Detecting Public Sentiment Over PM2.5 Pollution Hazards through analysis of Chinese Microblog

Yongzhong Sha
Lanzhou University, Gansu, China
Email: shayzh@lzu.edu.cn

Jinsong Yan
Lanzhou University, Gansu, China
Email: yanjs11@lzu.edu.cn

Guoray Cai
Penn State University, University Park, PA, USA
Email: cai@ist.psu.edu

ABSTRACT

Decision-making in crisis management can benefit from routine monitoring of the (social) media to discover the mass opinion on highly sensitive crisis events. We present an experiment that analyzes Chinese microblog data (extracted from Weibo.cn) to measure sentiment strength and its change in relation to the recent PM 2.5 air pollution events. The data were analyzed using SentiStrength algorithm together with a special sentiment words dictionary tailored and refined for Chinese language. The results of time series analysis on detected sentiment strength showed that less than one percent of the posts are strong-positive or strong negative. Weekly sentiment strength measures show symmetric changes in positive and negative strength, but overall trend moved towards more positive opinions. Special attention was given to sharp bursts of sentiment strength that coincide temporally with the occurrence of extreme social events. These findings suggest that sentiment strength analysis may generate useful alert and awareness of pending extreme social events.

Keywords

Sentiment analysis, air pollution, PM2.5, social media analysis, crisis, public opinion.

INTRODUCTION

The combination of high population density and rapid industrialization in China has inevitably led to the increase in emissions. These emissions have exacerbated the air pollution problems in large and medium size cities, resulting in visibility reduction and health concerns (He *et al.* 2001). Figure 1 is snapshot of the air quality measurement in Beijing, where PM2.5 pollution reached a high mark of 354 (hazardous).

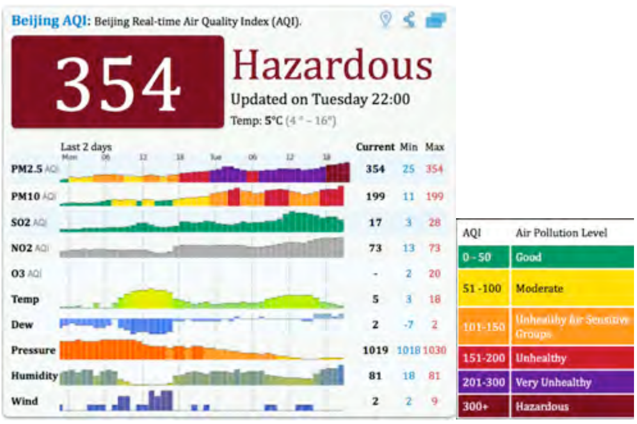


Figure 1 Air quality index
(Source: http://aqicn.org. Accessed on Nov. 12, 2013)

The nature of PM2.5-related hazards has been well documented (Yu 2010, Yuan *et al.* 2012). Particulate matter (PM) is made up of extremely small particles and liquid droplets containing acids, organic chemicals, metals, and soil or dust particles(Grahame and Schlesinger 2012). The seriousness of the PM2.5 pollution problems has been warned by both scientific evidence and public health concerns. The World Health Organization estimates that particulate matter (PM) air pollution contributes to approximately 800,000 premature deaths each year, ranking it the 13th leading cause of mortality worldwide. However, significant controversies remain to be resolved on whether there is a direct causal relationship between PM pollution and certain diseases and mortality, due to the lack o solid clinical evidence (Weinhold 2012). Such controversies create confusion on what actions ought to be taken. In China, PM2.5 related air pollution has been on the spotlight of public attention in microblog sites, which puts the government and regulatory bodies under pressure to implement administrative and policy measures to reverse the deteriorating trend. However, citizens were frustrated by the lack of actions and expressed strong impatience and mistrust towards their government. Such expressions through microblogs spread very fast and swayed the broad opinions to a dangerous mode. Early detection of extreme patterns of public opinions expressed through microblogs can provide timely alert to pending social crises and draw the attention of both the public and the decision-makers to act timely.

This paper presents a method and a case study for detecting the fluctuation of public opinions over PM2.5 pollution events through the coupling of sentiment strength analysis and time series analysis. Microblog data on PM2.5 related discussions was collected from Weibo.cn, covering 65 weeks of public comments. By running a sentiment analysis on the dataset, we were able to explore sentiment changes in very fine granularity. Sharp changes on sentiment polarizations and strength are detected and used to investigate the correlations between public sentiments and important social events.

