pISSN : 2508-3384 eISSN : 2508-3392
https://doi.org/10.22640/lxsiri.2017.47.1.53
Journal of Cadastre & Land InformatiX Vol.47 No.1 (2017)
한국국토정보공사「지적과 국토정보」제47권 제1호 2017년 6월 p. 53-79

Classification of Public Perceptions toward Smog Risks on Twitter Using Topic Modeling

Topic Modeling을 이용한 Twitter상에서 스모그 리스크에 관한 대중 인식 분류 연구

Kim, Yun-Ki*
김윤기

Abstract

The main purpose of this study was to detect and classify public perceptions toward smog disasters on Twitter using topic modeling. To help achieve these objectives and to identify gaps in the literature, this research carried out a literature review on public opinions toward smog disasters and topic modeling. The literature review indicated that there are huge gaps in the related literature. In this research, this author formed five research questions to fill the gaps in the literature. And then this study performed research steps such as data extraction, word cloud analysis on the cleaned data, building the network of terms, correlation analysis, hierarchical cluster analysis, topic modeling with the LDA, and stream graphs to answer those research questions. The results of this research revealed that there exist huge differences in the most frequent terms, the shapes of terms network, types of correlation, and smog-related topics changing patterns between New York and London. Therefore, this author could find positive answers to the four of the five research questions and a partially positive answer to Research question 4. Finally, on the basis of the results, this author suggested policy implications and recommendations for future study.

Keywords: Public Perceptions, Smog Risks, Topic Modeling, LDA, Stream Graphs

1. Introduction

Smog has become a worldwide problem. When the level of smog is high, people are likely to think its sources first. While some people may regard China as the main source of smog, others may think of cars or industries as the main causes of it. Smog risks seem to have serious effects on every corner of our daily lives. Many governments of the world have implemented a lot of policies to cope with smog disasters. Most of the anti-smog policies seem to have focused on the physical aspect of smog phenomenon. Smog risks have not only physical aspect but also psychological aspect. When the level

* 청주대학교 지적학과 교수 Professor, Department of Land Management Cheongju University(kim2875@cju.ac.kr)

of smog is severe in China, many people are said to escape from urban areas to stay with their relatives in countryside. Those people are called "smog refugees." Though smog risks have both physical aspect and psychological aspect, only a few studies have paid attention to the psychological aspect of smog risks. That is, some researchers have tried to detect public opinions toward smog risks (Bickerstaff et al. 2000; Semenza et al. 2008; Saksena 2011; Li et al. 2015; Yang et al. 2016; Cheng et al. 2017). Most of the studies as mentioned above utilized offline data to detect and classify public opinions smog risks. Using offline data collection methods such as face-to-face interview and questionnaire survey method is sure to cost researchers huge amount of time and money. And utilizing offline data collection methods prevents researchers from detecting rapidly changing public perceptions toward smog risks.

Social media such as Twitter has become a popular communication tool among citizens around the globe. Huge amounts of tweets are posted every day. Many people tweet regarding smog sources, smog level, smog-related government policies, the health effects of smog and so on. Even though a lot of people use social media to express their opinions concerning smog risks, only a few studies have used online data to detect public perceptions toward them (Sha et al. 2014; Chen et al. 2016; Chen et al. 2017). Most of the studies as reviewed above didn't utilize Twitter data to detect public perceptions toward smog risks. They didn't use topic modeling to classify public perceptions toward smog risks either. Topic modeling is looked upon as a very effective tool to detect and classify public perceptions toward a specific issue (Wang et al. 2011). Detecting and classifying public opinions toward smog risks is crucial to the successful implementation of government smog policies. Quite a few governments have failed to reflect public opinions regarding smog risks in their policy making processes as well. Hence, it is important for policy makers to correctly detect and classify public perceptions toward smog during the policy making processes.

Therefore, this study aims to detect and classify public perceptions toward smog risks on Twitter using topic modeling. This research is composed of seven sections. Followed by introduction, section 2 focuses on a literature review on topic modeling and public perceptions toward smog risks. Section 3 consists of research questions. Section 4 deals with methodology. Section 5 pays attention to the results of the study. Section 6 is composed of the discussions. The section 7 of this research is made up of a conclusion and recommendations for future research.

2. Review of Literature

2.1. Topic modeling

2.1.1. Definition of topic modeling

Topic modeling is a statistical method for detecting topics from a large amount of text documents without human supervision (Qin et al. 2016). The method has attracted lots of attention since it was utilized by both scholars and practitioners around the globe. Theoretically, it is based on hierarchical probabilistic models (Jiang et al. 2015). The main proposition of this method is that a text document is composed of different latent

topics, and each latent topic is represented as a probability distribution over the term (Blei 2012). Just like latent variables in structural equation modeling (SEM), each topic cannot be directly observed but be rather inferred (Paul et al. 2014).

There are different types of topic models. Among them are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet allocation (LDA) that are extensively utilized for social media data (Qin et al. 2016). LSA is a statistical method for extracting and representing the relations of expected contextual usage of words in text documents (Landauer et al. 1998). It doesn't use traditional methods such as semantic networks and lexicons. Probabilistic Latent Semantic Analysis is a useful statistical method for analyzing two-mode and co-occurrence text data. This technique utilizes a mixture decomposition derived from a latent class model (Hofmann 1999). LDA is the most widely used topic modeling method. LDA uses a probabilistic procedure to explain how text documents are created (Blei et al. 2003). This method is based on the Probabilistic Latent Semantic Analysis. That is, LDA regards text documents as bags of words created by topics. In this model, each document is assumed to have a multinomial distribution over topics, and each topic is in turn assumed to have a multinomial distribution over terms. Therefore, LDA can be interpreted as the extended version of PLSA (Thomas 2001).

2.1.2. Review of literature related to topic modeling

A lot of studies have been conducted to identify topics in many fields using unsupervised topic modeling. For example, Hyunh et al. (2008) attempted to introduce a new method for detecting daily routines from on-body data. They utilized LDA to identify activity patterns. Their research results show that people's activity patterns have significant correlation with daily routines. Hall et al. (2008) tried to discover the development patterns of idea in Computational Linguistics field. They used unsupervised topic modeling to analyze historical trends in that scientific field. They attempted to identify topics using LDA and examined how each topic changes over time. Their study results show that research topics in that field change greatly over time. Wang et al. (2011) tried to develop an algorithm to recommend academic papers to online community users. To achieve their research goal, they attempted to combine traditional text filtering methods with topic modeling. Their research results reveal that their recommendation method using topic modeling algorithm can provide services more effectively than traditional filtering system.

Bisgin et al. (2011) tried to find topics in drugs with similar safety concerns and/or effects together. They utilized the topic modeling approach with LDA to generate 100 drug related topics. Their study findings demonstrate that drugs clustered by topics have significant correlation with the same safety concerns and effects. Tuarob et al. (2013) made an attempt to create algorithms for automatic annotation of metadata. To achieve their research purpose, they utilized probabilistic topic modeling. They suggested a couple of algorithms for tag recommendation based on topic modeling method and TF-IDF (Term Frequency-Inverse Document Frequency) method. Regardless of the good

performance, their algorithms have some limitations such as scalability issue.

As we reviewed above, most of the topic models belong to unsupervised models. However, researchers revealed that unsupervised models may generate incoherent topics because those models can't reflect human judgement (Liu et al. 2014a). Therefore, some scholars proposed knowledge-based topic modeling methods to address this problem. For example, Liu et al. (2014a) tried to develop new topic model in which knowledge can guide the model inference. They compared their model with other topic modeling methods such as LDA (Blei et al. 2003), DF-LDA (Andrzejewski et al. 2009), GK-LDA (Chen et al. 2013), and AKL (Liu et al. 2014b). Then they tested their model using product reviews from 50 websites. Their study results show that their model is effective and reliable in discovering topics.

2.2. Public perceptions toward smog risks

2.2.1. Definition of public perceptions toward smog risks

Today, smog risks are getting serious day by day. Everybody talks about smog risks, but he soon realizes that there is nothing he can do to avoid them. In this way, smog disasters seem to have serious negative impacts on our daily lives. Identifying public perceptions toward smog risks forms the basis for the successful implementation of government anti-smog policies. While smog continues to pose major threats to citizen's health around the globe, very few studies have ever been performed to correctly detect and classify public perceptions of smog risks. Many governments have failed to reflect public perceptions of smog risks in their anti-smog policy making processes as well. Therefore, it is crucial for government authorities to correctly identify public perceptions toward smog before they make anti-smog policies.

