Journal of the Korean Data & Information Science Society
2015, **26**(5), 1047–1059

http://dx.doi.org/10.7465/jkdi.2015.26.5.1047
한국데이터정보과학회지

소셜 빅데이터를 활용한 담배 위험 예측[†]

송태민[1] · 송주영[2] · 천미경[3]

[13]한국보건사회연구원 정보통계연구실 · [2]펜실베니아주립대학 범죄학과

접수 2015년 8월 5일, 수정 2015년 9월 1일, 게재확정 2015년 9월 24일

요 약

본 연구는 국내의 블로그, 카페, SNS 등 인터넷을 통해 수집된 소셜 빅데이터를 데이터마이닝 분석 기법을 적용하여 우리나라 국민의 담배에 대한 위험요인을 예측하고자 하였다. 주요분석 결과는 다음과 같다. 첫째, 온라인상에 '담뱃값인상'이 언급될 경우 담배에 대한 일반군 (negative)이 58.6%에서 74.8%로 증가하며, '폐암'이 언급될 경우 73.1%로 증가하는 것으로 나타났다. 둘째, 담뱃값인상 이후 담배에 대한 위험군 (positive)은 5.6% 감소하고, 일반군은 6.1% 증가한 것으로 나타났다. 셋째, 'FCTC, 담뱃값인상, 금연관련법, 흡연규제, 금연광고, 금연사업'과 관련된 정책이 온라인상에 많이 언급될수록 담배에 대한 위험군이 감소하는 것으로 나타났다. 마지막으로 '금연약, 금연패치, 금연껌'이 온라인 상에 언급될수록 담배에 대한 위험군이 감소하나, '전자담배와 보조제'가 온라인상에 언급될수록 담배에 대한 위험군을 증가시키는 것으로 나타났다.

주요용어: 감성분석, 담배, 데이터마이닝, 소셜빅데이터, 연관분석.

1. 서론

우리나라 19세 이상 성인 남성 흡연율은 1998년 66.3%에서 2005년 51.6%, 2013년 42.1%로 감소추세이지만 (Ministry of Health and Welfare, 2014), 2012년 15세 이상 남성 흡연율은 OECD 평균 24.9% 보다 높은 37.6%로 세계에서 가장 높은 위치를 차지하고 있다 (OECD, 2014). 이와 같이 우리나라 남성 흡연율이 OECD 회원국 중 최고 수준에 달하는 상황에서 현 정부는 2015년 1월 1일부터 담뱃값을 2,000원 인상하는 등 범정부 차원의 금연 종합대책을 발표하였다 (Ministry of Health and Welfare, 2014 press release). 전 세계적으로 흡연으로 인해 매년 600만 명이 사망하고 있으며 (WHO, 2008), 전체 암 사망의 30.5%, 호흡기질환 사망의 19.8%, 심혈관질환 사망의 11.4%가 흡연으로 인해 사망한 것으로 예측되었다 (Zheng 등, 2014). 우리나라는 1985년 24,338명, 2003년 46,207명, 2012년 58,155명이 흡연으로 인한 사망자수로 보고되었고 (Jung 등, 2013), 2012년 기준 흡연에 의한 건강보험 진료비는 1조 8,466억 원으로 추정하고 있다 (Ji 등, 2014).

담배연기는 사람에게 치명적인 화학물질 7,000개 이상을 함유하고 있으며, 이로 인해 폐암을 비롯한 각종 암과 심혈관질환, 호흡기질환, 만성질환 등 다양한 질병과 관련있는것으로 알려져 있다 (Carter

[†] 본 논문의 일부내용은 한국보건사회연구원의 보건복지 ISSUE&FOCUS에 게재된 '송태민 (2015)의 소셜 빅데이터를 활용한 담배 위험 예측'의 내용임.

[1] (339-007) 세종특별자치시 시청대로 370 (반곡동), 세종국책연구단지 사회정책동 한국보건사회연구원 정보통계연구실, 빅데이터연구센터장.

[2] 교신저자: North Wales PA 19454, USA, 펜실베니아주립대학 범죄학과 조교수.
E-mail: juyoung81@gmail.com

[3] (339-007) 세종특별자치시 시청대로 370 (반곡동), 세종국책연구단지 사회정책동 한국보건사회연구원 정보통계연구실, 연구원.