METHODOLOGY

We monitor the changes of public opinions towards PM2.5 pollutions by detecting the sentiment polarity and strength observed in the content of Chinese microblog posts and associated comments. Next, we will describe our methodology for selecting and compiling data, steps for preprocessing the data, and the methods used in sentiment strength analysis.

Data Source and Characteristics. Data used in this study was extracted from Weibo.cn, which is the largest microblog system in China. Weibo.cn blogs are a hybrid of Twitter and Facebook. Microblog entries in Weibo.cn are downloaded using a custom-made Python program, based on predetermined topics related to PM2.5 events. The data covers 65 weeks of PM2.5 related microblog content during October 10, 2011 and January 6, 2013. Individual blogs are aggregated into weekly collections for the ease of analyzing sentiment in a small time scale. Selected microblog entries are limited to posts made by registered users. Each blog can either be a main post, a comment, or a forward of another post, with clear indication of the creation time and the author. A main post is likely to state some facts or observations, while a comment or repost is more likely to express sentiment. Total number of posts gathered is around 700 thousands. Due to the limitations of Weibo's microblog download utility, we were only able retrieve full text documents for the top 100 pages of main posts and all of their comments. This is a limitation of the source data, but we found that the top 100 pages are adequate for understanding sentiment strength.

Data Preprocessing. Before we can pass the data to SentiStrength algorithm, we cleaned up those data problems (such as duplicates and noise) to improve the integrity of the data. This involves the following three steps of processing: (1) we apply a language code conversion tool cconv (code.google.com/p/cconv) to convert any text in traditional Chinese to simplified Chinese; (2) we use VIM text editor (www.vim.org) to remove texts that have no real meaning or no effect on the sentiment of posts. VIM provides the best functions for text replacement, duplicate removal, and reordering of records; (3) we apply store posts into MySQL database and run database queries to delete duplicates and dummy records. After the above data cleaning process, we ended up a data set of 175 thousands items of posts.

Sentiment Strength Analysis. We apply sentiment analysis to characterize the fluctuation of public opinions over PM2.5 pollution events. Sentiment analysis is concerned with opinions that express or imply positive or negative sentiments. We follow Liu (2012) to define an *opinion* is a quadruple, (g, s, h, t), where g is the opinion (or sentiment) target, s is the sentiment about the target, h is the opinion holder, and t is the time when the opinion was expressed. Two approaches have been pursued in the literature for measuring opinion polarity. *Machine learning methods* train a feature-based classifier with a human-coded corpus, and then use the trained classifier to categorize sentiment polarity based on the extracted features from text(Ni *et al.* 2007, Thelwall *et al.* 2010). On the other hand, *lexical approach* uses only direct indicators (detected from sentiment-bearing words or phrases) to determine the opinion polarity. In the context of this paper, we choose to use lexicon-based methods to analyze PM2.5 related microblog data. Specifically, we adopted the SentiStrength version 2.2 algorithm developed by Thelwall and colleagues (Thelwall *et al.* 2010, Thelwall *et al.* 2012). SentiStrength has a lexicon-based classifier that uses an emotion word dictionary and rules to detect sentiment strength in short informal text. It is capable of analyzing both the polarity and the strength of the sentiment expressed.

Sentiment analysis for Chinese social media data presents a number of unique challenges (Zhang *et al.* 2009):(1) Chinese does not segment words by spaces in sentences; (2) the use of various adverbs in Chinese can lead to a higher degree of subtlety and ambiguity in sentences; (3) Chinese also shows a larger variety of word sense and syntactic dependency in sentences than English.

We constructed a Chinese Emotion Lookup Table (ELT) based on the Chinese Sentiment Word List developed by Tsinghua University (*www.datatang.com*). The list has 23419 Chinese sentiment words, each entry of the list is a <word, SentiValue> pair, where SentiValue may be an integer value from -5(extremely negative) to +5 (extremely positive). This Chinese ELT is further modified by incorporating a Negating Word List (NWT) compiled by Zhang(2008) and a Chinese BoosterWordList (with 212 booster words). The resulted sentiment word list is further refined by modifying word sentiment strength with human-coded emotion strength values. First, we randomly choose 1000 items from the Chinese ELT and ask three trained human coders to conduct independent sentiment strength judgment on perceived sentiment strength. The three human estimates are averaged to derive the modified sentiment strength for a particular word. Finally, we run the "optimize emotion dictionary weights" procedure of SentiStrength to created a new Sentiment Word Dictionary that is optimized for our data collection of microblogs. To confirm inter-coder consistency, we calculated Krippendorff's α measure of inter-coder reliability for the three independent coding results. The α value for positive sentiment coding is 0.58, and the α value for negative sentiment coding is 0.634. These values are comparable to those reported by Thelwall (2010) and are positive enough for us to conclude that *there is modest consistency in coder's judgment on sentiment strength, despite the differences*.

