International Journal of Pure and Applied Mathematics
Volume 120 No. 6 2018, 5793-5807
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

Analysis of Recognition of Climate Changes using Word2Vec

Do-Yeon Kim[1] and Sung-Won Kang[*2]
[1,*2]Korea Environment Institute,
370, Sicheong-daero, Sejong-si,
Republic of Korea, 30147
dykim@kei.re.kr, swkang@kei.re.kr

Abstract

In this study, we discovered significant differences between the content of two distinct media (academic research and newspaper) regarding climate changes by performing word2vec analysis. For the academic research text data, we digitized the research reports from the Korea Environment Institute; and for the newspaper text data, we collected environmental news articles from the Naver portal service using jsoup: JAVA HTML parser. We applied a two-step approach. In the first step, we selected three keywords (global warming, flooding, and drought) related to "climate change" by performing word2vec skip-gram analysis. Further, we performed word2vec skip-gram analysis again to find words closely related to these three keywords. Our analysis showed that words on human beings and domestic damage were the words closely related to climate change in the academic research report. In contrast, words on collateral damage (food and agriculture) and catastrophic events were the words closely related to climate change in the environmental news. Our result reveals subtle yet significant difference between these two media in their attitude toward climate change. The research papers appear to be more focused on the effect of climate change on the quality of life of human beings, whereas environmental news appears to be more focused on the environmental and financial damage

caused owing to extreme climatic events.

Key Words : Climate Change, Machine Learning, Natural Language Processing, Word Embedding, Word2Vec.

1 Introduction

The climate change phenomena, such as global warming, have emerged as a global issue and are expected to have significant effects on the quality of life of the current and future generations [1]. Policies regarding climate changes require more scientific knowledge as compared to that required for general environmental issues. However, the manner in which the public perceive these problems is a crucial factor that affects government policy-making [2]. Further, public misconceptions about risks related to climate changes may be another big obstacle, because it takes a long time for a climate change-related policy to take effect [3]. Therefore, it is important to understand the public awareness of climate change before preparing a climate change-related policy. Recently, there have been many active domestic and foreign studies on the public awareness of climate change [46]. In Korea, it was revealed that most of the people perceived the risks of climate changes seriously [7,8]. Despite a high level of public awareness of climate change, there is an absence of sufficient change in their practical behavior till now. The previous studies analyzing awareness of climate change have been conducted using conventional methods, such as surveys and interviews with experts. However, these conventional methods have limitations in understanding the public awareness of climate change in detail. Further, the number of samples for analysis using conventional methods is small, and it consumes too much time and money.

Thus, this study analyzed environmental news reports to investigate the stance on the demand for environment policies. For analyzing the stance on the supply of environmental policies, the reports published by an institute researching environment policies were analyzed. To closely examine the interests in climate change-related phenomena, such as “global warming”, “flooding”, and “drought” this study comparatively analyzed recognition of climate changes by two media: environmental policy research reports issued by the

Korea Environment Institute (KEI) and environmental news of the Naver portal service, by using the machine learning tool skip-gram algorithm-based word2vec. KEI is an environmental policy research institute sponsored by Korean government and Naver is the leading internet portal service in Korea.

2 Word2Vec Methods

In the field of natural language processing (NLP), various methods to digitize words for computers to understand human language have been studied. Conventionally, "one-hot encoding" has been used to put "1" into a digit corresponding to a relevant word and put "0" into the other digits when there were no words. For example, a vector expressing an apple is [0, 0, 1, and 0] when the words given are [potato, strawberry, apple, and water melon]. However, this method is not efficient spatiotemporally, because this has a sparse problem generating too many "0," and it is difficult to understand differences between words.

