only occurs in Wc at 4 times but in Wp it happens 26 times, which could reflect that Wp expresses a higher extent of air pollution than Wc. Likewise, people talk about 'mask' and 'smog' at the same time for 49 times in Wp while Wc talks about the topic at 25 times, which can make sense that higher degree air pollution can push people to pay more attention to protected measurements. What is more, health-related word 'breath' is mentioned with 'smog' at 14 times in Wc and at 20 times in Wp, from which it also represents a more serious extent of air pollution of Wp than Wc.

Table 4.6: co-word matrix of Wc

Wc	shanghai	overpeak	smog	people	serious	mask	pollution	air	breath	tendency
shanghai	106	2	103	23	43	8	13	14	4	19
overpeak	2	4	4	0	1	0	0	1	0	0
smog	103	4	310	63	112	25	24	29	14	55
people	23	0	63	64	22	6	6	8	3	17
serious	43	1	112	22	116	12	19	13	4	20
mask	8	0	25	6	12	26	5	1	1	8
pollution	13	0	24	6	19	5	25	4	2	7
air	14	1	29	8	13	1	4	29	5	7
breath	4	0	14	3	4	1	2	5	16	6
tendency	19	0	55	17	20	8	7	7	6	58

Table 4.7: co-word matrix of Wp

Wp	shanghai	overpeak	smog	people	serious	mask	pollution	air	breath	tendency
shanghai	101	11	99	26	16	17	8	16	10	22
overpeak	11	26	26	8	1	4	2	6	1	4
smog	99	26	311	92	41	49	12	41	20	72
people	26	8	93	95	16	17	9	21	11	26
serious	16	1	41	16	42	11	4	8	5	14
mask	17	4	49	17	11	49	3	8	5	12
pollution	8	2	12	9	4	3	12	4	2	5
air	16	6	41	21	8	8	4	41	6	18
breath	10	1	20	11	5	5	2	6	20	8
tendency	22	4	72	26	14	12	5	18	8	74

Based on aforementioned 10*10 matrixes, we imported them into NetDraw and got figure 4.5 and figure 4.6. From the graphical co-word networks, generally there are similar structures between Wc and Wp but Wp shows a more intensive relationship among those semantic elements. According to Hong, Shin and Kim (2016), “the closer relationship and the shorter distance they show”. It is quite clear to find that Wp shows that ‘over-peak’ has more links with other keywords than Wc. Simultaneously, the size of node ‘overpeak’ increases obviously, which means the word regarding to the degree of air pollution is involved with more times in Wp. That is to say, Wp reflect that the air pollution is more serious than Wc.