등, 2015; CDC, 2010; Thun 등, 2013). 우리나라는 1995년 국민건강증진법이 제정됨에 따라 본격적으로 담배 판매, 광고, 금연구역 확대 등을 추진하였고, 청소년 보호법, 학교보건법 등에서도 청소년 흡연과 관련하여 제도적으로 규제하고 있다. 또한 2005년 WHO 담배규제기본협약 (FCTC) 비준 이후 다양한 흡연 예방 및 담배규제정책을 시행하고 있다 (Kang과 Lee, 2011). 담배규제정책들은 선진국과 개발도상국의 차이가 있을지라도 실제 사례를 통해 효과가 입증되었다. 미국은 지속적으로 담뱃값이 인상됨에 따라 담배 소비량이 줄어들었고 (Campaign for Tobacco-Free Kids, 2013), 터키도 2008년에 비해 2012년 담뱃값이 42.1% 증가했을 때 흡연율은 14.6% 감소하였다 (CDC, 2014). 우리나라는 2004년 12월 2,000원에서 500원 인상된 후 10년 동안 추가적인 인상이 이루어지지 않아 흡연율의 상승과 하락을 반복하여 담뱃값 인상에 대한 금연효과는 크지 않은 것으로 나타났다 (Ministry of Health and Welfare, 2014). 담뱃갑 경고 그림은 2000년 12월 캐나다에서 제일 먼저 시작되었고, 흡연자의 63%는 담뱃갑 경고 그림을 통해 적어도 1번 이상의 금연효과를 경험했으며 (Hammond 등, 2004), 세계 여러 나라에서도 법안으로 정하여 시행되고 있다. 우리나라는 담뱃갑 경고 그림을 의무화하는 국민건강증진법 개정안이 '사실적 근거를 바탕으로 지나치게 혐오감을 주지 않는다'는 조건하에 통과되어 2016년 12월부터는 담뱃갑에 경고 그림이 의무적으로 표기된다.

최근 2015년 1월 1일 담뱃값 인상으로 건강증진 부담금 비중을 확대 (14.2%→18.7%)하였으며, 추가 확보된 재원을 금연 성공률이 가장 높은 약물·상담 치료에 지원하고 학교, 군부대, 사업장 등에 대한 금연지원을 대폭 확대하는 한편, 금연광고와 금연캠페인을 연중 실시하고 보건소 금연클리닉, 금연상담전화, 온라인 상담 등 1:1 맞춤형 금연상담서비스도 대폭 강화할 계획이다 (Ministry of Health and Welfare, 2014).

한편 모바일 인터넷과 소셜미디어의 확산으로 데이터양이 증가하여 데이터의 생산, 유통 소비체계에 큰 변화가 일어나면서 데이터가 경제적 자산이 될 수 있는 빅데이터 시대를 맞이하게 되었다. 세계 각국의 기업들이 빅데이터가 공공과 민간에 미치는 파급효과를 전망함에 따라 SNS를 통해 생산되는 소셜 빅데이터의 활용과 분석을 통하여 사회적 문제의 해결과 정부의 정책을 효과적으로 추진할 수 있을 것으로 예측하고 있다. 또한 SNS의 역할은 기업에서 마케팅 측면뿐만 아니라 학자들 간의 학문연구에서도 갈수록 중요해지고 있으며, 이러한 공동의 협력은 집단창의성 (swarm creativity)을 통해 혁신을 가져올 수 있을 뿐만 아니라 성공의 가능성도 더욱 커지게 하는 결과를 가져온다 (Chun, 2015). 우리나라는 정부 3.0과 창조경제의 추진과 실현을 위하여 다양한 분야에 빅데이터의 효율적 활용을 적극적으로 모색하고 있다. 정부 3.0은 공공 부문의 데이터 공개를 통해 행정의 효율성을 높이고, 국민의 참여를 활성화시키며 경제 활성화 등의 파급효과를 기대하고 있으며, 정부의 데이터 공개 정책은 정보화 시대에 소통과 공유, 협업 전략이 무엇보다 중요하다는 것을 의미한다 (Hong, 2014).

소셜 빅데이터의 분석은 사용자가 남긴 온라인 문서의 의미를 분석하는 것으로 자연어 처리 기술인 주제분석 (text mining)과 감성분석 기술인 오피니언마이닝 (opinion mining)을 실시한 후, 네트워크 분석 (network analysis)과 통계분석 (statistics analysis)을 실시해야 한다. 특히, 소셜 네트워크 데이터는 일반적인 테이블 형태의 분석 데이터와는 다른 노드와 노드의 연결을 나타내는 관계 데이터의 형태를 가진다((Chun과 Leem, 2014). 기존에 실시하던 횡단적 조사나 종단적 조사 등을 대상으로 한 연구는 정해진 변인들에 대한 개인과 집단의 관계를 보는 데에는 유용하나, 사이버 상에서 언급된 개인별 문서 (버즈: buzz)에 논의된 관련 정보 상호간의 연관관계를 밝히고 원인을 파악하기에는 한계가 있다 (Song 등, 2013). 소셜 빅데이터의 분석은 훨씬 방대한 양의 데이터를 활용하여 다양한 참여자의 생각과 의견을 확인 할 수 있기 때문에 기존의 오프라인 조사와 함께 활용하면 사회적 문제의 예측을 보다 정확히 할 수 있다. 본 연구는 우리나라 온라인 뉴스사이트, 블로그, 카페, SNS, 게시판 등에서 수집한 소셜 빅데이터를 바탕으로 우리나라 국민의 담배에 대한 위험 예측모형과 연관규칙을 파악한다.

