A Comparative study on Data Crawling and Extraction of Climate Change Issues Using Machine Learning Technique

Do-Yeon Kim, Dae-Yong Jin, Kuk-Jin Han, Seong-Taek Park

Abstract*: Many different problems are triggered in different fields such as society, culture, and economy due to constant climate change problems. As time goes by, its influences are mounting and national attention is increasing. Therefore, it is necessary to understand various issues and improve policies on climate change. If it is possible to analyze information from media outlet data created on real time by using text mining technique, various climate change issues can be understood. In this comparative study, therefore will collect news article data related to climate change, identify issues utilizing text mining, and see complex information through the detailed analysis considering the characteristics of the text. We crawled news related to climate change issues and analyzed related keywords in terms of cause, result (phenomenon), and response. First, we extracted news related to climate change by using keyword-based document extraction method and Latent Dirichlet Allocation (LDA)-based document extraction method. In addition, we propose four related keyword analysis methods using Word2Vec, which is one of word embedding methods, and keyword frequency based method. Methods proposed in this comparative study are expected to be used in extracting and analyzing data on other specific issues not upcoming climate change issues.

***Index Terms*: Climate Change, Machine Learning, Natural Language Processing, LDA, Word2Vec**

I. INTRODUCTION

The climate change issue is no longer a matter of particular region or country, but a global public material management issue that the entire world should take joint actions [1]. Accordingly, a variety of issues arisen from the climate change need to be understood and alternative policy is required [2]. Recently, with the development of ICT (Information and Communication Technology), many researches have been carried out to draw insights using big data in various fields. In particular, a variety of tools and methods for analyzing massive amounts of text data have been developed, reducing the time and costs involved.

As a result, numerous researches identify major issues and draw insights by analyzing the frequency and trend of keywords included in text data. However, research utilizing text mining in the environmental field remains at basic level. More research and foundation need to be conducted and set up because there are few study cases. In addition, strong points of text mining are that it automates data collection and analysis, reduce manual work, minimize manual effort needed for research through constant use, and continues to produce results. The previous analysis of environmental text primarily focused on analyzing results without considering how to build up text analysis algorithm, apply and use an array of methodologies. Accordingly, research is needed to analyze issues and specific contents according to the characteristics of each text by using text mining algorithm and utilization in order to draw various issues.

This comparative study explores text mining methods to extract key climate environmental issues. We collect environmental texts for analysis of environmental issues and understand and investigate what results can be achieved through the construction and use of analysis algorithms. In particular, we collect environmental texts from the archive relevant to climate change, analyze them according to text characteristics, and identify environmental problems and draw various insights.

II. MATERIALS AND METHODS

This comparative study consists of two main streams: (1) the construction of environmental text analysis framework available for analyzing climate and environmental issues constantly and (2) analyzing results by using them and drawing climate change issues. An environmental text analysis framework in this study is shown in [Figure 1] below.

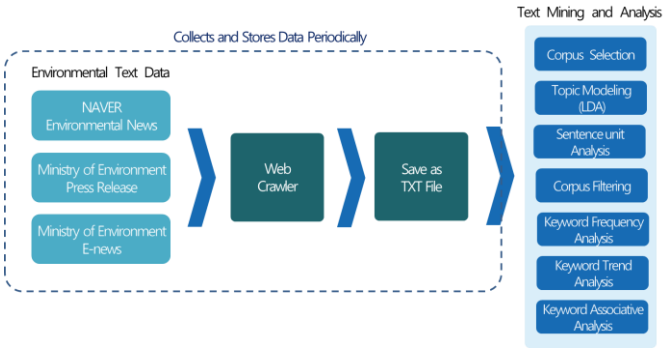

Fig 1. Environmental Text Data Analysis Framework

Revised Manuscript Received on May 22, 2019.

