International Journal of Engineering & Technology, 7(4.16) (2018) 5-9

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Study on the Decision-Making of Effective S/W Education based on Opinion Mining Analysis

Ji-Hoon Seo[1], Nam-Hun Park[2], Kil-Hong Joo[3]

[1,3]Dept. of Computer Education, Gyeongin National Univ. of Education, Korea
[2]Dept. of Computer Science, Anyang University, Korea
**Corresponding author E-mail:[1]sserz@naver.com, [2]nmhnpark@anyang.ac.kr(co-first Author), [3]khjoo@ginue.ac.kr(Corresponding Author)*

Abstract

The Currently, along with the advent of the web 2.0 era, due to the continuous expansion of social media service infrastructures, the shares of conventional public opinion evaluation functions have been gradually shifting from the existing mass media to social media. This phenomenon is attributable to the two-way communication and convenience unique to social media and social media are now in charge of an axis of public opinion evaluation standards. In particular, since diverse interests conflict in education policies and countless conflicts of opinions occur in the process of setting up policy agendas, in establishing education policies, accurately analyzing reputations among the public, who are the targets of education policies, in order to set up effective policy agendas, is the most important issue. Therefore, in this study, the resultant values of huge unstructured data on the positive and negative reputations of past policy agendas related to the mandatory software education that has been organized as a regular curriculum of middle/high schools from 2018 in Korea, which have been addressed by the Ministry of Education, the Ministry of Science, ICT and Future Planning, and the Korea Foundation for the Advancement of Science and Creativity, felt and judged by the general public on social media such as blogs and Twitter and on online media including portal news were visualized through opinion mining analysis techniques to derive more effective software education related policy agendas. In addition, based on the foregoing, a Korean style software education system that fits circumstances was constructed and the system is expected to become an important measure that provides guidelines for setting mid/long-term road maps for the fostering of creative and convergent talented persons equipped with international competitiveness and software education in Korea.

Keywords: *S/W Education Policy, Opinion Mining, Social Media, Unstructured Data*

1. Introduction

Recently, following the advent of the web 2.0 era, online media such as social media services, which have grown rapidly thanks to the popularization of smartphones and the development of IT technology, have been becoming a new field of public opinion evaluation standards. Unlike the existing mass media, social media can provide information using the Internet networks and above all, they have the swiftness and far-reaching power in information transmission thanks to interactive communication. Therefore, they have been used as an important tool for analysis of reputations among the public together with Internet news. In addition, since situations where social issues on social media are changed from public agendas to policy agendas by the public opinions on online media constantly occur, public opinions on online media including social media have become important reference points that can be no longer overlooked in the setting of policy agendas from the government's position [1]. In response to this trend, public institutions in the government are making efforts to communicate one-to-one, not one-to-many, via social media such as blogs and Twitter, and open and share national information assets such as public data. Therefore, the government should collect information from online media such as social media and analyze the information to set up new policy agendas or identify public opinions on particular policies [2]. Meanwhile, software education that has become mandatory in middle and high schools since 2018 is rising as one of hot issues in educational policies in South Korea and in relation to it, many education agendas such as improvement measures against problems or side effects occurring after software education became mandatory have been derived at symposiums or conferences where many experts gathered together. Among the problems, the one that is the most frequently discussed is the hours allocated to software education and hot brainstorming is continued on the appropriate number of hours of software education out of existing allocation of hours by subject [3]. In addition, the most indispensable stage before deciding and enforcing education policies is the formation of education agendas. The appropriateness of the agenda setting is one of the important elements that determine the success or failure of educational policies in the future [4]. Therefore, for effective derivation of education that sets the direction of education policies, a more scientific and systematic decision making support system must be equipped without fail. Therefore, for the derivation of the most effective software education agendas, this study is intended to make the unstructured data on the trends of public opinions on the software education policies that have been enforced by related organizations in the past, that is, in 2014 and 2015, collected from online media including social media and portal news into big data, analyze the big data through opinion mining analysis techniques, visualize the resultant values obtained through the analysis into graphs, apply the numerical data obtained from the graphs to the process of setting up education policy agendas in order to present a model for the derivation of opinion mining based education policy agendas for effective and successful enforcement of current software education policies in Korea [5][6].