2. 연구방법

2.1. 연구대상

본 연구는 국내의 SNS, 온라인 뉴스 사이트 등 인터넷을 통해 수집된 소셜 빅데이터를 대상으로 하였다. 본 분석에서는 200개의 온라인 뉴스사이트, 10개의 게시판, 1개의 SNS (트위터), 4개의 블로그 등 총 217개의 온라인 채널을 통해 수집 가능한 텍스트 기반의 웹문서 (버즈)를 소셜 빅데이터로 정의하였다. 담배 관련 토픽 (topic)의 수집은 2011~2015년의 1/4분기 기간 동안 (각 연도의 1~3월, 총 15개월간) 해당 채널에서 요일, 주말, 휴일을 고려하지 않고 매시간 단위로 수집하였으며, 수집된 총 1,091,958건 (2011년: 94,412건, 2012년: 229,322건, 2013년: 286,067건, 2014년: 181,713건, 2015년: 300,444건)의 텍스트 (text) 문서를 본 연구의 분석에 포함시켰다. 토픽은 소셜 분석 및 모니터링의 '대상이 되는 주제어'를 의미한다. 담배와 관련된 토픽이 포함된 모든 문서를 수집하기 위해 '담배'를 사용하였으며, 토픽과 같은 의미로 사용되는 토픽 유사어로는 '흡연, 담뱃값, 담배 피, 담배 추천, 담배가격, 훈녀생정담배, 중딩담배, 고딩담배, 중고딩 담배, 청소년 담배' 용어를 사용하였다. 본 연구를 위한 소셜 빅데이터의 수집은 크롤러 (crawler)를 사용하였고, 이후 주제분석을 통해 분류된 명사형 어휘를 유목화 (categorization)하여 분석요인으로 설정하였다.

2.2. 연구도구

담배와 관련하여 수집된 문서는 주제분석의 과정을 거친다. 이때 주제분석에 사용되는 사전은 '21세기 세종계획'과 같은 범용사전도 있지만 대부분 분석의 목적에 맞게 사용자가 설계한 사전을 사용하게 된다. 본 연구의 담배관련 주제분석은 '(주)SK텔레콤 스마트인사이트'에서 관련 문서를 수집 한 후 원시자료 (raw data)에서 나타난 상위 2,000개의 키워드들을 대상으로 유목화를 하여 사용자 사전을 구축하였다. 주제분석을 거친 후 다음과 같이 정형화 데이터로 코드화하여 사용하였다.

1) 담배 관련 감정

본 연구의 담배 감정 키워드는 문서 수집 이후, 주제 분석을 통하여 총 66개 (걱정, 고민, 고생, 고통, 깔끔, 다짐, 대단, 두려움, 만족, 믿음, 부담, 불가능, 불리, 불만, 불안, 불편함, 사랑, 스트레스, 실패, 어려움, 여유, 염려, 욕구, 위험, 유혹, 응원, 의지, 의지력, 자신감, 재미, 조심, 즐거움, 짜증, 창피, 최고, 최선, 충격, 치유, 편안, 포기, 피곤, 필요, 행복, 호기심, 파이팅, 활력, 후회, 희망, 힐링, 힘들다, 성공, 도움, 문제, 추천, 관심, 도전, 결심, 잘못, 혐오, 심각, 논란, 불편, 고발, 이해, 지적, 끔찍) 키워드로 분류하였다. 본 연구에서는 66개의 담배 감정 키워드 (변수)가 가지는 담배 감정 정도를 판단하기 위해 요인분석을 통하여 12개의 요인 (44개 변수)으로 축약을 실시한 후, 감성분석을 실시하였다. 일반적으로 감성분석은 긍정과 부정의 감성어 사전으로 분석해야 하나, 본 연구에서는 요인분석의 결과로 분류된 주제어의 의미를 파악하여 감성분석을 실시하였다. 요인분석에서 결정된 12개의 요인에 대한 주제어의 의미를 파악하여 '일반군, 잠재군, 위험군'으로 감성분석을 실시하였다. 따라서 본 연구에서 일반군은 23개 변수 (스트레스, 위험, 문제, 조심, 성공, 실패, 결심, 의지, 욕구, 논란, 지적, 부담, 불만, 염려, 걱정, 짜증, 창피, 불안, 끔찍, 충격, 불편, 파이팅, 응원), 위험군은 16개 변수 (믿음, 사랑, 희망, 행복, 최선, 추천, 깔끔, 만족, 고민, 최고, 즐거움, 여유, 대단, 피곤, 힐링, 치유)로 분류하였다. 그리고 일반군과 위험군의 감정을 동일한 횟수로 표현한 문서는 잠재군으로 분류하였다. 일반군은 담배가 위험하다고 생각하는 혐오적인 감정이고, 위험군은 담배가 위험하지 않다는 애호적인 감정이며, 잠재군은 담배의 위험을 보통으로 생각하는 감정을 나타낸다.