2.2.2. Review of literature related to public perceptions of smog risks

Quite a few studies have been conducted to examine public perceptions toward smog risks. For instance, Bickerstaff et al. (2000) attempted to measure public perceptions toward the problems of air pollution. In order to achieve their research goals, they utilized questionnaire survey and in-depth interviews with citizen in Birmingham in the UK. Their study results show that public perceptions toward air pollution is far from universal. Therefore, they placed emphasis on the localization of air pollution risks to achieve the objectives of government environmental policies. However, their research has some limitations because questionnaire survey method that they used to measure public perceptions toward air pollution risks cannot reflect rapidly changing public opinion correctly. Semenza et al. (2008) tried to measure citizen's perceptions and behavior changes with regard to air pollution. They measured air quality and took telephone interviews to collect research data. Their study results demonstrate that not advisories but public perceptions towards air pollution cause citizen's behavioral change. However, they paid only attention to the cross-sectional aspects of public perceptions toward air pollution. Saksena (2011)

made an attempt to review the literature related to public perceptions of air pollution risks including smog risks. He utilized a Pressure-State-Response framework to attain his research purpose. His study results reveal that there is a big gap with public perceptions of government policy response. However, he didn't try to fill the gap in the literature by conducting empirical studies. Li et al. (2015) tried to explore international tourists' perceptions toward smog in Beijing, China. They developed a scale to measure tourists' perceptions toward smog risks. Then they tested relationships among such latent variables as smog risks, satisfaction, destination royalty, and risk perceptions. Their research results reveal that smog concern has effects on destination royalty both directly and indirectly. However, they ignored important variables such as service quality and consumer prices that influence on destination royalty greatly. Yang et al. (2016) tried to find the chemical properties of smog and its influences on human health. They conducted questionnaire survey using stratified sampling to measure public perceptions toward smog. Their research results indicate that people perceive automobiles, factories, and kitchen hums as the main sources of smog in Ningbo city, China. They concluded that studying public perceptions toward smog is necessary to cope with smog disasters effectively. Cheng et al. (2017) tried to examine citizen's protective behavior in response to smog risks. They conducted survey in Heifei and Anhui, China to collect data regarding citizen's perception toward smog risks. They used PCA (the Principal Component Analysis) to analyze collected data. Their study results show that hazard related variables have strong effects on willingness to adopt protective actions. However, they ignored risk perceptions' role as a mediator in the model. That is, in this case structural equation modeling (SEM) should be utilized to measure the mediation effects of risk perceptions.

Some studies focused on people's willingness to pay (WTP) for reducing the level of smog. For example, Sun et al. (2016b) made an attempt to estimate citizen's willingness to pay for lowering smog risks in China. They utilized face-to-face survey method to collect data regarding WTP. They used the two-part model to measure citizen's willingness to pay. Their study results reveal that around 90% of those surveyed are willing to pay for reducing the level of smog. However, since their study used face-to-face survey method, they cannot measure rapidly changing public opinion correctly. Sun et al. (2016a) tried to estimate the value of WTP and analyze its determinants. They utilized a CV method framework to achieve their research goals. And they used the bivariate sample selection method to estimate WTP for reducing the level of smog. Their research results indicate that nearly 15% of respondents do not willing to pay for lowering the level of smog. However, they only focused on the cross-sectional aspects of public perceptions concerning smog risks.

2.3. Topic modeling and public perceptions on smog

2.3.1. Topic modeling on public perceptions of smog

While smog problem is getting worse every day, only a few studies have addressed public

perceptions toward smog related risks on social media. For example, Sha et al. (2014) tried to estimate public sentiment with regard to air pollution. They extracted social media data from Chinese Weibo. Then they analyzed the data using SentiStrength algorithm along with a Chinese sentiment lexicon. Their study results indicate that overall public sentiment regarding smog risks tend to converge toward positive territory over time. However, they didn't make an attempt to classify public perceptions toward smog risks using topic modeling. Chen et al. (2016) tried to predict smog disasters using social media data. They extracted smog related data from a Chinese social media called Sina Weibo. They used semantic reasoning to forecast smog risks. In that study, they tried to combine semantic reasoning with machine learning algorithms. Even though they used semantic reasoning to forecast smog risks in their analysis, they didn't use topic modeling to classify smog related topics.

Chen et al. (2017) tried to predict smog related risks. That is, they utilized both social media data and physical sensor data to forecast smog risks in its early stage. Then they developed a prediction model based on artificial neutral networks (ANNs) to measure relationships between social and physical smog data and smog risks. Their study results demonstrate that the model using both social media data and physical sensor data can provide greater predictability than that using only one of them. Even though they used social media data in their analysis, they didn't use topic modeling or sentiment analysis to measure public perceptions toward smog risks.

This author's literature review demonstrates that public perceptions toward smog risks play important parts in predicting smog related disasters. Most of the studies on public perceptions toward smog risks have been carried out using offline data. Many researchers used questionnaire methods or face-to-face interview methods to examine public perceptions toward smog risks. Those methods are likely to cost researchers a lot of time and money. Public perceptions toward smog tend to change over time and space. Today more and more people use social media to express their opinions regarding smog related risks and problems. It is very important for policy-makers to detect and classify public perceptions toward smog risks to make a reasonable anti-smog policy. However, few studies have ever been conducted to detect and classify public perceptions toward smog risks on Twitter. Few studies on perceptions toward smog have ever used topic modeling to detect smog related topics. Therefore, this research may be the first attempt to detect and classify public perceptions toward smog risks using Twitter data.

2.3.2. Topics in public perceptions of smog risks

Public perceptions toward smog risks can be categorized as (1) perceptions of smog sources (2) perceptions of smog level (3) perceptions of its health effects (4) perceptions of government smog policies and responses (Saksena 2011). When smog alerts are on, people tend to think its sources or causes first. While some people may perceive China as the main source of smog, others may view automobiles or factories as the main causes of it.

Public perceptions of smog risks are likely to change over time and space. Perceptions of smog level seem to have significant impact on activity patterns. For example, perceptions of smog level turned out to have significant effects on outdoor activities (Saksena 2011). When people think that smog level is high, they tend to refrain from doing outdoor activities. Otherwise, they are likely to do more outdoor activities. Perceptions toward the health effects of smog seem to play important roles in forming public perceptions toward smog in general. Whenever smog alerts are on, people tend to wear masks before going out. Everybody worries about the health effects of smog. Today many people tend to express their opinion toward the health effects of smog via Twitter frequently. Perceptions of government smog policies and responses are another area which draw public attention. When people begin to feel the health effects of smog, they tend to complain about inadequate government anti-smog policies and measures.

3. Research Question

This study aims to detect and classify public opinion toward smog on Twitter using topic modeling. To help achieve these goals and to identify gaps in the literature, this research carried out a literature review on public perceptions toward smog risks and topic modeling. This author's literature review demonstrates that there are considerable gaps in the existing literature. In order to fill the gaps in the literature, this study build the following five research questions.

Research Question 1: what are the most frequent terms with regard to smog? Are there any differences in the most frequent terms regarding smog between New York and London?

Most frequent terms are the foundation of the text analysis including word cloud analysis and topic modeling. Since most frequent terms are likely to reflect public perceptions toward smog risks and regional environmental characteristics, this author assumes that there may be big differences in the most frequent terms regarding smog between New York and London.

Research Question 2: how are the most frequent terms connected to one another? Are there any differences in the shapes of the terms network between New York and London?

Some of the most frequent terms can be used in many tweets at the same time. The network of terms reflects the co-occurrences of words in lots of tweets. Now that the most frequent terms that people use in their tweets are different from city to city, this author assumes that there may be big differences in the shapes of the terms network between New York and London.

Research Question 3: what type of correlation does exist between smog and other frequent terms? Are there any differences in the type of correlation between New York and London?

Correlation among the most frequent words can reflect the co-occurrences of those words in multiple tweets. Since the most frequent terms are likely to be influenced by geographical location, this author assumes that there may be big differences in correlation between New York and London.

Research Question 4: how can different terms be clustered into smog-related topics? Are there any

differences in the types of smog-related topics between New York and London?

Some words can be clustered into one topic, because they are used together in multiple tweets. Since the words that people use in their tweets are likely to reflect their perceptions toward smog risks and locational characteristics, this author assumes that there may be big differences in the types of smog-related topics between New York and London.

Research Question 5: how do different smog-related topics change over time? Are there any differences in smog-related topics changing patterns over time between New York and London?

As public perceptions toward smog risks change over time and space, so do smog-related topics. Now that smog-related topics are likely to reflect public perceptions toward smog risks, this researcher assumes that there may be big differences in smog-related topics changing patterns over time between New York and London.

