Korea Institute for Health and Social Affairs

Working Paper 2014-01

Social Risk Factor Prediction Utilizing Social Big Data

Taemin Song · Juyoung Song · Ji-Young An · Jong-Min Woo

KIHASA Korea Institute for Health and Social Affairs

Social Risk Factor Prediction Utilizing Social Big Data

Taemin Song, Head, Research Center for Psychosocial Health

Korea Institute for Health and Social Affairs
Jinhungro 235, Eunpyeong-gu, Seoul 122-705, Korea
http://www.kihasa.re.kr
ISBN: 978-89-6827-128-1 93510

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1 Introduction

The use of mobile Internet and Social Networking Service (SNS) has increased remarkably in Korea in line with the greater availability of smart phones. The Internet use rate of the population aged older than 3 years stood at 82.1% as of July 2013, of whom 55.1% of those 6 years old or older were found to have started using SNS within the preceding year (The Ministry of Science, ICT and Future Planning, and the Korean Internet Security Agency, 2013)[1]. As the volume of data transmitted through SNS has increased exponentially, a greater number of countries and corporations have endeavored to use and analyze big data in order to bring about new economic effects, create jobs, and resolve social issues. In the public sector, big data is utilized for disease prevention, prediction, and treatment and patient management through sharing genes and life research resources, and multinational IT (Information Technology) corporations and web search portal sites are producing various value information by analyzing big data saved in servers (Policy Exchange, 2012)[2]. SNS is where gloomy feelings,

1) The Ministry of Science, ICT and Future Planning, and the Korean Internet Security Agency(2013) 2013 Survey on the Internet Usage, Seoul, Korea: Author.
2) Policy Exchange (2012). The Big Data Opportunity: Making government faster, smarter and more personal.

stress and worries that adolescents have in their everyday lives are heard and behaviors thereof are understood, so the analysis of emotional expressions about suicide and psychologically risky behaviors appearing on SNS can bring about the positive effect of preventing suicide by detecting risk signs and meaningful patterns. The suicide rate of Korea, amid remarkable socioeconomic changes, has been the highest level since 2004 among the member nations of the Organization for Economic Cooperation and Development (OECD) and in particular, as adolescent suicide is emerging as a social issue, government-level, pro-active measures are urgently needed. In addition, as adolescents exposed to cyber bullying commit suicide or become the perpetrators of violence, cyber bullying is emerging as a serious social issue. Cross-sectional research and longitudinal research, adopted thus far to investigate the causes and relevant factors of the above-noted social risk, are useful in inquiring into the relationships between an individual and a group with respect to pre-determined factors, but have limitations in determining how and to what extent an individual buzz mentioned on cyberspace is related to a social phenomenon. In this sense, the decision tree analysis of data mining that utilizes social big data can be deemed as a useful tool to analyze effectively the relationships of interaction of various causes arising from a complex and dynamic phenomenon of human behavior such as a social risk factor, by identi-

fying a correlation or patterns according to decision rules without special statistical hypothesis. Accordingly, this research document is intended to propose a method to predict social risk factors of Korea through the decision tree analysis of data mining based upon social big data collected from online news sites, blogs, cafes, bulletin boards, etc.

Chapter 2

Healthcare & Welfare 3.0 and Big Data

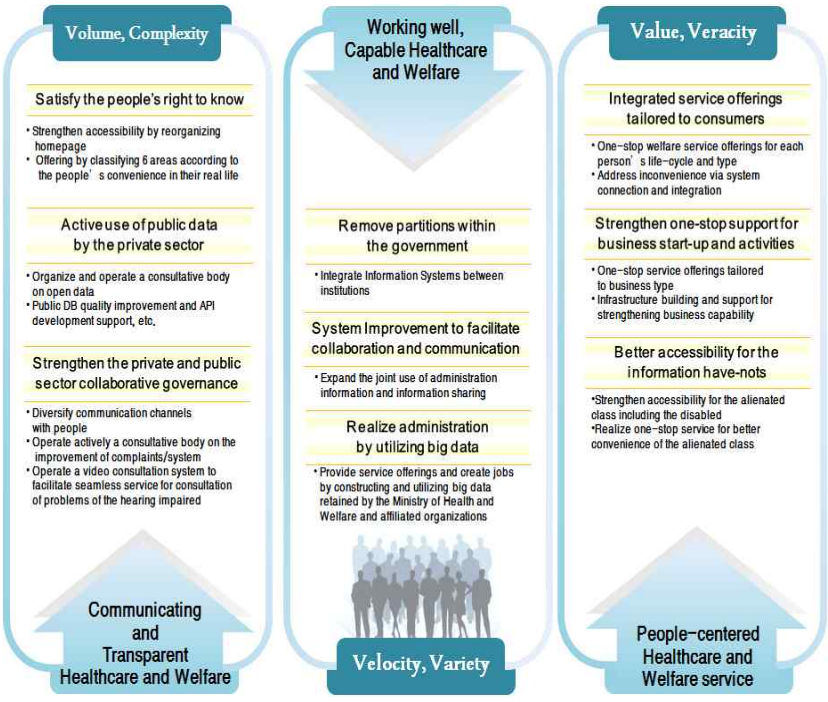
2 Healthcare & Welfare 3.0 and Big Data

1. Measures to Push Forward Big Data in Healthcare and Welfare

Measures to promote big data in healthcare and welfare areas were formulated to seek effective ways of utilizing big data there to effectively push forward Government 3.0, create healthcare and welfare tailored to a person's life-cycle, and realize the happiness of people. Government 3.0 refers to a new paradigm for government operation, designed to ensure a driving force with respect to government projects, provide services tailored to the people, as well as create jobs and support creative economy by pro-actively opening up and sharing public information, removing compartmentalization between government agencies and promoting communication and cooperation. Big data refers to high-volume (Volume), high-velocity (Velocity), high-variety (Variety) and highly complex (Complexity) forms of data sets as well as refers to technologies to extract information of veracity (Veracity) and value (Value) through the utilization and analysis of high-volume data and pro-actively respond to or predict changes based upon the extracted knowledge. Characteristics of big data (5V and 1C) as shown in Figure 1 are organically related to the strategies for pushing forward the

Ministry of Healthcare and Welfare 3.0. Under 'Communicating and Transparent Healthcare and Welfare' of the Ministry of Healthcare and Welfare 3.0, public data is opened proactively to facilitate the use of big data, and consequently the readily available material becomes highly complex (Complexity) and a vast amount of information (Volume). In addition, under 'Working well and Capable Healthcare and Welfare' of the Ministry of Healthcare and Welfare 3.0, the integration of information of variety (Variety) becomes possible by realizing a scientific administration through the utilization of big data, and the velocity of material accumulation is high through improvement of the government operation system. Last but not least, 'People-centered Healthcare and Welfare Service' of the Ministry of Healthcare and Welfare 3.0 creates information of veracity and new value by providing integrated services tailored to consumers based on the result of big data analysis(Figure 1).

[Figure 1] Big Data Characteristics & Strategies for the Ministry of Healthcare and Welfare 3.0



2. Effective Big Data Utilization Measures in Healthcare and Welfare Area

Big Data can provide a new paradigm for healthcare and welfare services as an engine to create new values. A large volume of big data in the domestic healthcare and welfare area has been saved and maintained due to the stable implementation of the existing legacy systems. In the healthcare

sector, the National Healthcare Insurance Corporation started to implement data warehouse from 2002 and currently stores and manages data generated from the wage management system, recuperation allowance payment system, health examination system, medical care system, and the system for eligibility, premium payment and post-management. Data warehouse of the National Healthcare Insurance Corporation provides premium simulation, premium and the estimates of premium allowance rises, etc. Information related to the application for reviews had expanded sharply ever since the implementation of the separation of dispensing and prescribing function in 2000, and the Health Insurance Review and Assessment Service began to construct data warehouse from 2002, and currently stores and manages data related to master information, sanatorium information, and payment information. The data warehouse of the Health Insurance Review and Assessment Service adopts various analysis methods such as statistical analysis, time series analysis, multidimensional analysis and trend analysis for each subject area in order to enable timely analysis of information. Also, review analysis data mart, assessment analysis data mart, and statistics analysis data mart are operated according to data utilization purpose.

The National Cancer Center implemented in 2002 and currently runs a data warehouse of cancer registration material for the purpose of trend analysis etc. to determine the cancer at-

tributable burden level and to lay a basis for cancer management policy by calculating cancer statistics (occurrence rate, death rate and survival rate). In the welfare sector, almost the whole welfare information is integrated and maintained in the integrated social welfare management network. The integrated social welfare management network stores and manages data in the 44 detail-level task areas under the six high-level categories including the integrated social welfare management network channel, hope-inspiring welfare, welfare administration, welfare allowance integration, Saeol [note: refers to new and upright] administration and external areas. In addition, the Korea Food and Drug Administration operates the database related to the current status of imported food and general food and the National Statistical Office, and the state policy-related research organizations constructed panel data for the purpose of producing various statistics relevant to healthcare and welfare. As described above, high-volume structured big data is already stored and managed in the public sector and moreover, a large volume of unstructured data is managed through an organization's homepage or SNS service.

Meanwhile, as the Personal Health Record (PHR) provides a wide variety of health information to healthcare consumers and offers a tool to control and manage one's own health by oneself, the construction of the PHR on public and private levels has consistently been pushed forward. The PHR can collect ob-

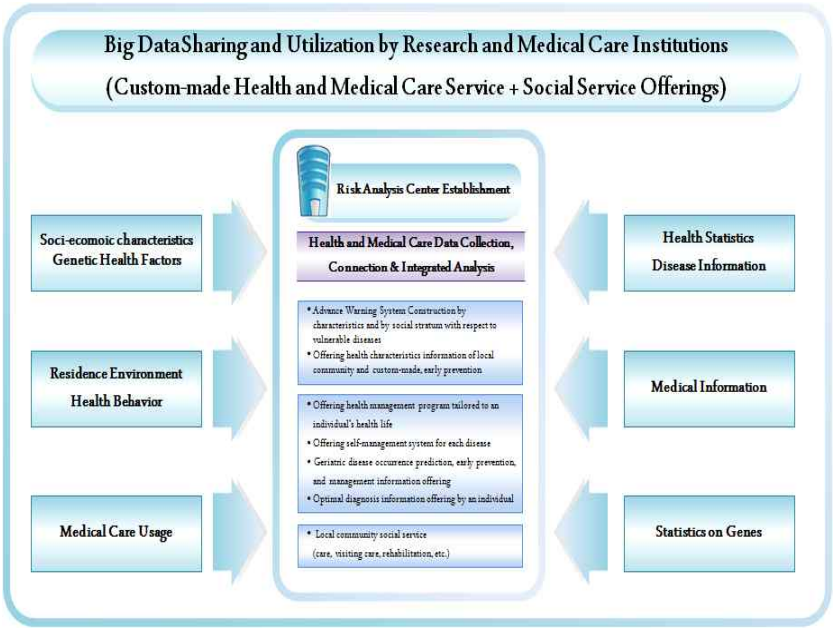
jective material such as blood pressure, which can be measured and entered manually by a patient or transmitted directly though a u-Health device. In other words, the PHR information, which is transmitted through a u-Health device for the purpose of 24-hour remote health management monitoring, can lengthen the healthy longevity of the people; improve life quality by forming healthy life habits (non-smoking, moderation in drinking, nutrition and exercise) and by fostering self-health management capability; provide a benefit of medical cost reduction, which can be expected through the prevention-centered health management away from the treatment with respect to chronic diseases (Song et al., 2011)[3]. As noted above, the relationships between the healthcare/welfare and big data is interrelated very closely. In the health and medical care sector, in order to provide healthcare service tailored to a person's life-cycle, it is imperative to derive a model of future prediction and policy decision by utilizing big data centered around health and medical care as well as pending issues of society or critical issues of the future and for this purpose, it is necessary to establish the 'Risk Analysis Center' to enhance value-added data as common assets of society.

The Risk Analysis Center, presumably, is expected to enable a state-level early response to diseases by predicting the dis-

3) Song Tae Min et al. (2011). u-Health: Current Status and Tasks Ahead. the Korean Institute for Health and Social Affairs.

tribution and trends of major diseases through disease management and prediction and through the utilization of disease-related statistics of various users (Go & Jeong, 2012).

