보건·복지 Issue & Focus

Korea Institute for Health and Social Affairs

ISSN 2092-7117
제238호(2014-17) 발행일 : 2014. 05. 02

KIHASA 한국보건사회연구원 Korea Institute for Health and Social Affairs

소셜 빅데이터를 활용한 사회위험 요인 예측: 청소년 자살과 사이버따돌림을 중심으로

SNS를 통하여 전송되는 데이터양이 기하급수적으로 증가하면서 많은 국가와 기업에서 새로운 경제적 효과와 일자리 창출, 그리고 사회적 문제의 해결을 위해 빅데이터의 활용과 분석을 적극적으로 시도함

SNS 상에서 나타나는 자살 등 사회위험 요인에 대한 감정표현이나 심리적 위기 행태들을 분석하게 되면 위험징후와 유의미한 패턴을 감지하여 사회위험 요인을 예측 할 수 있음

민간기관의 검색포털이나 SNS의 비정형 빅데이터의 수집·분류와 함께 정부나 공공기관의 정형 빅데이터와 연계한 후, 다변량 분석을 실시하여 사회위험 요인을 예측하고 대책을 수립할수 있음

보건복지 빅데이터의 부가가치를 높이고 사회위험과 불확실성에 효과적으로 대응하기 위해서는 국가차원의 사회위험관리 빅데이터 분석 센터의 설립이 필요함

송태민
사회정신건강연구센터장

1. 보건복지분야 빅데이터 추진방안

■ 정부 3.0의 효과적인 추진과 생애주기별 맞춤형 보건복지 및 국민 행복 실현을 위한 보건복지분야 빅데이터의 효율적 활용 방안 모색

○ 정부 3.0은 공공정보를 적극 개방·공유하고, 부처 간 칸막이를 없애고 소통·협력함으로써 국정과제에 대한 추진동력을 확보하고 국민 맞춤형 서비스를 제공함과 동시에 일자리 창출과 창조경제를 지원하는 새로운 정부운영 패러다임을 의미함

○ 빅데이터는 방대한 규모(Volume), 빠른 생성주기(Velocity), 다양하고(Variety), 복잡한(Complexity) 형태의 데이터를 뜻하며, 대용량의 데이터를 활용·분석하여 신뢰성있고(Veracity) 가치있는(Value) 정보를 추출하고, 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 기술을 의미함

■ 빅데이터의 특성(5V, 1C)와 보건복지부 3.0의 추진 전략은 유기적인 연관성이 있음[1)]

1) '오미애(2014). 정부 3.0과 빅데이터: 보건복지 분야 사례를 중심으로. 보건·복지 Issue & Focus, 제230호.'의 내용을 보완함.

○ 보건복지부 3.0의 '소통하는 투명한 보건복지'는 빅데이터의 이용 활성화를 위해 공공 데이터를 적극 개방함으로써 활용 가능한 자료가 복잡하고(Complexity), 양이 매우 방대해짐(Volume)

○ 보건복지부 3.0의 '일 잘하는 유능한 보건복지'는 빅데이터를 활용한 과학적 행정 구현으로 다양한(Variety) 정보의 결합이 가능하고, 정부운영시스템 개선으로 인한 자료의 축적 속도(Velocity)가 빠름

○ 보건복지부 3.0의 '국민중심 보건복지 서비스'는 빅데이터 분석결과를 기초로 수요자 맞춤형 서비스 통합을 제공함으로써 신뢰성있는(Veracity) 새로운 가치(Value)를 창출함

[그림 1] 빅데이터의 특성과 보건복지부 3.0 추진 전략

2. 소셜 빅데이터를 활용한 청소년 자살 위험예측

■ 우리나라는 최근 스마트폰 보급의 확산에 따라 모바일 인터넷과 SNS 이용이 급속히 증가함

○ 2013년 7월 현재 우리나라 만 3세 이상 인구의 인터넷 이용률은 82.1%이며 이중 만 6세 이상 인터넷 이용자의 55.1%가 1년 이내 SNS을 이용하고 있음[2)]

■ SNS를 통하여 전송되는 데이터양이 기하급수적으로 증가하면서 많은 국가와 기업에서 새로운 경제적 효과와 일자리 창출, 그리고 사회적 문제의 해결을 위해 빅데이터의 활용과 분석을 적극적으로 시도함