Recently, for overcoming this problem, a method of vectorization for words in a multidimensional space was designed to understand the meanings of words. In 2013, Google engineers including Tomas Mikolov proposed word2vec, a technique used for word embedding [9]. Word2vec, which is a method to transform words into vectors, is a machine learning technique that analyzes similarity among words [9]. For the similarity, this study evaluates how close the distance is between words, by deriving the cosine similarity from word vectors abstracted by the word2vec technique [9]. Using cosine similarity is a very useful method to measure the similarity between two documents in multidimensional space, if only the condition of + space is met [10]. For two vectors A and B in n-dimensional space, with an angle, it is defined by Tan et al., 2005 [11] as follows:

$$similarity(A, B) = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sum_{i=1}^{n} A_i^2 \times \sum_{i=1}^{n} B_i^2} \quad (1)$$

Among the various models that embody the word2vec technique, the continuous bag-of-words (CBOW) and skip-gram models are shown in Figure 1. The structures of word2vec models consist

of input, projection, and output layers. The structure of skip-gram contrasts with that of CBOW model. In the CBOW architecture, the model predicts the current word from a window of surrounding context words. In the skip-gram architecture, the model uses the current word to predict the window of surrounding context words [12,13]. This study put keywords related to climate change phenomena, such as "global warming", "flood", and "drought" into the input by using the skip-gram model.

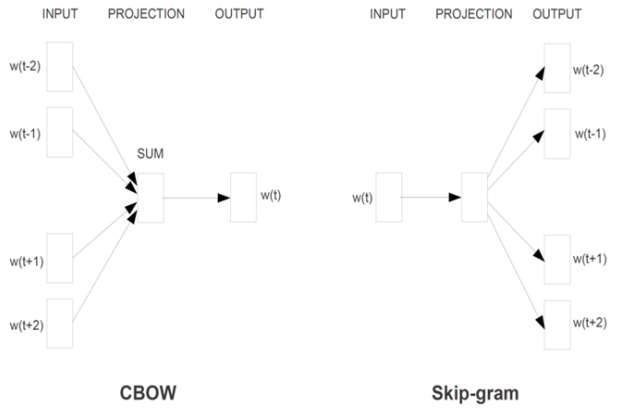

Figure 1: CBOW and Skip-Gram Architectures [12]

3 Results and Discussion

This study is focused on investigating "climate change" keywords in detail using word2vec. Word2vec is a machine learning technique used for analyzing the similarity between words by converting words to vectors [12]. In this study, the similarity between words is rated based on the distance between words after measuring the cosine similarity and cosine distance and word vectors extracted by word2vec. Cosine similarity is used to measure an angle between two vectors. A bigger value means there is a high similarity between words. When the range of value is between 01 and the angle is equal, the maximum value of similarity is 1.0. Cosine distance is calculated according to the formula "cosine distance = 1 cosine similarity." Therefore, contrary to cosine similarity, a lower value

means there is a close distance and a high similarity between words. In this study, word2vec analysis was performed using the skip-gram algorithm to find words around "climate change" keywords.

First, the skip-gram algorithm-based word2vec analysis of "climate change" keywords was performed, based on the abstracts and tables of contents of the research reports published by KEI between the period 1993-2016. As a result, a lot of climate change-related keywords, such as "sea level," "unusual weather," "warming," "flood," "typhoon," "cold wave," "heat island," and "drought" were discovered. Based on the result of this analysis and the researcher's judgment, three climate change-related phenomena, i.e., "warming," "flood," and "drought" were selected. Further, the skip-gram algorithm-based word2vec analysis on climate change-related phenomena was performed for each medium along with a comparative analysis of the same. The process of word2vec analysis with climate change-related keywords is shown in Figure 2.

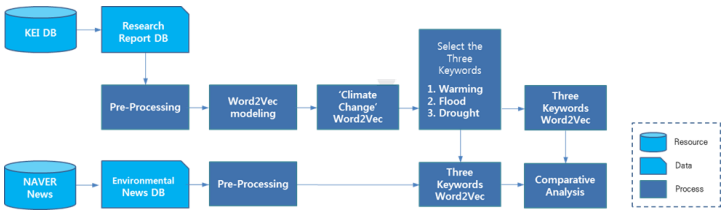

Figure 2: Analysis Process

3.1 Data Collection and Pre-Processing

3.1.1 KEI Research Reports

For analysis, a total of 1,697 KEI research reports (titles, tables of contents, summaries, and dates) of period between 1993-2016 were analyzed. For pre-processing to sort KEI research reports into natural language units, nouns were extracted first using a morphological analyzer package Korean Natural Language Processing (KoNLP) provided using R. Based on the result of this analysis, stop words including special characters and specific words were processed, and only those terms with two letters or more were extracted. Further, sparse terms with an extremely low frequency of appearance

were deleted, and those with low term frequency-inverse document frequency (TF-IDF) values were removed as well.