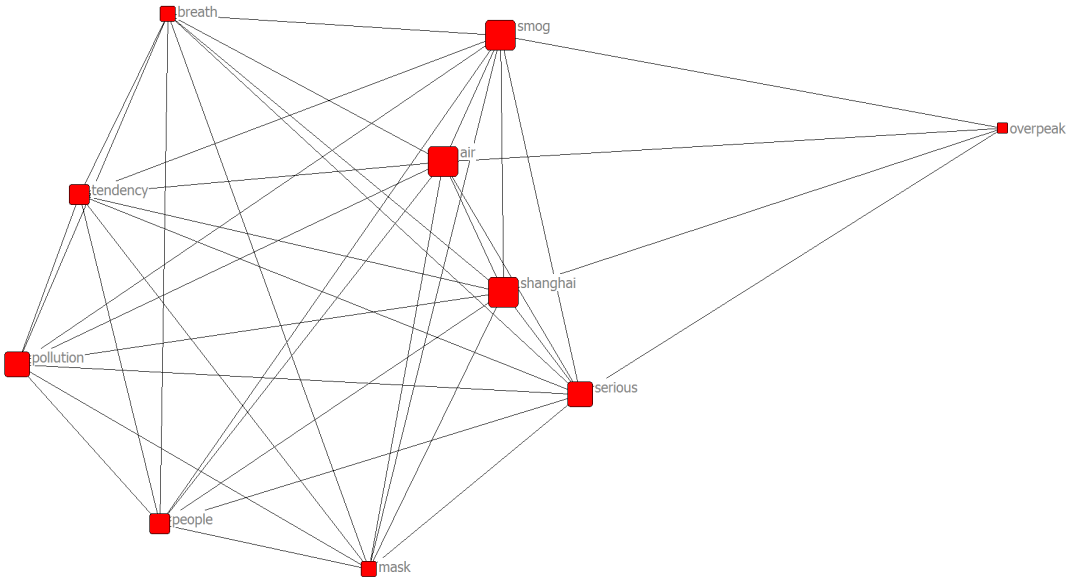

Figure 4.5: Co-word network of Wc

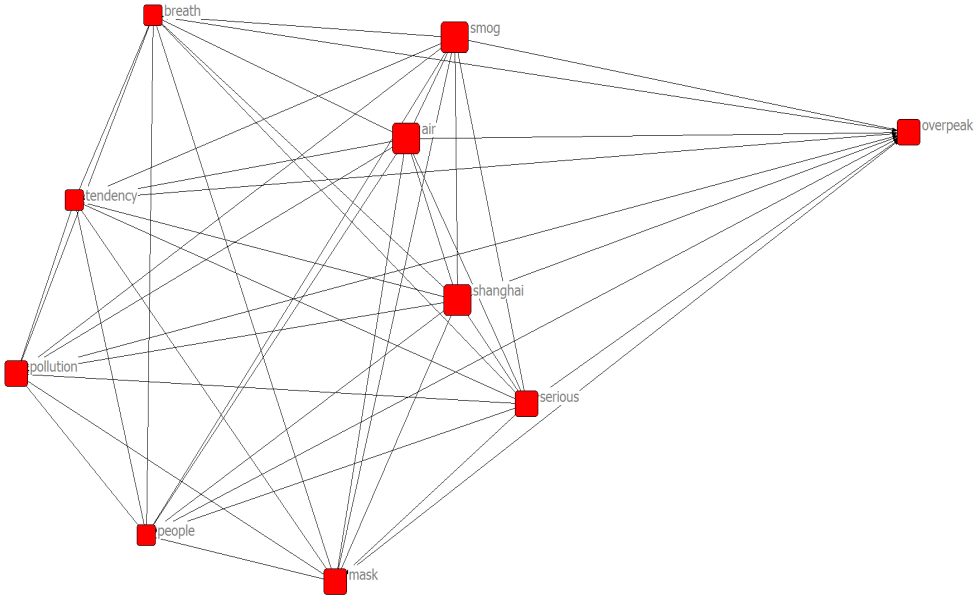

Figure 4.6: Co-word network of Wp

5 Discussion

We will discuss further the results in this section. We elaborated the reflections from correlation analysis and the implications from text analysis. Meanwhile, the result was compared with the two previous researches which is mentioned in 2.2.2. In addition, the potential contributions of this research was also discussed here from the perspective of users, social media, organizations and government. In the end, we highlighted the limitations of the research and talked about possible future directions of the following researches.

5.1 Reflections based on correlation analysis

Based on our research question that how the salient change of air quality can affect social media to reflect the extension of air pollution, we leveraged a time series model to statistically analyze the correlation between the messages released on Sina Weibo, Tencent Weibo and observed AQI value. Our correlation analysis shows that the serious polluted cities highlighted in the event 2013 China eastern smog, such as Shanghai, Nanjing, Hangzhou, Hefei, Shijiazhuang all have significant correlations between social media data and observed AQI. In other hand, our research has further verified previous researches that Jiang, et al., (2015) and Wang, Paul and Dredze (2015) did in terms of a specific event -2013 China eastern smog. The result of Pearson correlation can also preliminarily imply that there is a correlation between the social media data and observed AQI in a general way.