○ 공공부분에서 유전자와 생명연구자원 공유를 통한 질병 예방 및 예측, 치료, 그리고 환자 관리 등에 활용하고 있으며, 다국적 IT(Information Technology)기업들과 웹(web) 검색 포털(portal) 사이트들은 서버에 저장된 빅데이터를 분석함으로써 다양한 가치 정보를 생산함[3)]

○ SNS는 청소년들이 일상생활 속에서 갖는 우울한 감정이나 스트레스, 고민을 들을 수 있고 행태를 이해할 수 있는 장소로 SNS 상에서 나타나는 자살에 대한 감정표현이나 심리적 위기 행태들을 분석하게 되면 위험 징후와 유의미한 패턴을 감지하여 자살을 예방하는데 긍정적 효과가 발휘됨[4)]

■ 우리나라는 급격한 사회 · 경제적 변화속에 자살률이 2004년부터 OECD 국가중 최고의 수준이며, 특히 청소년계층의 자살 문제가 사회적 이슈로 대두되면서 정부차원의 적극적인 대책이 시급한 실정임

2) 미래창조과학부 · 한국인터넷진흥원(2013). 2013 인터넷 이용자 실태조사.
3) Policy Exchange (2012). The Big Data Opportunity: Making government faster, smarter and more personal.
4) 한국정보화진흥원(2012). 소셜 분석으로 살펴본 청소년 자살예방정책의 시사점. 7면.

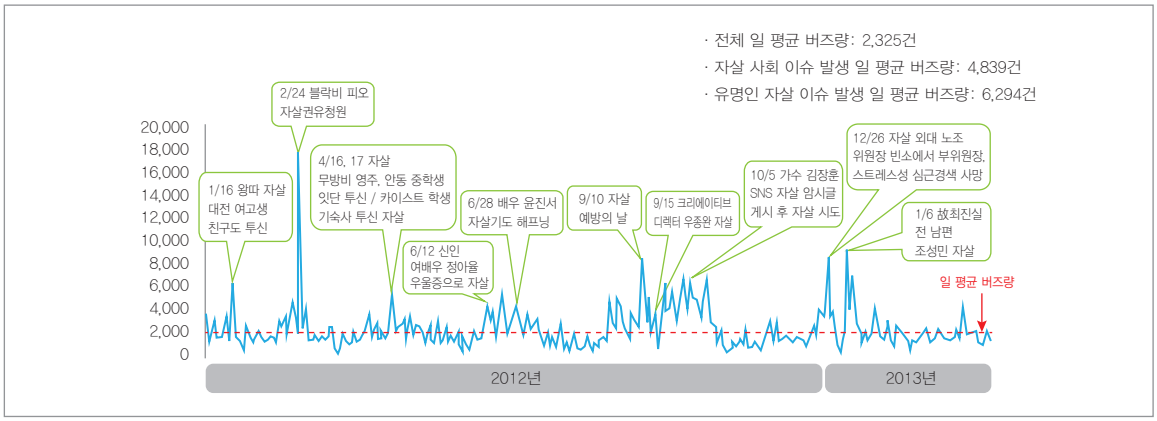
○ 청소년 자살의 원인과 관련 요인을 규명하기 위하여 기존에 실시하던 횡단적 조사나 종단적 조사 등을 대상으로 한 연구는 정해진 변인들에 대한 개인과 집단의 관계를 보는 데에는 유용하나 사이버상에서 언급된 개인별 버즈(buzz: 입소문)가 사회적 현상들과 어떻게 얼마나 연관되어 있는지 밝히는데는 한계가 있음

○ 본 연구는 2011. 1. 1 ~ 2013. 3. 31(821일) 동안 수집[5]된 자살관련 소셜 빅데이터를 활용하여 SNS상의 청소년 자살의 원인을 살펴보고 데이터마이닝 분석을 통해 한국의 청소년 자살 위험 예측모형을 제시함

■ '자살' 관련 버즈 일별 추이

○ 청소년 자살, 유명인 자살 등 자살과 관련된 사회적 이슈 발생 시에 자살과 관련한 커뮤니케이션이 급증하는 양상을 보이고 있으며 특히 연예인 관련 자살 이슈 발생 시 버즈량이 급증함