[Figure 2] Big Data Utilization Measures through Establishing the Risk Analysis Center

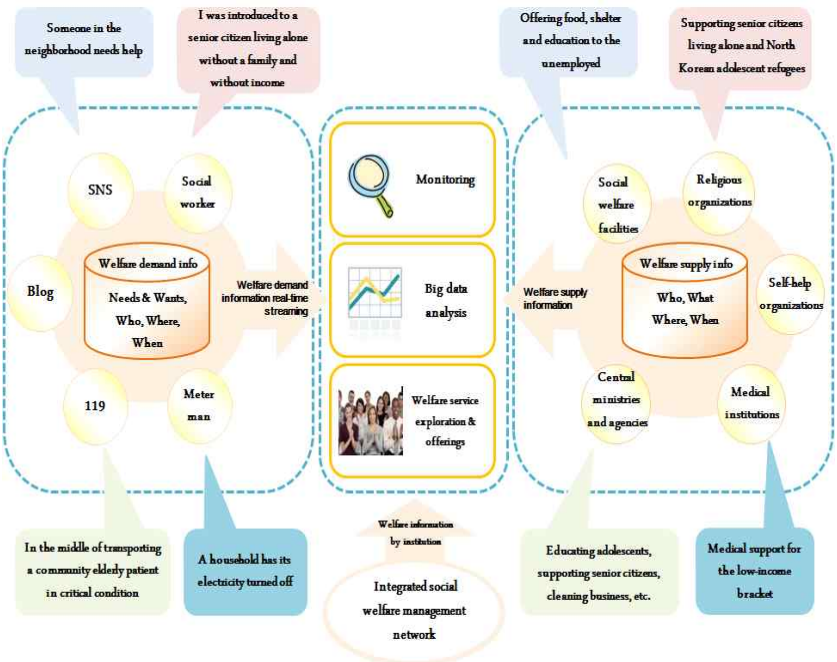


Material: Go Suk Ja and Jeong Young Ho (November, 2012), "National Health Future Prediction System Implementation Measures", 「Healthcare and Welfare Forum」, Serial No. 193, the Korean Institute for Health and Social Affairs.

In the welfare sector, the Integrated Social Welfare Management Network ('Happiness e-Connection'), which integrates and manages welfare businesses conducted by the government ministries and agencies including the Ministry of

Healthcare and Welfare, together with beneficiary information, has been up and running since February 2010. The current form of the Integrated Social Welfare Management Network should be expanded to national big data infrastructure by integrating with and linking to the information systems of the entire government ministries and agencies in order to remove blind spots of welfare and provide welfare services tailored to an individual and according to life-cycle (Hwang et al., 2013).

[Figure 3] Big Data Utilization Measures in Welfare Sector



Material: Hwang Seung Gu et al. (2013). 「Big Data Platform Strategies」, The Electronics Times Co., Ltd., p. 201.

The hereunder-described strategies shall be needed to effectively utilize big data in the healthcare and welfare sector.

First, it is required to operate the (tentatively named) Healthcare and Welfare Big Data Management Commission at the government ministries and agencies-wide level in order to manage big data in an integrated manner. At the moment, healthcare and welfare big data is managed and operated by a large number of the government ministries and agencies including the Ministry of Healthcare and Welfare, the Ministry of Employment and Labor, the formerly, Ministry of Knowledge Economy (currently, divided into the Ministry of Trade, Industry and Energy, and the Ministry of Science, ICT and Future Planning), the Korea Food and Drug Administration as well as public organizations including the National Health Insurance Corporation, the Health Insurance Review and Assessment Agency, and other state-level research organizations, etc. A government-wide level organization should be established in order to connect and share the information currently operated independently by each organization. Second, it is imperative to establish a cooperative system with private organizations that maintain unstructured big data related to healthcare and welfare. Considering that the unstructured big data related to healthcare and welfare is stored and maintained through search portals or SNS in the private sector, a close cooperative system (tentatively named, 'Healthcare and Welfare Big Data

Forum') should be established. Third, the open API (Application Programming Interface) at the national level should be made available. Most of healthcare and welfare-related big data is owned exclusively by the public sector. It is important to make information open on the web in real time, simultaneous with information collection and analysis, but more importantly, it is necessary to proactively consider the opening of API at the government level in order to utilize healthcare and welfare big data effectively and efficiently. As of February 2014, the number of open APIs made public in the open data portal on shared resources (www.data.go.kr) is 503 and among them, healthcare merely accounts for 79 and welfare for 24. The Presidential Council on Information Society proposes that the government proactively utilize big data and build a national knowledge platform primarily because now is the era when the explosively increasing data becomes an economic asset. Therefore, with respect to the opening of big data related to healthcare and welfare, healthcare and welfare big data can be categorized according to the needs of the government and the people with the participation of relevant organizations and big data professionals, and the target for open data can be stored in the national knowledge platform with strict security of personal information confidentiality. Fourth, it is necessary to develop the technologies related to the analysis and processing of healthcare and welfare big data. In the smart era, such technologies

as the storage and analysis of non-relational, unstructured data, expansion of cloud service, semantic search service, inference-based situation recognition service, etc. will become critical technologies. Accordingly, the development of technologies to enable the big data 'collection → storage → analysis → inference as well as the technology standardization should be pushed forward on a preferential basis. Fifth, it is critical to train data scientists who can detect information hidden in the large-volume data. In the Big Data era, human resources who are capable of maintaining and analyzing data are very crucial. Global IT corporations have already devoted great efforts to secure talented data scientists and strengthen competences.[4] Therefore, strategies to train data scientists in healthcare and welfare should be formulated in cooperation with the Ministry of Education. Last but not least, security policies should be prepared to deal with personal information and confidential information related to healthcare and welfare big data. Healthcare and welfare big data contains almost all information about a person, but legal and institutional systems are not ad-

4) eBay has 5,000 employees dedicated to customer data analysis and processing; EMC operates the 'Analytics' Lab composed of doctorate degree talents majored in economics, statistics, psychology, etc.; and at IBM, USA, more than 200 mathematicians within the company focus their efforts on studying 'analytics'. A forecast says that the US will be short of about 140,000 up to 190,000 professional experts and as many as 1.5 million data managers and analysts by the year of 2018 (McKinsey Global Institute (2011). Big data: The next frontier for innovation, competition, and productivity)

equately prepared nor discussed. Utilization of big data is crucial, but the leak of excessive personal information can invade a person's privacy and violate human rights in the cyber world or can be used wrongfully in crimes. One of the most critical factors in protecting a person from big data is data anonymity, meaning making a person unrecognizable, and the control of information access and processing. However, as the control of information access and processing becomes stricter, the utilization of information becomes inactive, and thus, effective policies with respect to healthcare and welfare big data 'utilization and protection' should be developed.

Chapter 3

Adolescent Suicide Risk Prediction Factors by Using Social Big Data: With Application of Decision Tree of Data Mining

3 Adolescents Suicide Risk Prediction Factors by Using Social Big Data: With Application of Decision Tree of Data Mining[5)]

The death rate resulting from suicide in Korea (suicide rate) is 29.1 per 100,000 persons, which is the highest rate among OECD member nations, compared with the average suicide rate of 12.5 persons among those countries. Moreover, as many as 1,632 out of the total death toll of 14,160 in 2012 are less than 30 years old, and suicide was the biggest cause of death in Koreans in their teens and twenties [1]. It was found in 2012 that 6 out of 10 adolescents aged between 13 and 24 years old are stressed in daily and school life, and moreover, 11.2% of teenagers were tempted at least once to become suicidal in the past year [2]. Major causes of adolescents' suicidal feelings are poor grades and proceeding to university (28.0%), followed by economic hardship (20.5%), loneliness and solitude (14.1%), family troubles (13.6%), employment problems (6.7%), and others (17.1%), which include problems related to romantic relationships, disease, disability and friendship troubles. As suicide becomes the major reason attributable to youth death, there is

5) This manuscript was originally written by Tae Min Song, Juyoung Song, Ji-Young An, and Jong-Min to prepare the draft of a paper to be submitted to an international journal.

a growing need across the world to prevent and intervene in youth death in multidimensional manner and as such multilateral efforts are made to prevent youth death at the moment.

Research on suicide has been conducted so far to investigate the causes of suicide, focusing on factors such as psychiatric, biological and medical, social and environmental, etc. [3]. Preceding research on causes of suicide behavior has discovered various factors and in particular, reported as personal factors melancholy [4], hopelessness depression [5], impulsiveness [6], lack of capacity to deal with stress [7], and low self-esteem [8], etc. Suicide can be attributed to various and complex causes and can also be affected by various external stimuli like stress [9]. Exposure to extreme stress may threaten a person's well-being beyond a person's ability or resources and is reported as a dangerous preclude to a suicide incident and suicide behavior [10, 11]. Self-killing is not a fragmentary act but rather an act of a series of processes: self-killing is entailed with illnesses such as depression and alcoholism through the interplay of social and psychological resources and stress factors [12], and over time intervals, such illnesses worsen and eventually, result in a suicide [13]. One study shows that those who have depression in ordinary life, when being exposed to external stimulus like a suicide in their vicinity, tend to imitate the suicide [14].

In particular, the causes of adolescent suicide are varied, and

an example is how Korean high schoolers choose suicide as a last resort to escape as confusion caused by frequently changing entrance examination policies and schoolwork stress creates unbearable tension for them [11, 15]. Most of self-killing adolescents choose suicide as they experience temporary impulsiveness, anger and reduced self-control rather than cognitively perceiving reality [16]. A study has discovered that a social network such as participating in a sport team at school can boost a sense of belonging and thus reduce suicide behavior in young people [17]. It was reported that self-esteem was a representative protective factor against adolescent suicide, and as such the lower the self-esteem, the higher the suicide risk [18]. It was also found that familial factors, which have an effect on youth suicide, such as parents' separation and divorce and poor bonding between parents and child, can induce stress and as such is reported as a suicide risk factor [19, 20]. An experience of being a victim of violence occurring in friendships evokes negative feelings such as depression, exerting an effect on suicidal thoughts [21]. Also, stress related to appearance and allowance are found to influence suicidal impulse [22]. The fact that a family member had committed suicide can be a factor to increase the suicide risk of a young person [23]. Parents' interest in children have the positive effect of preventing them from having suicidal feelings [24]. Teenagers lacking interaction with peers are disposed to have gloomy feelings, consequently be-

coming suicidal [25]. The less ability for self-control, the higher the suicide rate [26]. Adolescents who continually undergo stress due to low household income are highly likely to have suicidal feelings [27, 28]. In cases where an exact account of a suicide and the method thereof are described in detail in newspapers, media and the Internet, copycat suicides may ensue [29, 30, 31].

In the meantime, with the expansion of smart phone availability recently in Korea, the use of mobile internet and SNS is increasing sharply. The Internet use rate of the population aged older than 3 years stands at 82.1% as of July 2013, and among them, 55.1% of Internet users 6 years or older started to use SNS within 1 year [32]. SNS, or the Social Networking Service, has characteristics of real time and acceleration, and for this reason, issues related to society-wide problems are spread via SNS. In particular, SNS is a place where gloomy feelings, stress and worries that the adolescents have in everyday life are heard and behaviors thereof are understood, thus the analysis of emotion expressions about suicide and psychologically risky behaviors appearing on SNS can bring about the positive effect of preventing suicide by detecting risk signs and meaningful patterns [33].

With the volume of data transmitted through SNS increasing exponentially, big data is being valued as an economic asset. Its utilization and analysis are expected to become a source of

new economic value that can determine the success or failure of a country or a company, and in this sense, attempts are made to actively utilize big data in various sectors. In Korea, although a massive volume of big data is managed and stored by the portals of the government, public or private organizations, or by the SNS, the utilization and analysis of big data have not made sufficient progress due to problems related to information access and analysis methods [34]. Cross-sectional research and longitudinal research, adopted thus far to determine the causes and relevant factors, are useful in inquiring into the relationships between an individual and a group with respect to pre-determined factors, but have limitations in determining how and to what extent an individual-level buzz mentioned on cyber space is related to a social phenomenon [35].

The decision tree analysis of data mining utilizing social big data in order to overcome such limitations identifies new correlations or patterns according to decision rules based on the analysis process and without statistical hypothesis, which can be deemed as a useful tool to analyze effectively the relationships of interaction of various causes arising from a complex and dynamic phenomenon of human behavior like suicide [36].

Accordingly, this research document is intended to present a model to predict suicide risk in youth and relevant rules based on social big data collected from online news sites, blogs, cafes, SNS, bulletin boards, etc. The purpose of this research docu-

ment is to present a model to predict suicide risk in Korean youth by utilizing social big data through the decision tree analysis of data mining, and its specific purposes are as follows. First, determine factors affecting adolescent suicide risk. Second, develop a decision tree that can predict the risk factors of adolescent suicide.