Then, we build time series model so that we can compare the figure of social media data and observed AQI during the whole period in a more visualized way compared with the numerical data. Based on the figures of these six cities, we can conclude that observed AQI is more fluctuant that social media data, which means that the trend of social media data is easier to observe compared with the complex changes of observed AQI. Besides, the change of social media data is more significant no matter rise or decrease. The figures of social media data in these cities are more simple, which are like “mountain” during the whole period. However, considering the whole rise and decrease trend, the observed AQI and social media data illustrate similar figures in spite of the drop of social media data occurs one day before or after the decline of observed AQI. In addition to the visual analysis, we introduce ARIMA model to fit the time series data. The result indicates there is a proportional relation between the social media data and inferred AQI in the target cities, which can prove the correlation between the social media data and observed AQI in a more specific way. In conclusion, through the analysis of the frequency in 2013 China Eastern smog, we can find that as air pollution becomes more serious , the relevance between social media data and AQI is more evident.

Considering the difference between the trends of social media data and observed AQI, we have come up with some assumptions to discuss. For example, when we pay attention to the frequency and observed AQI figures of city Shijiazhuang, we find that the significant rise and decrease of

social media data are both one day later than the trend of observed AQI. We assume that the reason is that the lag of information dissemination in social media. The time lag of information dissemination has been discussed in the social media analytics for Disaster Pre-Warning (Zhang et al, 2014). Since the air pollution becomes serious, people realize and post in the social media, which needs some reaction time. So the change of social media data is later than observed AQI. Furthermore, we try to explain the reason that the change of social media data is more significant than observed AQI. We focus on the case--2013 China Eastern smog itself. This is one of the most serious smog that China has suffered, which cause an extensive attention in society. Therefore, when air pollution becomes serious, the discuss in the social media also becomes intense. Many social media users post their opinions about air pollution not only because of the serious pollution but also affected by others' opinions in social media. So the number of social media data increases rapidly at the beginning of 2013 China Eastern smog.

This part answers the research question that whether the salient change of air quality can affect social media to reflect the extent of air pollution in a quantitative and visualized way, which focuses on the whole period of this event. Throughout the whole period, we observed the salient change of air quality at a more precise time point. At this point, the contents of messages derived from Sina Weibo and Tencent Weibo reflect further the air quality compared with numeric data. The following implication of text analysis will discuss more about the social media how to reflect the air pollution when there is a salient change of air condition, which will focus on the specific time point.

5.2 Comparison with previous researches

Our research is inspired by previous research that try to use social media to detect air pollution. Wang, Paul and Dredze (2015) investigated the value of Chinese social media for monitoring air quality trends and related public perceptions and response based on a country level data. They also use a qualitative method to try to understand the content of air quality-related messages. Another important previous research is that Jiang, et al., (2015) researched the relationship between real air condition and three kinds of tweets, retweet messages, mobile app messages and original individual messages in Beijing.

The similarity of our research with the two researches is part of the research method. Inspired by the aforementioned two researches, we also analyse the correlation between the social media and air condition. We all use frequency of tweets about air pollution in social media as social media data. And we use observed AQI from the China Meteorological Administration to indicate air condition. The primary analysis of correlation we use is Pearson correlation analysis, which has been applied in the two researches to determine the relevance of social media data and observed AQI.

However, there are many differences between our research and the two researches. Firstly, the data source is different. the two researches only pay attention to one of the most popular social media in China—Sina Weibo. Whereas, there are many kinds of social media which have large

quantity of users in China. In case of the one-sidedness of Sina Weibo, we comprehensively consider Sina Weibo and Tencent Weibo as our data source to make the data more convincing.

Secondly, we choose a specific event—2013 China Eastern smog. In this case, we can compare the social media data with observed AQI in a normal level and a serious polluted period. We can observe of the salient change of observed AQI and social media easily. Compared with Jiang, et al.(2015),we select six most polluted cities not only Beijing to make our research more representative. Regarding Wang, Paul and Dredze (2015), our research is more detailed and specific.

Thirdly, we concern more about the salient change of air condition. the two researches focus on the whole year data in 2013. But we focus on the specific period to research the relevance of social media data and AQI when air condition changes significantly to optimize the two researches.

Finally, the most important difference is our research method. the two researches only use Pearson correlation analysis to determine the relevance of social media data and observed AQI, which is not totally convincing. So in order to optimize the two researches, we use visualized time series analysis and introduce ARIMA model to determine the relevance of social media data and inferred AQI further. And aiming at the salient change of air condition, we do text analysis to understand the content of air quality-related messages.

