Analyzing Public Opinions on Particulate Matter Contents using Comments of News Article

Hyun-Seok Hwang[1]

Abstract--- *PM (Particulate Matter) is a mixture of solid particles and liquid droplets found in the air. PM10 and PM2.5 mean PM whose diameter is smaller than 10 micrometers and 2.5 micrometers or less respectively. PM contains microscopic solids or liquid droplets that can be inhaled and can cause lung disease or invade the blood or brain. Since 2013, South Korea has published official PM statistics and informed the public about how to act.*

With the development of artificial intelligence, there is an increasing number of studies that analyze texts' emotions and public opinions embedded in texts. Many people comment on news related to fine dust, and the comments contain words that can be used to understand what news readers think about fine dust using opinion mining.

This study aims to analyze people's perception by analyzing comments expressed on PM news. After reviewing related researches, we will present three research questions and provide answers to them through an empirical analysis using web crawling, basic sentiment analysis, and multiple linear regression. We will also provide a concluding remarks with research limitation and future research directions.

Keywords--- *Particulate Matter, News Article, Text Mining, Sentiment Analysis, Linear Regression*

I. Introduction

Many countries have problems such as global warming, water pollution, and fine plastics as the economy develops. One of these problems is air pollution due to PM. PM is a mixture of solid particles and liquid droplets originated from dust, dirt, or smoke. Although PM occurs naturally, it is caused by artificial factors such as fossil fuel power plants, automobile smoke, and factory emissions. Many researchers have found that PM is a causative agent of respiratory and vascular diseases. However, the only way for individuals to cope with PM is through manual means such as wearing a mask, frequent ventilation, and running an air purifier. People who are interested in PM often report their opinions about news articles through comments when news related to PM is published in the media.

In this study we aim to analyze the contents of comments on PM news articles. Searching for popular news from Korea's largest news portal, we collected articles containing the word ‘PM’ in the headlines and comments posted on them. We counted the number of PM article comments that matched the words in the four dictionaries. We examined whether the polarity of comments on PM articles changed over time. We also analyzed which categories of words the commenters recommend/unrecommend.

II. LITERATURE REVIEW

Big data has been actively researched recently, since the volume and the variety of research data increase. Big data research includes not only engineering but also marketing, communication and healthcare. One of the big data fields is sentiment analysis, which analyzes the emotions of people in texts to identify their dispositions.

[1] Hyun-Seok Hwang, Professor of Dept. of Business Administration, College of Business, Hallym University, Chun-Cheon, Gang-Won Do, Republic of Korea
E-mail: hshwang@hallym.ac.kr

2.1 Big data analysis

Big data, gathered from various sources, can be used for read customers' hidden minds [1]. IIoT (Industrial Internet of Things) was proposed as a tool for monitoring indoor air quality [2]. Li, & Mariappan suggested a predictive IoT sensor algorithm to predict future behaviors, outcomes, and trends [3]. A web-based real-time personal PM 2.5 exposure monitoring system was proposed to get big data of personal PM 2.5 exposure efficiently [4]. Another research built a spatio-temporally weighted model to improve the estimation of PM 2.5 exposure by integrating annual data from multiple sources [5].

2.2 Sentiment Analysis

Many studies have conducted sentiment analysis in the areas including healthcare [6], trip reviews [7], and financial text [8]. Hyun aimed to investigate political polarities of news producers / news organizations by analyzing editorials from Korean newspapers [9]. Park & Oh performed an emotional analysis using the environmental color of the lobby in a general hospital [10].

In order to understand the emotions of people embedded in the comments of PM articles, we derived the following research questions:

RQ1: Will the content of comments on PM news articles vary over time?

RQ2: What kinds of lexicon words do the commenters recommend or not recommend?

III. METHODS AND EMPIRICAL STUDY

3.1 Data Gathering

Data was collected for six years from March 2013 to February 2019. March 2013 is the time when the Korean government officially announced the level of PM. The data was collected by web crawling, and among the 30 news published daily in Naver's ranking news - the biggest news portal site in Korea, we collected articles containing the word PM in the headline. Along with the articles, sections of the news, comments on the collected news articles, posting dates, the number of recommendations and non-recommendations for each comment were also collected. The open source software R and R packages, RSelenium and rvest, were used for crawling and about a million comments were collected.