1. Research Methods

A. Research Target

This research was conducted targeting social big data collected on the Internet such as domestic online new sites, blogs, cafes, SNS, bulletin boards, etc. According to this analysis, social big data is defined as text-based buzz that can be collected from a total of 228 online channels including 215 online news sites, 4 blogs (Naver, Nate, Daum, Tistory), 3 cafes (Naver, Daum, Ppomppu), 2 SNS's (Twitter, me2day), and 4 bulletin boards (NaverKnowledgeiN, NateKnowledge, NateTalk, NatePann). Adolescent suicide topics were collected from a relevant channel during a period between January 1, 2011 and March 31, 2013 (a total of 821 days) on an hourly basis regardless of weekdays, weekends and holidays. Out of a total of 221,691 cases that were collected, the 83,657 buzz cases (37.7%) commenting on the causes of youth suicide were included in the analysis of this

research.

Crawler was adopted to collect social big data required for this research and the text mining technique was used to classify topics. Crawler collects documents by generating a copy of all pages of a site visited by following an Internet link. Text mining means the finding of useful information of social big data by deriving useful information through the use of national language processing technology or by determining, categorizing or grouping connectivity. With respect to adolescent suicide-related topics, 15 synonyms such as 'girl student suicide, male middle schooler suicide, male high schooler suicide, male student suicide, femal middle schooler suicide, middle schooler suicide, female high schooler suicide, high schooler suicide, Goding (Korean slang for a high schooler) suicide, Jungding (Korean slang for a middle schooler) suicide, Choding (Korean slang for an elemetary student) suicide, elementary schooler suicide, middle school suicide, high school suicide, elementary school suicide' were used to cover all adolescents. Moreover, 46 stopwords unrelated to suicide such as 'own goal, NLL, Kim Jang Hoon, Mun Jai In, Noh Mu Hyun, Ahn Chul Su, battle, senior presidential secretary for civil affairs, President, Kim Ju Ik, crocodile bird, suspicion, Park Jung Hee, Park Guen Hye, Kim Dae Jung, candidate, grabbing public opinion, war of nerves, independent, presidential candidate, the ruling party, the opposition party, Arang (name of unmarried female ghost),

Arangsattogeon (folk tale about Arang and a district magistrate), Gang Mun Young, Shin Min Ah, Mu Young, Mu Yeon, Lee Jun Ki, Gwang Hae, Lee Byung Hyun, Werther, Movie Gwang Hae, Alain De Botton, full-length novel, novel, movie, living alone, live alone, Lee Myung Bak, dolphin show, alone living, animal show (of dolphin) were used for collection during the collection period.

B. Research Tools

Buzz collected with respect to adolescent suicide went through the process of text mining and opinion mining, and then coded to structured data before the use. Opinion mining refers to the analysis of psychological expressions-- that is, whether collected suicide-related buzz is positive ('will commit a suicide', 'like suicide', etc.) or negative ('suicide pitiful', 'suicide bad', etc.).

1) Suicide Causes

Causes of adolescent suicide are classified into 18, such as 'death, college scholastic ability test, depression, sexual assault, pains, shock, grades, worries, stress, discharge, bullying, violence, hardship of life, divorce, inferiority feeling, face, disease, and school violence,' which are determined as suicide

causes in the theoretical background and in 2013 adolescents statistics [2]. In addition, a value of '1' is coded if a target is applicable, and '0' if not applicable.

2) Suicide Methods

Suicide methods are classified into 18 such as 'jumping, self-burning, joint suicide, electrical cord, small-size briquette, suicide by taking poison, self-harm, suicide by disembowelment, sleeping pill, briquette, rope, necktie, agricultural pesticide, gas, high caffeine, steel wire, glue, and suffocation' and in addition, a value of '1' is coded if a method is applicable and '0' if not applicable.

3) Suicide Opinion

In suicide opinion, opinion mining codified buzz (1 to positive, 2 for medium and 3 for negative) depends on whether suicide-related expressions included in the buzz are positive (e.g., 'will commit suicide', 'choose suicide', 'suicide is good', 'suicide is easy'), or negative (e.g., 'suicide is pitiful', 'suicide is serious', 'suicide is shocking') or medium, (the mixture of positive and negative expressions). Suicide opinion factors used in this research were coded to suicide risk (a value of 1 is for documents recognizing suicide positively or as medium) and

suicide protection (a value of 0 for documents negatively recognizing suicide).

C. Analysis Methods

This research study adopted the decision tree analysis method of data mining, not requiring special statistical hypothesis, in order to create the most effective predictive model that can explain the risk factors of Korean adolescent suicide. The decision tree analysis of data mining can easily determine risk factors of adolescent suicides attributable to different causes, by automatically generating a predictive model that best explains dependent variables amid the enormous amount of material.

Exhaustive CHAID was used as an analysis algorithm to develop the decision tree of this research primarily because Exhaustive CHAID has the highest prediction rate of models among CHAID (Chi-squared Automatic Interaction Detection), Exhaustive CHAID, CRT (Classification and Regression Tree), and QUEST (Quick, Unbiased, Efficient Statistical Tree) growing methods [37]. Exhaustive CHAID uses chi-square (χ^2-test) as separation standard of discrete-type dependent variables and determines the most optimal separation by exploring all possible combinations [37]. As there are sufficient observable numerical numbers with stopping rules, the minimum number of cases at the higher node (parent node) is set as 100 and that of

lower node (child node) as 50, with the depth of tree defined at 3 levels [38]. Moreover, the rates of training data and test data are set as 70 and 30, respectively, for feasibility study according to data partitioning. SPSS v. 20.0 was adopted for technology analysis, multiple logistic regression analysis, and decision tree analysis.

D. Ethical Considerations of Research

In due consideration of ethics, this research was begun after receiving approval (No.2014-1) from the Institutional Review Board (IRB) of the Korea Institute for Health and Social Affairs. The secondary material that the Korea Institute for Health and Social Affairs and SKT collected in May 2013 was used as research material, and the collected social big data was unable to recognize personal information, ensuring anonymity and confidentiality in the study. This research was performed through budget support from the Office of the Prime Minister of the Korean Government, (Project Name: Research on Effective Management Measures of Healthcare and Welfare Big Data).

2. Research Results

A. Descriptive Statistics of Major Factors

A total of 221,61 buzz cases commented on suicide topics and among them, 37.7% of buzz mentioned the causes of suicide. School violence was ranked first as the cause of adolescent suicide searches, followed by gloomy feelings, grades, appearance, feeling of inferiority, shock, and hardship of life. Adolescent suicide methods include jumping, suicide by hanging or suffocation by applying pressure (necktie, steel wire), pesticide poisoning and exposure (agricultural pesticide), and other means (briquette, glue, self-burning, caffeine) (Table1).

〈Table 1〉 Descriptive Statistics of Factors

Causes of Suicide				Methods of Suicide			
Cause	Buzz (%)	Cause*	Buzz (%)	Method	Buzz (%)	Method†	Buzz (%)
Dying	10309 (6.4)	School Violence	28665 (25.3)	Jumping	3404 (38.2)	Glue	1017 (12.0)
SAT	2768 (1.7)			Posse	172 (1.9)		
Melancholia	8518 (5.3)	Dying	13809 (12.2)	Double suicide	810 (9.1)	Coal briquette	1139 (13.5)
Sexual Assault	2977 (1.8)			Cords	199 (2.2)		
Pain	9195 (5.7)	Inferiority	9726 (8.6)	Briquette	56 (0.6)	Pesticide	1031 (12.2)
Impact	9057 (5.6)			Overdose	121 (1.4)		
Test scores	16234 (10.0)	Melancholy	17921 (15.8)	Self-harm	1130 (12.7)	Wire	293 (3.5)
Worry	8548 (5.3)			Disembowelment	62 (0.7)		
Stress	10833 (6.7)	Appearance	15213 (13.4)	Sleeping pills	668 (7.5)	Jumping	3577 (42.3)
Firing	643 (0.4)			Coal briquette	345 (3.9)		
Bullying	15282 (9.4)			Tether	126 (1.4)		

Causes of Suicide				Methods of Suicide			
Violence	28665 (17.7)	Economy	1275 (1.1)	Tie	111 (1.2)	Posse	172 (2.0)
Economy	742 (0.5)			Pesticide	300 (3.4)		
Divorce	6249 (3.9)	SAT	17791 (15.7)	Gas	841 (9.4)	Tie	1227 (14.5)
Inferiority	1234 (0.8)			Caffeine	7 (0.1)		
Appearance	8393 (5.2)	Impact	9057 (8.0)	Wire	45 (0.5)	Caffeine	7 (0.1)
Disease	2142 (1.3)			Glue	247 (2.8)		
School Violence	20127 (12.4)			Suffocation	256 (2.9)		

* Cause analysis (unique value of 1 or greater) was conducted on 18 causes and cause factors were reduced to 8.

†Cause analysis (unique value of 1 or greater) was conducted on 18 methods and method factors were reduced to 8.

B. Cause Factors Affecting Adolescent Suicide Risk

Except for death factor, all factors exert greater influence on suicide risk than on suicide protection. Factors affecting suicide risk include appearance, feelings of inferiority, melancholia, shock, school violence, hardship of life, and grades in the order named.

〈Table 2〉 Multinomial Logistic of Suicide Cause Factors

Variable	B	S.E.	P	OR(95%CI)†
Intercept	-1.219	0.014	.001	
School Violence	0.179	0.017	.001	1.196(1.16~1.24)
Dying	-0.240	0.022	.001	0.786(0.75~0.82)
Inferiority	0.272	0.023	.001	1.313(1.25~1.37)
Melancholia	0.255	0.019	.001	1.290(1.24~1.37)
Appearance	0.514	0.019	.001	1.673(1.61~1.74)
Economy	0.159	0.062	.01	1.172(1.04~1.32)
SAT	0.128	0.019	.001	1.136(1.09~1.18)
Impact	0.204	0.024	.001	1.227(1.17~1.29)

†Adjusted odds ratio (95% Confidence interval)

C. Predictive Model of Adolescent Suicide Risk Factor

A model with the highest predictive performance was selected by verifying predictive performance (correct percentage) of model according to tree-structure classification model by using node separation rules. According to the result of comparison and analysis of correct percentage, which indicates the accuracy level of tree separation, both Exhaustive CHAID and CRT algorithms showed the highest rates of 72. 1% and 72.0% in training data and test data, respectively. However, CRT algorithm produced lower correct percentage in training data than Exhaustive CHAID in the 2nd data mining, and Exhaustive CHAID algorithm, which showed highest correct percentage both in training data and test data, was eventually selected (Table 3).

〈Table 3〉 Predictive Performance according to Modeling Methods

Modeling methods	Training data		Test data	
	Correct (%)	Wrong (%)	Correct (%)	Wrong (%)
CHAID	72.1	27.9	71.9	28.1
Exhaustive CHAID	72.1	27.9	72.0	28.0
CRT	72.1	27.9	72.0	28.0
QUEST	72.1	27.9	71.7	28.3

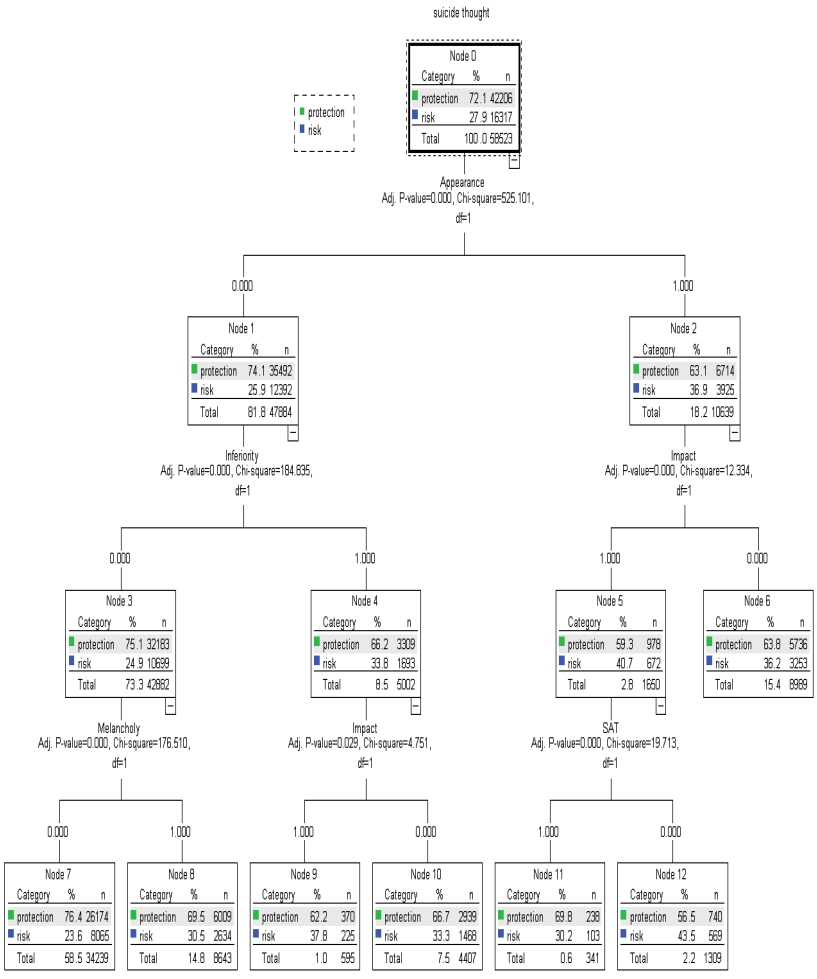
Analysis result of decision tree with respect to the predictive model of Korean adolescent suicide risk is illustrated in the Figure 4. The highest-level rectangle of the tree structure of the

training data is a root node, indicating the frequency of dependent variables (adolescent suicide risk) wherein predictive variables (independent variables) are not inputted. At the root node, 27.9% (16,317 cases) is shown for adolescent suicide risk and 72.1% (42,206 cases) for suicide protection. The factor at the top among the lower part of the root node has the greatest (the most related) effect on youth suicide risk prediction, and it was found that 'appearance factor' has the largest influence. In other words, if the risk of 'appearance factor' was high, the youth suicide risk increased from 27.9% to 36.9%, whereas the youth suicide protection decreased from 72.1% to 63.1%. Moreover, if the risk of 'appearance factor' was high and 'shock factor' was high, the youth suicide risk rose from 36.9% to 40.7%, whereas the youth protection declined from 63.1% to 59.3%. As compared therewith, if the risk of 'grade factor' was high, even if that of 'shock factor' was high, the youth suicide risk dropped from 40.7% to 30.2%, whereas the youth suicide protection grew from 59.3% to 69.8%.