Do-Yeon Kim, Researcher, Korea Environment Institute, 370, Sicheong-daero, Sejong-si, 30147, Republic of Korea
***Dae-Yong Jin**, Research Fellow, Korea Environment Institute, 370, Sicheong-daero, Sejong-si, 30147, Republic of Korea
Kuk-Jin Han, Researcher, Korea Environment Institute, 370, Sicheong-daero, Sejong-si, 30147, Republic of Korea
Seong-Taek Park, Professor, Sungkyunkwan University, 25-2, Seonggyungwan-ro, Seoul, 03063, Republic of Korea

First, it represents the procedure of automating the collection of text data. Crawling for NAVER Environmental News, press released by the Ministry of Environment and e-Environmental News released by the Ministry of Environment is built up and collects daily updated Environmental news articles are regularly collected by using this device. It aims to use them constantly in other studies on environmental texts through the automation of collecting environmental text data. Then it moved on to construct various algorithms performing text analysis based on collected data. Specifically, an environment text analysis framework is built containing diverse functions such as document filtering through date or content, topic modeling that understands document themes, analysis of keyword frequency, analysis of keyword trend, Word Embedding, and extracting sentences containing keyword. Issues, trends, and specific contents associated with the climate change are analyzed by utilizing a variety of text mining techniques. A variety of insights can be drawn by utilizing various texts, considering text characteristics, and applying the algorithm. Lastly, organize results of text analysis, summarize major climate environmental issues, and propose use cases as a policy.

A. LDA Topic Modeling

Topic Modeling is a technique for grouping similar words and used for identifying themes from many documents. LDA model, which is one of the most commonly used topic modeling methods is a statistical algorithm for identifying potential themes embedded in extensive and non-standardized documents [3]. The LDA model is unsupervised learning algorithm that does not need to label documents. And it's a bayesian model of three layers of document-topic-word. These LDA model outperform better than traditional topic modeling algorithm [4].

LDA is a method that calculates document-topic matrix and topic-word matrix capable of describing the distribution of words in documents, namely document-word matrix by setting up the number of topics (K). It represents the probability that each document is included in a topic and the probability that each word is included in a topic based on this matrix. Final results are words contained in each topic. This algorithm determines topic based on the subjective judgment of researchers by seeing contained words in the topic. In addition to this, Cao et al(2009) [5] proposed a method of minimizing the association with topics by calculating the distance between topics in order to determine the number of appropriate topics. Zhao et al. (2015) [6] contributes to various studies by proposing methods that use perplexity and cross-validation to select the appropriate number of topics.

As shown Figure 2 below, it represents a process for LDA, aiming at identifying latent variables such as θ, z, β through observed variable prospect value. In this paper, LDA was used to extract documents pertaining to climate change clusters from the collected entire environment news data.

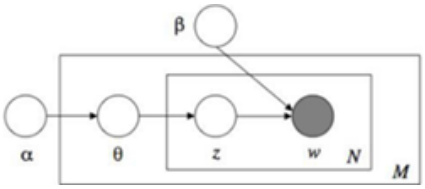

Fig 2. LDA Model [3]

B. Word2Vec

Word2Vec is a method that represents all the keywords of the learning data as a vector considering the surrounding context. Using these characteristics, the meaning of the context can be quantified and vectorized as shown in Figure 3 below [7]. Then, from the word vector shown as a Word2Vec result for each word, use cosine similarity to calculate the similarity between the words as in formula (1).

$$\text{Cosine Similarity}(u,v) = \frac{\sum_{i=1}^{n}(u_i \cdot v_i)}{\sqrt{\sum_{i=1}^{n}(u_i)^2}\sqrt{\sum_{i=1}^{n}(v_i)^2}} \quad (1)$$

u_i : i element of word vector u, v_i : i element of word vector v

The figure below is a two-dimensional vector space that reflects semantic information on the relationship between words using t-SNE (Stochastic Neighbor Embedding) dimensionality reduction techniques [8].