2. Related Works

2.1. Opinion Mining Analysis Method

Opinion mining is an analysis technique that distinguishes among positive, negative and neutral preferences for structured or unstructured texts collected from online media such as social media and portal news, and is appropriately used for market size forecasting, consumer reaction and word of mouth analyses for particular services and products [7]. This technique extracts vocabulary information that expresses affirmation or negation and recognizes sentences composed of opinions about objects to measure positive and negative words with the sum of patterns including certain opinion articles based on the sentiment dictionary entered into the server, and can visualize the resultant values obtained through the measurement into graphs and extract more meaningful information data from the unstructured data composed of the opinions of many unspecified users. In this respect, opinion mining can be called sentiment analysis, and can be interpreted as the broad sense of text mining, a technique used in natural language processing and computer linguistic analysis [8][9].

2.2. Text Mining Analysis Method

Text mining can extract meaningful information data based on the natural language processing technology, figure out linkages with other information data, and obtain more resultant values than normal information retrieval such as finding out the categories of texts. Document summarization and feature extraction are the core research fields of text mining and the fields of application of text mining are quite diverse [10]. Extracting structured information from documents to make the information into databases or find out rules from the information is the most common application and helping finding documents on the user's web, creating and analyzing user profiles, identifying natural languages used in documents, and classifying and clustering documents in massive databases, document reanalysis using document classification information, document summarization, translation, and exploration, market and risk analysis through the acquisition of time series information, document indexing, document filtering and recommendation, and representative keyword and topic extraction can be said to be the most representative fields of application [11][12].

3. Proposed Method

As part of attempts to derive reputations in and trends of public opinions regarding software education policies included in writings written on social media or portal news comments based on unstructured data, the opinion mining analysis process proposed in this study randomly collects unstructured data including the words "making software education mandatory", which are a core keyword in relation to the Ministry of Education, the Ministry of Science, ICT and Future Planning, and the Korea Foundation for the Advancement of Science and Creativity from the three items, blogs, Twitter, and portal news distributed on online media to classify the polarity of sentiment words according to affirmation and negation and select sentiment data. The configuration of the overall system for the development of the sentiment dictionary, which is one of the important stages of this study, is classified into three models, a storage server for data collection, storage, and pre-processing, a natural language processing learning model for natural language processing and sentiment word morpheme analysis, and a sentiment dictionary construction server conversion stage for opinion extraction.

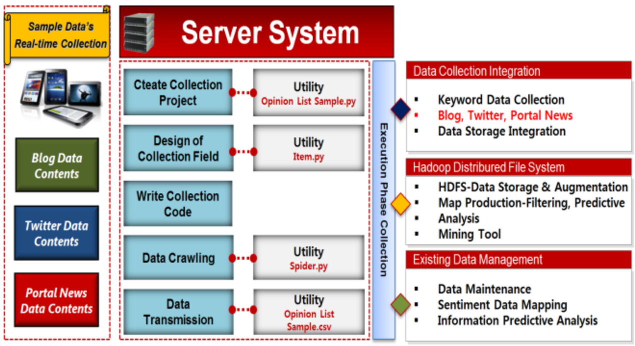


Fig. 1: Opinion analysis model

3.1. Data Collection

To conduct opinion mining analysis, a sentiment dictionary must be constructed. In order to improve the accuracy and reliability of this sentiment dictionary, the collection of unstructured data is presented as an important element. First, the times and scales of the collected data should be clear and the larger the quantity of collected data, the higher the achievable accuracy of the sentiment dictionary. In this study, in relation to software education, unstructured data during 2014 and 2015 were collected from three types of online media, that is, blogs, Twitter, and portal news.

3.2. Data Classification

In this study, after setting "making software education mandatory" as a related core search keyword and setting the Ministry of Education, the Ministry of Science, ICT and Future Planning, and the Korea Foundation for the Advancement of Science and Creativity as related search keywords by part, unstructured data were randomly collected and individual categories were created and classified. Thereafter, the entire classified data were stored in the main server.