In case 'appearance factor' at the bottom of the root node had a low risk, the youth suicide risk decreased from 27.9% to 25.9%, whereas the youth suicide protection increased from 2.1% to 74.1%. In case 'inferiority feeling factor' was high, even if 'appearance factor' was low, then the youth suicide risk rose from 25.9% to 33.8% whereas the youth suicide protection fell from 74.1% to 66.2%. As compared therewith, in case

'inferiority feelings factor' was high and 'shock factor' was high, the youth suicide risk grew from 33.8% to 37.8%, whereas the youth suicide protection declined from 66.2% to 62.2%.

[Figure 4] Predictive Model of Suicide Risk Factors



D. Profit Chart of Predictive Model of Adolescent Suicide Risk Factors

This research paper shows that the highest rate of adolescent suicide risk is found in the combination of a high risk of 'appearance factor', a high risk of 'shock factor' and a low risk of 'grade factor'. That is, the index of the no. 12 node is 155.9%, and compared with the knotroot, the adolescent group satisfying the conditions of the no. 12 node has 1.56 times greater risk of youth suicide. By contrast, the highest rate of youth suicide protection is found in the combination of a low risk of 'appearance factor', a low risk of 'inferiority feeling' and a low risk of 'melancholy factor'. In other words, the index of the no. 7 node is 106.0% and compared with knotroot, the adolescent group satisfying the conditions of the no. 7 node has 1.06 times lower risk of youth suicide protection (Table 4). Training data and test data were compared to perform feasibility evaluation according to data partitioning, and the result shows that the risk estimate of training data is 0.279 (standard error: 0.002) and that of test data is 0.280 (standard error: 0.003), indicating that the generalization of this predictive model of adolescent suicide risk factor holds water.

〈Table 4〉 Profit chart of Predictive models of Suicide

Type	Node	Profit index				Cumulative index			
		node: n	node: %	gain (%)	index (%)	node : n	node : %	gain (%)	index (%)
Risk	12	1309	2.2	3.5	155.9	1309	2.2	3.5	155.9
	9	595	1.0	1.4	135.6	1904	3.3	4.9	149.6
	6	8989	15.4	19.9	129.8	10893	18.6	24.8	133.3
	10	4407	7.5	9.0	119.5	15300	26.1	33.8	129.3
	8	8643	14.8	16.1	109.3	23943	40.9	49.9	122.1
	11	341	0.6	0.6	108.3	24284	41.5	50.6	121.9
	7	34239	58.5	49.4	84.5	58523	100.0	100.0	100.0
Protection	7	34239	58.5	62.0	106.0	34239	58.5	62.0	106.0
	11	341	0.6	0.6	96.8	34580	59.1	62.6	105.9
	8	8643	14.8	14.2	96.4	43223	73.9	76.8	104.0
	10	4407	7.5	7.0	92.5	47630	81.4	83.8	102.9
	6	8989	15.4	13.6	88.5	56619	96.7	97.4	100.6
	9	595	1.0	0.9	86.2	57214	97.8	98.2	100.5
	12	1309	2.2	1.8	78.4	58523	100.0	100.0	100.0

3. Discussion and Policy Proposal

This research is purported to develop a predictive model for Korean adolescent suicide risk factors through analysis of social big data by applying the decision tree analysis technique of data mining. Results of this research are summed up as follows. First, the causes of youth suicide on SNS include school violence, melancholy, grade, appearance, death, feelings of inferiority, shock, hardship of life, etc. It makes a contrast with the causes of suicide investigated in the 2012 adolescent statistics, i.e., grade, economic hardship, loneliness, family trouble, employment problem, and others (opposite sex, disease, friendship trouble).

Second, youth suicide methods on SNS include jumping (42.3%), suffocation (18.0%), pesticide (12.2%), glue (12.0%), etc. in the order named. This result shows similarities with the actual suicide methods of adolescents of 19 years or younger investigated in the 2012 Suicide Causes Statistics by the National Statistical Office, i.e., jumping (57.1%), suffocation (27.5%), glue (7.4%), pesticide (2.1%) and others, but a slight difference between the two results is that apparently, Korean adolescents tend to choose jumping, which is the most violent, impulsive and fatal method of self-killing. The fact that adolescents' suicide methods searched on SNS are similar to the suicide methods actually used by the adolescents who killed themselves gives support to the research studies [29, 30, 31] that the exact account of a suicide incident and suicide methods can trigger copycat suicides and as such, giving too much detail of a suicide should be regulated autonomously by media.

Third, the youth suicide risk factors include appearance, feelings of inferiority, melancholia, shock, school violence, hardship of life, and grade factors in the order named. In other words, it affirms the research that shows appearance [22], self-esteem [18], melancholia [14], impulse [16], school violence [21], hardship of life [27, 28], schoolwork stress [11, 15] have an effect on suicidal impulse.

Fourth, one of the most critical factors to predict adolescent suicide risk is appearance. The higher the appearance factor,

the higher the suicide risk and the lower suicide protection, affirming the existing research [22] that says that appearance stress affects suicidal impulse. In addition, in case of a high risk of appearance factor and a high risk of impulse factor, the suicide risk increases and the suicide protection declines, confirming the existing research [16] that says that most of the adolescents who chose to kill themselves were affected by impulsiveness, anger, reduced self-control, etc. rather than cognitive perceiving of reality. Also, it seems that the above-mentioned is related to the searches of suicide on the heels of reporting of a suicide incident as a consequence of a failed plastic surgery [39]. As compared therewith, in case both appearance factor and impulse factor are high and grade factor is high as well, suicide risk declines whereas suicide protection rises, supporting existing research [11, 15] that says confusion over the frequently changing entrance examination policies and schoolwork stress affect suicide.

Fifth, if the inferiority feeling factor is high, even if the risk of appearance factor is low, the suicide risk rises and the suicide protection declines, upholding the existing research [18] that says that self-esteem is a representative protection factor against adolescent suicide. In addition, if the risk of melancholia factor is high, even if that of appearance factor and inferiority feeling factor are low, suicide risk increases and suicide protection decreases, supporting the existing research [14]

that says when those who have depression are exposed to a suicide of someone close to them, copycat suicides may rise.

Last but not least, the highest risk of adolescent suicide is found in a combination of a high risk of appearance factor, a high risk of shock factor and a low risk of grade factor, with the suicide risk of this group increasing by 1.56 times compared with the knotroot. By contrast, the highest risk of suicide protection is found in a combination of a low risk of appearance factor, a low risk of inferiority feeling factor and a low risk of melancholia factor. The suicide protection risk of this group, compared with knotroot, grew by 1.06 times, indicating that the causes of suicide commented on online have a greater affect on suicide risk than suicide protection.

Based on this research, a policy-level agreement can be reached with respect to suicide prevention in Korea. First, appearance, shock, grade, melancholia, and inferiority feeling factors have a complex effect with respect to suicide risk factors, and thus, measures to prevent and address complex factors should be prepared within the government's student suicide prevention and risk management framework. Second, as adolescents exchange suicide-related buzz online, such comments may reveal psychological and behavioral characteristics related to actual suicide and thus, it will be necessary to develop an application enabling intervention real-time as soon as any risk symptom is predicted according to the predictive mod-

el of suicide in this research [34]. Third, suicide-relevant social buzz exchanged on SNS is, presumably, an online psychological postmortem report wherein gloomy feelings and worries that a person has in everyday life are registered. Therefore, as seen in Finland where the Finland government reduced suicide rates based on the government-level offline report, Korea should come up with government-level suicide prevention measures through analysis of social big data [35].

In this research, analysis was conducted with the materials of a group of members, not by using characteristics of each and every individual and for this reason, if being applied to an individual, may cause an ecological fallacy. In addition, it must be noted that the factors (terms) related to suicide causes were defined as a frequency of words appearing within buzz, so the meaning thereof may differ from that of suicide causes used in theoretical models through the existing investigation. It seems that verification will be required in a subsequent research study. Despite these limitations, this research is meaningful in terms of analysis methodology in that it categorized suicide causes, methods, etc. through text mining and opinion mining of suicide buzz collected from social big data and presented a predictive model for adolescents suicide risk factors through data mining analysis. This research also carries significance from the perspective of analysis methodology in that the value of big data was confirmed as a new research method that can

supplement the limitations of social statistics by quickly and effectively grasping actual details.

4. Conclusion

This research proposes a model and its relevant rules that can predict adolescent suicide risk based on social big data collected from domestic online news sites, blogs, cafes, SNS, bulletin boards, etc. and specific purposes of this research are provided as follows. First, determine factors affecting adolescent suicide risk. Second, develop the decision tree, which can predict adolescent suicide risk factors. This research used the decision tree analysis technique of data mining, which does not require special statistical hypothesis, in order to develop the most effective prediction model, which explains the Korean adolescent suicide risk factors. In this research, out of 221,691 cases of buzz collected from 228 online channels during a period between January 1, 2011 and March, 31, 2013 (821 days), the 83,657 cases of buzz that commented on the causes of adolescent suicide became a target of analysis. The highest risk of adolescent suicide was found in a combination of a high risk of appearance factor, a high risk of shock factor and a low risk of grade factor, with the suicide risk of this group growing about 1.56 times, compared with knotroot. As adolescents exchange suicide-related buzz online, such comments may reveal psy-

chological and behavioral characteristics related to actual suicide, and thus, it will be necessary to develop an application enabling intervention real-time as soon as any risk symptom is predicted according to a predictive model of suicide in this research.

References

[1] Statistics Korea. Annual report on the causes of death statistics 2012. Seoul, Korea: Statistics Korea; 2013.

[2] Statistics Korea. 2013 Youth statistics. Seoul, Korea: Statistics Korea; 2013.

[3] Bae SB, Woo JM. Suicide prevention strategies from medical perspective. J Korean Med Assoc 2011;54(4):386-91.

[4] Konick LC, Gutierrez PM. Testing a model of suicide ideation in college students. Suicide and Life Threatening Behavior 2005;35(2):181-92.

[5] Goldston DB, Daniel SS., Reboussin DM, Frazier PH, Harris AE. Cognitive risk factors and suicide attempts among formerly hospitalized adolescents: A prospective naturalistic study. J am Acad Child Adolsec Psychiatry 2001;40(1):91-9.

[6] Turecki G. Dissecting the suicide phenotype: The role of impulsive-aggressive behaviours. J psychiatry Neurosci 2005;30(6):398-408.

[7] Rudd MD, Rajab MH. Treating suicidal behavior: An effective time-limited approach. New York (NY): Guilford Press; 2001.

[8] Wilburn VR, Smith DE. Stress, self-esteem, and suicidal ideation in late adolescents. Adolescence 2005;40:33-45.

[9] Hong YS. Jeon SY. The effects of life stress and depression for adolescent suicidal ideation. Mental Health & Social Work 2005;19:125-49.

[10] Izadinia N, Amiri M, Jahromi RG, Hamidi S. A study of relationship between suicidal ideas, depression, anxiety,

resiliency, daily stresses and mental health among Teheran university students. Procedia Soc Behav Sci 2010;5:1515-19.