'자살' 관련 버즈 일별 추이

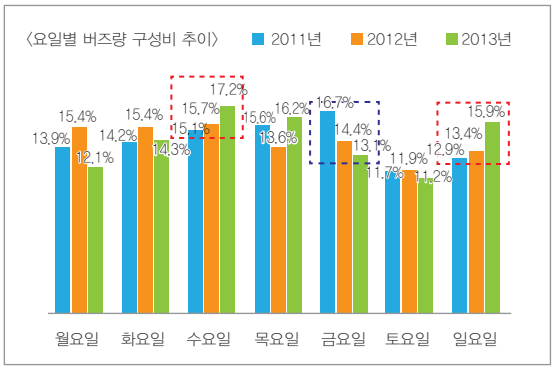


■ '자살' 관련 버즈 요일별 추이

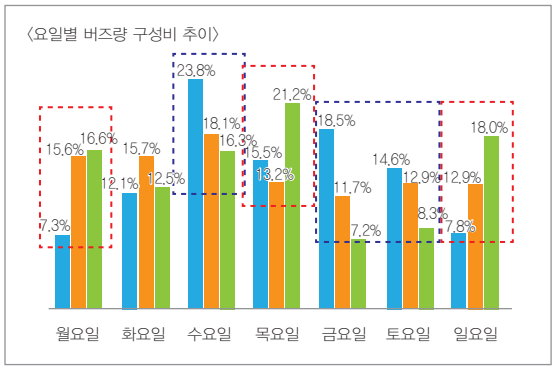
○ 최근 3년 간 전체 '자살' 관련 버즈량은 수요일과 일요일에 지속적으로 증가한 반면, 금요일에는 감소 추이를 보임

○ '청소년 자살' 관련 버즈량은 월, 목, 일요일에 증가 추이를 보인 반면에 수, 금, 토요일에는 감소 추이를 보임

요일별 버즈량 – 전체



요일별 버즈량 – 청소년



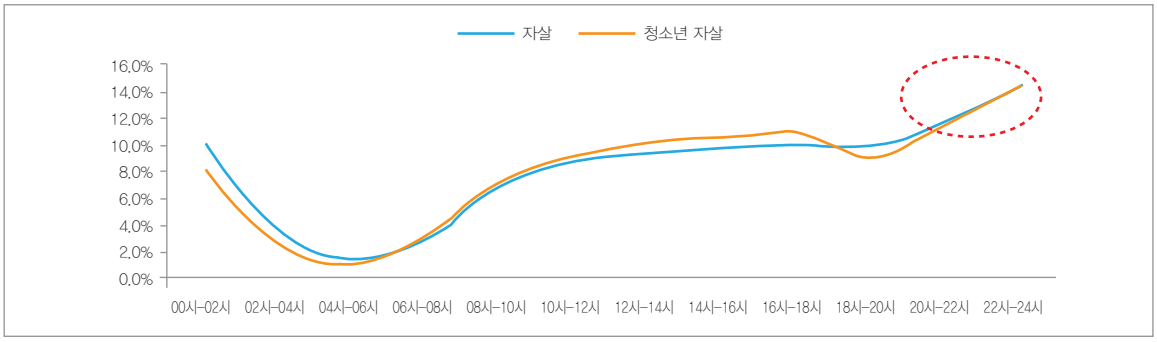
5) 본 연구를 위한 소셜 빅데이터의 수집 및 토픽 분류는 '(주) SK텔레콤 스마트 인사이트'에서 수행함