To compute sentiment polarity and strength of the PM2.5 related blog data, we performed the following tasks. *First*, in order to detect sentiment words the microblog posts, we must first segment the documents into words. Word segmentation is quite difficult a task in Chinese language due to lack of explicit separation and boundaries between words. We used the Chinese word segmentation tool called ICTCLAS (ictclas.org) developed by the Institute of Computing, Chinese Academy of Science. *Second*, we call the SentiStrength algorithm (using the optimized sentiment Strength Lookup Table) to calculate the sentiment measures of each microblog post. The positive sentiment of a post is represented as $S^+_i = S^+(P_i)$ and is assigned a value equal to the most positive of its sentence emotions. The negative sentiment of the post P_i is represented as $S^-_i = S^-(P_i)$ and is assigned a value as the most negative of its sentence emotions. The output of the SentiStrength analysis is a table describing each post with the associated positive and negative sentiment, (P_i, S^+_i, S^-_i).

For the purpose of validating the outcome, we applied the "10-fold cross-validation assessment" procedure (of SentiStrength tool), which allows us to compare the correlation and accuracy between human judged sentiment strength with machine predicted sentiment strength in both the positive and negative dimensions.

Corr+	Corr-	Acc+	Acc-
0.4214	0.4646	80.80%	76.40%

Table 1 Accuracy assessment of computed sentiment strength

Corr+ and Corr- are measures of positive / negative sentiment strength correlation between human-coded and computed values. Acc+ and Acc- are measures of accuracy of the predicted sentiment values (compared with human-coded values) in positive and negative dimensions. As shown in Table 1, both Corr+ and Corr- are greater than 0.3, indicating that *there is a positive correlation between computed and human-coded sentiment strength values*. Furthermore, Acc+ and Acc- are above 75%, indicating a high degree of accuracy. These measures established our confidence on the applicability of SentiStrength algorithm to Chinese blog texts.

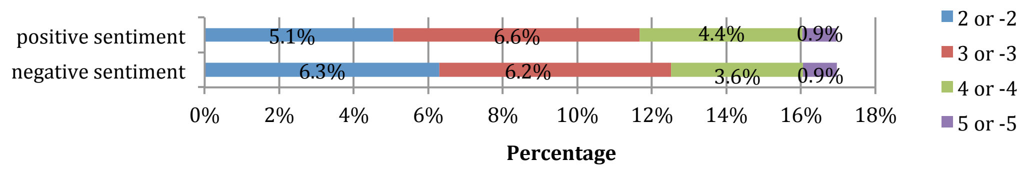


Figure 2. The proportion of sentiment strengths in data set

THE SENTIMENT ANALYSIS RESULTS: FINDINGS AND DISCUSSION

Overall Sentiment Distribution. Our data shows that 68.5% of the microblog posts have no obvious sentiment orientation. Among those blogs that do have clear sentiment orientation (about 31.5%), we categorized them by their strength (slight:±2; moderate:±3; high:±4; and extrme:±5) and plotted the distribution of sentiment strength on both positive and negative dimensions.

As Figure 2 shows, (1) approximately 17% of the posts expressed positive sentiment (i.e., S^+ >= 2); (2) approximately 17% of the posts expressed negative sentiment (i.e., S^-_i =< -2); (3) there are only 0.9% of the posts that expressed extreme positive or extreme negative sentiment. We noticed that there are 2.4% of the posts that expressed both positive and negative sentiment (i.e., S^+ >= 2 and S^- <= -2), which explains why the percentages of all categories do not add up to exactly 100%.

Temporal Dynamics of Sentiment polarization. We explored how sentiment polarizations change over time. For each week (of the 65-week period), we calculate two numbers: (1) ϕ^+ the percentage of posts each week that expressed positive sentiment (i.e., $S^+_i > 1$), and (2) ϕ^- the percentage of posts each week that expressed negative sentiment (i.e., $S^-_i < -1$). By plotting ϕ^+ and ϕ^- in Figure 3, we can observe the trend that the percentage of negative posts was decreasing while the percentage of positive posts was increasing. This negative correlation between the percentage of positive and percentage of negative posts was confirmed by a Pearson correlation analysis, where we found that the Pearson Correlation of the two curves is - 0. 313 under $\propto$ = 0.01. This overall trend of sentiment change during our observation period is interesting, since it suggests that public opinion was moving towards more positive direction. Discovering such trends of polarization is useful for understanding the effect of certain social processes or policy measures