4. Methodology

4.1. Research design

In this research, this author developed five research questions to fill the gaps in the related literature. And then this study performed the following research procedures to answer those research questions. First, this research extracted 10,000 tweets concerning smog risks from the Twitter server on March 17, 2017 using Twitter API and preprocessed them. Second, this author conducted word cloud analyses on the cleaned data to derive the most frequent terms. Third, this study built the network of terms to explain the co-occurrences of the most frequent terms in multiple tweets using the Rgraphviz package of R (Zhao 2013). Fourth, this author conducted correlation analysis to measure the association between smog and other frequent terms. Fifth, this study conducted hierarchical cluster analyses on the cleaned data to find how different terms are clustered into smog-related topics. Sixth, this author used topic modeling with the LDA to identify smog-related topics using the topicmodels packages of R. Finally, this research utilized stream graphs to demonstrate time series variation of smog-related topics.

4.2. Data collection

4.2.1. Authentication

To extract tweets regarding smog risks, this author created his Twitter account and Twitter application. Two R packages such as ROAuth and TwitteR were used to get authentication from Twitter. ROAuth package allows us to get authentication from the Twitter server and TwitterR package enables us to use an interface to the API. We can use the API to extract Twitter data on a certain topic in a specific location and language (Aldayel et al. 2016).

4.2.2. Extracting tweets

In this stage, this author used Twitter API to extract tweets concerning smog risks from the Twitter server. The desired tweets extraction procedures were conducted on March 17, 2017. This

author utilized searchTwitter function of TwitterR package to get Twitter data on smog risks. This study chose New York and London as its sample. For each city, the number of tweets extracted was set to 5,000. This author utilized coordinates and radius as the strings of geocode to extract 5,000 tweets regarding smog risks posted from each city. For example, to get 5,000 tweets regarding smog risks from New York, its coordinates (40.730610, -73.935242) and radius (100 mile) were utilized as its geocode strings. To extract 5,000 tweets concerning smog risks posted from London, this author used its coordinates (51.508530, -0.076132) and radius (100 mile) as its geocode strings as well.

4.3. Preprocessing data

In this stage, this author preprocessed the data for word cloud analysis, hierarchical cluster analysis, and topic modeling. That is, this research removed all numbers, all hashtags, all html links, all retweets, all tabs, all punctuations, all blank spaces at the beginning, all blank spaces at the end, all speech effects, all mentions(@ID), all stop words, and all newline characters to improve the quality of extracted data. And then this study converted all terms to lower case by using the tolower function of R. Finally, stemming procedures were carried out to get rid of all affixes using R'S SnowballC package.

4.4. Data analysis

4.4.1. Word cloud analysis

This research built a term-document-matrix to derive the most frequent terms from each city in the sample. A term-document-matrix describes words frequencies which occurs in documents. This author used R's tm package to build the term-document-matrix. Then this research created word clouds for public perceptions toward smog risks using wordcloud package of R.

4.4.2. Creating the network of terms

This study built the network of terms to explain the co-occurrences of the most frequent terms in multiple tweets using the Rgraphviz package of R. In the network of terms, thick line shows that two connected terms co-occur in multiple tweets.

4.4.3. Correlation analysis

This research conducted correlation analysis to measure the association between smog and other frequent terms. To measure the correlation between smog and other frequent terms, this author built a document-term-matrix. Then findAssocs function of R was utilized to calculate the correlation between smog and other frequent terms.

4.4.4. Hierarchical cluster analysis

This author carried out hierarchical cluster analyses on the cleaned data to discover how different terms are clustered into smog-related topics. This research removed sparse terms using the removeSparseTerms function of R. Then this study measured the distances with dist() function. The hclust() function was utilized to cluster terms and rect.hclust was used to draw dendrograms. The Ward agglomeration method was used to cluster

terms into topics.

4.4.5. Topic modeling

This author used topic modeling with the LDA to identify smog-related topics. Then this research counted the number of tweets in each topic and created stream graphs to demonstrate time series variation of smog related topics (Zhao 2013). Such R packages as data.time package and topicmodels package were utilized to identify smog-related topics and create stream graphs.

5. Result

5.1. Research Question 1

Research question 1 is about whether there are big differences in the most frequent terms between New York and London. Table 1 demonstrates the most frequent terms for New York and London.

The word cloud analysis shows that the smog-related terms people in New York used most frequently on Twitter are rt, smog, air, pollute, this, EPA, new, new, bad, china and place. New Yorkers' perceptions toward smog risks seem to cluster around such topics as smog sources (china, air, smog), smog level (pollute, bad), and government policies and responses (EPA). However, Londoners use different kinds of the most frequent terms to express themselves regarding smog risks on Twitter. The smog-related words people in London use most frequently on Twitter turn out to be airpocalypse, cap, event, ice, link, melt, research, reveal, smog, and rt. The results of this study demonstrate that Londoners perceptions toward smog risks seem to center on such topics as smog sources (smog), health effects (airpocalypse), and weather effect (ice, melt, research, reveal).

Judging from the above results, there appear to be big differences in the most frequent terms between New York and London.

5.2. Research Question 2

Research question 2 is about whether there are big differences in the shapes of the terms network between New York and London. Figure 1 and Figure 2 show the network of terms for New York and London respectively.

As can be seen in Figure 1, there are many thick lines in the network of terms for New York. Thick lines in the network of terms indicate that two connected terms co-occur in multiple tweets (Zhao 2013). There are thick lines between bad and EPA, between this and EPA, between bad and environment, between EPA and protect, between protect and place, between protect and environment, between need and protect, between need and

Table 1. The most frequent terms regarding smog risks for New York and Laondon

Rank	New York	London
1	rt	airpocalypse
2	smog	cap
3	air	event
4	pollute	ice
5	this	link
6	EPA	melt
7	new	research
8	bad	reveal
9	china	smog
10	place	rt

environment, and between protect and this. The above results demonstrate that there seem to be many tweets regarding bad EPA, this EPA, protecting environment, protecting place, and protecting this place. Figure 1 shows that New Yorkers' perceptions toward smog risks are centered around government environmental policies.

Figure 2 indicates that there are several thick lines in the network of terms for London. There seem to be thick lines between research and reveal, between research and link, between research and china, between research and event, between research and aipocalypse, between research and smog, between reveal and smog, between reveal and

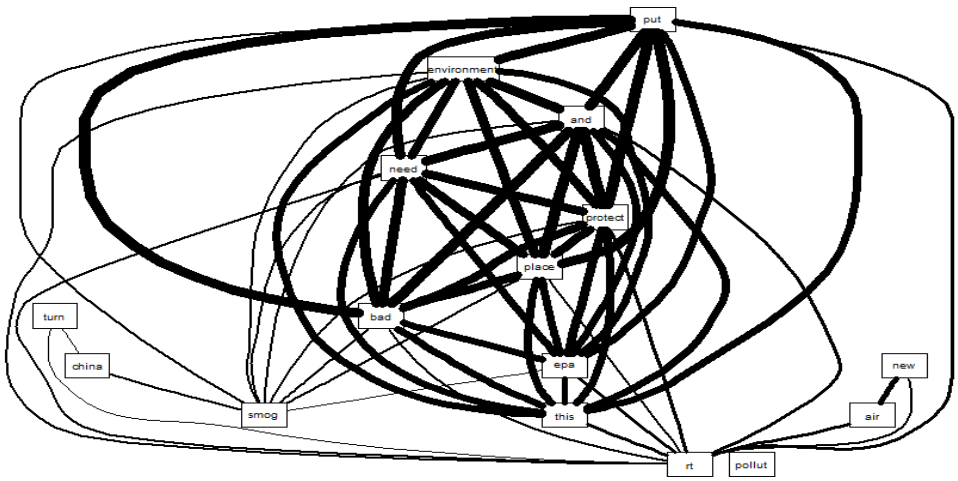

Figure 1. The network of terms for New York

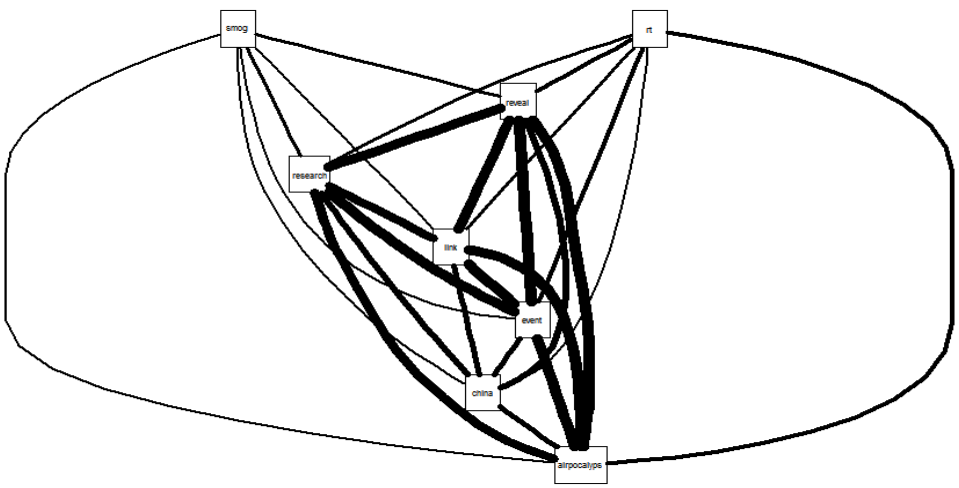

Figure 2. The network of terms for London

Table 2. Correlation coefficients between smog and other frequent terms

Frequent terms for New York	Correlation coefficients	Frequent terms for London	Correlation coefficients
kill	0.25	cap	0.26
bad	0.25	ice	0.26
environment	0.25	research	0.26
remember	0.24	reveal	0.26
place	0.24	event	0.21
need	0.24	airpocalypse	0.20
protect	0.21	melt	0.20