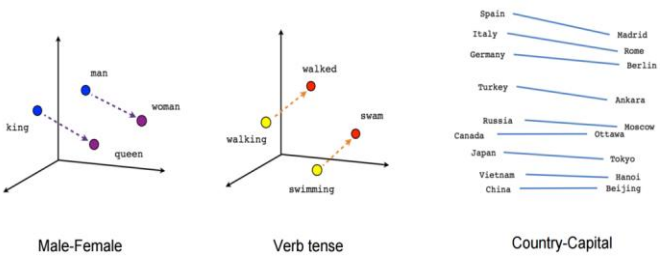

Fig 3. Examples of vector expressions between words [8]

Learning method of Word2Vec include skip-gram and CBOW. The window of adjoining keywords and dimension of vector space should be defined when it comes to learning. CBOW is a model for predicting certain words using words that appear around them. And Skip-gram is a model for predicting words that appear around a certain word[9]. CBOW is suitable for smaller data set, and Skip-gram is suitable for lager data set. This paper used cosine relevance on the vector space by using CBOW method.

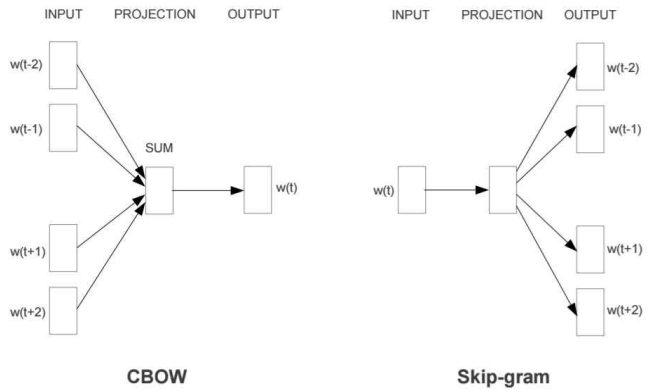

Fig 4. CBOW and Skip-gram model structure [9]

III. RESULTS AND DISCUSSION

A. Analysis of Environmental News

1) Collection of Climate Change News

In this study 220,098 news on environment from 2005 to 2017 were collected provided at Naver website (http://www.naver.com) by utilizing Python-Beautifulsoup for text mining analysis on the climate change issues. Overlapped articles were eliminated and eventually 197,757 news articles were used. As shown in Figure 5, it used documents pertaining to climate change clusters resulting from LDA and documents containing keywords associated with climate change in order to extract only documents related with climate change.

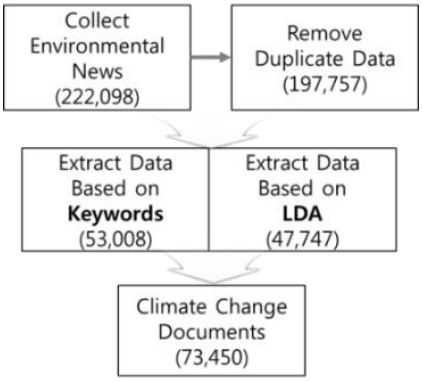

Fig 5. Process of Extracting Documents on Climate Change : Environmental News

2) Analysis of Related Keywords on Climate Change

This study analyzed climate change issues in terms of cause, result (phenomenon), and response Specifically, 'Greenhouse gases' corresponding to the causes of climate change, and 'Heat wave', 'Cold wave' and corresponding to the results were analyzed. We also analyzed ‘energy’ and ‘Eco-friendly’ which related to climate change response. This study used four methods to analyze related keywords [Figure 6]. First, it used word2vec’s Cosine Similarity and the other one is a method based on the probability of Word2Vec. When specific keywords are given, it identifies keywords with the high probability of finding the keywords nearby. Third, it extracts a sentence containing a specific keyword, perceives a noun, and ranks it, utilizing sentence unit analysis. The last one uses document unit analysis.