■ '자살' 관련 버즈 시간대별 추이

○ '자살'과 '청소년 자살' 관련 버즈 모두 20시부터 24시 사이에 버즈량이 많음. 특히 22시부터 24시에 집중적 발생

○ '자살'과 '청소년 자살'의 시간대별 버즈량 추이는 유사한 패턴으로 나타남

시간대별 버즈량



■ '청소년 자살' 버즈 원인

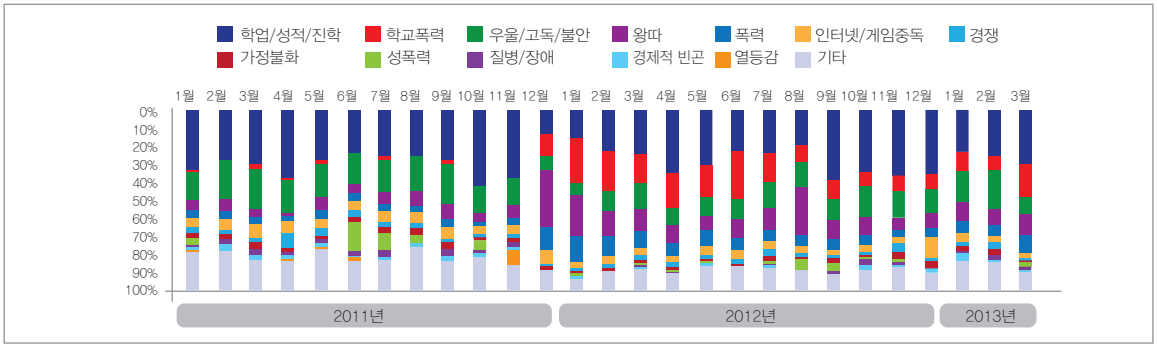
○ 거의 모든 기간에서 '학업/성적/진학'이 청소년 자살 버즈 원인 1위로 나타남

– 2012년 통계청 사회조사에서 13~19세 청소년은 '학교성적/진학문제'가 39.2%로 자살충동 이유 1위로 나타남

○ 2011년 12월 이후 '학교폭력'과 '왕따'가 주요 청소년 자살 버즈 원인으로 지속 등장

○ '우울/고독/불안'은 청소년 자살에서 지속적으로 주요 자살 원인으로 나타나고 있음

주요 청소년 자살 원인 월별 추이

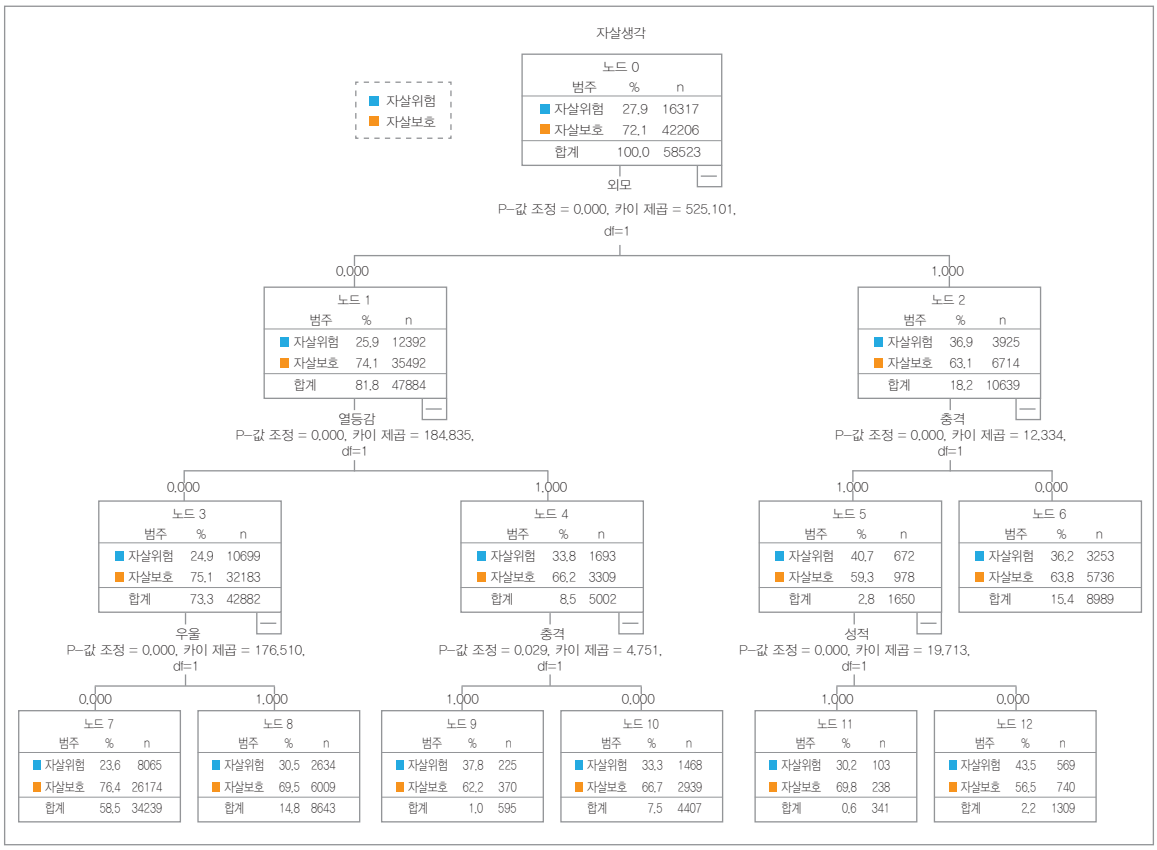


■ '청소년 자살' 위험 예측

○ 청소년자살 위험 예측에 가장 영향력이 높은 요인은 '외모요인'으로 '외모요인'의 위험이 높은 경우 청소년 자살 위험은 이전의 27.9%에서 36.9%로 증가하고, '외모요인'이 높고 '충격요인'이 높으면 청소년 자살 위험이 이전의 36.7%에서 40.7%로 증가함