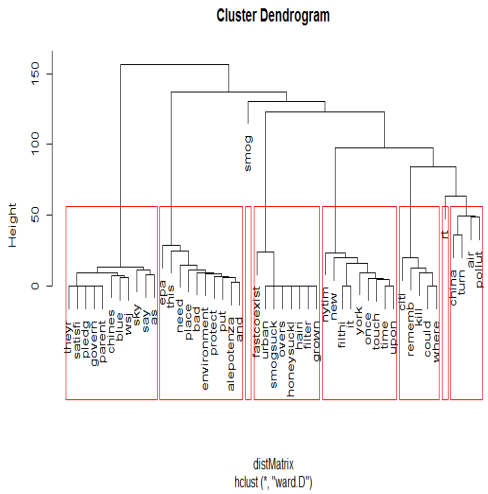

Figure 3. Cluster dendrogram for New York

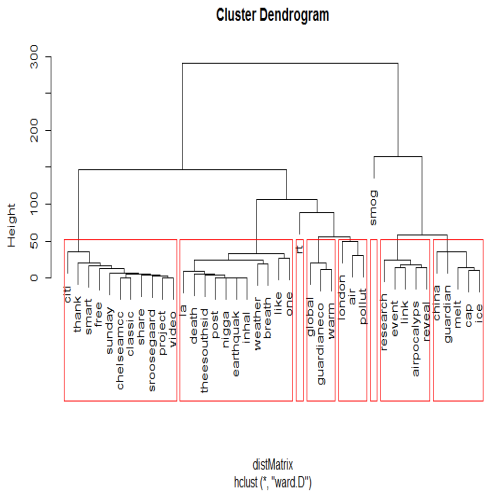

Figure 4. Cluster dendrogram for London

china, between reveal and event, between reveal and link, between reveal and airpocalypse, between link and china, between link and event, between link and airpocalypse, between event and china, between event and airpocalypse, and between china and airpocalypse. The results of this study demonstrate that there appear to be a lot of tweets regarding airpocalypse. Figure 2 indicates that Londoners' perceptions toward smog risks are centered around smog sources and its health effects. Judging from the results in Figure 1 and Figure 2, there seem to be big differences in the shapes of the terms network between New York and London.

5.3. Research Question 3

Research question 3 is about whether there are big differences in the type of correlation between New York and London. Table 2 demonstrates Correlation coefficients between smog and other frequent terms for New York and London.

All of the seven most frequent terms seem to have somewhat high correlation with smog($r>0.20$). The top correlated terms with smog for New Yorkers

turn out to be kill, bad, environment, remember, place, need, and protect. That is, many people in New York tend to use such phrases as "smog kills", "bad smog", "remember smog", and "need to protect environment or place" in the same tweet.

However, Londoners use quite different terms along with smog in same tweet. Top seven correlated terms with smog for Londoners prove to be cap, ice, research, reveal, event, airpocalypse, and melt (r〉 0.20). Therefore, the results of this study indicate that there are big differences in the type of correlation between New York and London.

5.4. Research Question 4

Research question 4 is about whether there are big differences in the types of smog-related topics between New York and London. Figure 3 and Figure 4 shows cluster dendrograms for New York and London.

Figure 3 demonstrates eight smog-related topics for New York. Terms such as theyr, satisfi, pledge, govern, parent, Chinese, blue, sky, say, WSJ, and as are clustered into the first topic, indicating that there are multiple tweets regarding Chinese government's pledge of blue sky. Therefore, the first topic can be labeled as public perceptions toward smog source. Such words as EPA, this, need, place, bad, environment, protect, put, and, alepotenza are clustered into the second topic. This cluster of terms implies that there are many tweets in which people want EPA to protect their places and environment. Thus, this topic seems to be related with public perceptions toward government policies and actions. Figure 3 reveals that there is only one term (smog) in the third topic.

Terms such as nytimes, new, filthi, it, york, once, touch, time, and upon are clustered into the fifth topic. This topic seems to be associated with the New York Times's coverage about smog effects on city landscapes. Thus, this cluster can be labeled as public perceptions toward smog's effects on urban landscapes. Terms such as citi, remember, kill, could, and where belong to the sixth cluster. Terms in this cluster imply that there are many tweets in which people perceive that smog can kill people everywhere. Therefore, this cluster can be named as public perceptions toward smog's health effects as well. Figure 3 shows that there is only one term (rt) in seventh topic. Words such as china, turn, air, and pollute are clustered into the eighth topic. Judging from this research results, there seem to be many tweets in which people perceive china as the main source of smog. Thus, this cluster can be named as public perceptions toward smog source.

Figure 4 reveals eight smog-related topics for London. Terms such as citi, thank, smart, free, Sunday, chelseamcc, classic, share, sroosegaard, project, and video are clustered into the first topic. This cluster of terms implies that there are quite a few tweets regarding smog free project in china. This cluster of words is closely related with public perceptions regarding anti-smog policies and actions. Hence, this cluster of terms can be labeled as public perceptions regarding anti-smog policies and actions. Words such as la, death, theesothsid, post, nigga, inhale, weather, breath, like and one are grouped into the second topic. Terms in this cluster reveal that there are multiple tweets concerning public perceptions on smog's health effects. That is,

words in this cluster imply that inhaling smog can have serious negative effect on people's health. Figure 4 shows that there is only one term (rt) in the third topic. Such terms as global, warm, and guardianeco are clustered into the fourth topic. Terms in this cluster imply that there are multiple tweets regarding global warming. This cluster is associated with public perceptions toward global warming. Words such as London, air, and pollute are grouped into the fifth topic. Terms in this topic indicate that there are many tweets regarding polluted air. Therefore, this cluster can be named as public perceptions toward the level of smog risks. Figure 4 shows that there is only one term (smog) in the sixth topic. Such terms as research, link, event, and airpocalypse are clustered into the seventh topic. Words in this cluster imply that there are many tweets regarding research results of smog risks (airpocalypse). This cluster can be named as public perceptions toward smog research results. Words such as China, guardian, melt, ice, and cap are grouped into the eighth topic. Terms in this topic imply that many people perceive that smog is closely associated with melting ice cap (global warming). Judging from the results in Figure 3 and Figure 4, there seem to be big differences in the types of smog-related topics between New York and London.

5.5. Research Question 5

Research question 5 is about whether there are big differences in smog-related topics changing patterns over time between New York and London.

Table 3 shows the results of the topic modeling with LDK for New York. In this research, this author used the topic modeling with LDK to detect and classify public perceptions toward smog risks on Twitter. That is, as can be seen in Table 3 and Table 4, this author classified public perceptions toward smog risks into 8 topics. Terms such as smog, rt, remind, pollute, air, problem, get, u, LA, and not are classified into topic 1.