[11] Yoon MS, Lee HS. The Relationship between depression, Job preparing stress and suicidal ideation among college students: moderating effect of problem drinking. Kor J Youth Studies 2012;19(3):109-37.

[12] Alloy LB, Hartlage S, Abramson LY. Testing the cognitive diathesis-stress theories of depression: Issues of research design, conceptualization, and assessment. In Alloy LB(Ed.). Cognitive Processes in depression. New York, NY (NY): The Guilford Press; 1988;31-73.

[13] Park JY, Lim YO, Yoon HS. Suicidal impulse caused by stress in Korea: Focusing on mediation effects of existent spirituality, family support, and depression. Kor J Soc Welfare Studies 2010;41(4):81-105.

[14] Peruzzi N, Bongar B. Assessing risk for completed suicide in patients with major depression: Psychologists' views of critical factors. Professional Psychology; Research and Practice 1999;30(6):576-80.

[15] Seo SJ, Jung MS. Effect of family flexibility on the idea of adolescents suicide: the senior year of high school boys. J Kor Contents Assoc 2013;13(5):262-74.

[16] Kim MK. Relationship between negative emotions, family Resilience, self-esteem and suicide ideation in university students. Kor Assoc Family Welfare 2012;17(1): 61-83.

[17] Brown DR, Blanton CJ. Physical activity, sport participation, and suicidal behavior among college students. Med Sci Sports Exerc 2012;34(7):1087-96.

[18] Heather LR, Michael AB, Nishad K, Youth Net Hamilton. Youth engagement and suicide risk: Testing a mediated model in a Canadian community sample. J Youth Adolescence 2010;39(3):243-58.

[19] Beautrais AL. Suicide and serious suicide attempts in youth: A Multiple-Group comparison study. Am J Psychiatry 2003;160(6):1093-99.

[20] Meltzer H, Harrington R, Goodman R, Jenkins R. Children and adolescents who try to harm, hurt or kill themselves. National Statistics London Office;2001

[21] Kaufman JM. Gendered responses to serious strain: The argument for a general strain theory of deviance. Justice Q 2009;26(3):410-44.

[22] Kim KM, Youm YS, Park YM. Impact of school violence on psychological well-being Korean students happiness and suicidal impulse. J Kor Contents Assoc 2013;13(1):236-47.

[23] Bridge JA, Brent DA, Johnson BA, Connolly J. Familial aggregation of psychiatric disorders in a community sample of adolescents. J Am Acad child Adolesc Psychiatry 1997;36:628-37.

[24] Fergusson DM, Lynskey MT. Childhood circumstances, adolescent adjustment, and suicide attempts in a New Zealand birth cohort. J Am Acad child Adolesc Psychiatry 1995;34(5):612-22.

[25] Kendel DB, Ravis VH, Davies M. Suicidal ideation in adolescence: Depression, substance abuse, and other risk factor. J Youth Adolesc 1991;20(2):289-309.

[26] Kim HJ. Effect factors of Adolescences suicide risk. J Kor soc child welfare 2008;27:69-93.

[27] Toprak S, Cetin I, Guven T, Can G, Demircan C. Self-harm, suicidal ideation and suicide attempts among college. Psychiatry Res 2011;87:140-44.

[28] Kim SA. Effects of childhood stress, depression, and social support on middle school adolescent suicidal ideation. Kor J Family Welfare 2009;14(3):5-27.

[29] Kim BS, Yun DH. Prevention of suicide infection on part of ad\-vance guard in media. Newspapers and Broadcasting 2011;7:22-6.

[30] Stack S. Suicide in the media: a quantitative review of studies based on non-fictional stories. Suicide Life Threat Behav 2005;35: 121-33.

[31] Tousignant M, Mishara BL, Caillaud A, Fortin V, St-Laurent D. The impact of media coverage of the suicide of a well-known Quebec reporter: the case of Gaëtan Girouard. Soc Sci Med 2005;60:1919-26.

[32] The Ministry of Science. ICT and future planning·Korea Internet & Security Agency. 2013 Survey on the Internet Usage. Seoul, Korea: The Ministry of Science; 2013.

[33] National Information Society Agency. Implications for suicide prevention policy of youth described in the social Analysis. Seoul, Korea: National Information Society Agency; 2012.

[34] Song TM, Song J, An JY, Jin D. Multivariate Analysis of Factors for Search on Suicide Using Social Big Data. Korean Journal of Health Education and Promotion 2013;30(3):59-73.

[35] Song TM, Song J, An JY, Hayman LL, Woo JM. Psychological and Social Factors Affecting Internet Searches on Suicide in Korea: A Big Data Analysis of Google Search Trends. Yonsei Med J 2014;55(1):254-263.

[36] Lee JL. A Forecast Model on High School Students' Suicidal Ideation: The Investigation Risk Factors and Protective Using

Data Mining. Journal of the Korean Home Economics Association 2009;47(5):67-77.

[37] Bigg DB, Ville B, Suen E. A Method of choosing multiway partitions for classification and decision trees, Journal of Applies Statistics, 1991;18:49-62.

[38] SPSS Inc. AnswerTree 1.0 User's Guide, Chicago: SPSS Inc 1998.

[39] Failure pessimism suicide of plastic surgery. Korea New Network. [accessed on 2014 February 8. Available at: http://www.knn.co.kr/news/todaynews_read.asp?ctime=20130625064242&stime=20130625064625&etime=20130625064227&userid=skkim&newsgubun=accidents.

http://www.naver.com/

http://www.nate.com/

http://www.daum.net/

http://www.tistory.com/

http://www.ppomppu.co.kr/

https://twitter.com/

http://me2day.net/

http://kin.naver.com/index.nhn

http://ask.nate.com/

http://pann.nate.com/talk/

http://pann.nate.com/

Chapter 4

Predictive Model of Risk Factors by Korean Cyber Bullying Types:

With Application of Data Mining Using Social Big Data

4 Predictive Model of Risk Factors by Korean Cyber Bullying: With Application of Data Mining Using Scocial Big Data[6)]

Nowadays, smart media has been widely made available across the world and the use of mobile Internet and SNS has soared rapidly in daily lives. SNS, which connects the relationships among an individual, group and society in the network, has characteristics of real-time and acceleration, and for this reason, the speed of propagating an issue is faster than in any other media. The Internet and SNS have positive effects such as information search and online chatting as well as negative effects including cyber bullying, Internet addiction and preoccupation with games. In particular, SNS is utilized as a venue where gloomy feelings, stress, and worries that adolescents feel in daily lives are expressed and relieved, but this has emerged as a serious social problem as teenagers exposed to cyber bullying on SNS chose to kill themselves or became a bully offender inflicting harm on a victim. It was found in a study that as of November 2013 in Korea, 29.2 % of the youth and 14.4% of ordinary people cyber-bullied other persons, while 30.3% of the youth and 30.0% of ordinary people fell victim to cyber bullying [1]. Considering that cyber bullying can be committed

6) This manuscript was originally written by Tae Min Song, Juyoung Song to prepare the draft of a paper to be submitted to an international journal.

persistently on cyberspaces like SNS regardless of time and place, victims of cyber bullying, as with the victims of traditional bullying in the psychological distress perspective, are inclined to become violent [2] or lethargic [3] or become emotionally unbalanced, and eventually the victim can sustain psychological injuries such as self-harm and suicidal impulse [4].

Traditional bullying generally refers to a situation where a student is exposed to repeating and persistent negative behavior by one or more students and that student is a bully victim [5], with negative behavior including psychological and physical harrassment and so on. [6]. Cyber bullying is an aggressive act or behavior that is carried out using electronic means by a group or an individual repeatedly and over time against a victim who cannot easily defend himself or herself [7:p.26]. Cyber bullying is categorized into seven types [8]. Flaming refers to online fights using electronic messages with angry and vulgar language. Cyber harrassment refers to an act of repeatedly sending nasty, mean, and insulting messages. Denigration means dissing someone online by sending or posting gossip or rumors about a person to damage his or her reputation or friendships. Impersonation is an act of pretending to be someone else and sending or posting material to get that person in trouble or danger or to damage that person's reputation or friendships. Outing means sharing someone's secrets or embarrassing information or images online. Exclusion refers to an act

of intentionally and cruelly excluding someone from an online group. Cyber-stalking means repeated, intense harassment and denigration that includes threats or creates significant fear. As such, the term of cyber bullying is used in terms of definition as a comprehensive meaning, as with cyber violence.

It was found in some studies that those who participated in traditional bullying are disposed to perpetrate cyber bullying [9], and those who fell victim to traditional bullying are likely to sustain injuries of cyber bullying [10]. Self-harm and suicidal impulse, which are discovered in traditional bullying, are closely related to cyber bullying as well [4]. It was reported that the student-victims of traditional bullying are highly impulsive [11], and the psychological characteristics of student-offenders are related to impulsiveness [5]. Traditional bullying is related to the stress that an individual perceives [12, 13]. It was discovered that traditional bullying is a phenomenon of cultural clash that is unavoidable due to cultural discrepancies between an individual and a group [14], and that traditional bullying may occur in the process of imitating, learning and practicing the group culture or a phenomenon of group bullying that is depicted in a movie, TV, a play, etc. [15]. It was found that traditional bully offenders are positively related to dominance propensity [5,16]. Reportedly 37% of children of multi-cultural families in Korea fall victim to traditional bullying [17]. It was reported that they are sometimes victimized by traditional bul-

lying due to physical characteristics such as being fat and not fitting with others [18], and that traditional bullying is related to a victim student's appearance [19] as well as relationships and social skill [20].

Traditionally, bullying in one class classifies participants as a bully offender, a bully victim and a bystander(or outsider), depending on roles an individual student plays [21]. An offender child is the main actor of bullying and actively (sometimes intentionally) instigates other children to follow suit; a victim child is a target of bullying and experiences physical and psychological injuries; and the majority of children, except for the offender and victim, are bystanders and are intrigued or watch the bullying but are reluctant to intervene by dissuading an offender from bullying or extending a helping hand to a victim [22]. As such, the fact that a majority of bystanders contributed to preservation of the status quo or worsening of peer bullying as a matter of fact was discovered [23], and thus, greater attention is paid to a bystander. In a domestic research paper, traditional bullying is classified into the following: a group of offenders who have higher tendency to inflict harm and lower tendency to be victimized; a group of victims who have lower tendency to inflict harm and higher tendency to be victimized; a group of offenders/victims who have higher tendency to inflict harm and be victimized; a general group who have lower tendency to cause harm and fall victim [24,25,26].

A study classified the proportion of bully-related actors into 14.1% for a group of offenders, 4.8% for a group of victims, 6.8% for a group of offenders/victims, and 74.2% for a general group [25]. In a related study, 61.8% of all students belong to a general group and the remaining ratio of 38.2% are ascribed to a group of offenders, a group of victims and a group of offenders/defenders [26]. In traditional bullying, a victim has more feelings of insecurity than an offender [27], and in addition thereto, as anger persists longer and increases, such feelings pile up and the victim sometimes evolves to become an offender [28]. Reportedly, a victim who sustains injuries of the most extreme bullying may sometimes become the most violent offender [29]. Despite the gravity of cyber bullying as described so far, related research studies were not conducted sufficiently [7]. Up until now, studies on cyber bullying are limited to the definition of traditional bullying and cyber bulling [5,7,8] and the relevance thereof [30]. However, to date, there has been no research done to empirically verify the relationships between cyber bullying and traditional bullying by analyzing actually occurring bullying behaviors on cyberspace.

Meanwhile, the volume of data transmitted on SNS soared exponentially and the value of data is recognized as an economic asset. Against this backdrop, attempts are made to actively utilize big data in various areas. In Korea, an enormous volume of big data is managed and stored by portals and SNS's

of the government and public or private organizations, but the utilization and analysis of big data have not made progress sufficiently largely due to the difficulties of information access and analysis methods. In particular, a study utilizing existing cross-sectional research or longitudinal research, which are used to determine the causes of cyber bullying and its related factors, is useful in determining the relationships between an individual and a group with respect to pre-determined variables but has limitations in how and to what extent an individual-level buzz commented on cyberspace is related to a social phenomenon. In this sense, the decision tree analysis of data mining, which utilizes social big data, is a useful tool to effectively analyze the interaction relationships of various causes arising from a complex and dynamic phenomenon of human behaviors like cyber bullying by finding a new correlation or pattern according to decision rules without special statistical hypothesis. Accordingly, this study proposes a predictive model and its related rules that can explain causes of cyber bullying according to types thereof based on big social data collected from domestic news sites, blogs, cafes, SNS, bulletin boards, etc. The purpose of this research is to propose a predictive model of risk factors by cyber bullying type in Korea through decision tree analysis of data mining by utilizing social big data, with its specific purposes provided as follows. First, classify the types of cyber bullying and determine affecting fac-

tors by type. Second, develop a decision tree that can determine risk factors for each cyber bullying type.