3.2 Preprocessing Procedure

Before analyzing the data, we preprocessed the crawled data. Figure 1 shows a preprocessing procedure. After eliminating the stop words and replacing slangs and emotional icons with usual words, we extracted words related to PM. We choose 20000 comments for classifying the comments and create 4 lexicons. In previous studies related to sentiment analysis, the polarity of a given text is analyzed using a lexicon divided into two polarities, positive and negative. So we just need to create and analyze two lexicons [11][12].

Personal lexicon denotes the words related to personal respondences to PM and healthcare issues while Internal lexicon focuses on the words related to domestic political issues. Some commenters used a number of political words unrelated to fine dust to express their feelings for the Korean government.

Sample comments are categorized into the following dimensions: Personal dimension, Internal dimension, External dimension, and Sentimental dimension. Unlike the general emotional word composition, we constructed the emotional dimension with only negative words. We counted the number of PM article comments that matched the words in the four lexicons. Table 1 shows the four lexicon categories

External lexicon is composed of words mainly used by commenters who consider the cause of fine dust as an adjacent country. The words in this lexicon mainly included dissatisfaction, responsibility and emotional dissatisfaction with neighboring countries. Emotional lexicon includes the words that express subjective and negative feelings about the risk of fine dust.

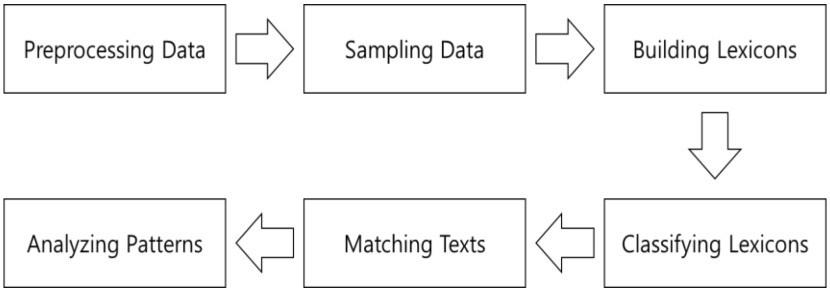


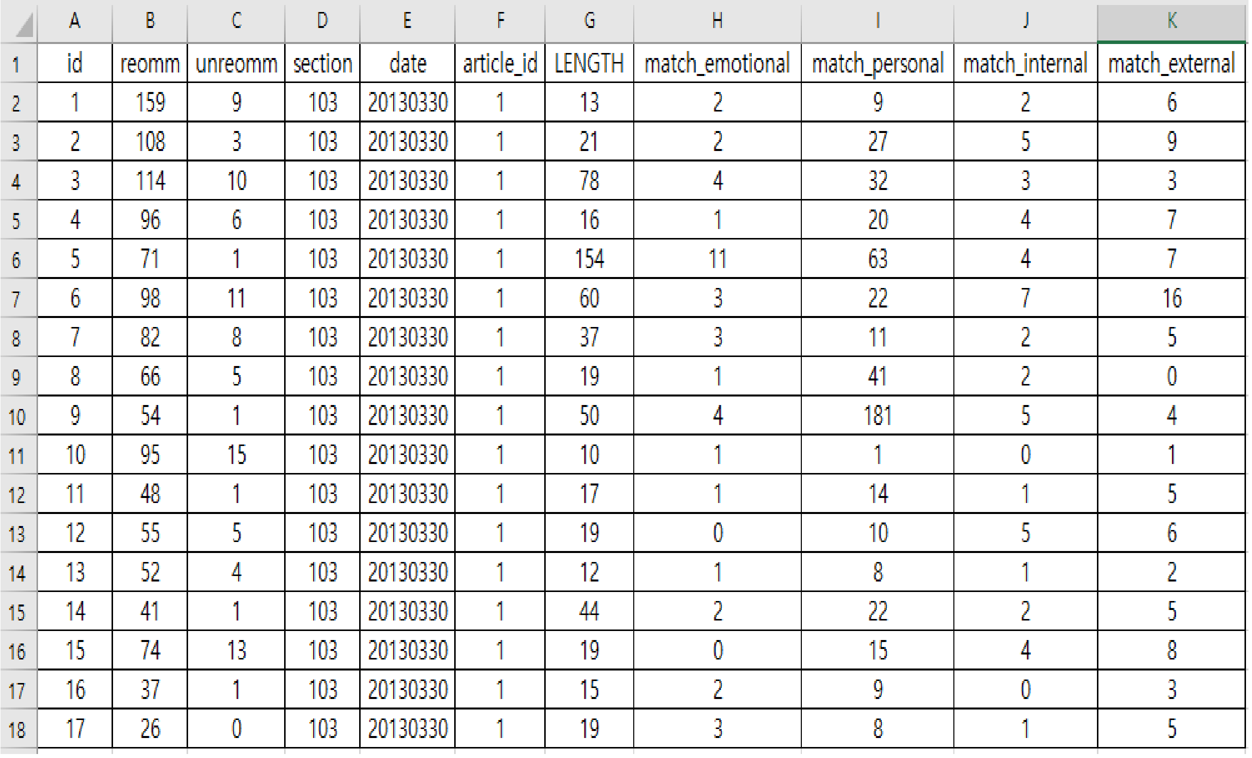
Figure 1. Preprocessing procedure

As shown in Table 1, the Words in the personal lexicon make up 58.8% of the whole words. People often refer to words that describe personal responses (masks, air purifiers, ventilation, etc.) rather than expecting a national response to fine dust or making a rational response.

Table 1. Four lexicon categories

Category (# of words)	Focus	Keywords
Personal (4428)	Personal responses	Personal respondence and Healthcare
Internal (1135)	Domestic issues including political opinions, emotions	Government, Domestic pollutants
External (1213)	Thoughts and Feelings for Adjacent PM Discharge Countries	Adjacent Country