TF-IDF is a statistical digitization to show how important some words are in a specific document. TF in TF-IDF represents how often words appear in a document. DF refers to the number of documents with words, and IDF is an inverse number of DF. In other words, TF-IDF means a value multiplying TF by IDF. Finally, each data set was transformed to a document term matrix (DTM) that is appropriate for the analysis, using the tm package provided by R. DTM refers to the word frequency in a document and is expressed as an $N \times V$ matrix, as indicated in Table 1.

Table 1: DTM

	Term 1	Term 2	…	Term V
Document 1	0	2	…	1
Document 2	2	1	…	1
…				…
Document N	0	1	…	5

3.1.2 Naver Environmental News

In this study, environmental news articles provided by Naver(www.-naver.com) were collected using jsoup: Java HTML parser to analyze the environment news trends. The environment news data collected include news article titles, dates of publication, and press data from the society (environment) channel on the news section of Naver for the period between 2004-2016. A total of 193,636 data provided by 101 press companies were gathered (Table 2). Naver news started with a system to search news articles of 15 newspapers and press agencies in May 2000, and now, over 453 press companies are posting tens of thousands of articles each day. Therefore, the environment news with poor data posted on Naver before 2004 were excluded from this research.

Prior to text mining, for pre-processing to separate the environment news article data collected from Naver into natural language units, nouns were extracted primarily using a morphological analyzer package (KoNLP) provided by R. Based on the morpheme

Table 2: Naver Environmental News Data Range

	Content
Channel	Naver (www.naver.com)
Tool	Java HTML parser : jsoup
Period	2004-01-01 00:00:00–2016-12-12 23:59:59
Area	Title, Date (year, month, day, hour), Press
Volume	193,536
Collection Condition	Naver News → Social field → Environmental field
Press (Total of 100)	EBN, EPA yonhapnews, JTBC, KBS, MBC IMTV, MBC, MBN, OSEN, SBS, CNBC, SBS funE, TVReport, TVChosun, Y-STAR, YTN, YTN Live broadcast, ZDNet Korea, Gangwon Daily, Trend newspaper, Gwangju Dream, Kookmin Daily, Government briefing, tomorrow's newspaper, Nokot News, News1, Newsis, Daejeon Daily, Daily Surprise, Daily, Dong-A Ilbo, Digital Daily, Digital Times, Radio Korea, Lady Kyunghyang, My Daily, Maekyung Economy, Maeil Business, Daily Newspaper, Money S, Money Turday, Media Daily, Media Today, Blotter, Seoul Economy, Seoul Business Daily, Seoul Times, World Daily Korea Daily, Star News, Sports Dong-A, Sports Seoul, Sports datkeom, Sports Chosun, sports Korea, current affairs journal, Shindonga, Asian economy, Ai News 24, Up Korea, ek Sports News, Yonhap News Agency, Yonhap News TV, Oh My TV, Omai News, Edaily, Economy 21, Economic Review, Interview 365, Daily Sports (OLD), Ilda, e-newspaper, Jeju Ilbo, Chosun Biz, Chosun Ilbo, Tax Daily, Weekly Kyunghyang, Weekly Donga, Weekly Korea, Central SUNDAY, JoongAng Ilbo, The Weekly World, Newscham, Newschamvod, Culture news, Comedy datkeom, Cookie news, Financial News, Pop News, Prime Economics, Pressian, Prometheus, Hankyoreh, Hankyoreh 21, Korea economy, Korean Economic TV, Hankook Ilbo, Herald POP, Herald Business, Health Chosun

analysis results, stop words including special characters and specific words were processed, and only those terms with two letters or more were extracted. Further, sparse terms with an extremely low frequency of appearance were deleted, and those with low TF-IDF values were removed as well. Finally, each data set was converted to DTM that is appropriate for the analysis using the tm packaged provided by R.

3.2 Skip-Gram Analysis of "Climate Change" Keyword

In this study, a word2vec model was developed using R for the analysis. The skip-gram algorithm was applied to the word2vec model, and it was designed to learn a variety of parametric values. The window size to set how many words on both sides of specific keywords to consider with parametric values was 10 and a worker threads value determining the processing speed was set to 3. The bigger the worker threads, the faster the speed. Finally, 100 word vector dimensionalities were designated to learn the model. When word vector dimensionality gets bigger, it is possible to obtain a

detailed result. Usually, vector dimensionalities are set between 100500 words when there are over 100,000 data. In Table 3, red letters refer to climate-related keywords and blue letters refer to human-related keywords.