3.3. Data Pre-processing

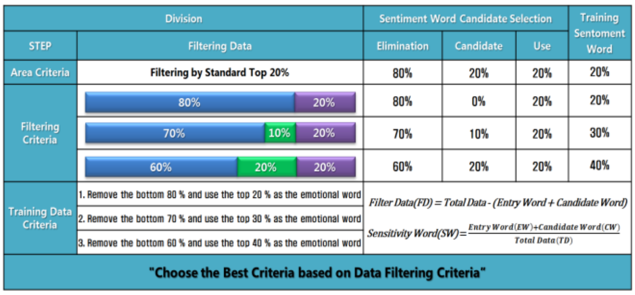
The data pre-processing process is one of important processes to select candidates for sentiment words. In this study, no missing value, outlier, or wrong value, which can appear in numerical data, exists because pre-processing is conducted based on the collected unstructured data. However, filtering work was performed to extract word with high weighted values in relevant sentences or important words that can sufficiently become sentiment words. In addition, although "SentiWordNet" (SWN) can be used to construct sentiment dictionaries in English grammar, SWN has a shortcoming that its accuracy can be reduced in Korean syntax because of the grammatical structure. Therefore, to extract meaningful words, filtering was carried out under the following rules.

Table 1: Word Stemming Filter Process

Application of word filtering rules to construct sentiment dictionaries
1. Remove special characters, English, and stop words. 2. Remove meaningless terms and one-character texts. 3. Distinguish the same words in the form of conjunctions to classify the essence of sentiment words. 4. Distinguish between homonyms and synonyms. 5. In the case of abbreviations and coined words, reflect only those that have been registered in the Wikipedia and Korean dictionaries.

3.4. Construction of Sentiment Dictionary

The words derived through the preprocessing of meaningful words are classified into sentiment word candidate groups. In this study, to improve the accuracy and reliability of the resultant values of opinion mining analysis, the top 20% of the words selected as the sentiment word candidate groups were extracted to finally select them as sentiment words. Although the data distributed in the lower groups are composed of the sets of words with high weighted values or meaningful words, most of them fall under neutral words and words close to other polarities in cases where they are tagged as positive, negative, neutral, and other polarities and polarity classification is carried out. Therefore, the word sets that are less useful as such were removed in advance. On the other hand, words that correspond to neutral and other polarities also exist among the top 20% words finally selected as sentiment words but they have very high weighted values or correspond to higher word groups with very high frequencies in documents.

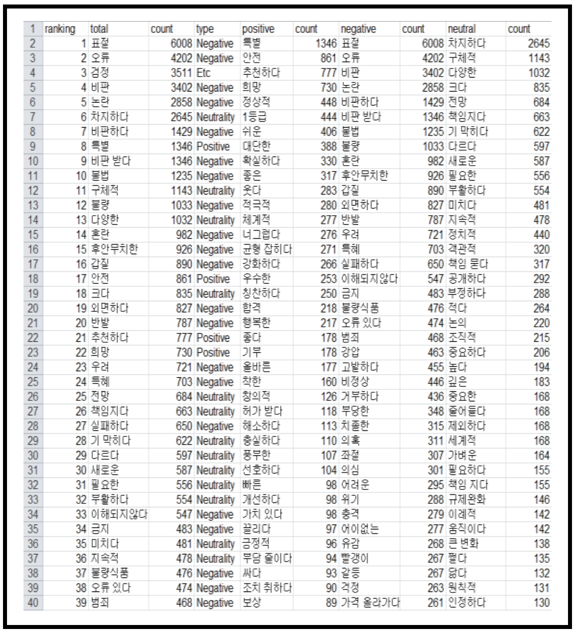


Division		Sentiment Word Candidate Selection			Training Sentoment Word
STEP	Filtering Data	Elimination	Candidate	Use	
Area Criteria	Filtering by Standard Top 20%	80%	20%	20%	20%
Filtering Criteria	80% 20%	80%	0%	20%	20%
	70% 10% 20%	70%	10%	20%	30%
	60% 20% 20%	60%	20%	20%	40%
Training Data Criteria	1. Remove the bottom 80 % and use the top 20 % as the emotional word	*Filter Data(FD) = Total Data - (Entry Word + Candidate Word)*			
	2. Remove the bottom 70 % and use the top 30 % as the emotional word	$Sensitivity\ Word(SW) = \frac{Entry\ Word(EW)+Candidate\ Word(CW)}{Total\ Data(TD)}$			
	3. Remove the bottom 60 % and use the top 40 % as the emotional word				
"Choose the Best Criteria based on Data Filtering Criteria"					

Fig. 2: Opinion analysis model

3.5. Sentiment Word Tagging & Opinion Mining Analysis

Sentiment word tagging classifies the words into positive, negative, neutral, and other ones. Sentences recorded in one document are compared with the sentiment dictionary data to calculate the frequencies of positive and negative words and the resulting reputation data are classified into individual categories and stored as monthly data. An example of the frequencies and types of sentiment words derived from the document is shown in Fig. 3. Time series analysis was conducted using the stored reputation data.