Emotional (751)	Subjective and negative emotional expressions	Negative emotions

Counting the number of matches of one million comments and four Lexicon words generated the data shown in Figure 2.



	A	B	C	D	E	F	G	H	I	J	K
1	id	reomm	unreomm	section	date	article_id	LENGTH	match_emotional	match_personal	match_internal	match_external
2	1	159	9	103	20130330	1	13	2	9	2	6
3	2	108	3	103	20130330	1	21	2	27	5	9
4	3	114	10	103	20130330	1	78	4	32	3	3
5	4	96	6	103	20130330	1	16	1	20	4	7
6	5	71	1	103	20130330	1	154	11	63	4	7
7	6	98	11	103	20130330	1	60	3	22	7	16
8	7	82	8	103	20130330	1	37	3	11	2	5
9	8	66	5	103	20130330	1	19	1	41	2	0
10	9	54	1	103	20130330	1	50	4	181	5	4
11	10	95	15	103	20130330	1	10	1	1	0	1
12	11	48	1	103	20130330	1	17	1	14	1	5
13	12	55	5	103	20130330	1	19	0	10	5	6
14	13	52	4	103	20130330	1	12	1	8	1	2
15	14	41	1	103	20130330	1	44	2	22	2	5
16	15	74	13	103	20130330	1	19	0	15	4	8
17	16	37	1	103	20130330	1	15	2	9	0	3
18	17	26	0	103	20130330	1	19	3	8	1	5

Figure 2. Preprocessed data

IV. RESULTS

In this section, we illustrate the answer to the two research questions.

4.1 Changes in the contents of comments on PM news articles (RQ1)

As shown in Figure 3, it can be seen that the number of comments is increasing steeply with periodic changes. PM concentration, however, is decreasing slowly. These results indicate that even though PM is decreasing, comments on the PM news are increasing due to an increase in awareness of the seriousness of PM and interest in health. It is understood that interest in fine dust varies according to seasons, and this change of interest is reflected in the number of comments.

In the winter, when fossil fuels are increasing and the concentration of fine dust is absolutely high, the number of comments is increasing periodically compared to other seasons. However, the increase in the number of fine dust comments in April and May, when the concentration of fine dust decreases, may be different from common sense. These results can be interpreted that the interest in the fine dust also increased due to the increase in outdoor activities in spring after winter. Overall, the increase in comments on fine dust can be said to be when the concentration of fine dust is absolutely high or subjectively affected by fine dust.