○ '외모요인'의 위험이 낮더라도 '열등감요인'의 위험이 높으면 청소년 자살위험은 이전의 25.9%에서 33.8%로 증가하였으며, '열등감요인'이 높고, '충격요인'의 위험이 높으면 청소년 자살위험은 이전의 33.8%에서 37.8%로 증가함

[그림 2] 청소년 자살 위험 예측모형



3. 소셜 빅데이터를 활용한 사이버따돌림[6] 위험예측

■ 사이버따돌림에 노출된 청소년들이 자살을 선택하거나 폭력의 가해자가 됨에 따라 심각한 사회문제로 떠오르고 있음

○ 우리나라는 2013년 11월 현재 청소년의 29.2%, 일반인의 14.4%가 타인에게 사이버 따돌림을 가한 경험이 있으며, 청소년의 30.3%, 일반인의 30.0%가 사이버 따돌림의 피해를 경험한 것으로 나타남[7]

○ 사이버따돌림은 '개인 혹은 집단이 자기 자신을 스스로 방어하기 힘든 피해자를 대상으로 반복적으로 전자기기를 통해 이루어지는 공격적 행동 혹은 행위'[8]로 우울증, 자해, 자살과 같은 심각한 심리적 상해를 가져올 수 있음[9]

6) 본 연구의 사이버따돌림은 '사이버언어폭력, 사이버명예훼손, 사이버스토킹, 사이버성폭력, 신상정보유출, 사이버왕따'를 포괄하는 사이버폭력의 의미로 사용함

7) 방송통신위원회 · 한국인터넷진흥원(2013). 2013년 사이버폭력 실태조사.

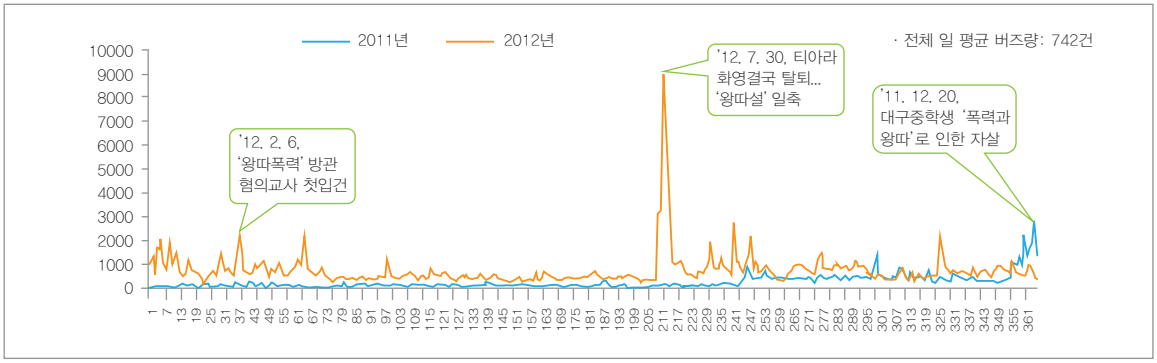
8) Slonje, R., Smith, P. K. and Frisén, A. (2013). The nature of cyberbullying and strategies for prevention. Computers in Human Behavior, 29(1), pp.26~32.

9) Erentaitė, R., Bergman, L. and Zukauskienė, R. (2012). Cross-contextual stability of bullying victimization: a person-oriented analysis of cyber and traditional bullying experiences among adolescents. Scandinavian Journal of Psychology, 53(2), pp.181~190.