Topic 1 seems to be closely associated with public perceptions toward the internal sources of smog risks. Therefore, this topic can be labeled as the internal sources of smog risks. Words in this topic imply that there are a lot of tweets regarding the internal smog sources such as polluted air. Such words as smog, rt, China, pollute, diamond, design,

Table 3. The results of the topic modeling with LDK for New York

Topics	Terms
Topic 1	smog / rt / remind / pollute / air / problem / get / u / LA / not
Topic 2	smog / rt / China / pollute / diamond / design / change / us / turn / Dutch
Topic 3	smog / rt / air / turn / Beijing / pollute / jewelries / purifi / nytimes / the
Topic 4	smog / citi / remember / kill / could / where / I / rt / EPA / NYC
Topic 5	say / as / sky / Chinese / blue / turn / govern / parent / pledge / satisfied
Topic 6	EPA / smog / need / bad / protect / this / place / rt / environment / put
Topic 7	air / rt / fastcoexist / this / filter / grown / hairi / honeysuckle / overs / smogsuck
Topic 8	air / new / rt / york / once / time / upon / filthi / it

Table 4. The results of the topic modeling with LDK for London

Topics	Terms
Topic 1	smog / rt / London / across / Avant garde / French / jazz / lounge
Topic 2	smog / rt / China / world / coal / social / London / heavi
Topic 3	smog / pollute / live / get / set / the listen / now play
Topic 4	smog / like / one / rt / breath / weather / death / LA
Topic 5	smog / climate change / mani / global / air quality / problem
Topic 6	smog / air / rt / hand / pollute / busi / change / climate
Topic 7	link / airpocalypse / event / research / reveal / rt / smog / warm
Topic 8	China / ice / cap / melt / guardian / smog / rt / air

change, us, turn, and Dutch are clustered into topic 2. Topic 2 appears to be related with public perceptions toward the external sources of smog risks. Hence, this topic can be named as the external sources of smog risks. In light of the above results, this author can conjecture that there are multiple tweets regarding air pollution and smog in China. Words such as smog, rt, air, turn, Beijing, pollute, jewelries, purifi, nytimes, and the are grouped into topic 3. Now that most of the terms in this topic are closely related with public perceptions toward the external sources of smog risks, this topic can be labeled as the external sources of smog risks as well. Terms in this topic imply that many New Yorkers perceive China as the main sources of Smog. Such terms as smog, citi, remember, kill, could, where, I, rt, EPA, and NYC are clustered into Topic 4. Topic 4 seems to be closely associated with public perceptions toward the health effects of smog. Thus, this topic can be named as the health effects of smog. Considering the above results, it is clear that there are a lot of tweets in which people worry about the health effects of smog.

Such words as say, as, sky, Chinese, blue, turn, govern, parent, pledge, and satisfied belong to topic 5. Topic 5 appears to be closely associated with public perceptions toward foreign government smog policies and responses. Therefore, this topic can be labeled as public perceptions toward foreign government smog policies. In light of the results in Table 3, this author can conjecture that people in New York are very interested in foreign government anti-smog policies, because smog can affect global air quality. Terms such as EPA, smog, need, bad, protect, this, place, rt, environment, and put are classified into topic 6. Topic 6 seems to be related with public opinion on the government smog policies and responses. Hence, this topic can be named as public perceptions toward the government smog policies and responses. Considering the results in Table 3, this author can assume that there are a lot of tweets in which people want EPA to protect New York and its environment from smog risks. Such words as air, rt, fastcoexist, this, filter, grown, hairi, honeysuckle, overs, and smogsuck are clustered into topic 7. Topic 7 appears to be closely associated with public perceptions toward the level of smog. Thus, this topic can be labeled as public perceptions toward the level of smog. Terms in this topic imply that there are many

tweets complaining about smog risks. Such words as air, new, rt, york, once, time, upon, filthi, and it are classified into topic 8. This topic seems to be related with public perceptions toward the level of smog as well. Hence, this topic can be named as public perceptions toward the level of smog. Considering the results in Table 3, this author can assume that there are many tweets in which people complain about the seriousness of smog risks.

Figure 5 shows stream graph for New York. Stream graph is very useful for showing the fluctuation of each topic over time. As can be seen in Figure 5, some topics have grown very rapidly on a particular date. For example, there was a big fluctuation in the number of tweets regarding topic 6 between March 8 and March 10. While topic 3 had been a dominating topic from March 9 to March 11, topic 10 had been a winner from March 11 to March 13. There was a big change in the number of tweets with regard to topic 4 between March 11 and March 17. In the case of topic 1, the number of tweets has rapidly changed from March 6 and March 18.

Table 4 shows the results of the topic modeling with LDK for London. Such terms as smog, rt, London, across, Avant garde, French, jazz, and lounge are classified into topic 1. Topic 1 doesn't seem to be associated with public perceptions toward smog risks. It is related with the British rock star Smog Veil. Terms such as smog, rt, China, world, coal, social, London, and heavi are clustered into topic 2. Now that topic 2 seems to be closely associated with public perceptions toward smog sources, this topic can be named as the sources of smog. That is, Londoners seem to regard China and coal as the sources of smog. Such words as smog, pollute, live, get, set, the listen, and now play are classified into topic 3. Words in this topic imply that there are multiple tweets regarding public perceptions toward the level of smog. Therefore, this topic can be labeled as public perceptions toward the level of smog. Such terms as smog, like, one, rt, breath, weather, death, and LA belong to topic 4. Topic 4 appears to be quite different from topic 3. In light of the above results, this author can

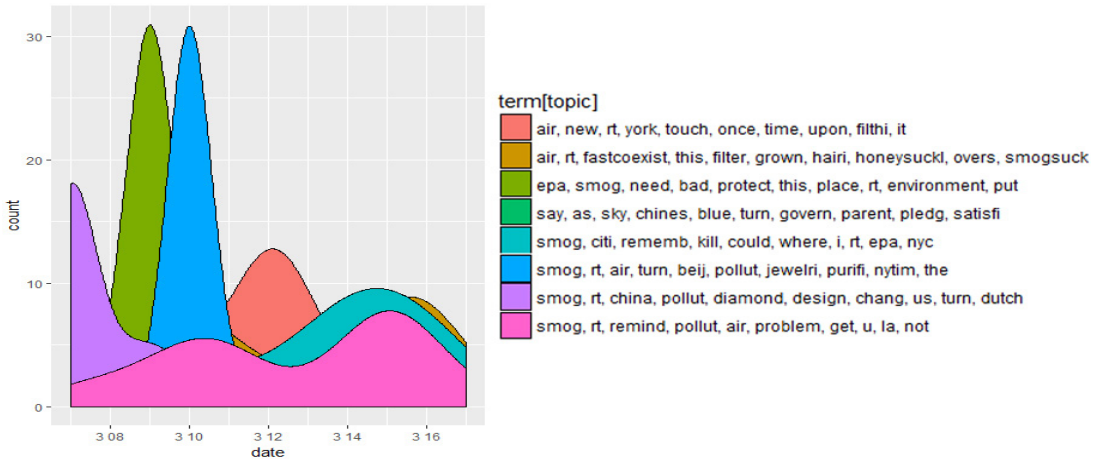

Figure 5. Stream graph for New York

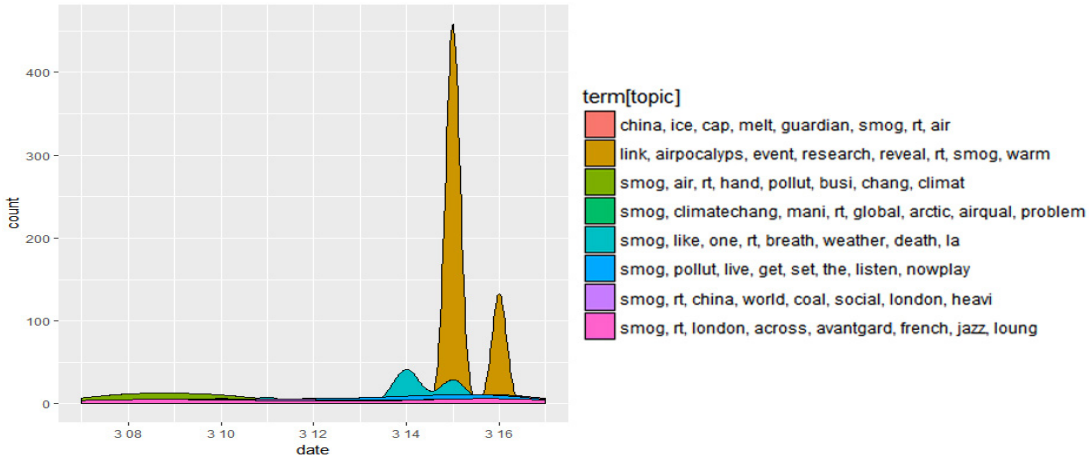

Figure 6. Stream graph for London

conjecture that there a lot of tweets in which people worry about the health effects of smog. Therefore, this topic can be labeled as public perceptions toward the health effects of smog.

Words such as smog, climate change, mani, global, air quality, and problem are classified into topic 5. Topic 5 seems to be closely associated with smog's effects on global climate. That is, terms in this topic imply that many people in London perceive smog as the cause of global climate change. Hence, this topic can be labeled as smog's effects on global climate. Such terms as smog, air, rt, hand, pollute, busi, change, and climate are clustered into topic 6. Topic 6 appears to be related with public perceptions toward smog's effects on climate change as well. Considering the results in Table 4, this study can assume that there are multiple tweets in which people complain about smog and climate change. Therefore, topic 6 can be labeled as smog's effects on climate change as well. Terms such as link, airpocalypse, event, research, reveal, rt, smog, and warm are grouped into topic 7. Topic 7 seems to be closely associated with public perception toward smog's effects on global warming. Terms in this topic suggest that many Londoners perceive smog as the main cause of global warming. Thus, this topic can be named as public perception toward smog's effects on global warming. Such words as China, ice, cap, melt, guardian, smog, rt, and air are classified into topic 8. Words in this topic imply that there are many tweets in which people perceive China as the main source of smog. Hence, this topic can be labeled as public perceptions toward the main source of smog.