Using the skip-gram-based word2vec model consisting of these parametric values, analysis was conducted. As data for analysis, titles, abstracts, and tables of contents of the research reports (1993-2016) published by KEI were used, and the keyword was "climate change." Table 3 indicates cosine distance and 100 words were arranged in order of the closest cosine distance to "climate change." Numbers are cosine distances between "climate change" and relevant words. Shorter cosine distances mean that words are close to "climate change" and highly relevant. As a result, many keywords associated with climate change detailed phenomena, such as sea level, unusual weather, warming, flood, typhoon, cold wave, heat island, and drought appeared. Moreover, many human-related keywords, such as demography, opinion, work, individual, low income, health, food, and dwelling were extracted as well.

Table 3: "Climate Change" Word2VecResults

Weakness	Adaptive	External force	Risk	Phenomenon	Minutes	Sea level	Unusual weather	KMA	Summer
0.270	0.291	0.396	0.403	0.405	0.415	0.428	0.430	0.435	0.439
Black bird	Scale	Butane	Limit	Influence	Idea	Flood position	Downpour	Stratum	Discharge
0.455	0.455	0.457	0.457	0.458	0.460	0.462	0.469	0.469	0.469
Aspects	Forest	Demography	Suspicion	Opinion	Chronic	Precipitation	Temperature	Envelope	Position
0.471	0.473	0.481	0.485	0.486	0.489	0.489	0.491	0.494	0.495
Future	Precipitation	A blind eye	Go to extremes	Work	Warm	Face	Flood	Typhoon	Impact
0.500	0.501	0.501	0.501	0.503	0.503	0.503	0.511	0.512	0.514
Response	Ultraviolet rays	Cold wave	Acceleration	Emergency	Invasion	Extreme	Heat island	Response	Drought
0.514	0.515	0.516	0.518	0.519	0.520	0.520	0.525	0.526	0.528
Dynamics	Sea	Heat wave	Ethiopia	Differ	Symposium	Individual	Water temperature	Cold	Low income
0.528	0.529	0.529	0.530	0.531	0.532	0.532	0.532	0.532	0.537
Help	Domestic demand	Masanman	Standing water	United Nations	Priority	Write	Tide level	Defense	Ability
0.537	0.54	0.541	0.541	0.542	0.542	0.544	0.547	0.547	0.547
Time and space	Eight	Health	River discharge	Play	Uprise.	A water-controlled place	Salt	Search	Scenario
0.547	0.548	0.550	0.551	0.553	0.558	0.558	0.558	0.560	0.561
Food	South Korea	Frequency	Rainfall	Nerve	Evapotranspiration	Maximum	Bad influence	Ground	Weather
0.561	0.563	0.565	0.566	0.568	0.569	0.569	0.569	0.570	0.571
Sensitive	Leader	Research Institute	Humidity	Heart	Warming	Saliva	Dwelling	Capabilities	Climb
0.571	0.571	0.572	0.572	0.572	0.574	0.574	0.574	0.575	0.575

Figure 3 below shows the visualized word2vec analysis results of keywords related to "climate change." The visualized "climate change" keywords were expressed focusing on climate change-related

keywords (16) and human-related keywords (9). The numerical values of the analysis results are cosine distances between "climate change" and relevant words. Lower values mean words are close to "climate change" and very relevant. Among climate change-related keywords, "sea level" was the most associated keyword with "climate change" and it was followed by "unusual weather," "flood position," "downpour," and "precipitation." This reveals that the KEI has conducted many studies on water disasters. Further, among human-related keywords, "demography" was the most associated keyword with "climate change," and it was followed by "opinion," "work," and "individual." This demonstrates that environment research is focused on the quality of lives of the people.