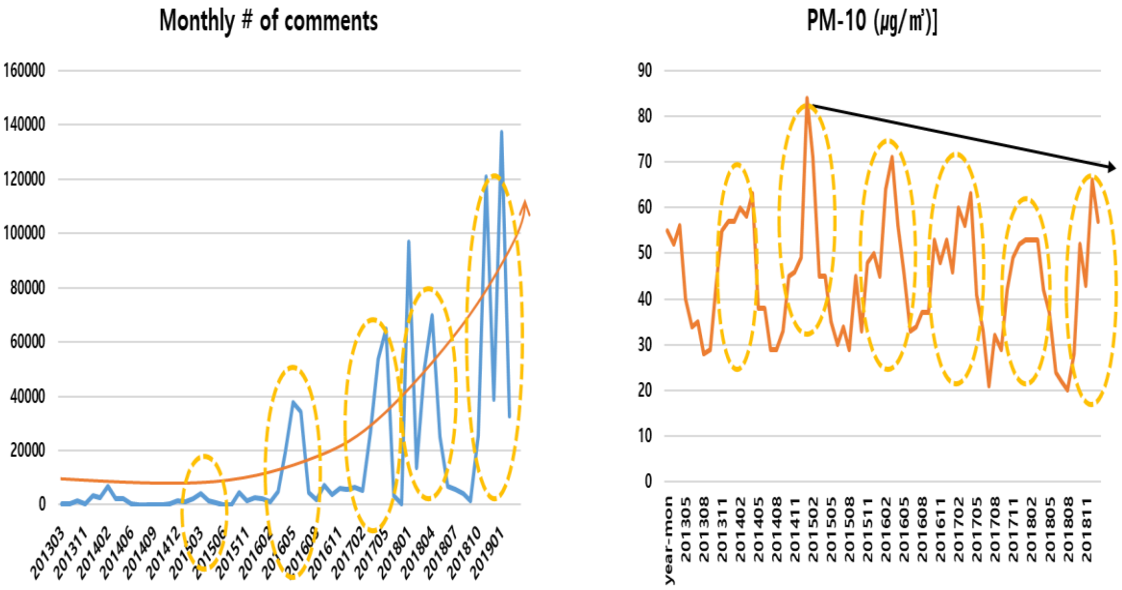


Figure 3. Monthly trends in the number of comments and PM concentration

4.2 Relationship between lexicons and recommendation/non-recommendation (RQ2)

We performed linear regression analysis with the frequency of 4 lexicon words, the length of comments as independent variables and the recommendation number / non-recommendation number as dependent variables. The results are shown in Table 2 and Table 3, respectively.

Table 2. Linear regression result (recommendation)

Model		Unstandardized		Standardized			Multi-collinearity	
		B	Std. B	Beta	t	Sig.	Tolerance	VIF
Dep. Var. : # of Recommen-dation	(Constant)	14.359	.307		46.849	0.000		
	LENGTH	-.026	.007	-.007	-3.816	.000	.330	3.027
	Emotional	.398	.108	.008	3.676	.000	.236	4.230
	Personal	.013	.002	.008	5.635	.000	.458	2.182
	Internal	-.107	.071	-.003	-1.501	.133	.284	3.515
	External	.991	.090	.023	11.006	.000	.249	4.016
F: 161.768(<0.0001), adjusted-R^2: 0.001, d.f. : 958013								

All variables, excluding ‘Internal Lexicon’, were significant to predict the number of recommendations. No multi-collinearity problem is found.

Table 3. Linear regression result (non-recommendation)

Model		Unstandardized		Standardized			Multi-collinearity	
		B	Std. B	Beta	t	Sig.	Tolerance	VIF
Dep. Var. :	(Constant)	1.432	.033		43.896	0.000		

# of Non recommendation	LENGTH	.006	.001	.015	8.531	.000	.330	3.027
	Emotional	-.017	.012	-.003	-1.509	.131	.236	4.230
	Personal	.002	.000	.011	7.165	.000	.458	2.182
	Internal	.102	.008	.026	13.424	.000	.284	3.515
	External	-.056	.010	-.012	-5.860	.000	.249	4.016
F: 216.383(<0.0001), adjusted-R^2: 0.001, d.f. : 958013								

All variables, excluding ‘Emotional Lexicon’, were significant to predict the number of recommendations. No multi-collinearity problem is found. As a result of multiple regression analysis using the number of comments recommended and not recommended as the dependent variable, we found the followings:

- Longer comments reduce the number of recommendations and increase the number of non-recommendations.
- Mentioning words with emotional factors increases the number of recommendations.
- The more words you mention internal dimension words, the more the number of non-recommendations increases.
- The more external words are mentioned, the more the number of recommendations and the less the number of recommendations