5.3 Implications based on co-word network analysis

Based on Figure 4.7, we can see that the co-word network analysis denotes that Wp reflects more serious air pollution than Wc. It implies that a salient change of air quality does not interrupt social media to reflect a peak of air pollution through a text analysis.

In this case, what an interesting result of co-word network model is that though the observed AQI of Wc is much lower than Wp, the text analysis shows that Wc's topics about 'serious air pollution' are reflected more times than Wp. A possible explanation is that when the air quality has a significant fluctuation, such the salient change may result in people overreacted on social media and attract higher public awareness than a peak of air pollution occurred afterwards. In spite of a salient change of surrounding outdoor air quality can make people overstressed, it is not enough to influence the reflection of an air pollution-related peak from a semantic perspective. The matrixes indicate that people have significantly strong cognitions on the topic about serious smog in Shanghai during the day with airy circumstance changing than the following day with the worst air quality, but there are more words that could implicate higher degree of air pollution highlighted in Wp, such as 'overpeak', 'breath', 'mask'. Furthermore, as for the words with health concerns, higher frequency they occur, stronger the extent of air pollution they imply. If health concerned words are in the keywords list, they would be identified key indicators of poor air quality (Wang, Paul and Dredze, 2015). Physical feeling about air condition is the most authoritative. However, air pollution harms public health seriously. When people express the topic regarding physical health like '呼吸' (breath), it has already implied a serious pollution in air quality. Also, given that the semantics of nature Chinese language is complicate, the word 'overpeak' that is translated from Chinese word '爆表', namely observed AQI value is super high

and over a standard of serious pollution measurement. Based on above, the over discussion on social media caused by the sudden change of air quality can be taken as an alert for air pollution outbreak.

Furthermore, the knowledge related to the extent of air pollution is presented excessively in Wc. From the table 1 we can see an evident fluctuation occurs from 4th to 5th December, 2013 in Shanghai in terms of observed AQI and a very high value happens in 6th. However, the frequency of talking about 'serious smog' in Wc is much more than Wp, which implies that people fail to manage accurately knowledge presentation on social media. When people see an obvious change of air quality in someday, they would be over-stressed and post complaints about air condition on social media. The comment may exaggerate the degree of the air pollution. After all, users do not need to take responsibility to their own words on the social media so that it is pointless to think whether or not the words are fit to the reality before posting. In the other hand, when the circumstance condition goes worse slightly till a vertex, people gradually get used to the situation and will not put any more emphasises on the air condition. Thus it makes sense why the theme 'serious smog' in Wp shown less than Wc. Nevertheless, people's excessive expression perhaps reflects dissatisfactions on government, of which related departments did bad in controlling and preventing air pollution.

Besides, we also discover that the keyword 'tendency' has more links with other keywords in the Wp, which implicates that when a real serious pollution occurs people become more concerned about the tendency of air quality. Although people exaggerated air condition in the day of significant change, they did not emphasize on how the air pollution was going until air pollution peaked. The reason why people talk about air quality-related trends may be because they care about them and consider following behaviours as well, which just likes we think about tomorrow's weather to pick up suitable clothes. In contrast, if we do not plan to go outside or intend to dress up myself, we will not take tomorrow's weather into consideration. Therefore, the result may imply harmful behaviors, from public health's perspectives, that people could not realize the extent of outdoor air pollution and protect themselves timely. That is also to say, when it comes to outdoor air pollution, health issues are not highly valued by individual.

5.4 Contributions

This study aims to explore whether the public air quality-related perceptions on social media can be sensors of air pollution in China. In this part, we will discuss the contributions of our research to the users, social media, organizations and government.