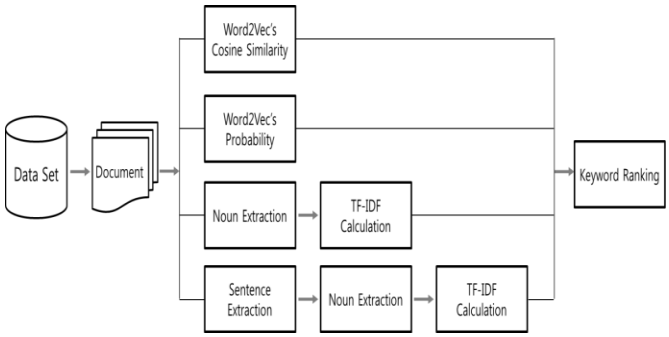

Fig 6. Four Related Keywords Analysis Process

It cannot be said that neither of them is excellent in terms of performance. They can vary according to the size and characteristics of corpus. Therefore, this study used all four methods and presented results. The table below shows parts of results from analyzing related words.

Table 1. Keywords related with Greenhouse gas

Rank	Method 1	Method 2	Method 3	Method 4
1	Carbon dioxide	Reduction	Climate change	Green
2	Carbonic acid	Emission amounts	Emission amounts	Eco-friendly
3	Greenhouse gas	Emission	Energy	Reduction goal
4	Full shot	Reduction amounts	Reduction goal	Trading
5	Target quantity	Generating amounts	Carbon dioxide	China
6	Hot air	Emission trading system	Green	Advanced country

Table 2. Keywords related with Heat Wave

Rank	Method 1	Method 2	Method 3	Method 4
1	Cold wave	Special weather report	Advisory	Daegu
2	Heat wave	Advisory	Special weather report	Gwangju
3	Dryness	Warning	Daegu	Tropical night
4	Heavy snow	Killing	Tropical night	Damage
5	Heavy rain	First aid	Gwangju	Gangwon
6	Storm surge	Tropical night	Gyeongbuk	Gyeongbuk

Table 3. Keywords related with Cold Wave

Rank	Method 1	Method 2	Method 3	Method 4
1	Heat wave	Special weather report	Special cold wave report	Gwangju
2	Cold	Advisory	Gangwon	Special weather report
3	Heavy snow	Attack	Lowest	Jeju
4	Recurrence of cold	Scholastic ability test	Gyeonggi	Daegu
5	Jack Frost	Peak	Gyeongbuk	Chuncheon
6	Heavy rain	Strong	Gangwon	Gyeonggi

Table 4. Keywords related with Energy

Rank	Method 1	Method 2	Method 3	Method 4
1	Renewable energy	Saving	Greenhouse gas	Green
2	Fuel	Renewable	Renewable energy	Carbon
3	Electric power	Self-reliance	Saving	Greenhouse gas reduction
4	Alternative energy	Consumption	Climate change	Goal
5	Ascon	Fossil	Green	Eco-friendly
6	Energy policy	Clean	Eco-friendly	Carbon dioxide

Table 5. Keywords related with Eco-friendly

Rank	Method 1	Method 2	Method 3	Method 4
1	Green	Product	Energy	Climate change
2	Innovation	Driving	Green	City
3	Eco	Goods	Carbon	Automobile
4	Idea	Certification	Greenhouse gas	Carbon dioxide
5	Local	Architecture	Automobile	Fine dust
6	Food	Town	City	Greenhouse gas

B. Analysis of Ministry Environment News

1) Collection of Climate Change News

In this study, ministry of environment press release from 2005 to 2017 and e-Environmental News from 2010 to 2017 were collected provided at ministry of environment website (www.me.go.kr) by utilizing Python-Beautifulsoup for text mining analysis on the climate change issues. As shown in Figure 7-8, it used documents pertaining to climate change clusters resulting from LDA and documents containing keywords associated with climate change in order to extract only documents related with climate change.