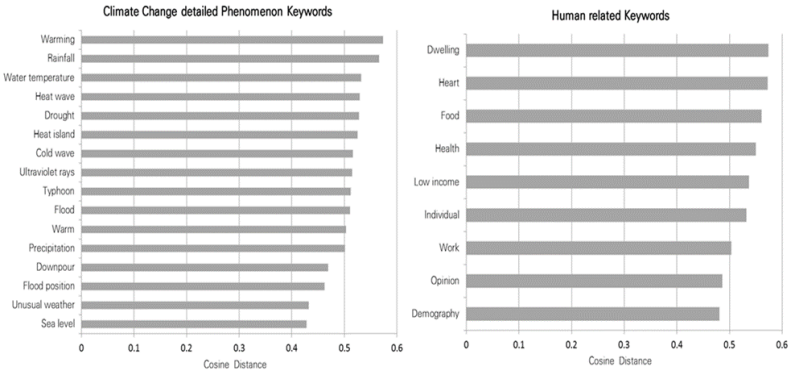

Figure 3: Words Associated with "Climate Change"

3.3 Skip-Gram Analysis of Climate Change-Related Keywords

According to the skip-gram analysis of "climate change" keywords, many climate change-related keywords, such as "sea level," "unusual weather," "warming," "flood," "typhoon," "cold wave," "heat island," and "drought" appeared. Based on the results of this analysis and the researcher's judgment, three climate change-related phenomena, i.e., "warning," "flood," and "drought" were selected. Further, the skip-gram algorithm-based word2vec analysis of climate change-related phenomena was conducted for each medium,

and comparative analysis was performed for the same.

3.3.1 Results of "Global Warming" Word2Vec

Figure 4 shows the results of word2vec analysis for the keyword "global warming." The analysis of policy research report exhibits that human-related keywords, such as "dwelling," "human being," "life," "respiratory organs," and "heart's blood" appear more often. The analysis of environmental news suggests that biology- and food-related keywords, including "plankton," "coral reef," and "Korean blockish cicada" appear more often.

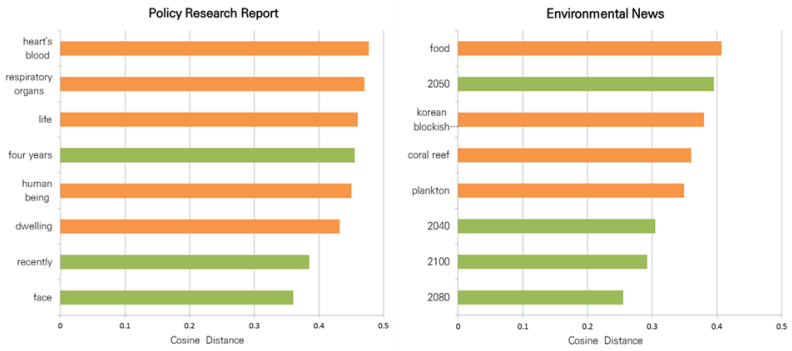

Figure 4: Words Associated with "Global Warming"

3.3.2 Results of "Flood" Word2Vec

Figure 5 shows the results of word2vec analysis for the keyword "flood." The analysis of policy research report exhibits that keywords related to Korean provinces, such as "Geoncheon," "Masan-bay," and "basin" appear more often. The analysis of environmental news shows that keywords related to Chinese provinces, including "Hunan," "Hwangha," and "Sichuan" appear more often.

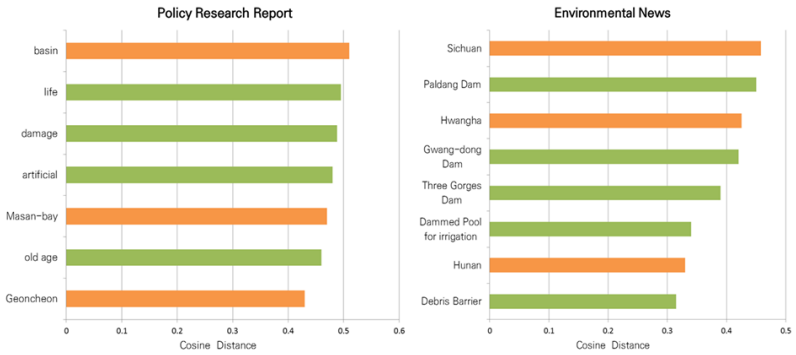

Figure 5: Words Associated with "Flood"

3.3.3 Results of "Drought" Word2Vec

Figure 6 shows the results of word2vec analysis for the keyword "draught." The analysis of policy research report exhibits those human-related keywords, such as "water for living," "disaster," and "life" appear more often. The analysis of environmental news shows that some keywords related to agriculture, including "upland-field crop," "ritual for rain," "agricultural water," and "rice crop" appear more often.