1	ranking	total	count	type	positive	count	negative	count	neutral	count
2	1	표절	6008	Negative	특별	1346	표절	6008	차지하다	2645
3	2	오류	4202	Negative	안전	861	오류	4202	구체적	1143
4	3	검정	3511	Etc	추천하다	777	비판	3402	다양한	1032
5	4	비판	3402	Negative	희망	730	논란	2858	크다	835
6	5	논란	2858	Negative	정상적	448	비판하다	1429	전망	684
7	6	차지하다	2645	Neutrality	1등급	444	비판 받다	1346	책임지다	663
8	7	비판하다	1429	Negative	쉬운	406	불법	1235	기 막히다	622
9	8	특별	1346	Positive	대단한	388	불량	1033	다르다	597
10	9	비판 받다	1346	Negative	확실하다	330	혼란	982	새로운	587
11	10	불법	1235	Negative	좋은	317	후안무치한	926	필요한	556
12	11	구체적	1143	Neutrality	웃다	283	갑질	890	부활하다	554
13	12	불량	1033	Negative	적극적	280	외면하다	827	미치다	481
14	13	다양한	1032	Neutrality	체계적	277	반발	787	지속적	478
15	14	혼란	982	Negative	너그럽다	276	우려	721	정치적	440
16	15	후안무치한	926	Negative	균형 잡히다	271	특혜	703	객관적	320
17	16	갑질	890	Negative	강화하다	266	실패하다	650	책임 묻다	317
18	17	안전	861	Positive	우수한	253	이해되지않다	547	공개하다	292
19	18	크다	835	Neutrality	칭찬하다	250	금지	483	부정하다	288
20	19	외면하다	827	Negative	합격	218	불량식품	476	적다	264
21	20	반발	787	Negative	행복한	217	오류 있다	474	논의	220
22	21	추천하다	777	Positive	좋다	178	범죄	468	조직적	215
23	22	희망	730	Positive	기부	178	강압	463	중요하다	206
24	23	우려	721	Negative	올바른	177	고발하다	455	높다	194
25	24	특혜	703	Negative	착한	160	비정상	446	깊은	183
26	25	전망	684	Neutrality	창의적	126	거부하다	436	중요한	168
27	26	책임지다	663	Neutrality	허가 받다	118	부당한	348	줄어들다	168
28	27	실패하다	650	Negative	해소하다	113	치졸한	315	제외하다	168
29	28	기 막히다	622	Neutrality	충실하다	110	의혹	311	세계적	168
30	29	다르다	597	Neutrality	풍부한	107	좌절	307	가벼운	164
31	30	새로운	587	Neutrality	선호하다	104	의심	301	필요하다	155
32	31	필요한	556	Neutrality	빠른	98	어려운	295	책임 지다	155
33	32	부활하다	554	Neutrality	개선하다	98	위기	288	규제완화	146
34	33	이해되지않다	547	Negative	가치 있다	98	충격	279	이례적	142
35	34	금지	483	Negative	끌리다	97	어이없는	277	움직이다	142
36	35	미치다	481	Neutrality	긍정적	96	유감	268	큰 변화	138
37	36	지속적	478	Neutrality	부담 줄이다	94	빨갱이	267	짧다	135
38	37	불량식품	476	Negative	싸다	93	갈등	267	닮다	132
39	38	오류 있다	474	Negative	조치 취하다	90	걱정	263	원칙적	131
40	39	범죄	468	Negative	보상	89	가격 올라가다	261	인정하다	130

Fig. 3: Example of sentiment word tagging & derivation of frequencies

4. Experimental Result

In this study, "making software education mandatory" was set as a core search keyword, a search period was set as January 1, 2014 through September 30, 2015, and unstructured data provided by the Ministry of Education, the Ministry of Science, ICT and Future Planning and the related organization the Korea Foundation for the Advancement of Science and Creativity were collected from a total of three media, that is, blogs, Twitter, and portal news. Opinion mining analyses were conducted with the collected unstructured data after dividing the ranges of analysis by medium to calculate yearly reputation values and finally, the reputation values were summed up by organization to calculate resultant values.