■ '사이버따돌림' 관련 버즈 일별 추이

○ 사이버따돌림과 관련한 온라인 커뮤니케이션은 일 평균 742건이 발생하였으며, 2012년 7~8월에 유명 걸그룹의 왕따설이 사회적 이슈가 되면서 SNS상에서 이에 대한 커뮤니케이션이 매우 활발했음

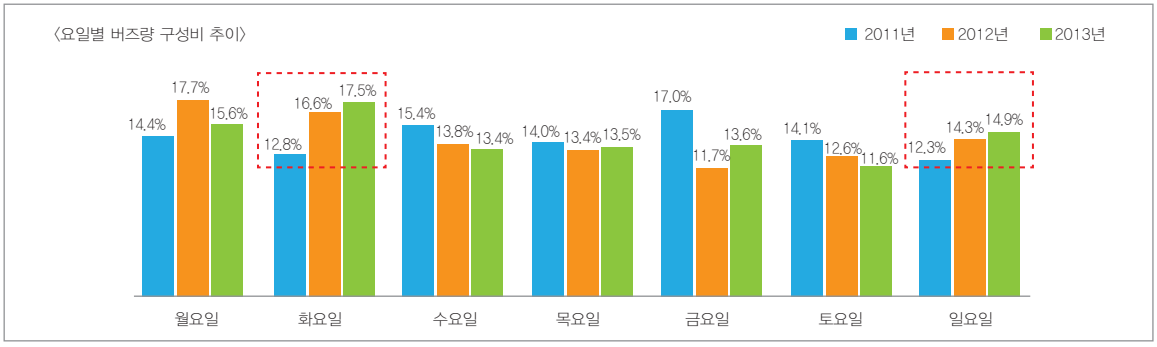
'사이버따돌림' 관련 버즈량 일별 추이



■ '사이버따돌림' 관련 버즈량 요일별 추이

○ 최근 3년 간 전체 '사이버따돌림' 관련 버즈량은 화요일과 일요일에 지속적으로 증가한 반면, 금요일과 토요일에는 감소 추이를 보임

요일별 버즈량

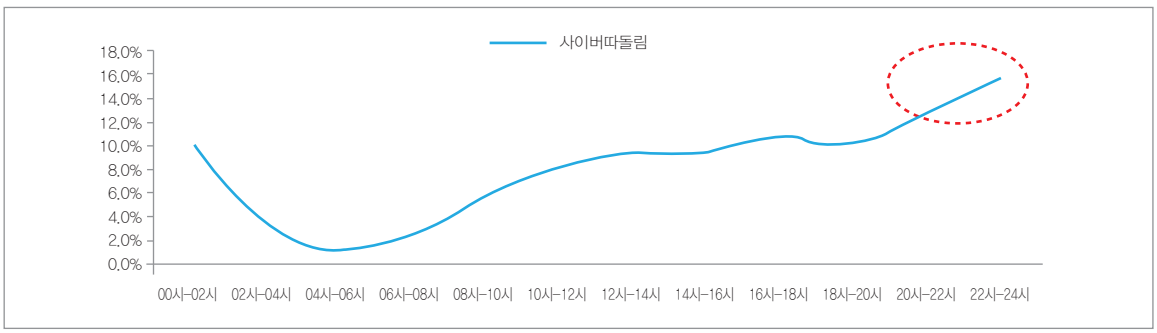


■ '사이버따돌림' 관련 버즈 시간대별 추이

○ '사이버따돌림' 관련 버즈량은 20시부터 24시 사이에 주로 발생함. 특히 22시부터 24시에 집중적 발생

○ '사이버따돌림' 관련 시간대별 버즈량 추이는 '자살'과 유사한 패턴을 보임

시간대별 버즈량



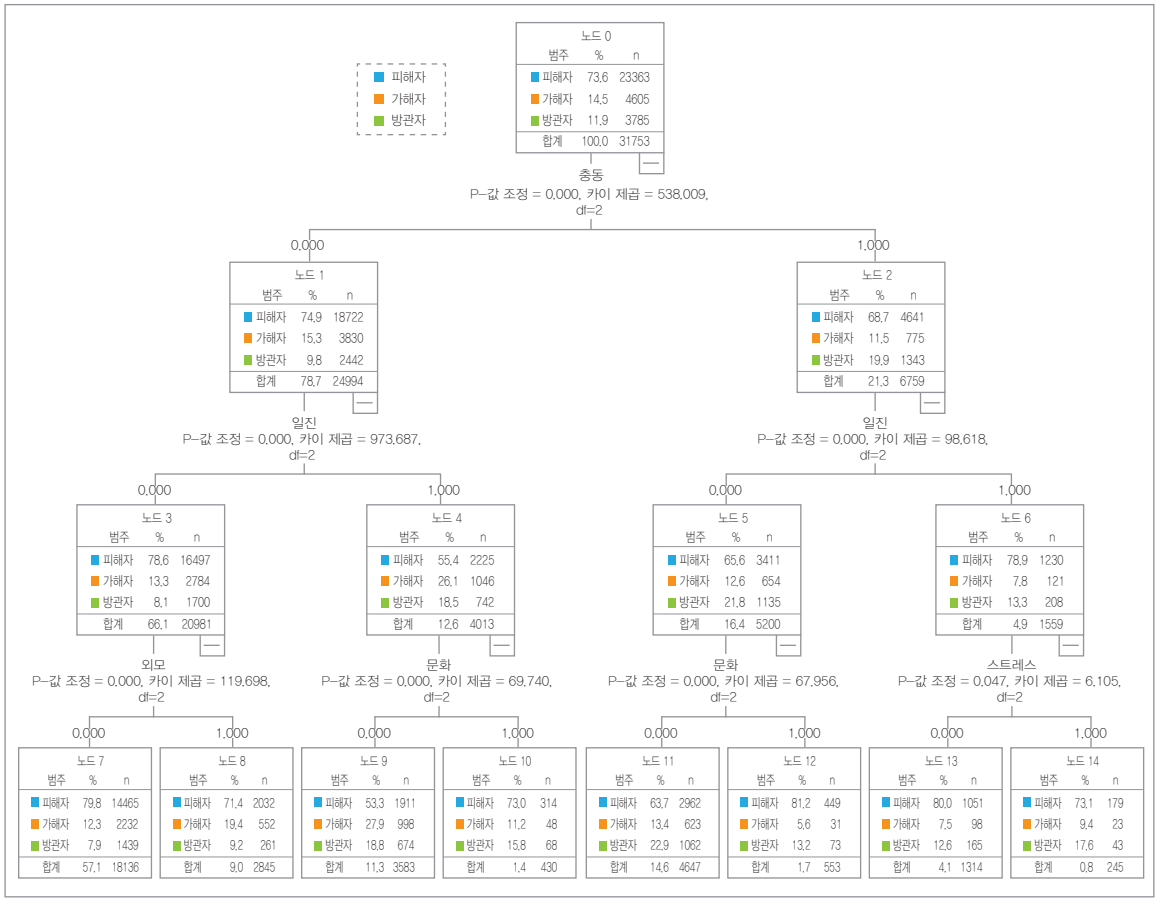
■ ‘사이버따돌림’ 위험 예측

○사이버따돌림 위험 예측에 가장 영향력이 높은 요인은 ‘충동요인’으로, ‘충동요인’의 위험이 높은 경우 피해자의 위험이 이전의 73.6%에서 68.7%, 가해자의 위험이 이전의 14.5%에서 11.5%로 감소한 반면, 방관자의 위험은 이전의 11.9%에서 19.9%로 크게 증가함