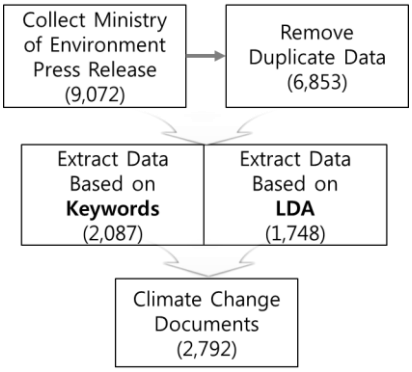

Fig 7. Process of Extracting Documents on Climate Change : Ministry of Environment Press Release

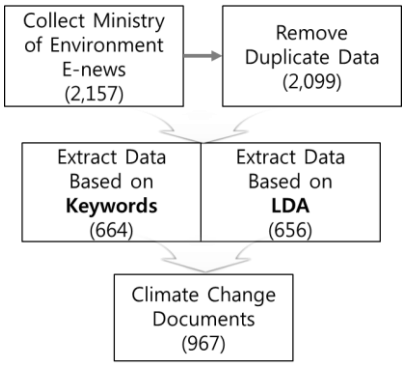

Fig 8. Process of Extracting Documents on Climate Change : Ministry of Environment e-Environmental News

2) Analysis of Related Keywords on Climate Change

Similar to Naver environmental news analysis, the four key word analysis methods developed in this study were used to perform associated keyword analysis on 'Greenhouse gas', 'Heat wave', 'Cold wave' 'Energy' and 'Eco-friendly' respectively. The table below shows parts of results from analyzing related words.

Table 6: Keywords related with Greenhouse gas

Rank	Method 1	Method 2	Method 3	Method 4
1	Coefficient	Reduction	Greenhouse gas emission	Certification
2	Carbon	Emission quantity	Goal	Convention
3	Model	Management system	Energy	Car
4	Offset	Energy	Carbon	Purchase
5	Goal	Emission trading system	Green	Reduction
6	Transaction	Reduction volume	Climate change	Green Growth

Table 7: Keywords related with Heat Wave

Rank	Method 1	Method 2	Method 3	Method 4
1	Typhoon	Temperature	Health	Report
2	Summertime	Summertime	Typhoon	National park
3	Meteorological Disasters	Reduction	Flood	Visit
4	Sewol ferry	Damage	Heavy rain	Safety
5	Crop	Danger	Cold wave	Life
6	Humidity	Flood	City	Vulnerability

Table 8: Keywords related with Cold Wave

Rank	Method 1	Method 2	Method 3	Method 4
1	Heavy snow	Temperature	Heavy snow	Goat
2	Heavy rain	Summertime	Vulnerable	Climate Change
3	Freezing ice	Wintertime	Winter	Saemangeum
4	Mating	Springtime	Heavy snow	Hiking
5	Below zero	Breeding	Food	Accident
6	Summertime	Heavy snow	Goat	Safety

Table 9: Keywords related with Energy

Rank	Method 1	Method 2	Method 3	Method 4
1	Renewable energy	Reduction	Greenhouse gas	Certification
2	Reduction	Greenhouse gas	Resource	Car
3	Transportation	Eco-friendly	Green	Life
4	Saving	Efficiency	Waste	Gas
5	Electricity	Comsumption	Eco-friendly	Agreement
6	Biomass	Waste	Reduction	Green

Table 10: Keywords related with Eco-friendly

Rank	Method 1	Method 2	Method 3	Method 4
1	Green	Driving	Environment-friendly product	Certification
2	Consumer	Product	Eco-friendly driving	Car
3	Habit	Building	Green	Living
4	Promotion	Car	Certification	Gas
5	Living	Living	Energy	Agreement
6	Green Revolution	Consumption	Consumption	Green

IV. CONCLUSION

This comparative study collected data on various environmental news, extracted data associated with climate change from collected data, and conducted text mining analysis. The analysis results showed that NAVER Environmental News showed contents associated with causes and response to the climate change such as "energy", "greenhouse gas", "carbon" and "environment-friendly" and specific phenomenon (disaster and catastrophe) from the climate change such as "typhoon", "heat wave", "drought", and "flood".

However, press release and e-Environmental News from the Ministry of Environment mainly involve causes and responses to the climate change. For example, Naver Environmental News contains an abundance of documents associated with "heat wave". Also, "Daegu", "Gwangju", "Gyeongbuk", "Gangwon" and "Jeju" were frequently mentioned in terms of region. A variety of terms such as tropical night and simmering heat are covered in articles.