5.4.1 Users

From the aspect of users, if the social media such as Sina Weibo can be used to monitor air condition based on the tweets that users post, users will also benefit themselves. This kind of surveillance of air pollution will be closer to these social media users. Because more and more people would like to take part in the hot issues' discussion in social media (Agichtein et al ,2008).For those people who will browse the social media frequently, they are more likely to notice the air condition in social media rather than the special websites or applications. In another word, if the social media can automatically reflect the air condition and push it to the users, they

will promptly prevent the air pollution. For instance, users can wear dust masks and choose public transport if they notice today's air pollution warning in social media.

In addition ,if users know what they post in social media can be used to surveil air pollution, it will inspire their enthusiasm to discuss more about air pollution. Users will pay more attention to the daily air condition, which will increase the public awareness of air pollution. Users will post more opinions and suggestions about how to control air pollution in social media. These opinions and suggestions can help related organizations control and manage air pollution. The surveillance of air pollution in social media can make more and more people involve in the air pollution management.

5.4.2 Social media

With regard to the contribution to the social media, our research can broaden the application of social media. In the theoretical framework, we have shown the application of social media analytics in emergency management, such as Queensland floods in Australia during 2010 to 2011 (Ahmed and Sinnappan,2013) and Mitchell factory fire in Canberra (Eustace and Alam ,2012). Our research further argued that the application of social media in air pollution management based on the previous research.

For the social media itself, the surveillance of air pollution is a public welfare program, which will benefit all sectors of society. As a result, the implementation will improve the corporate image of social media enterprises and attract more users, which will bring business profits for social media enterprises. Besides, as the discussion about air pollution increase, social media can collect more and more users' data. According to the data, social media can do user analysis further to know the users' preference so that they can improve themselves to offer more service and function that users want.

5.4.3 Organizations

In terms of organizations related to the air protection, the the most direct contribution of our research is that we propose a new air pollution surveillance way to them. These Non-Governmental Organizations such as NACAA (National Association of Clean Air Agencies) and Greenpeace East Asia can be inspired by our research. The organizations can detect the air condition in a cost-effective way base on the social media data. In this way, they will pay more attention to the topics and opinions about air pollution in social media. As a result, they can have a better understanding of public opinions so that they can implement their measures to control and manage air pollution more effectively and get more support of public.

Besides, if NGOs want to monitor the air pollution according to the social media data, the organizations need to work with social media. They not only get the social media data about air pollution, but also popularize their measures to control air pollution in social media. This kind of collaboration will cause NGOs get more attention and feedback of public, which can promote they do better in air pollution management.

5.4.4 Government

From the aspect of government, this study could give fresh knowledge for whether social media enable to be 'sensors' to surveille air pollution. The following researches will contribute to save government finance instead of building costly air quality monitor stations in small towns. In many big cities, there are air quality monitor stations, while in small cities and towns the monitor stations are lacking because of the financial limitation. But the access to the social media is low-cost. So air quality can be detected according to the people's subjective perceptions with the help social media in these areas.

On the other hand, it is beneficial for government to learn and measure public awareness and response based on social media (Wang, Paul and Dredze, 2015), which cannot be captured by monitor stations or other physical equipment. Especially health perceptions, government could get to know about public health issues from social media and based on that can take corresponding measurements to prevent a turbulent situation due to serious air pollution outbreak. In another word, government can take immediate measures to control air pollution from the view of public in social media. For example, if people in a specific area post their strong displeasure with air pollution, government can take measures such as shutting down a factory or traffic control to relieve air pollution and placate public. So this can help government understand public opinions and act promptly.

5.5 Limitations

5.5.1Data collection

We use web crawlers to collect data. And all the data is related to the keywords based on the previous research. But the keywords are not enough. There are still some extra keywords related to air pollution, which will cause part data missing.

5.5.2Region

Our purpose is to explore whether the social media can be 'sensors' to monitor air pollution in China. And the event-2013 Eastern China smog is as a case study, thereby original data is collected from Chinese social media platforms in terms of our research question. Given that Shanghai is the biggest city within the event and gets the most serious air pollution as well, data analysis was focused on Shanghai for a salient result. The rest cities like Nanjing, Hangzhou, Shijiazhuang, Tianjin and Hefei are as a group to contrast. Thus there could be a limitation to generalize the knowledge because of a limited region to be researched. Besides, the population of each city is different. The population of Shanghai is much larger than Shijiazhuang, which will affect the accuracy of result.