Figure 6 shows stream graph for London. As can be seen Figure 6, there seemed to be fluctuations in the number of tweets regarding some topics. For example, there was a couple of fluctuations in the number of tweets regarding topic 7 between March 15 and March 16. That is, on March 15 the number of tweets regarding topic 7 had rapidly increased and declined very sharply. On March 15 there seem to be a big change in the number of tweets regarding topic 7 again. These changes in the number of

tweets regarding topic 7 appear to be largely due to the research report on smog released on the same day. There seemed to be some changes in the number of tweets concerning topic 4 between March 13 and March 15. However, there were little changes in the number of tweets regarding other topics. Judging from the results in Figure 5 and Figure 6, there appear to be big differences in smog-related topics changing patterns over time between New York and London.

6. Discussions

6.1. Research Question 1

The most frequent terms are thought to be the basis of the text analysis. They are likely to reflect public perceptions toward specific issues. In this study, this author assumed that there may be big differences in the most frequent terms regarding smog risks between New York and London. As expected, the results of this research show that there are big differences in the most frequent terms concerning smog risks between New York and London. For instance, New Yorkers' perceptions toward smog risks appear to be clustered around such topics as smog sources (china, air, smog), smog level (pollute, bad), and government policies and responses (EPA).

However, Londoners perceptions toward smog risks seem to center on such topics as smog sources (smog), health effects (airpocalypse), and weather effect (ice, melt, research, reveal). From the results of this research, it can be implied that every city in the world has its own smog-related problems which are certain to have effects on public perceptions toward smog risks. To overcome smog risks effectively, it is crucial for environmental policy makers to examine public perceptions toward smog risks exactly. Topic modeling and word cloud analysis can provide the effective means of measuring public perceptions toward smog risks.

6.2. Research Question 2

Research Question 2 is regarding the network of term. Terms network is likely to reflect the co-occurrences of terms in multiple tweets. In this research, this author assumed that there may be big differences in the shapes of terms network between New York and London. As expected, the results of this study demonstrate that there are big differences in the shapes of terms network between New York and London. For example, New Yorkers' perceptions toward smog risks are centered around government environmental policies. That is, there appear to be thick lines between bad and EPA, between bad and environment, between EPA and protect, between protect and place, between protect and environment, and between need and protect. However, Londoners' perceptions toward smog risks are centered around smog sources and its health effects. That is, there appear to be thick lines between research and reveal, between research and china, between research and smog, between reveal and china, between reveal and airpocalypse, and between link and china. From the results of this study, it can be suggested that the shapes of terms network can be affected by geographical location. Each term in the network

seems to reflect public perceptions toward smog risks. Hence, the shape of terms network can be the epitome of public perceptions of smog risks. Therefore, it is crucial for policy makers to use the network of terms to detect public perceptions toward smog risks.

6.3. Research Question 3

Association among the most frequent terms can reflect the co-occurrences of those words in multiple tweets. Thick lines in the network of terms means high correlation between terms. In this study, this author supposed that there may be big differences in the type of correlation between New York and London. As can be seen in Table 2, the results of this study reveal that there are big differences in type of association between New York and London. For instance, the top seven correlated words with smog for New Yorkers prove to be kill, bad, environment, remember, place, need, and protect ($r > 0.20$). However, people in London tend to use very different words along with smog in same tweet. Top correlated terms with smog for Londoners turn out to be cap, ice, research, reveal, event, airpocalypse, and melt ($r > 0.20$). When people tweet regarding a specific issue, they usually use multiple terms in their tweets. Some terms are likely to co-occur in multiple tweets. Co-occurred terms in multiple tweets tend to reflect public perceptions toward a specific issue. Smog risks is no exception. Therefore, it is important for us to find top correlated terms with smog to detect public perceptions toward smog risks exactly.

6.4. Research Question 4

Public perceptions toward smog risks can be clustered into different topics. Each topic is composed of a group of related words which reflect public perceptions toward a specific aspect of smog risks. In this research, this author assumed that there may be big differences in the types of smog-related topics between New York and London. However, the results of this study partially support this author's assumption. For instance, New Yorkers' perceptions toward smog risks appear to be composed of perceptions toward smog sources, perceptions toward smog policies and actions, perceptions toward the health effect of smog 1, perceptions toward smog's effects on urban landscape, and perceptions toward the health effect of smog 2. However, Londoners' perceptions toward smog risks seem to be made up of perceptions toward smog policies and actions, perceptions toward the health effect of smog, perceptions toward global warming 1, perceptions toward the level of smog, perceptions toward research results, and perceptions toward global warming 2. In light of the above results, this author can conjecture that even though there are some differences in certain aspect of public perceptions, there exist considerable similarities in perceptions toward smog policies and perceptions toward the health effect of smog between New York and London. As can be seen from the results of this study, every city in the world seems to have both locally-oriented smog problems and globally-oriented smog problems which are sure to influence on public

perceptions on smog risks. Therefore, it is necessary that we should use both micro approaches and macro approaches to detect public perceptions toward smog risks exactly.

6.5. Research Question 5

In this research, this author assumed that there may be big differences in smog-related topics changing patterns over time between New York and London. As expected, the results of this study indicate that there are big differences in smog-related topics changing patterns between New York and London. In the case of New York, some topics have grown sharply on a particular date. For example, there was a big change in the number of tweets regarding topic 6 between March 8 and March 10. As for London, there appeared to be fluctuations in the number of tweets regarding some topics. For instance, there was a couple of fluctuations in the number of tweets regarding topic 7 between March 15 and March 16. As can be seen above, public perceptions toward smog risks are likely to change over time. When a smog-related incident happens, public perceptions toward smog tend to fluctuate. Both global incidents and local incidents are likely to have impacts on public perceptions toward smog risks, which can in turn influence on smog-related topics changing patterns. Therefore, it is important that government policy makers and scholars should analyze smog-related topic changing patterns exactly to understand the nature of public perceptions toward smog risks.

7. Conclusion

This study aimed to detect and classify public opinions toward smog on Twitter using topic modeling. To help attain these goals and to find gaps in the literature, this study performed a literature review on public perceptions toward smog risks and topic modeling. This author's literature review revealed that there are considerable gaps in the related literature. In this study, this author developed five research questions to fill the gaps in the existing body of knowledge. And then this study carried out such research procedures as data extraction, word cloud analysis on the cleaned data, building the network of terms, correlation analysis, hierarchical cluster analysis, topic modeling with the LDA, and stream graphs to answer those research questions.

In Research Question 1, it was presumed that there may be big differences in the most frequent words concerning smog between New York and London. As assumed, the results of this research showed that there exist huge differences in the most frequent words regarding smog risks between New York and London. In Research Question 2, this author assumed that there may be big differences in the shapes of terms network between New York and London. As presumed, the results of this study demonstrated that there exist huge differences in the shapes of terms network between New York and London. In Research Question 3, this author supposed that there may be big differences in the type of correlation between New York and London. As expected, the results of this study revealed that

there exist huge differences in the type of correlation between New York and London. In Research Question 4, this author assumed that there may be big differences in the types of smog-related topics between New York and London. However, the results of this study partially supported this author's assumption. In Research Question 5, this author assumed that there may be big differences in smog-related topics changing patterns over time between New York and London. As assumed, the results of this study indicated that there exist huge differences in smog-related topics changing patterns between New York and London. As can be seen from the results of this study, this author could find positive answers to the four of the five research questions and a partially positive answer to Research question 4.

There are some fields in which this research can contribute to the existing body of knowledge. Many of the studies on public perceptions toward smog have been performed using offline data. Previous studies utilized questionnaire surveys or interview methods to measure public opinions toward smog. Traditional techniques are sure to cost researchers lots of time and money to collect data. Today a lot of people use Twitter or Facebook to express their perceptions toward smog risks. Public opinions toward smog are likely to change over time. Traditional techniques can't detect the fluctuations of public perceptions toward smog risks rapidly. However, few studies have ever been carried out to detect and classify public opinions toward smog risks on Twitter. Few studies have ever utilized topic modeling with LDK to detect and classify public opinions toward smog risks. Therefore, this study may be the first attempt to detect and classify public opinions toward smog risks using Twitter data. And this research may be the first try to analyze smog-related topics changing patterns over time.