4.1. Blog Reputation Analysis in Reputation to Making S/W Education Mandatory

In the case of blog reputation analysis, negative reputation values were identified to have increased remarkably because blogs are close to private spaces where personal emotions or opinions are freely expressed in contrast to portal news that belong to mass media from the objective viewpoint of the public. Consequently, whereas sentiment word data from portal news showed relatively more positive reputation values, sentiment word data from blogs showed more negative reputation values.

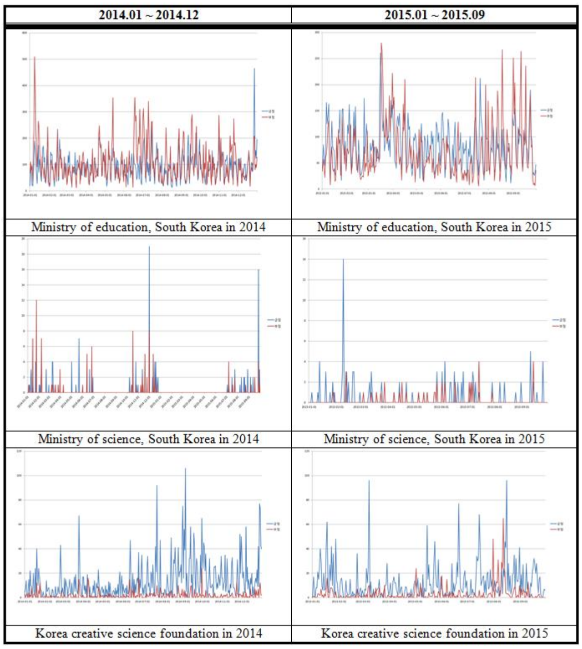


Fig. 4: Blog reputation analysis in relation to policies to make s/w education mandatory by organization

4.2. Twitter Reputation to Making S/W Education Mandatory

In the case of Twitter reputation analysis, the ratio of negative reputations was shown to be high in 2014 because exposure on online media related to making software education mandatory began before 2013 but the ratio of positive reactions was shown to have increased drastically from December 2014. Although the amount of data was shown to be small when swearing words, slangs, and abbreviations among Korean words were excluded in the process of filtering of words, when all of swearing words, slangs, and abbreviations were included, the frequencies of contents close to negative words were shown to be high. Therefore, it can be said that the ratio of negative views is high in fact and because reputation analysis values were calculated when the respondents' Twitter accounts were exposed because of the nature of Twitter, which is one of social media, the results can be interpreted as indicating that the values of positive reputations were shown to be relatively higher as the date of full-scale enforcement of "mandatory software education" became closer from the position of those who directly receive the benefits of "mandatory software education."

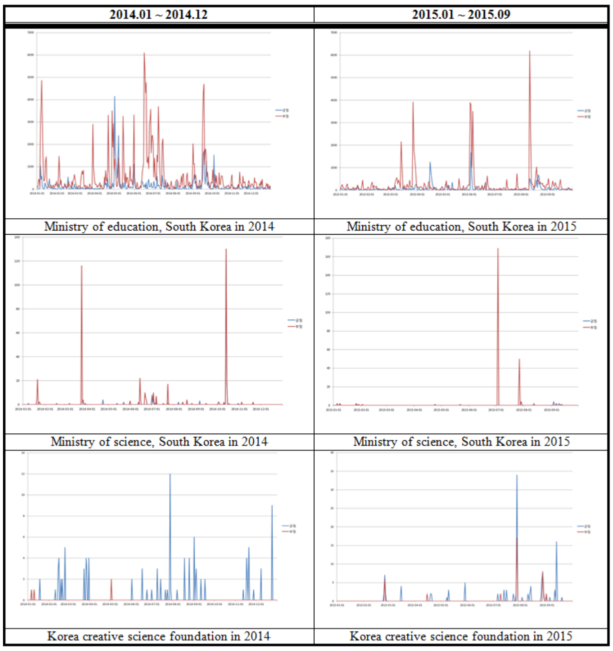


Fig. 5: Twitter reputation analysis in relation to policies to make s/w education mandatory by organization

4.3. Portal News Reputation Analysis in Relation to Making S/W Education Mandatory

With regard to portal news reputation analysis, in the case of the portal news contributions retrieved with the core keyword "making software education mandatory", the media reputations of the government's education policy agendas was shown to indicate positive resultant values in most cases.