5.5.3 Interpretation

We have to admit that the semantics of nature Chinese language usually is ambiguity. In spite of using a co-word network approach to analyze text content in Chinese to strengthen the correlation with the topic air pollution within the case, it is difficult to prevent Chinese polysemy, symptoms

and web words, which could affect the result of text analysis especially when extracting keywords. Additionally, the interpretation from Chinese to English is also limited. For instance, we translated '将要', '趋势', '要' that were filtered in the list of topic-related keywords via TextRank algorithm into a keyword 'tendency' and summed up their frequency, because in Chinese all of them express the same meaning 'tendency'. However, as mentioned in 1.4, we did not take linguistic syntax into account in this research.

5.6 Suggestions for future researches

Based on aforementioned limitations in this study, we would like to give some suggestions for future researches as below.

a. Discuss more about the difference between the social media data and AQI. As mentioned in 5.1 and 5.2, although the trend of AQI and social media is similar, there are still some differences about the research results. The further research can figure out the reasons of these differences so that the surveillance can be more accurate.

b. Verify small towns. To cover more regions without monitor stations to achieve air pollution surveillance, this research and previous researches (Jiang et al. 2015; Wang, Paul and Dredze, 2015) explore the monetization capability of social media to sniff air pollution by leveraging the data from microblog in big cities. However, it is still far away from that goal. Thus in the following researches, small towns need to be verified especially specific areas under serious air pollution frequently.

c. Gain more data from other social media platforms. Wechat (微信) and some forum like Tianya (天涯), Zhihu (知乎) are currently popular social media platforms in China. There are not many literatures to investigate by using social media to monitor air pollution in China up to now. Hence taking advantage of more social media data sources could help build a sound theory and provide wider insights.

d. Generalizing the models. The models we selected in this research that are time series analysis model and co-word network model could be applied into the researches that studying other emergency managements or exploring the both of content and volume of social media messages by using social media analyst. Besides, for following researches, there should be better models to facilitate and construct the social media analyst.

6. Conclusion

Overall, the discussions on social media about air quality reflect the level of air pollution when the air quality changes saliently, according to the analysis of frequency and content of air quality-related Weibo messages. Furthermore, as air pollution becomes more serious, the relevance between social media data and AQI is more evident.

This study aims to investigate whether social media can be qualified 'sensors' to monitor air pollution in China. Based on the purpose, we examine an assumption that mentioned in the end of a previous study (Jiang et. al, 2015), which is taken as our research question that how does a salient change of air quality expressed on social media discussions reflect the extent of air pollution? To answer the question, we focus on an event 2013 Eastern China smog as the case study and take advantage of Sina Weibo and Tencent Weibo, that are two of the most popular microblogs in China, as our social media data source. Through a correlation analysis with a time series model, we determine a positive correlation between the volume of individual messages and observed AQI value. Also, it illustrates the interaction between the salient change of air quality and the reflection of level of air pollution. Then we go deeper to explore the research question, a co-word network model is leveraged to analyse the content of the messages posted on Sina and Tencent Weibo. Via analysing the two different significantly critical points based on observed AQI, which are the day with a salient change and the day with a peak of air pollution, it shows that an evident change of air quality cannot affect social media to reflect a peak of air pollution but the change still attracts quite high attentions. Thus, the impact of the salient change can reflect that the surveillance of air pollution outbreak is timely and the occurrence of the peak reflects the most serious moment.

To determine social media 'sensors' to surveil air pollution precisely, this study verifies a potential error that a phenomenon that air circumstance changes obviously would lead to a temporal virtual peak which could misunderstand the extent of air pollution. We compare with prior literatures that researched in the same field and used two different approaches to analyze a case. Besides, we provide extra insights on public health perceptions and online users' behaviors based on our text analysis. Hence, this study could give more fresh insights to researcher who are willing to develop further on this field with social media and outdoor air pollution surveillance.