Based on the results of this study, this author suggested the effective means of measuring public perceptions toward smog risks. First, it was advised that topic modeling and word cloud analysis should be used to detect public perceptions toward smog risks. Topic modeling and word cloud analysis can be a reasonable alternative to traditional techniques such as questionnaire survey and interview method which cost researcher huge money and time. Second, it is suggested that policy makers should use network of terms to detect public perceptions toward smog risks. Words in the network are likely to reflect public opinions toward smog. Hence, the shape of words network can be the epitome of public opinions of smog risks. Third, it is advised that government policy makers should find top correlated terms with smog to detect public perceptions toward smog risks exactly. Some words tend to co-occur in many tweets. Co-occurred words in many tweets are likely to reflect people's opinions toward smog risks. Forth, it is suggested that policy makers use both micro approaches and macro approaches to measure public opinions toward smog exactly. Every city seems to have both locally-oriented smog problems and globally-oriented smog problems which are sure to have effects on public opinions toward smog risks. Finally, it is advised that government policy makers should detect smog-related topic changing patterns exactly to cope with smog risks effectively. Accidents are likely to have impacts on public

opinions toward smog risks, which can in turn affect smog-related topics changing patterns. When a smog-related incident happens, public perceptions toward smog tend to fluctuate. Both global incidents and local incidents are likely to have impacts on public perceptions toward smog risks, which can in turn influence on smog-related topics changing patterns. Therefore, it is important that government policy makers and scholars should analyze smog-related topic changing patterns exactly to understand the nature of public perceptions toward smog risks.

This study has some limitations which need to be resolved in future research. First, this study only used Twitter data to detect and classify public perceptions toward smog risks. Today many people in the world use social media such as Facebook and Instagram to express their own opinions regarding a specific topic. Only using Twitter data can cause the representativeness problem of sample. Therefore, in future study, other social media data should be utilized to detect and classify public perceptions toward smog risks correctly. Second, this research analyzed only the ten-day period changing patterns of public perceptions toward smog risks. Therefore, it was difficult for this author to detect public perceptions changing patterns exactly. Hence, in future research, long-period data should be used to detect the changing patterns of public perceptions toward smog risks correctly.

참고문헌

References

Aldahawi HA. 2015. *Mining and analysing social network in the oil business: Twitter sentiment analysis and prediction approaches*. [dissertation]. Cardiff University.

Akerlof K, DeBono R, Berry P, Leiserowitz A, Roser-Renouf C, Clarke KL., Maibach EW. 2010. Public perceptions of climate change as a human health risk: surveys of the United States, Canada and Maltaz. I*nternational journal of environmental research and public health*. 7(6):2559-2606.

Alghamdi R. & Alfalqi K. 2015. A Survey of Topic Modeling in Text Mining. I*nternational Journal of Advanced Computer Science and Applications*. 6(1).

Anandkumar A, Kakade SM, Foster DP, Liu YK, Hsu D. 2012. *Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation* (No. arXiv: 1204.6703).

Andrzejewski D, Zhu X, Craven M. 2009. *Incorporating domain knowledge into topic modeling via Dirichlet forest priors*. In Proceedings of the 26th Annual International Conference on Machine Learning. p. 25-32.

Arora R, Ravindran B. 2008. *Latent dirichlet allocation based multi-document summarization*. In Proceedings of the second workshop on Analytics for noisy unstructured text data. p. 91-97.

Arora S, Ge R, Halpern Y, Mimno DM, Moitra A, Sontag D, Zhu M. 2013. *A Practical Algorithm for Topic Modeling with Provable Guarantees*. In ICML. p. 280-288.

Asuncion HU, Asuncion AU, Taylor RN. 2010. *Software traceability with topic modeling*. In Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1 . p. 95-104.

Bicalho P, Pita M, Pedrosa G, Lacerda A, Pappa GL.. 2017. A general framework to expand short text for topic modeling. *Information Sciences*. 393:66-81.

Bickerstaff K, Walker G. 2001. Public understandings of air pollution: the 'localisation'of environmental risk. *Global Environmental Change*. 11(2):133-145.

Bisgin H, Liu Z., Fang H, Xu X, Tong W. 2011. Mining FDA drug labels using an unsupervised learning technique-topic modeling. *BMC bioinformatics*. 12(10): S11.

Blei DM. 2012. *Probabilistic topic models*. *Communications of the ACM*. 55(4):77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*. 3(Jan):993-1022.

Brechin SR. 2003. Comparative public opinion and knowledge on global climatic change and the Kyoto Protocol: the US versus the world?. *International Journal of Sociology and Social Policy*. 23(10): 106-134.

Brody SD, Zahran S, Vedlitz A, Grover H. 2008. Examining the relationship between physical vulnerability and public perceptions of global climate change in the United States. *Environment and behavior*. 40(1):72-95.

Chen J, Chen H, Pan JZ. 2016. *Semantic Reasoning for Smog Disaster Analysis*. In Description Logics.

Chen J, Chen H, Wu Z, Hu D, Pan JZ. 2017. Forecasting smog-related health hazard based on social media and physical sensor. *Information Systems*. 64:281-291.

Chen Z, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R. 2013. *Discovering coherent topics using general knowledge*. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management. p. 209-218.

Cheng P, We J, Marinova D, Guo X. 2017. Adoption of Protective Behaviours: Residents Response to City Smog in Hefei, China. *Journal of Contingencies and Crisis Management*. 1468-5973

Cody EM, Reagan AJ, Mitchell L, Dodds PS, Danforth CM. 2015. Climate change sentiment on twitter: an unsolicited public opinion poll. *PloS one*. 10(8):e0136092.

Crowe MJ. 1968. Toward a "definitional model" of public perceptions of air pollution. *Journal of the Air Pollution Control Association*. 18(3):154-157.

Dunlap RE. 1998. Lay perceptions of global risk: Public views of global warming in cross-national context. *International sociology*. 13(4):473-498.

Elliott SJ, Cole DC, Kruege P, Voorberg N, Wakefield S. 1999. The power of perception: Health risk attributed to air pollution in anurban industrial neighbourhood. *Risk analysis*. 19(4):621-634.

Foulds JR, Kumar SH, Getoor L. 2015. *Latent Topic Networks: A Versatile Probabilistic Programming Framework for Topic Models*. p. 777-786.

Fu G, Wang X. 2010. C*hinese sentence-level sentiment classification based on fuzzy sets*. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics. p. 312-319.

Gretarsson B, O'donovan J, Bostandjiev S, Höllerer T, Asuncion A, Newman D, Smyth, P. 2012. Topicnets: Visual analysis of large text corpora

with topic modeling. *ACM Transactions on Intelligent Systems and Technology*. 3(2):23.

Hall D, Jurafsky D, Manning CD. 2008. *Studying the history of ideas using topic models*. In Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics. p. 363-371.

Hofmann T. 1999. *Probabilistic latent semantic analysis*. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc. p. 289-296.

Hofmann T. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*. 42(1-2):177-196.

Hoffman M, Bach FR, Blei DM. 2010. *Online learning for latent dirichlet allocation*. In advances in neural information processing systems. p. 856-864.

Hong L., Davison BD. 2010. *Empirical study of topic modeling in twitter*. In Proceedings of the first workshop on social media analytics. p.80-88.

Hu Y, Boyd-Graber J, Satinoff B, Smith A. 2014. Interactive topic modeling. *Machine learning*. 95(3):423-469.

Huynh T, Fritz M, Schiele B. 2008. Discovery of activity patterns using topic models. In Proceedings of the 10th international conference on Ubiquitous computing. p. 10-19.

Iacus SM, Porro G, Salini S, Siletti E. 2015. S*ocial networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being*. arXiv preprint arXiv:1512.01569.

Ji X, Chun SA, Wei Z, Geller J. 2015. Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*. 5(1):1-25.

Jiang H, Lin P, Qiang M. 2015. Public-opinion sentiment analysis for large hydro projects. *Journal of Construction Engineering and Management*. 142(2): 05015013.

Jiang Y, Meng W, Yu C. 2011. *Topic sentiment change analysis*. In International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer Berlin Heidelberg. p. 443-457

Koltsova O, Koltcov S. 2013. Mapping the public agenda with topic modeling: The case of the Russian livejournal. *Policy & Internet*. 5(2):207-227.

Landauer TK, Foltz PW, Laham D. 1998. An introduction to latent semantic analysis. *Discourse processes*. 25(2-3): 259-284.

Lee H, Kim J, Choo J, Stasko J, Park H. 2012. *iVisClustering: An interactive visual document clustering via topic modeling*. In Computer Graphics Forum. Blackwell Publishing Ltd. 31(3):1155-1164.