○‘충동요인’이 높더라도 ‘일진(지배욕)요인’이 높으면 피해자의 위험은 이전의 68.7%에서 78.9%로 증가 한 반면, 가해자의 위험은 이전의 11.5%에서 7.8%, 방관자의 위험은 이전의 19.9%에서 13.3%로 크게 감소함

○‘일진요인’이 높더라도 ‘스트레스요인’이 높으면 피해자의 위험은 이전의 78.9%에서 73.1%로 감소한 반면, 가해자의 위험은 7.8%에서 9.4%로 증가하였고, 방관자의 위험도 13.3%에서 17.6%로 증가함

[그림 3] 사이버따돌림 위험 예측모형



4. 사회위험 요인 예측을 위한 빅데이터 분석방안

■ 대상 소셜 빅데이터 수집

○해당 버즈분석 모델링을 통해 수집대상(검색포털이나 SNS의 비정형 빅데이터)과 수집범위를 설정한 후, 대상채널(뉴스, 블로그, 카페, 게시판, SNS 등)에서 크롤러 등 수집엔진(로봇)을 이용하여 수집

■ 수집한 비정형 빅데이터의 분석

○비정형 빅데이터 분석은 버즈분석, 키워드분석, 감성분석, 계정분석 등으로 진행

○수집한 비정형 데이터를 텍스트마이닝(text mining), 오피니언마이닝(opinion mining), 네트워크 분석(network analysis)을 통하여 분석