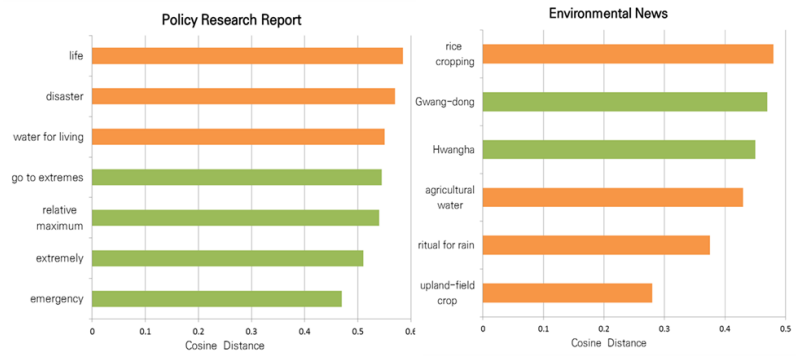

Figure 6: Words Associated with "Drought"

4 Conclusion

In this study, we discovered differences between the content of two distinct media regarding climate change by performing word2vec analysis. For the analysis, we collected environmental news data provided by Naver using jsoup (Java HTML parser) and performed word2vec skip-gram analysis on the collected data. From the results of the word2vec analysis, it was observed that environmental policy research works focused on the life quality of people, and environmental news gave much attention to the seriousness of damage caused due to climate changes. This type of analysis may reflect the differences in media between environmental policy research works and news. Environmental policy research works usually focus on political performance and press usually focuses on factual reports. This discovery suggests that interests in one issue, "environment" in this study, can appear differently according to the media. It is expected that complex demands for policy research works can be met and understood by comprehensively analyzing various media sources through which people express their various interests in the environmental research works, given the fact that the consumers of these environmental policies are people. Thus, further analysis should be performed by extending the research to social media websites, such as Facebook and Twitter, and treatises and materials issued by public agencies. In addition, to detect trends related to a specific topic, it is necessary to perform trend analysis by classifying the data on a yearly basis.

5 Acknowledgment

This study was conducted following the research work Big Data Analysis: Application to Environmental Research and Service (GP 2017-14) and was funded by the Korea Environment Institute.

References

[1] Boussalis, C., & Coan, T. G. (2016). *Text-mining the signals of climate change doubt.* Global Environmental Change, 36, 89-100.

[2] Sterman, J. D., & Sweeney, L. B. (2007). *Understanding public complacency about climate change: Adults' mental models of climate change violate conservation of matter.* Climatic Change, 80(3-4), 213-238.

[3] Kim, M. S., Ko, J. K., & Kim, J. H. (2007). *A Study on Factors Determining the Public Support of Climate Policy.* Country Plan, 42(4), 233-247.

[4] Whitmarsh, L. (2008). *Are flood victims more concerned about climate change than other people? The role of direct experience in risk perception and behavioural response.* Journal of risk research, 11(3), 351-374.

[5] Weber, E. U. (2010). *What shapes perceptions of climate change?.* Wiley Interdisciplinary Reviews: Climate Change, 1(3), 332-342.

[6] Kahan, D. M., Wittlin, M., Peters, E., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. N. (2011). *The tragedy of the risk-perception commons: culture conflict,* rationality conflict, and climate change.

[7] Chae, H. M., Lee, S. S., & Lee, H. J. (2011). *A Study on Development of Climate Change Adaptation Strategies through County Recognition.* Journal of The Korean Society of Hazard Mitigation, 11(6), 131-138.

[8] Kim, J. H., & Koh, J. K. (2012). *A Study on Local Public Officials'Perception on Climate Change Adaptation: the case of Gyeonggi-Do.* GRI REVIEW, 14(1), 319-343.

[9] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* In Advances in neural information processing systems (pp. 3111-3119).

[10] Singhal, A. (2001). *Modern information retrieval: A brief overview.* IEEE Data Eng. Bull., 24(4), 35-43.

[11] Tan, P. N. (2006). *Introduction to data mining.* Pearson Education India.

[12] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781.

[13] Rong, X. (2014). *word2vec parameter learning explained.* arXiv preprint arXiv:1411.2738.