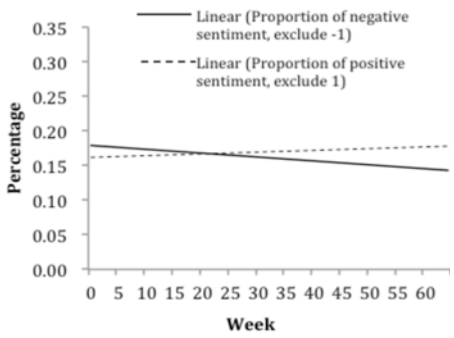
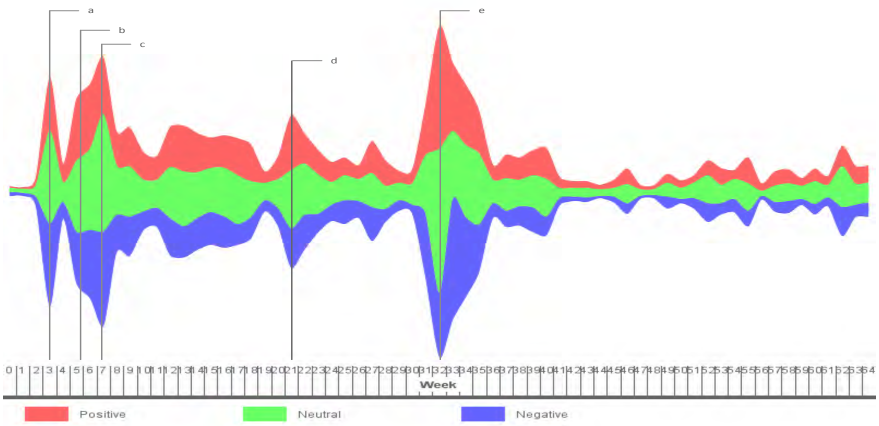


Figure 3. Trends on sentiment polarity

Temporal Dynamics of Sentiment Strength. Changes in sentiment strength provide more details on the dynamics of the public opinions. We calculated three sentiment strength measures (*weighted sum*, *weighted sum ratio*, and *weighted average* sentiment) for each week on three classes of sentiment: *positive* (S^+ between [2, 5]), *neutral* (S^+=1 and S^-= -1), and *negative* (S^- between [-5, -2]). Due to space limits, we present only the results of weighted sum sentiment strength measures. First, we divide all posts into 65 weeks. Each week is assigned four sentiment volume measures: Σ^{++} (Positive) , Σ^{+}(Positive neutral), Σ^{-} (Negative neutral), Σ^{--}(negative). For each week *k*, we calculate weighted sum of sentiment by the following algorithm:

Set the initial value of Σ^{++}, Σ^{+}, Σ^{-}, Σ^{--} to zero
For each post P_{jk} in week# k,
  If P_{jk} has positive sentiment ($S^+(P_{jk})$ between [2, 5]), then set $\Sigma^{++} \mathrel{+}= S^+(P_{jk})$
  If P_{jk} has positive neutral sentiment ($S^+(P_{jk})$ =1), then set $\Sigma^{+} \mathrel{+}= 1$
  If P_{jk} has a negative neutral sentiment ($S^-(P_{jk})$ = -1), then change the value of Σ^{-} by -1: $\Sigma^{-} \mathrel{+}= -1$
  If P_{jk} has negative sentiment ($S^-(P_{jk})$ between [-5, -2]), then set $\Sigma^{--} \mathrel{+}= S^-(P_{jk})$
End for loop



a. GOC: *the U.S. Embassy was questioned for the hype; normalized PM2.5 measurement standards*; Public: *decreased public trust in government; contamination anxiety.*
b. GOC: *Ambient Air Quality Standards (second draft) was released; One person in Nanjing was punished for the data release of PM2.5.* Public: *Nanshan Zhong(an opinion leader) appealed publishing of PM2.5 values.*
c. GOC: *Beijing Environmental Protection Bureau refused to disclose PM2.5 data.* Fact: *Burst limit of PM2.5.*
d. GOC: *Ministry of Public Health of China began to research specific protective measures for PM2.5.*
e. Public: *Shiqi Pan(an opinion leader) initiated the Governance PM2.5*