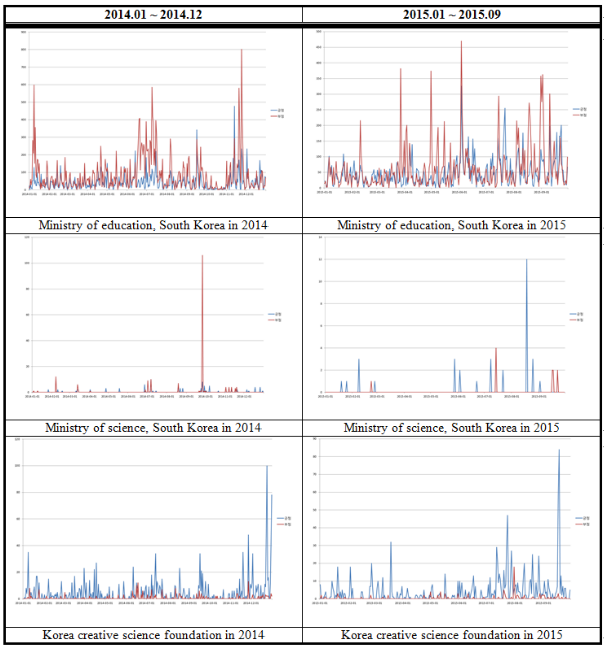


Fig. 6: Portal news reputation analysis in relation to policies to make s/w education mandatory by organization

However, because of the structural problem that the contributions of the professional commentators employed by the press, which are mass media, cannot but lack objectivity to some extent, in the process of opinion mining analysis in this study, not so much weight was given to the resultant values of portal news reputation analysis. In the process of calculating the reputation analysis values, the frequency of the search keyword gradually increased from January 2014, and the negative reputation values began to appear in portal news articles as the year 2018, which is the year of full-scale introduction of software education, became closer. On interpreting this numerically, it can be seen that positive reputation values were predominant until the end of 2014, but negative reputation values began to increase again from 2015. This can be interpreted as the result of the side effects such as the high cost of private education appeared due to the uncertainty of software education.

4.4. Opinion Mining Analysis Result Including Reputation Analysis Values of All Item by Organization

According to the results of opinion mining analyses conducted including reputation analysis values for all items, that is, blogs, Twitter, and portal news by organization, in the case of the Ministry of Education with the largest amount of collected data, negative reputations against the agenda "making software education mandatory" were shown to be at least two times larger in number than positive reputations and in the case of the Ministry of Education, the Ministry of Science, ICT and Future Planning, negative reputations were also predominant over positive reputations. Finally, in the case of the Korea Foundation for the Advancement of Science and Creativity, interestingly, the resultant values were derived so that positive reputations are overwhelmingly predominant over negative reputations.

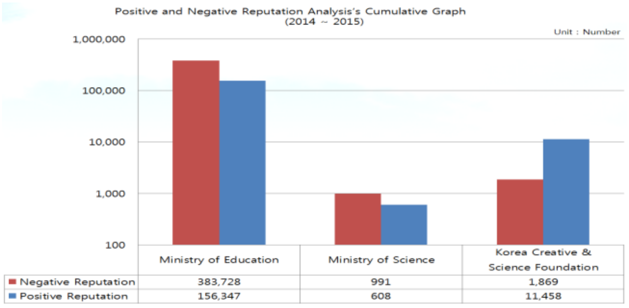


Fig. 7: Results of opinion mining including reputation analysis values for all items by organization

Among the reasons why the ratios of negative reputations of the Ministry of Education and the Ministry of Science, ICT and Future Planning, which are government organizations, were identified to be relatively high, the distrust of the public in the current educational policies that are frequently changed according to the trend of reputations presented by the press without consistency can be interpreted to be a fundamental reason and the public's concern about increases in private education expenses of households due to the addition of mandatory education can be interpreted to be an additional reason. On the contrary, in the case of the Korea Foundation for the Advancement of Science and Creativity, which is a related organization in relation to education, since the nature of this organization is close to that of pure academic research organizations not related to the establishment of government's education policies, software education related academic research activities or the expression of positions in online media participated by the Korea Foundation for the Advancement of Science and Creativity can be interpreted to be more positively accepted from the position of the public in Korea.