1. Research Methods

A. Research Target

This research targeted social big data collected from the Internet such as domestic online news sites, blogs, cafes, SNS, bulletin boards, etc. In this research, social big data is defined as text-based web documents (buzz) that can be collected from a total of 227 online channels, i.e., 214 online news sites, 4 blogs (Naver, Nate, Daum, Tistory), 3 cafes (Naver, Daum, Ppomppu), 2 SNS's (twitter, me2day), 4 bulletin boards (NaverKnowledgeiN, NateKnowledge, NateTalk, NatePann) etc. A total of 435,563 cases of cyber bullying-related topics were collected from each channel during the period between January, 1, 2011 to March 31, 2013 (821 days) and thereafter analyzed. In this study, cyber bullying is used as a term of having comprehensive meaning, as with cyber violence defined by Willard (2007) [8]. Crawler was used to collect social big data for this research, and the text mining technique was adopted to classify topics. Cyber bullying related topics were collected using 33 synonyms as listed hereunder: 'euntta (bullying in secret), youngtta (bullying forever), youngwonhi tta (bullying forever), jeontta (bullying by the whole school), bantta (bullying by

class or small group), ttungtta (bullying for being obese), bullying for obesity, jiktta (workplace bullying), workplace bullying, workplace ttadollim (Korean slang for bullying), ijime (Japanese word for bullying), kakaotalk bullying, kakaotalk ttadolim, Kakaostory bullying, kakaostory ttadollim, cas (slang for kakaostory) bullying, cas ttadollim (*Twitter: collection using a "catta (bulling on Twitter)"), group ttadollim, group harrassment, bullying, cyber bullying, ccyber bulling (ccyber indicates Korean consonant spelling for English word, cyber), peer violence, mass violence, peer maltreatment, group maltreatment, cigarette shuttle (Korean slang for bullying a victim for an errand of buying cigarette), bread shuttle (Korean slang for bullying a victim for an errand of buying bread), online ttadollim, cyber ttadollim', which were adopted from traditional bullying terms, and in addition 9 stopwords such as 'Top 3 of Jump (Japan's weekly comics magazine), Hoya wangtta (Koran slang for bullying), wangtta novel, wangtta comics, cafe wangtta, Hindi, jeontta novel' were used for collection purposes. In addition, cyber bullying topics were collected on an hourly basis regardless of weekdays, weekends and holidays. Analysis of this data targeted 103,212 cases of buzz wherein causes of cyber bullying were commented, out of a total of 435, 565 cases of cyber bullying buzz.

B. Ethical Considerations of Research

In due consideration of ethics, this research was begun after receiving approval (No.2014-1) from the Institutional Review Board (IRB) of the Korea Institute for Health and Social Affairs. The secondary material that the Korea Institute for Health and Social Affairs and SKT collected in May 2013 was used as research material and the collected social big data was unable to recognize personal information, ensuring anonymity and confidentiality of target persons in the research.

C. Research Tools

Collected buzz related to cyber bullying were coded to structured data through the process of text mining and opinion mining, and thereafter used.

1) Cyber Bullying Causes

Causes of cyber bullying are classified into 118: 'physical immaturity, disability, obesity, child obesity, undersized, appearance, face, autism, lack of sociality, lack of sociability, school maladjustment, insufficient number of friends, overprotection, princess syndrome, prince syndrome, conceit, talebearing, reclusive loner, personality, complex, wit, mischief, flattering,

fear of being around people, pretending to be good, otaku, superiority display, egoism, lacking tolerance, aggressiveness, assault, stress relief, wrath, self-esteem, will to dominate, disregard, impulse, impulsive, envy, jealousy, delinquency experience, running away from home, aberration, conflict, frustration, inferiority feeling, disobedience, influence, insufficient conversation with parent, parent violence, parental surveillance, fight between husband and wife, parental apathy, parent bonding, parental affection, parent discipline, parent divorce, family relationships, family dissolution, friend violence, friend bullying, friend harrassment, friend bonding, contact with a delinquent friend, friend, opposite sex friend, teacher violence, teacher scolding, teacher apathy, teacher bonding, school expansion, closed class organization, teacher's role weakening, nuclearization of family, violence-tolerance climate, collectivistic culture, factionalism, loss of human respect, value system reversal, social inequality, economic bankruptcy, mindset of valuing school background the most, excessive competition for entrance exams, wangtta (Korean slang for bullying) culture, military barracks culture, gangster culture, competitive education, income inequality, violent society, multicultural household, mass media violence, online game, smart phone, entertainer, movie, novel, entry into entertainment establishment, smoking, cigarette, drinking, alcohohol, iljin (Korean slang for bullies), harmful business establishment, drug, crime, yangachi (Korean slang for bullies or gang-

sters), Nallari (Korean slang for bullies or gangsters), hobby, taste, don (Korean word for money), geumjeon (another Korean word for money), thing, item, clothes, bag', and a value of '1' is coded if cause is available and '0' if not available.

2) Cyber Bullying Methods

Cyber bullying methods are divided into 91 methods, i.e., 'harmful act, harsh act, rape in confinement, isolation, disregard, torture, intimidation, attack, interrupting study, overkindness, harrassment, ill-treatment, extortion of money and valuables, avoiding, Ggolttong (Korean slang for knucklehead) treatment, slave, teasing, not playing with, disdain, extortion of money, robbing money video, talking behind one's back, bullying, beating, rumour, witch-hunt, rough words, reproof, contempt, insult, slander, group beating, ignoring, extortion of things, crime, accusation, mocking, ridicule, criticism, sarcasm, bread shuttle (Korean slang for bullying a victim into an errand, e.g., bread buying), picture-taking, homicide, murder, sexual insult, sexual molestation, sexual violence, sexual assault, sexual harrassment, shuttle, gossip, gossiping, alienation, quarrel, physical abuse, errand, send one on an errand, curse, malicious reply, jeering, being sidelined, extortion of clothes, undressing another person, robbing clothes, tearing clothes, wifi shuttle (Korean slang for bullying a victim into offering

wireless internet, e.g. hotspot for an offender), turn one's face away, swear word, menace, mischief, sneering, mob violence, look at another from the tail of eyes, corporal punishment, spitting, tiara game (Korean slang, referring to a game of acting a role of a bully victim), violence, violence punishment, violent act, disclosure, assault, assault and battery, persecution, school violence, abuse, trap, threat, blackmail, tyranny', and the value of '1' is coded if a method is available and '0' if not available.

3) Cyber Bullying Types

Cyber bullying types are classified through opinion mining according to the frequency of different buzz types, i.e., buzz in which an offender's psychological expression (e.g., 'deserve bullying', 'bullying is easy', 'bullying is just') is included, buzz in which a victim's psychological expression (e.g., 'being bullied, 'not playing with', 'scary bullying') is included, and buzz in which a bystander's psychological expression (e.g., 'experienced bullying', 'a bully victim needs help', 'have a bullying-victim friend) is included.

D. Material Analysis Methods

This research adopted the decision tree analysis method of data mining, not requiring special statistical hypothesis, in order to establish the most effective predictive model that can

explain the risk factors for each type of cyber bullying. The decision tree analysis of data mining can easily identify risk factors for each type of cyber bullying, which is associated with different causes, by generating automatically a predictive model that best explains dependent variables amid an enormous amount of material. Exhaustive CHAID was used as an analysis algorithm to develop the decision tree of this research primarily because Exhaustive CHAID has the highest prediction rate of model among CHAID, Exhaustive CHAID, CRT, and QUEST growing methods. Exhaustive CHAID explores all possible interaction effects and uses chi-square (χ^2-test) as separation standard of discrete-type dependent variables. As stopping rules, the minimum number of cases at the higher node (parent node) is set as 100 and that of lower node (child node) as 50, with the depth of the tree 3 levels. In addition, the rates of training data and test data are set as 70 and 30, respectively, for feasibility study according to data partitioning. SPSS 20.0 was adopted for technology analysis, multiple logistic regression analysis, and decision tree analysis.

2. Research Results

A. Descriptive Statistics of Major Factors

Out of a total of 435,565 buzz cases wherein cyber bullying topics were commented, the rate of buzz wherein the causes of cyber bullying were commented accounts for 23.7% (103,212

cases). With respect to the buzz wherein the causes of cyber bullying were commented, the types of cyber bullying are classified into ordinary people of 56% (57,817 cases) who did not express any emotion, victims at 32.3% (33,361 cases), offenders at 6.4% (6,587 cases), and bystanders at 5.3% (5,447 cases) in the order named. Impulse factor ranked first among cause factors of cyber bullying, followed by iljin (Korean salng for bullies) factor, appearance factor, culture factor, etc. Methods of cyber bullying include violence, shuttle, mocking, collective violence, etc. in the order named (Table 5).

〈Table 5〉 Descriptive Statistics

Cause factors of Cyber bullying				Cyber bullying methods	
Cause*	Buzz (%)	Cause†	Buzz (%)	Method‡	Buzz (%)
Impulse	294 (0.4)	Impulse	19,848 (29.9)	Not play	120 (0.2)
Obesity	279 (0.4)			Violence	18,600 (31.2)
Princess	160 (0.2)	Obesity	279 (0.4)	Insult	2,090 (3.5)
Stress	1,998 (2.9)			Shuttle	9,542 (16.0)
Drug	3,791 (5.4)	Stress	5,841 (8.8)	Mock	296 (0.5)
Deviation	967 (1.4)			Collective violence	5,510 (9.2)
Appearance	12,729 (18.3)	Appearance	12,729 (19.2)	Rear discourse	1,323 (2.2)
Money	8,413 (12.1)			Harassment	6,030 (10.1)
Envy	3,879 (5.6)	Lack of sociality	298 (0.4)	Harsh action	581 (1.0)
Lack of sociability	147 (0.2)			Booing	232 (0.4)
Assault	8,568 (12.3)	Culture	8,955 (13.5)	Blame	4,110 (6.9)
Smoking	3,246 (4.7)			Mischief	3,593 (6.0)
Culture	798 (1.1)	Multicultural	1,127 (1,7)	Malicious writing	2,042 (3.4)
Ruling	5,886 (8.5)			Intimidation	2,766 (4.6)
Multicultural	1,127 (1.6)	Iljin	17,327 (26.1)	Abuse	792 (3.3)
Iljin	17,327 (24.9)			Murder	1,992 (3.3)

* Cause analysis (unique value of 1 or greater) was conducted on 118 causes and cause factors were reduced to 16.
† 16 cause factors, scaled back as a result of the 1st cause analysis, were reduced to 8.
‡ Cause analysis (unique value of 1 or greater) was conducted on 91 methods and method factors were reduced to 16.

B. Factors Affecting the Types of Cyber Bullying

Impulse factor has an effect on a bystander and a victim. That is, impulse factor does not affect an offender, but affects a victim and a bystander. Obesity factor has a positive effect on a victim and a negative effect on an offender, consequently exerting a greater effect on a victim. It was found that the stress factor has a lesser effect on a victim, an offender and a bystander than an ordinary person, and thus cyber bullying can't be directly ascribed to stress factor. Appearance factor has the greatest effect on an offender, followed by a victim and a bystander in the order named. That is, appearance factor acts as a cause of all types of cyber bullying. Lack of sociability factor influences a bystander but not an offender and thus, lack of sociability becomes a cause for a cyber-bullying victim and bystander. It was found that cultural factor affects a victim and a bystander but not a victim and thus, cultural factor becomes a cause for a bully victim and a bystander. Multi-cultural factor has an effect solely on a victim and thus multi-cultural factor becomes a cause for a cyber-bullying victim. It was found that iljin factor has a negative effect on a victim and positive effect on an offender and a bystander, and thus, iljin factor has a bigger effect on an offender and a bystander and a lesser effect on a victim (Table 6).