Li J, Pearce PL, Morrison AM, Wu B. 2015. Up in Smoke? The Impact of Smog on Risk Perception and Satisfaction of International Tourists in Beijing. I*nternational Journal of Tourism Research*. 10:2055.

Liu B, EDU U. 2014. *Topic modeling using topics from many domains*. lifelong learning and big data.

Lu Y, Zhai C. 2008. *Opinion integration through semi-supervised topic modeling*. In Proceedings of the 17th international

conference on World Wide Web. p. 121-130.

Macnaghten P, Grove-White R, Jacobs M, Wynne B. 1995. *Public perceptions and sustainability in Lancashire. Indicators, Institutions, Participation*. A report by the Centre for the Study of Environmental Change commissioned by Lancashire County Council.

Mehrotra R, Sanner S, Buntine W, Xie L. 2013. *Improving lda topic models for microblogs via tweet pooling and automatic labeling*. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. p. 889-892.

Mei Q, Cai D, Zhang D, Zhai C. 2008. *Topic modeling with network regularization*. In Proceedings of the 17th international conference on World Wide Web. p. 101-110.

Min Z, Jianping W. 2015. Visualization Analysis on Contemporary Youth's Haze Sentiment. *Youth Studies*. 4:006.

Montague JJ. 2016. *Using Visual Communication Design To Optimize Exploration of Large Text-Mining Datasets*. [dissertation]. University of Alberta.

Nguyen AT, Nguyen TT, Nguyen TN, Lo D, Sun C. 2012. Duplicate bug report detection with a combination of information retrieval and topic modeling. In Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering. p. 70-79.

Pang B, Lee L, Vaithyanathan S. 2002. T*humbs up?: sentiment classification using machine learning techniques*. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Association for Computational Linguistics. 10:79-86.

Paul MJ, Dredze M. 2014. Discovering health topics in social media using topic models. *PloS one*. 9(8):e103408.

Pingclasai N, Hata H, Matsumoto KI. 2013. *Classifying bug reports to bugs and other requests using topic modeling*. In Software Engineering Conference (APSEC), 2013 20th Asia-Pacific. 2:13-18.

Ponweiser M. 2012. L*atent Dirichlet allocation in R*.

Qin Z, Cong Y, Wan T. 2016. Topic modeling of Chinese language beyond a bag-of-words. *Computer Speech & Language*. 40:60-78.

Ramage D, Hall D, Nallapati R, Manning CD. 2009. *Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora*. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. Association for Computational Linguistics. p. 248-256.

Ritter A, Etzioni O. 2010. *A latent dirichlet allocation method for selectional preferences*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. p. 424-434.

Saksena S. 2007. *Public perceptions of urban air pollution with a focus on developing countries*.

Saksena S. 2011. Public perceptions of urban air pollution risks. *Risk, Hazards & Crisis in Public Policy*. 2(1):1-19.

Sang ETK. 2014. *Using tweets for assigning sentiments to regions*. In Proc. of the International Workshop on Emotion, Social Signal, Sentiment & Linked Open Data.

Semenza JC, Wilson DJ, Parra J, Bontempo BD, Hart M, Sailor DJ, George LA. 2008. Public perception and behavior change in relationship to hot weather and air pollution. *Environmental research*. 107(3):401-411.

Sha Y, Yan J, Cai G. 2014. *Detecting public sentiment over PM2. 5 pollution hazards through analysis of Chinese microblog*. In ISCRAM: The 11th International Conference on Information Systems for Crisis Response and Management. p. 722-726.

Shatnawi S, Gaber MM, Cocea M. 2014. Text stream mining for Massive Open Online Courses: review and perspectives. Systems Science & Control Engineering: *An Open Access Journal*. 2(1):664-676.

Sluban B, Smailovic J, Juric M, Mozetic I, Battiston S. 2014. *Community sentiment on environmental topics in social networks*. In Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on. p. 376-382.

Sun C, Yuan X, Yao X. 2016a. Social acceptance towards the air pollution in China: Evidence from public's willingness to pay for smog mitigation. *Energy Policy*. 92:313-324.

Sun C, Yuan X, Xu M. 2016b. The public perceptions and willingness to pay: from the perspective of the smog crisis in China. *Journal of Cleaner Production*. 112:1635-1644.

Sun L, Yin Y. 2017. Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C: Emerging Technologies*. 77:49-66.

Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. 2016. Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. *Journal of Medical Internet Research*. 18(8):e232.

Tan S, Li Y, Sun H, Guan Z, Yan X, Bu J, He X. 2014. Interpreting the public sentiment variations on twitter. *ieee transactions on knowledge and data engineering*. 26(5):1158-1170.

Tang J, Jin R, Zhang J. 2008. *A topic modeling approach and its integration into the random walk framework for academic search*. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference. p. 1055-1060.

Titov I, McDonald R. 2008. *Modeling online reviews with multi-grain topic models*. In Proceedings of the 17th international conference on World Wide Web. p. 111-120.

Tuarob S, Pouchard LC, Giles CL. 2013. *Automatic tag recommendation for metadata annotation using probabilistic topic modeling*. In Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. p. 239-248.

Wallach HM. 2006. *Topic modeling: beyond bag-of-words*. In Proceedings of the 23rd international conference on Machine learning. p. 977-984.

Wang C, Blei DM. 2011. *Collaborative topic modeling for recommending scientific articles*. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. p. 448-456.

Weber EU, Stern PC. 2011. Public understanding of climate change in the United States. *American Psychologist*. 66(4):315.

Yan J, Zeng J, Liu ZQ, Yang L., Gao Y. 2016. Towards

big topic modeling. *Information Sciences*. 390:15-31.

Yang S, Shi L. 2016. Public Perception of Smog: A Case Study in Ningbo City, China. *Journal of the Air & Waste Management Association*. (just-accepted).

Yousefpour A, Ibrahim R, Hamed HNA, Hajmohammadi MS. 2014. A comparative study on sentiment analysis. *Advances in Environmental Biology*. 53-69.

Yoon HG, Kim H, Kim CO, Song M. 2016. Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling. *Journal of Informetrics*. 10(2):634-644.

Yu X. 2016. *Noise Levels Associated with Sentiment Analysis on Twitter: A Case Study of New York City* [dissertation]. Tufts University.

Zhang D, Guo B, Yu Z. 2011. The emergence of social and community intelligence. *Computer*. 44(7):21-28.

Zhai K, Boyd-Graber J, Asadi N, Alkhouja ML. 2012. *Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce*. In Proceedings of the 21st international conference on World Wide Web. p. 879-888.

Zhao Y. 2013. *Analysing twitter data with text mining and social network analysis*. In Proceedings of the 11th Australasian Data Mining and Analytics Conference.

Zhao W, Zou W, Chen JJ. 2014) Topic modeling for cluster analysis of large biological and medical datasets. *BMC bioinformatics*. 15(11):S11.

Zhou Y, Lu T, Zhu T, Chen Z. 2016. *Environmental Incidents Detection from Chinese Microblog Based on Sentiment Analysis*. In International Conference on Human Centered Computing Springer International Publishing. p. 849-854.

2017년 4월 13일 원고접수(Received)
2017년 6월 07일 1차심사(1st Reviewed)
2017년 6월 20일 게재확정(Accepted)

초 록

본 연구의 주된 목적은 토픽 모델링(topic modeling)을 이용하여 트위터 상에서 스모그 리스크(smog risks)에 관한 대중 인식(public perceptions)을 측정하고 분류하는 것이다. 선행연구에 있어서 연구 갭(research gap)을 확인하기 위하여 본 연구는 스모그 리스크와 토픽 모델링에 대한 선행연구를 검토하였다. 그 결과 본 저자는 기존의 연구에서 상당한 연구 갭이 존재하고 있음을 확인하였으며, 이러한 연구 갭을 메우기 위해 다섯 개의 연구 질문을 설정하였다. 연구 질문들에 답을 구하기 위하여 본 연구는 10,000개의 트위터 자료를 추출하였고, 이에 대하여 워드 클라우드 분석(word cloud analysis), 상관분석, LDA를 이용한 토픽 모델링, 스트림그래프(stream graph), 위계적 집락분석(hierarchical cluster analysis)을 실시하였다. 분석 결과 자주 언급되는 단어들(the most frequent terms), 단어네트워크(terms network)의 형태, 상관관계의 유형, 스모그 관련 주제의 변동패턴에 있어서 뉴욕과 런던 사이에 큰 차이가 있음을 확인하였다. 그리하여 본 저자는 다섯 개의 연구 질문 중 네 개에 대하여 긍정적인 답을 구할 수 있었고, 이를 토대로 몇 가지 정책적 시사점을 제시하고, 향후 연구를 위한 제안들을 하였다.

주요어 : 대중인식, 스모그 리스크, 토픽모델링, LDA, 스트림그래프