5. Conclusion

Exiting utilization of big data for education was a way to derive implications through the analysis of mutual relations between diverse variables such as the way to use the big data for learning, related data inquiry, and frequency relationships. However, such fragmentary and uniform analysis models have limitations in analyzing performance factors for the setting of education policy agendas and suggesting guidelines for mid/long-term road map establishment. Thus far, the analysis techniques for education policies enabled the related analyses only when the policy agendas were presented in advance and accordingly, education policies have been promoted through subjective keywords. Therefore, if the opinion mining analysis data for derivation of effective education policy agendas for software education presented in this study are utilized in the establishment of policy agendas, the extensive and huge unstructured data on online media can be managed and analyzed and the range of errors in the resultant values from numerical big data analysis can be minimized so that improvement in accuracy and reliability can be expected. In addition, thanks to the derivation of policy agendas based on the scientific logic as such, it is expected that the necessity of the construction of Korean style software education programs suitable for circumstances will be recognized and a momentum to greatly contribute to the fostering of creative and convergent talents equipped with international competitiveness indispensable in the age of the fourth industrial revolution can be prepared. When seen from the technical aspect, the data are expected to be a good guide in the foreign fields of research into the aspects of modern unstructured data and methods to analyze and manage the data thereby being quite helpful for academic technology development in Korea.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2017R1D1A1B03029292).

References

[1] Sun, Y. and K. Jia. 2009. Research of word sense disambiguation based on mining association rules, In: Third International Symposium on Intelligent Information Technology Application workshops, November 21-22, NanChang, China, pp. 86-88.B. Sklar, Digital Communications, Prentice Hall, pp. 187, 1998.J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[2] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, "Data Mining with Big Data", Vol.26, No.1, pp. 97-107, 2014, January.)

[3] Tang, C. and C. Liu.2008. Method of Chinese grammar rules automatically access based on association rules, In: Proceedings of the. Computer Science and Computational Technology volume, 1 pp. 265-268 (ISCSCT, Shanghai, Dec. 20-22, 2008).

[4] Irfan Ajmal Khan, Jin Tak Choi, International Journal of Software Engineering and Its Applications, "An Application of Educational Data Mining (EDM) Technique for Scholarship Prediction" , Vol. 8, No. 12 (2014), pp. 31-42

[5] Xu, Yue, Li, Yuefeng, & Shaw, Gavin, Reliable representations for association rules. Data & Knowledge Engineering, Volume 70 Issue 6, pp. 555-575. June, 2011.

[6] Bo pang, Lillian Lee and Shivakumar Vaithyanathan, 2002, "Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 Conference on Empirical methods in Natural Language Processing, Vol.10, pp.79-86

[7] Seo Ji Hoon, "Design of Opinion Sensitivity Dictionary Model for Big Data Management", 2015.

[8] Khan, I.A. and J.T. Choi. 2015. An application of educational data mining (EDM) technique for scholarship prediction. International Journal of Software Engineering and its Applications, Vol.8 No.12 [2014], pp 31-42.

[9] Ghose, P. G. Ipeirotis and A. Sundararajan, 2007, "Opinion Mining Using Econometrics: A Case Study on Reputation System" Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp.416-423.

[10] Pang and L. Lee, 2008, "Opinion Mining and Sentiment Analysis", Foundation and Trends in Information Retrieval, 2(1-2), pp.1-135.

[11] S. Shin, Read Emotions in the Article! Understanding Emotional Analysis, IDG Korea, pp. 1-11, 2014.

[12] E. Courses and T, Surveys, (2008), "Using Sentiment SentiWord-Net for multilingual sentiment analysis", IEEE 24th International Conference on Data Engineering Workshop (2008), Cancun, Mexico, pp.507-512