〈Table 6〉 Multinomial Logistic Regression Analysis

Type* Causes	Victim			Bully			Outsider		
	B	p	Odd ratios	B	p	Odd ratios	B	p	Odd ratios
Intercept	-0.59	.001		-2.31	.001		-2.75	.001	
Impulse	0.17	.001	1.18	0.01	.725	1.01	0.99	.001	2.69
Obesity	0.59	.001	1.80	-0.63	.083	0.53	-0.20	.536	0.82
Stress	-0.43	.001	0.65	-0.53	.001	0.59	-0.19	.002	0.83
Appearance	0.19	.001	1.21	0.42	.001	1.52	0.11	.016	1.11
Lack of sociability	1.10	.001	3.01	0.35	.202	1.42	0.62	.010	1.86
Culture	0.25	.001	1.29	0.02	.615	1.02	0.12	.016	1.13
Multicultural	0.11	.096	1.11	-0.22	.111	0.80	-0.14	.316	0.87
Iljin	-0.13	.001	0.88	0.54	.001	1.71	0.56	.001	1.75

* base category: Public

C. Predictive Model of Risk Factor by Cyber Bullying Type of Korea

A model of the highest predictive capacity was selected by verifying the predictive rates (correct percentage) of the model according to the tree-structure classification model by using the rules of node separation. As a result of comparison and analysis of correct percentage that shows the accuracy rate of tree branching, QUEST algorithm earned the highest correct percentage of 73.8% in training data, but the corresponding number decreased to 72.8% in test data and therefore, CHAID algorithm was selected for the reason of a slight differential of accuracy rate between training data and test data while the correct percentage in training data is high (Table 7).

〈Table 7〉 Predictive Performance of Modeling Methods

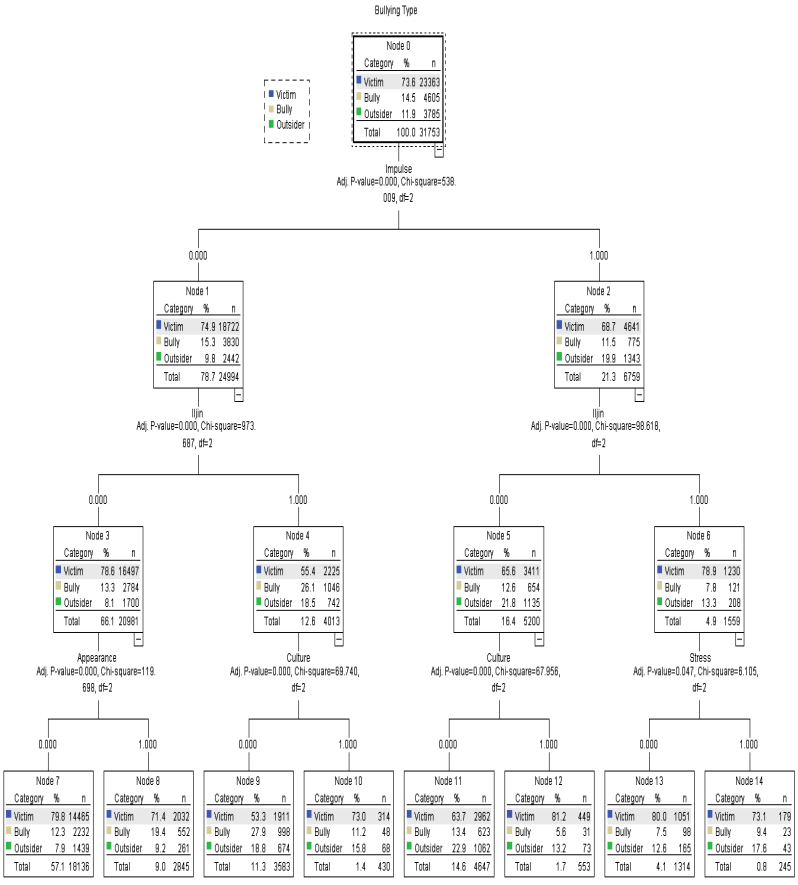
Modeling methods	Training data		Test data	
	Correct (%)	Wrong (%)	Correct (%)	Wrong (%)
CHAID*	73.5	26.5	73.6	26.4
Exhaustive CHAID	73.4	26.6	73.6	26.4
CRT†	73.1	26.9	74.5	25.5
QUEST‡	73.8	26.2	72.8	27.2

* Chi-squared Automatic Interaction Detection
† Classification and Regression Tree
‡ Quick, Unbiased, Efficient Statistical Tree

Result of decision tree analysis with respect to predictive model of risk factors for each type of cyber bullying in Korea is shown as in Figure 5. Knotroot at the highest level of tree structure means dependent variables (types of cyber bullying) in which predictive variables (independent variables) are not inputted. As for the percentage of cyber bullying types of knot-root, cyber bullying victim accounts for 73.6%, offender 14.5%, and bystander 11.9%. A factor standing at the highest-level of the bottom part of knotroot has the largest (deeply related) effect on dependent variables, and in this research, it was found that 'impulse factor' has the largest effect on risk prediction of cyber bullying types. That is, if 'impulse factor' is high, the risk of victim decreased from 73.6% to 68.7%, and that of offender from 14.5% to 11.5%, whereas that of bystander soared from 11.9% to 19.9%. In addition, in case of high 'iljin factor', even if 'impulse factor' is high, the risk of victim grew from 68.7% to 78.9%, whereas that of offender dropped sharply from 11.5% to 7.8% and that of by-stander from 19.9% to 13.3%. Moreover, in

case of high 'stress factor', even if 'iljin factor' is high, then the risk of victim lowered from 78.9% to 73.1%, whereas that of offender rose from 7.8% 9.4% and that of bystander from 13.3% to 17.6%. As compared therewith, in case of a group with low risk of 'impulse factor', the risk of victim increased from 73.6% to 74.9% and that of offender from 14.5% to 15.3%, whereas that of bystander decreased from 11.9% to 9.8%. In case of high 'iljin factor', even if 'impulse factor' was low, the risk of victim declined sharply from 74.9% to 55.4%, whereas that of offender soared from 15.3% to 26.1% and that of bystander from 9.8% to 18.5%. In case of high culture factor, even if 'iljin factor' was high, the risk of victim rose from 55.4% to 73.0%, whereas that of offender dropped from 26.1% to 11.2% and that of bystander from 18.5% to 15.8%.

[Figure 5] Decision tree of CHAID model



D. Predictive Mode-related Profit Chart of Risk Factor by Korean Cyber Bullying Type

In this research study, the highest risk of a cyber bullying victim was found in the combination of a high risk of 'impulse factor', a low risk of 'iljin' factor and a high risk of 'culture fac-

tor'. That is, the index of the no. 12 node is 110.4 and the group satisfying the conditions of the no. 12 node, compared with the knotroot, has 1.10 times greater risk of cyber bullying harm. The highest risk of a cyber bullying offender was found in the combination of a low risk of 'impulse factor', a high risk of 'iljin factor' and a low risk of 'culture factor. That is, the index of the no. 9 node is 192.1 and the group satisfying the conditions of no. 9, compared with the knotroot, has 1.92 times greater risk of cyber bullying perpetration. The highest risk of a cyber bullying bystander was found in the combination of a high risk of 'impulse factor', a low risk of 'iljin factor' and a low risk of 'culture factor'. That is, the index of the no. 11 node is 191.7 and the group satisfying the conditions of the no. 11 node, compared with the knotroot, has 1.92 times greater risk of being a bystander of cyber bullying. Training data and test data of this research were compared to perform feasibility evaluation according to data partitioning, and the result shows that the risk estimate of training data is 0.264 (standard error: 0.002) and that of test data is 0.267 (standard error: 0.004), indicating that the generalization of this predictive model of the cyber bullying types-based risk factor holds water (Table 8).

〈Table 8〉 Profit Chart of Predictive Models

Type	Node	Profit index				Cumulative index			
		node : n	node : %	gain (%)	index (%)	node : n	node : %	gain (%)	index (%)
Victim	12	553	1.7	1.9	110.4	553	1.7	1.9	110.4
	13	1314	4.1	4.5	108.7	1867	5.9	6.4	109.2
	7	18136	57.1	61.9	108.4	20003	63.0	68.3	108.5
	14	245	.8	.8	99.3	20248	63.8	69.1	108.4
	10	430	1.4	1.3	99.2	20678	65.1	70.4	108.2
	8	2845	9.0	8.7	97.1	23523	74.1	79.1	106.8
	11	4647	14.6	12.7	86.6	28170	88.7	91.8	103.5
	9	3583	11.3	8.2	72.5	31753	100.0	100.0	100.0
Bully	9	3583	11.3	21.7	192.1	3583	11.3	21.7	192.1
	8	2845	9.0	12.0	133.8	6428	20.2	33.7	166.3
	11	4647	14.6	13.5	92.4	11075	34.9	47.2	135.3
	7	18136	57.1	48.5	84.9	29211	92.0	95.7	104.0
	10	430	1.4	1.0	77.0	29641	93.3	96.7	103.6
	14	245	.8	.5	64.7	29886	94.1	97.2	103.3
	13	1314	4.1	2.1	51.4	31200	98.3	99.3	101.1
	12	553	1.7	.7	38.7	31753	100.0	100.0	100.0
Outsider	11	4647	14.6	28.1	191.7	4647	14.6	28.1	191.7
	9	3583	11.3	17.8	157.8	8230	25.9	45.9	177.0
	14	245	.8	1.1	147.2	8475	26.7	47.0	176.1
	10	430	1.4	1.8	132.7	8905	28.0	48.8	174.0
	12	553	1.7	1.9	110.7	9458	29.8	50.7	170.3
	13	1314	4.1	4.4	105.3	10772	33.9	55.1	162.4
	8	2845	9.0	6.9	77.0	13617	42.9	62.0	144.5
	7	18136	57.1	38.0	66.6	31753	100.0	100.0	100.0

In this research study, the decision tree analysis was performed by using social big data to verify the predictive model related to risk factors for each cyber bullying type. The decision tree analysis model, used in this research, unlike regression analysis or structural equation, performs analysis and prediction in a tree structure diagram according to decision rules without special statistical hypothesis, and as such is useful in determining patterns or relationships between factors that

have great effect on dependent variables from a number of independent variables relative to social big data. In particular, CHAID, an analysis algorithm adopted in this research, uses 'p-' value of χ^2- test volume relative to discrete target variable as separation standard, making prediction easy. In this research study, the types of cyber bullying are defined as three groups-- a group of victims, a group of offenders and a group of bystanders-- through opinion mining and text mining.

Main results of the analysis are summarized as follows. First, online communication related cyber bullying becomes very vigorous when a social issue related to cyber bullying comes up. Second, according to the result of multiple logistic regression analysis related to risk factors for each cyber bullying type, the impulse factor affects both the victim and bystander, while the obesity and appearance factors have a greater effect on the three groups mentioned than ordinary persons. Also, it was found that stress factor has less effect on those three groups than ordinary persons. Both the lack of sociability factor and culture factor have effects on victim and bystander. The multi-culture factor has an effect solely on the victim. Iljin factor has a greater effect on the offender and bystander than ordinary persons, but less effect on victim. Third, according to the decision tree analysis result related to the predictive model of cyber bullying type-based risk factors, the highest risk of a cyber bullying victim was found in the combination of a high risk

of 'impulse factor', a low risk of 'iljin factor', and a high risk of 'culture factor'. The highest risk of a cyber bullying offender was found in the combination of a low risk of 'impulse factor', a high risk of 'iljin factor' and a low risk of 'culture factor'. The highest risk of a cyber bullying bystander was found in the combination of a high risk of 'impulse factor', a low risk of 'iljin factor' and a low risk of 'culture factor'. The following discussions are provided revolving around the result of this research.

First, the types of cyber bulling in Korea are classified into three groups, such as victim, offender and bystander, and these 3 groups account for 44% of the whole group, specifically, 32.3% for victims, 6.4% for offenders, and 5.3% for bystanders. This rate is similar to 38.2% of the three groups (a group of those who inflicted the injuries of bullying, a group of those who sustained injuries, a group of those who inflicted and sustained injuries) in the result of the 'latent classification with respect to the changing patterns of experiences of inflicting and sustaining injuries of bullying' research [26], which used the existing research material (Korean Child/Youth Panel Study), thus confirming that the classification and prediction of cyber bullying types through text mining and opinion mining targeting social big data hold water to some extent. In other words, the existing classification method of traditional bullying was based on the standard deviation and median value from the

categorized responses material, or alternatively, compared and classified the number of latent class(es) with respect to the experiences of inflicting and sustaining the injuries. In the opinion mining part of this research study, if the expressions of being a victim are greater in frequency than those of being an offender or a bystander, the relevant person is classified into a group of victims. If the expressions of being an offender are greater in frequency than those of being a victim or a bystander, that person is categorized into a group of offenders. If the expressions of being a bystander are greater in frequency than those of being a victim or an offender, that person is classified into a group of bystanders. Therefore, the fact that there exist similarities between the prediction of cyber bullying types through opinion mining and that of traditional bullying types is evident that it is valid to apply opinion mining in classifying unstructured, social big data.