Appendix 1 -- Content data

https://docs.google.com/spreadsheets/d/1soOt32F233uVMuY78QIWQytk5Ik5ZgnkmE_NpTISjZI/edit

Appendix 2-- Frequency data

https://docs.google.com/spreadsheets/d/1OD-ejuZq_ykSG7pLLkCA4Gwc_zhHaiRPFbDn1nyeeyQ/edit

References

Abdullah, N.A., Nishioka, D., Tanaka, Y. and Murayama, Y., 2017, January. Why I Retweet? Exploring User's Perspective on Decision-Making of Information Spreading during Disasters. *In Proceedings of the 50th Hawaii International Conference on System Sciences*.

Agichtein, E., Castillo, C., Donato, D., Gionis, A. and Mishne, G., 2008, February. Finding high-quality content in social media. *In Proceedings of the 2008 international conference on web search and data mining* (pp. 183-194). ACM.

Ahmed, A. and Sinnappan, S., 2013. The role of social media during Queensland floods: An empirical investigation on the existence of multiple communities of practice (MCoPs). *Pacific Asia Journal of the Association for Information Systems*, 5(2).

Apostolato, I.A., 2013. An overview of Software Applications for Social Network Analysis. *International Review of Social Research*, 3(3), pp.71-77.

Asur, S. and Huberman, B.A., 2010, August. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492-499). IEEE.

Bhattacherjee, A., 2012. *Social science research: principles, methods, and practices*.pp108-281

Bisgaard, S. and Kulahci, M., 2011. *Time series analysis and forecasting by example.* John Wiley & Sons.

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

Conway, M. and O'Connor, D., 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9, pp.77-82.

Ding, Y., Chowdhury, G.G. and Foo, S., 2001. Bibliometric cartography of information retrieval research by using co-word analysis. *Information processing & management,* 37(6), pp.817-842.

Eustace, J. and Alam, S., 2012. Tweeting from the danger zone: *The use of Twitter by emergency agency during Mitchell factory fire in Canberra*. MCIS 2012 Proceedings, pp.1-9.

Fuchs, C., 2017. From digital positivism and administrative big data analytics towards critical digital and social media research! *European Journal of Communication*, p.0267323116682804.

Gil de Zúñiga, H., 2015. Citizenship, social media, and big Data: Current and future research in the social sciences. *Social Science Computer Review*, p.0894439315619589.

Hellsten, I., Dawson, J. and Leydesdorff, L., 2010. Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science*, 19(5), pp.590-608.

Hevner, A. and Chatterjee, S., 2010. *Design science research in information systems* (pp. 9-22). Springer US, p109-140.

Hu, M. and Liu, B., 2004, August. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

Jiang, W., Wang, Y., Tsou, M. H., & Fu, X. (2015). Using social media to detect outdoor air pollution and monitor air quality index (AQI): a geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter). *PloS one,* 10(10), e0141185.

Kaplan, A.M. and Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, 53(1), pp.59-68.

Kietzmann, J.H., Hermkens, K., McCarthy, I.P. and Silvestre, B.S., 2011. Social media? Get serious! *Understanding the functional building blocks of social media. Business horizons,* 54(3), pp.241-251.

Long, R., Wang, H., Chen, Y., Jin, O. and Yu, Y., 2011, September. Towards effective event detection, tracking and summarization on microblog data. In *International Conference on Web-Age Information Management* (pp. 652-663). Springer Berlin Heidelberg.

Mihalcea, R. and Tarau, P., 2004, July. *TextRank: Bringing order into texts*. Association for Computational Linguistics.

Miner, G., Elder IV, J. and Hill, T., 2012. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A., 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129), pp.1-2.

Pallant, J., 2001. SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS for Windows (Versions 10 and 11): SPSS Student Version 11.0 for Windows. Open University Press.

Power, R. and Kibell, J., 2017, January. The Social Media Intelligence Analyst for Emergency Management. *In Proceedings of the 50th Hawaii International Conference on System Science*s.

Recker, J., 2012. *Scientific research in information systems: a beginner's guide*. Springer Science & Business Media.