■ 정형 빅데이터 변환

○비정형 빅데이터를 정형 빅데이터로 변환 즉, 자살버즈 각각의 문서는 ID로 코드화하여야 하고, 버즈내 키워드나 방법 등도 모두 코드화함

■ 정형 빅데이터와 정부나 공공기관의 오프라인 통계(조사) 자료 연계

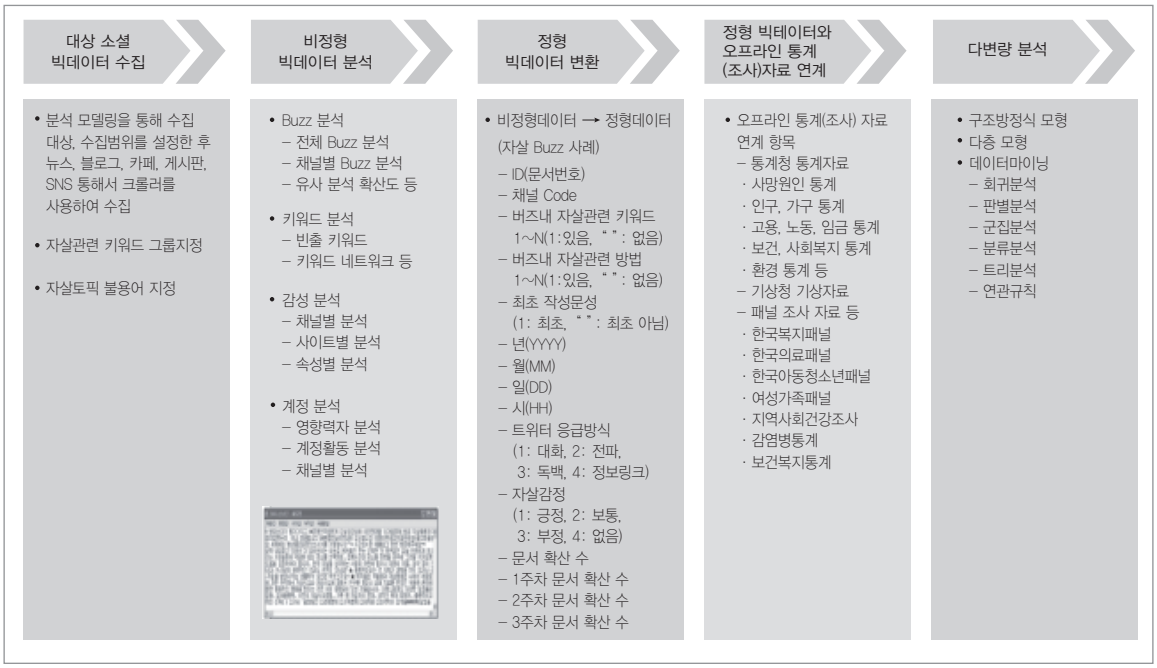
○사회현상과 연계해 분석하기 위하여 정형화된 빅데이터를 공공기관의 정형 빅데이터와 연계함

○연계 가능한 ID(일별 · 월별 · 연별 · 지역별)를 확인한 후, 공공기관의 빅데이터(오프라인 통계)와 연계함

■ 다변량 분석

○오프라인 통계(조사) 자료와 연계된 정형화된 빅데이터의 분석은 요인 간의 인과관계나 시간별 변화 궤적을 분석할 수 있는 구조방정식모형이나 일별(월별 · 연별), 지역별 사회현상과 관련된 요인과의 관계를 분석할 수 있는 다층모형, 그리고 수집된 키워드의 분류과정을 통해 새로운 현상을 발견할 수 있는 데이터마이닝 분석을 실시할 수 있음

[그림 4] 소셜 빅데이터 분석 절차 및 방법(자살버즈 분석 사례)[10)]



10) 송태민 · 송주영(2013). 빅데이터 분석방법론. 한나래아카데미.

집필자 | 송태민 (사회정신건강연구센터장) 문의 | 02-380-8201

발행인 | 최병호 **발행처** | 한국보건사회연구원
(122-705)서울특별시 은평구 진흥로 235 | TEL 02)380-8000 | FAX 02)352-9129 | http://www.kihasa.re.kr
한국보건사회연구원 홈페이지의 발간자료에서 온라인으로도 이용하실 수 있습니다. http://www.kihasa.re.kr/html/jsp/publication/periodical/focus/list.jsp