V. CONCLUSION

In this study, the comments on PM news articles are crawled and analyzed to investigate people’s emotions about PM. We can find that the interest in PM increases though the concentration of PM decreases. The response to PM was mainly emotional one. We also found that some seasonality in the number of comments and its correlation of PM10 density. People recommend short comments and are not preferred long posts. Comments that use a lot of emotional words or external words are ironic, with many people recommending and declining at the same time. This study has the limitation that the result can be different according to the classification of words belonging to lexicons. Therefore, the accuracy of lexicons has an important effect on the research results, and it is necessary to use objective lexicons built by other researchers to secure objectivity of research.

We expect to use methods such as Word embedding such as Word2Vec and Topic Modeling, which are being actively studied recently, to fully utilize the comments of collected fine dust articles. Through Word2Vec, we can map fine dust words to 200 ~ 300 dimension space to grasp word inference or similarity between words and infer the main subject of fine dust article comments through Topic Modeling.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A3A2074932)

REFERENCES

[1] Y.-I. Kim, S.-S. Yang, S.-S. Lee, and S.-C. Park, Design and Implementation of Mobile CRM Utilizing Big Data Analysis Techniques, The Journal of The Institute of Internet, Broadcasting and Communication, Vol. 14, No. 6 (pp. 289-294), 2014. https://doi.org/10.7236/JIIBC.2014.14.6.289

[2] T. Park, and J. Cha, A study on BEMS-linked Indoor Air Quality Monitoring Server using Industrial IoT,

International Journal of Internet, Broadcasting and Communication, Vol. 10, No. 4 (pp. 65-69), 2018. http://dx.doi.org/10.7236/IJIBC.2018.10.4.65

[3] V. Li, and V. Mariappan, Temperature Trend Predictive IoT Sensor Design for Precise Industrial Automation, International Journal of Advanced Smart Convergence, Vol. 7, No. 4 (pp. 75-83), 2018. http://dx.doi.org/10.7236/IJASC.2018.7.4.75

[4] Q. Sun, J. Zhuang, Y. Du, D. Xu, and T. Li, Design and application of a web-based real-time personal PM2.5 exposure monitoring system, Science of The Total Environment, Vol. 627 (pp. 852–859), 2018. https://doi.org/10.1016/j.scitotenv.2018.01.299

[5] Y. Ben, F. Ma, H. Wang, M. A. Hassan, R. Yevheniia, W. Fan, and Z. Dong, A spatio-temporally weighted hybrid model to improve estimates of personal PM2.5 exposure: Incorporating big data from multiple data sources, Environmental Pollution, Vol. 253 (pp. 403–411), 2019. https://doi.org/10.1016/j.envpol.2019.07.034

[6] M. T. Khan, and S. Khalid, Sentiment Analysis for Health Care, International Journal of Privacy and Health Information Management, Vol. 3, No. 2 (pp. 78–91), 2015. https://doi.org/10.4018/IJPHIM.2015070105

[7] Valdivia, M. V. Luzon, and F. Herrera, Sentiment Analysis in TripAdvisor, IEEE Intelligent Systems, Vol. 32, No. 4 (pp. 72–77), 2017.

[8] S. W. K. Chan and M. W. C. Chong, Sentiment analysis in financial texts, Decision Support Systems, Vol. 94 (pp. 53–64), 2017. https://doi.org/10.1016/j.dss.2016.10.006

[9] B. Hyun, The Study on Political Stances based on Editorials of Korean Newspapers, The Journal of the Convergence on Culture Technology, Vol. 4, No. 3 (pp.87-92), 2018. https://doi.org/10.17703/JCCT.2018.4.3.87

[10] H. Park, and J. Oh, Emotional Evaluation on the Environmental Color of the General Hospital's Lobby, The Journal of the Convergence on Culture Technology, Vol. 5, No. 3 (pp. 79-84), 2019. https://doi.org/10.17703/JCCT.2019.5.3.79

[11] Liu B., Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Vol. 5, No. 1 (pp. 1-167), 2012. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[12] Nasukawa, T. and Yi J., Sentiment analysis: Capturing favorability using natural language processing, in proc. of the 2nd international conference on Knowledge capture (pp. 70–77), Oct. 2003. https://doi.org/10.1145/945645.94565