Seaton, A., Godden, D., MacNee, W., & Donaldson, K. (1995). Particulate air pollution and acute health effects. *The lancet*, 345(8943), 176-178.

Shumway, R.H. and Stoffer, D.S., 2010. *Time series analysis and its applications: with R examples*. Springer Science & Business Media.

Shumway, R.H. and Stoffer, D.S., 2011. Time series regression and exploratory data analysis. *Time Series Analysis and Its Applications*, pp.47-82.

Tang, D. and Hoshiko, E., 2013. *Shanghai smog hits extremely hazardous level*s. The World Post.

Tsay, R.S., 2005. *Analysis of financial time series* (Vol. 543). John Wiley & Sons.

Van der Meer, T.G. and Verhoeven, P., 2013. Public framing organizational crisis situations: Social media versus news media. *Public Relations Review*, 39(3), pp.229-231.

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

Wang, X., Cheng, Q. and Lu, W., 2014. Analyzing evolution of research topics with NEViewer: a new method based on dynamic co-word networks. *Scientometrics,* 101(2), pp.1253-1271.

Wang, Y.D, Jin, T., Jiang, W., Wang, P., Fu, X.K., 2017. Modelling urban air quality trend surface using social media data. *Geomatics and Information Science of Wuhan University*, 42(1), 14-20

Wasserman, S. and Faust, K., 1994. Networks, relations and structure. *Social network analysis: Methods and applications* (Vol. 8), 4-10. Cambridge university press.

Watson, H.J., 2014. Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, 34(1), pp.1247-1268.

Wei, W.W.S., 1994. *Time series analysis*. Reading: Addison-Wesley publ.

Xiang, Z. and Gretzel, U., 2010. Role of social media in online travel information search. *Tourism management*, 31(2), pp.179-188.

Yanfang, L. and Nan, D., 2014, September. A study of Chinese culture consumption based on co-words analysis and social network. *In Advanced Research and Technology in Industry Applications (WARTIA)*, 2014 IEEE Workshop on (pp. 551-554). IEEE.

Yang, M., Kiang, M. and Shang, W., 2015. Filtering big data from social media–Building an early warning system for adverse drug reactions. *Journal of biomedical informatics*, 54, pp.230-240.

Yates, D. and Paquette, S., 2011. Emergency knowledge management and social media technologies: *A case study of the 2010 Haitian earthquake. International journal of information management*, 31(1), pp.6-13.

Zafarani, R., Abbasi, M.A. and Liu, H, 2014. Social Media Mining: An Introduction. Available at:<https://www-cambridge-org.ludwig.lub.lu.se/core/books/social-media-mining/D1050E05E04E082BF2AED1E8D4BB8656 >

Zafeiropoulou, S., Sarker, S. and Carlsson, S.A., 2015. What's Trending in Social Media Analytics Area? A Retrospective. *In 21st Americas Conference on Information Systems, AMCIS 2015.* Americas Conference on Information Systems (AMCIS).

Zissis, D., Xidias, E.K. and Lekkas, D., 2016. *Real-time vessel behavior prediction. Evolving Systems*, 7(1), pp.29-40.

Zhang, J., Xie, J., Hou, W., Tu, X., Xu, J., Song, F., Wang, Z. and Lu, Z., 2012. Mapping the knowledge structure of research on patient adherence: knowledge domain visualization based co-word analysis and social network analysis. *PloS one*, 7(4), p.e34497.

Zhang, N., Huang, H., Su, B., Zhao, J. and Zhang, B., 2014. Information dissemination analysis of different media towards the application for disaster pre-warning. *PloS one*, 9(5), p.e98649.

Zhang, Q.R., Li, Y., Liu, J.S., Chen, Y.D. and Chai, L.H., 2017. A dynamic co-word network-related approach on the evolution of China's urbanization research. *Scientometrics*, pp.1-20.

Zhang Y. *"Severe winter smog shrouds eastern China. Global Times*. 2013.

阚海东, 2012. 《环境空气质量标准》(GB3095-2012) 细颗粒物 (PM2. 5) 标准值解读. 中华预防医学杂志, (2012 年 05), pp.396-398.