Second, the result of multiple logistic regression analysis shows that the risk of impulse factor has a much greater effect on the victim, supporting the existing research that traditional bully victims are highly impulsive [11]. The risk of obesity factor has a greater effect on the victim, corroborating the result of existing result [18] that there are cases of people victimized by traditional bullying for the reason of obesity. The stress factor has negative effects on all three groups, contrary to the results of studies [12, 13] claiming that traditional bullying has a

positive correlation to the stress a person perceives. It is attributable to the fact that the stress factor on cyber space is defined differently from that in the relevant research material of traditional bullying and presumably, verification will be required in subsequent studies. The risk of appearance factor has effects on all three groups, supporting the result of existing research [19] saying that traditional bullying is attributed to the appearance of a victim. The lack of sociability factor is greatest for a victim, affirming the existing research [20] saying that traditional bullying is imputable to interpersonal relationships and the social skills of a victim. The risk of culture factor has an effect on a victim, supporting the existing research claims that traditional bullying is imputed to the cultural difference between an individual and a group [14] and that traditional bullying may occur through the imitation of a group culture [15]. The multi-culture factor has an effect solely on the victim, showing that multicultural families of Korea are seriously victimized by cyber bullying, as already reported in the news media [17]. The risk of iljin factor is negatively related with a victim but positively related with an offender, supporting the existing research studies [5, 16] saying that traditional bullying has positive relationships with the dominance disposition of an offender. That is, it was found that factors having the greatest effects on the types of cyber bullying are the lack of sociability factor, appearance factor and impulse factor in the cases of

victim, offender and bystander, respectively. Accordingly, since a victim tends to show excessive anxiety and responds in a highly sensitive manner in interpersonal relationships, various communication channels should be prepared at the school level to improve sociability and at the same time, social skill training, I.e. how to express oneself proactively and how to make an argument, should be offered. In addition, the types of bullying are not fixed to an offender, victim and bystander but are fluid, and the relationships among victim, offender and bystander may change depending on the surrounding environment. For this reason, research will be needed to determine how the environment and the changes of surroundings to which a person is exposed are related to the types of bullying [15].

Third, one of the most critical factors that can predict risk factors for each type of cyber bullying in Korea is the impulse factor. If the impulse factor is high, the risk of victim and offender decreases, whereas that of bystander increases. This finding supports the existing research [22, 23] that says that a majority of bystanders prolongs or worsens the behavior of peer bullying. With respect to the above result, it was found in this research study that the risk of a bystander's impulse factor is the highest among the three groups, presumably, making it necessary to provide a program tailored to bystanders to lower their impulse, in addition to the programs for victim and offender. In the research outcome saying that impulse factor is

the most critical factor in the prediction of risk factors for each cyber bullying type, the iljin factor was found to affect all 3 groups. That is, if impulse factor is low, the risk of offender rises slightly, but in contrast, if the iljin factor with dominance tendency is high, the risk of offenders soars to 1.71 times (15.3%→26.1%). Therefore, a group with high propensity to offend has individuals that tend to think oneself as superior and a special being, dominating others offensively. In response thereto, consultation programs and parent training programs presumably are urgently needed to deal with such groups properly. Following after impulse factor and iljin factor, culture factor and stress factor affect the types of cyber bullying. In order to prevent cyber bullying imputed stress and culture factors, various psycho-education programs should be developed, such as a program that can improve empathy through discussion for a better understanding of cyber bullying and to promote intimacy, as well as a stress handling program and others. Fourth, this research study utilized social big data in the analysis of cyber bullying, and if social big data dispersed and existing in varied forms across various sectors of society is effectively utilized, it will eventually contribute to the development of preventive measures against risk factors at the country level. In order to predict real-time complex social issues like cyber bullying and respond thereto, it will be imperative to attain technology development and standardization for analysis of big

data and to train talents like data scientists capable of detecting information hidden in the vast volume of data. In this research study, the analysis was conducted not by using characteristics of each and every individual but rather the material of the entire group to which members belong, and thus an ecological fallacy may arise if the outcome of analysis is applied to an individual. Also, the cyber bullying-relevant factors (terms) defined in this research refer to the frequency of opinions or words, which were discovered in the buzz, and for this reason, the factors of this study may differ in concept from the corresponding factors of cyber bullying in the existing models, so further verification will be necessary in subsequent studies. Despite these limitations, this study is significant from the analysis methodology perspective in that it intended to verify the validity of the classification and prediction of cyber bullying types in Korea by performing text mining and opinion mining of buzz collected from social big data; and it proposed a predictive model of risk factors for each cyber bullying type in Korea through data mining analysis. In addition, this research study holds significance from the analysis methodology perspective in that it confirmed the value of social big data as a new study method that can supplement the limitations of social statistics by determining fast and effectively the actual details.

3. Conclusion

This research study was performed to classify cyber bullying types through a data mining technique based on cyber bullying related buzz and collected social big data in Korea, as well as determine causes affecting each type and develop a decision tree capable of predicting risk factors. The result of the decision tree analysis shows that impulse factor is the greatest predictor of risk in cyber bullying types, followed by iljin factor. In particular, impulse factor has the greatest effect on a bystander, and iljin factor affects an offender the most. According to this research study, since a majority of bystanders can impulsively aggravate an act of cyber bullying, it will be necessary to develop programs for bystanders, in addition to those for victims and offenders, to lower impulse. Also, consultation programs or parent education programs will be required to bring about changes to the understanding of the dominant disposition or superior feeling of an offender.

Lastly, factors related to cyber bullying, proposed in this research study, may differ from the concept of factors in the existing theoretical model, and thus further verification will be necessary.

References

[1] Korea Communications Commission·Korea Internet & Security Agency. 2013 Survey on Cyber Bullying. Seoul, Korea: Author; 2013.

[2] Şahin M, Aydin B, Sari SV. Cyber Bullying, Cyber Victimization and Psychological Symptoms: A Study in Adolescents. Cukurova University Faculty of Education Journal. 2012;41(1):53-59.

[3] Butler D, Kift S, Campbell M. Cyber bullying in schools and the law: Is there an effective means of addressing the power imbalance. eLaw Journal: Murdoch University Electronic Journal of Law. 2009;16:84-114.

[4] Moon B, Hwang H, McCluskey JD. Causes of school bullying: Empirical test of a general theory of crime, differential association theory, and general strain theory. Crime & Delinquency. 2011;57:849-877.

[5] Olweus D. Bullying at school: Long-term outcomes for the victims and an effective school-based intervention program. In L. R. Huesmann (Ed.), Aggressive behavior: Current perspectives. New York: Plenum;1994.

[6] Song J. Examining bullying among Korean Youth: An empirical test of GST and MST. Hanyang Law Review. 2013;24(2):221-246.

[7] Slonje R, Smith PK, Frisén A. The nature of cyberbullying and strategies for prevention. Computers in Human Behavior. 2013;29:26-32.

[8] Willard N. Educator's Guide to Cyberbullying and Cyberthreats; 2007.p.1-16.

[9] Smith PK, Mahdavi J, Carvalho M, Fisher S, Russell S, Tippett N.

Cyberbullying: Its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry, 2008;49:376-385.

[10] Katzer C, Fetchenhauer D, Belschank F. Cyberbullying: Who are the victims? A comparison of victimization in internet chatrooms and victimization in school. Journal of Media Psychology. 2009;21(1):25-36.

[11] Kim YS, Koh YJ, Noh JS, Park MS, Sohn SH, Suh DH, et al. School Bullying and Related Psychopathology in Elementary School Students. J Korean Neuropsychiatr Assoc. 2001;40(5):876-884.

[12] Coie JD. "Toward a theory of peer rejection." In S. R. Asher & J. D. Coie (Eds). Peer rejection in childhood. New York: Cambridge University Press;1990.p.365-401.

[13] Chon JY, Lee EK, Yoo NH, Lee KH. A Study on the Relation between Conformity in Group Bullying and Psychological Characteristics. Korean Journal of Psychology: School. 2004;1(1):23-35.

[14] Park JK. A Socio-Cultural Study on Peer Rejection Phenomenon of Adolescent Group. Korean Journal of Youth Studies. 2000;7(2):39-71.

[15] Lee JG. Bullying and Alternative of Criminal Policy. Victimology. 2007;15(2):285-309.

[16] Pellegrini AD, Bartini M, Brooks F. School bullies, victims and aggressive victims; factors relating to group affiliation and victimization in early adolescence. Journal of Educational Psychology. 1999;91(2):216-224.

[17] Kim Y-j. 37% of children of multicultural families bullying. The Chosun Daily. 2012 January 12; http://news.chosun.com/site/data/html_dir/2012/01/10/201201

1000131.html

[18] Noh SH. Some Aspects of Bullying in School. Victimology. 2011;9(2):5-29.

[19] Roland E. School Influences on Bullying. Durharn: University of Durham;1988.

[20] Randoll P. Adult bullying: perpetrators and victims. London: Routledge;1997.

[21] Gini G, Albiero P, Benelli B, Altoè G. Determinants of adolescents' active defending and passive bystanding behavior in bullying. Journal of Adolescence. 2008;31: 91-105.

[22] Jang SJ, Choi YK. Development and validation of choldren's reactions scale to peer bullying. The korean Journal of School Psychology. 2010;7(2):251-267.

[23] Hawkins DL, Pepler DJ, Craig WM. Naturalistic observation of peer interventions in bullying. Social Development. 2001;10:512-527.

[24] Choi YJ, Jhin HK, Kim JW. A Study on the Personality Trait of Bullying & Victimized School Childrens. Journal of Child & Adolescent Psychiatry. 2001;12(1):94-102.

[25] Noh KA, Baik JS. A Discriminant Analysis on the Determinants of Adolescent Group Bullying Type. Youth Facility & Environment. 2013;11(3):113-124.

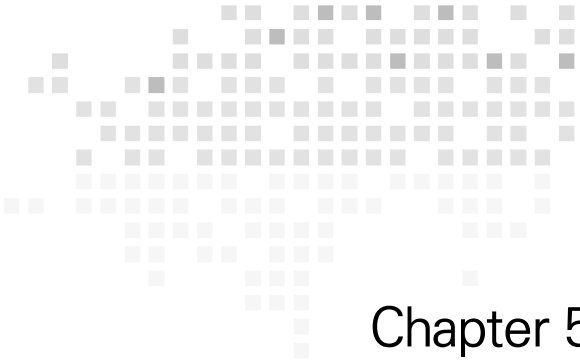
[26] No U, Hong S. Classification and Prediction of early Adolescents' Bullying and Victimized Experiences Using Dual Trajectory Modeling Approach. Survey Research. 2013;14(2):49-76.

[27] Bond L, Carlin JB, Thomas L, Rubin K, Patton G. Does bullying cause emotional problems? A prospective study of young teenagers. British Medical Journal. 2001;323:480-484.

[28] Kim WJ. Wang-Ta: A review on its significance, realities, and cause. Korea Journal of Counseling. 2004;5(2)451-472.

[29] Perry, DG, Kusel SJ, Perry LC. Victims of peer aggression. Developmental Psychology. 1998;24(6):807-814.

[30] Smith PK. Cyberbullying and cyber aggression. In S. R. Jimerson, A. B. Nickerson, M. J. Mayer, & M. J. Furlong (Eds.), Handbook of school violence and school safety: International research and practice. New York, NY: Routledge;2012.p.93-103.



Chapter 5

Methods of Social Risk Factors Prediction Utilizing Social Big Data

5 Methods of Social Risk Factors Prediction Utilizing Social Big Data

In the social big data sector, social analytics quickly analyzes unstructured data collected from Facebook, Twitter, SNS, etc. How to extract information from social media and perform information analysis and methods are illustrated in Figure 6. First, target social big data is collected. Collection target (unstructured big data on search portal or SNS) and scope are defined, and collection is performed through collection engines such as a crawler (robot). Second, the collected unstructured data is analyzed. Analysis of unstructured data is performed in the order of buzz analysis, keyword analysis, opinion analysis, and account analysis. The collected unstructured data is then applied with text mining and opinion mining. Third, the collected unstructured data is classified with network analysis. The unstructured big data is then converted to structured data. The conversion from unstructured to structured data means to codify each document of suicide buzz to ID and codify keywords and method within buzz. Fourth, the structured data is connected to offline statistics (research) of the government and public organizations. To perform analysis related to social phenomena, the structured big data is connected to the structured big data of public organizations. ID (by date/month/year/re-

gion) that can be connected is checked and then connected to big data (offline statistics) of public organizations. Lastly, the analysis of the structured big data connected to the offline statistics (researches) can be performed through the structural equation model, which enables the cause-effect analysis between factors or the tracking of time-based trajectory changes; the multi-level model, which enables the analysis between factors related to social phenomena by date/month/year and by region; and data mining analysis, which allows the finding of a new phenomenon through classification of the collected keywords.

[Figure 6] Social Big Data Analysis Process and Method (Suicide Buzz Analysis (Sample))