Figure 4. Temporal patterns of sentiment volumes

We generated a ThemeRiver visualization of the changes of sentiment volume over 65 weeks using the algorithm of Havre *et al* (2002). Figure 4 uses three "currents" to represent positive sentiment volume Σ^{++}, neutral sentiment volume (Σ^{+} and Σ^{-} combined), and negative sentiment volume Σ^{--}. We can see that the total volumes of the three classes of sentiment (positive, neutral, and negative) showed similar patterns of changes over the 65 weeks period. This suggests that *increase in positive sentiment is likely to be accompanied by similar increase in negative sentiment*, and *vise* versa. We also observed five sharp changes in sentiment (see legends in Figure 4) and asked how those changes coincide temporally with the occurrence of other critical events. Our findings revealed that the unusual burst of sentiment volume in week 3 was associated with "the release of PM2.5 air quality measure by the US embassy of Beijing", which caused deep concerns and a drop in public trust towards government's capacity to treat air pollution. Similarly the burst changes in week 5, 7, 21, and 32 were all matched by top events in the news of policy changes. This is an interesting observation for understanding how sentiment fluctuation is associated with the announcement of new regulations or government actions. This observation suggests that the happening of such events is likely to receive broad attention and emotional responses.

CONCLUSION

Managing social crises requires capabilities for detecting and monitoring public opinions and understanding the patterns and implications of extreme changes. In this study, we demonstrated how the analysis of sentiment strength in microblog data reveals both expected and unusual changes in public sentiments related to critical air pollution events. We found that sentiment strength fluctuates in zig-zag course and shows symmetric changes in positive and negative strength, although less than one percent of the posts are strong-positive or strong negative. Longitudinal sentiment change exhibits a trend towards more positive opinions (Figure 3), suggesting some effects of treatment in controlling PM2.5 pollution. The most interest result of this study is that those sharp bursts of sentiment strength were found to coincide temporally with the occurrence of social actions such as new environmental policies and new pollution-sensing capabilities. We are working on developing various visualization techniques that can present the detected sentiment changes to crisis managers (or decision-makers) for expert judgment. We do want to point out that the capability of detecting early sign of pending social crises can creative positive social effect (in the sense of alerting the government to act on redesigning environmental policies) or negative social effect (in the sense that it may empower the governments to exercise unfair control over public outrage). Cautions must be taken to ensure that the capability we develop is made equally available to the public and the decision-makers.

ACKNOWLEDGMENTS

This work is partially supported by a grant from the National Science Foundation of China under the award# NSFC-71373108, and a grant from the National Science Foundation (US) under the award IIS-1211059.

REFERENCES

1. Grahame, T. & Schlesinger, R., 2012. Is ambient pm2.5 sulfate harmful? *Environmental health perspectives,* 120 (12), A454.
2. Havre, S., Hetzler, E., Whitney, P. & Nowell, L., 2002. Themeriver: Visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on,* 8 (1), 9-20.
3. He, K., Yang, F., Ma, Y., Zhang, Q., Yao, X., Chan, C.K., Cadle, S., Chan, T. & Mulawa, P., 2001. The characteristics of pm2.5 in beijing, china. *Atmospheric Environment,* 35 (29), 4959-4970.
4. Liu, B., 2012. *Sentiment analysis and opinion mining* San Rafael, Calif. (1537 Fourth Street, San Rafael, CA 94901 USA): Morgan & Claypool Publishers.
5. Ni, X., Xue, G.-R., Ling, X., Yu, Y. & Yang, Q., 2007. Exploring in the weblog space by detecting informative and affective articles. *Proceedings of the 16th international conference on World Wide Web.* Banff, Alberta, Canada: ACM, 281-290.
6. Thelwall, M., Buckley, K. & Paltoglou, G., 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology,* 63 (1), 163-173.
7. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. & Kappas, A., 2010. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology,* 61 (12), 2544-2558.
8. Weinhold, B., 2012. Epa proposes tighter particulate air pollution standards. *Environmental health perspectives,* 120 (9), A348-A349.
9. Yu, T.-Y., 2010. Characterization of ambient pm2.5 concentrations. *Atmospheric Environment,* 44 (24), 2902-2912.
10. Yuan, Y., Liu, S., Castro, R. & Pan, X., 2012. Pm2.5 monitoring and mitigation in the cities of china. *Environmental science & technology,* 46 (7), 3627.
11. Zhang, C., Zeng, D., Li, J., Wang, F.-Y. & Zuo, W., 2009. Sentiment analysis of chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology,* 60 (12), 2474-2487.
12. Zhang, J., 2008. *Algorithms for recognizing semantic orientation of chinese texts*. Beijing University of Post and Telecommunication Engineering.