In addition, casualties such as "disease", "heatstroke", "sunstroke" and "mortality rate", "heat wave warning", "alert", "excessive heat warning" and fundamental causes of heat including "climate change", "global warming", "abnormal climate" were frequently mentioned.

On the contrary, press release from the Ministry of Environment included only 65 sentences with "heat wave" which contain small amounts of data. It is inappropriate for analyzing and keyword frequency and trends. Therefore, it adopted algorithm extracting sentences containing certain keyword and the function of summarizing documents for analysis. This analysis showed that "heat wave" was considered as just a one of phenomenon in the climate change and described with its fundamental contents such as "climate change" and "greenhouse gas" in the press release.

In addition to this, an analysis of associated keywords on "greenhouse gas," "environment-friendly," and "energy" pertaining to causes and responses to the climate change showed that distinct important issues could be understood both from NAVER Environmental News and press release. In other words, they covered news in various standpoints regarding the topic of "climate change" according to media outlets. Thus, it is better for environmental policy makers to use the results of environmental issue analysis derived from various media when they establish policies. If further studies analyze various medium and perspectives such as Twitter, Facebook, academic paper, and research report based on the algorithm and framework constructed in this comparative study, more efficient environmental policy are expected to be established.

ACKNOWLEDGMENT

This study was conducted following the research work 「Big Data Analysis: Application to Environmental Research and Service (GP2018-13)」and was funded by the Korea Environment Institute.

REFERENCES

1. R. H. Moss, J. A. Edmonds, K. A. Hibbard, M. R. Manning, and S. K. Rose, D. P. Van Vuuren, et al. “The next generation of scenarios for climate change research and assessment,” Nature. 2010, 463(7282), 747.
2. A. Giddens, “Politics of climate change,” Polity; 2009.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of machine Learning research. 2003 3(Jan), pp. 993–1022.
4. C. G. Chiru, T. Rebedea, and S. Ciotec. Comparison between LSA-LDA-Lexical Chains. InWEBIST (2). 2014, pp. 255–262.
5. J. Cao, T. Xia, J. Li, Y. Zhang, and S. A. Tang. “density-based method for adaptive LDAmodel selection,” Neurocomputing. 2009, 72(7), pp. 1775-1781.
6. W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, and Y. Ding, et al. “A heuristic approach to determine an appropriate number of topics in topic modeling,” BMC bioinformatics. 2015, 16(13), p.S8
7. A. Juyoung, B. Junghwan, H. Namgi, and S. A. Min. “Study of ‘Emotion Trigger’ by Text Mining Techniques,” Korea Intelligent Information System Society, 2015, pp. 59-81.
8. T. Mikolov, W. T. Yih, G. Zweig. “Linguistic regularities in continuous space word representations,” In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp.746-751.
9. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.

AUTHORS PROFILE

Do-Yeon Kim received Master’s Degree of Business Administration from Chungbuk National University, South Korea. She is currently Researcher at Korea Environment Institute (KEI). Her present research interests include: Big Data, Machine Learning, Text Mining, Artificial Emotion Intelligence, Technology Innovation, MIS

Dae-Yong Jin received Ph.D. Degree in Electrical Engineering and Computer Science from Gwangju Institute of Science and Technology, South Korea. He is currently Research Fellow at Korea Environment Institute (KEI). His present research interests include: Big Data, Text Mining, Machine Learning, Bioinformatics, Computational Biology and Statistical Analysis.

Kuk-Jin Han received Master's Degree in Software Engineering from Sogang University, South Korea. He is currently Data Engineer at Korea Environment Institute (KEI). His present research interests include: Big Data, AI, Automation and Optimization, Computer Science and Software Engineering.

Seong-Taek Park received Ph.D. Degree in Management Information Systems from Chungbuk National Univeristy, South Korea. He is currently a Professor at The Korea Association of Software Manpower. His present research interests include: Big Data, Cloud Computing, Technology Innovation, MIS, Approriability Mechanism, Patent Valuation.