2) 담배와 관련된 정책

담배와 관련된 정책의 정의는 주제 분석 과정을 거쳐 '담뱃값인상, FCTC (담배규제기본협약 등), 금

연관련법 (국민건강증진법, 학교보건법 등), 흡연규제 (금연구역, 벌금부과 등), 금연광고 (공익광고, 금연캠페인 등), 금연사업 (금연상담전화, 금연클리닉 등)' 6개 정책으로 정책이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

3) 담배와 관련된 질환

담배와 관련된 질환의 정의는 주제 분석을 거쳐 '가래, 간암, 감기, 동맥경화, 고혈압, 구토, 뇌혈관질환, 당뇨병, 대장암, 두통, 마비, 만성질환, 발기부전, 불면증, 사망, 식도암, 심혈관질환, 염증, 우울증, 위암, 유방암, 폐암, 치매, 후두암, 구강암'의 25개로 질환이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

4) 담배에 대한 금연도구

담배에 대한 금연도구의 정의는 주제 분석 과정을 거쳐 '금연껌 (금연껌, 니코틴로렌즈, 니코틴껌, 니코틴엘로젠즈, 사탕, 트로키), 금연약 (금연약, 약물, 니코엔, 니코스텝, 챔픽스, 니코피온, 니코그린, 니코레스, 부프로피온, 흡연욕구저하제, 챔픽스정, 바레니클린, 웰부트린), 전자담배 (전자담배, 스모키전자담배, 애니스틱, 라스트스틱), 금연패치 (니코레트, 니코틴패치, 패치, 금연패치, 니코틴보조제, 금연보조제, 보조제, 금연침), 보조제 (물담배, 파이프담배, 리엔파이프, 롤링토바코, 금연파이프, 금연초, 건향초)'의 5개 금연도구가 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

5) 담배에 대한 치료

담배에 대한 치료의 정의는 주제 분석 과정을 거쳐 '금연클리닉, 금연상담전화, 병원, 금연교실'의 4개로 해당 치료가 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

6) 담배와 관련된 폐해

담배와 관련된 폐해의 정의는 주제 분석을 거쳐 '간접흡연, 알코올, 중독, 기억력, 담배꽁초, 도박마약, 이혼, 정신건강, 폭력'의 9개 폐해로 해당 폐해가 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

7) 담배에 대한 유해물질

담배에 대한 유해물질의 정의는 주제 분석 과정을 거쳐 '니코틴, 발암물질, 유해물질, 일산화탄소, 타르, 화학물질, 노폐물'의 7개 유해물질로 해당 유해물질이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

8) 담배에 대한 장소

담배에 대한 장소의 정의는 주제 분석 과정을 거쳐 'PC방, 가정, 금연건물, 아파트, 공공장소, 흡연구역, 직장, 술집, 식당, 학교'의 10개 장소로 해당 장소가 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

9) 담배에 대한 관련기관

담배에 대한 관련기관의 정의는 주제 분석 과정을 거쳐 '청와대, 국회, 보건복지부, 여성가족부, 기획재정부, 지방자치단체, 공공기관, 세계보건기구, 금연단체 (한국금연운동협의회, 한국건강관리협회, 한국보건의료연구원 등), 담배회사'의 10개 기관으로 해당 기관이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

3. 분석 방법

본 연구에서 우리나라 담배의 위험을 설명하는 가장 효율적인 예측모형을 구축하기 위해 특별한 통계적 가정이 필요하지 않은 데이터마이닝의 연관분석 (association analysis)과 의사결정나무 (decision

tree) 방법을 사용하였다. 소셜 빅데이터 분석에서 연관분석은 하나의 온라인 문서 (transaction)에 포함된 둘 이상의 단어들에 대한 상호관련성을 발견하는 것으로 동시에 발생한 어떤 단어들의 집합에 대해 조건과 연관규칙을 찾는 분석방법이다. 전체 문서에서 연관규칙의 평가 측도는 지지도 (support), 신뢰도 (confidence), 향상도 (lift)로 나타낼 수 있다. 지지도는 자주 발생하지 않는 규칙을 제거하는데 이용되며 신뢰도는 단어들의 연관성 정도를 파악하는데 이용할 수 있다. 향상도는 연관규칙 ($X \rightarrow Y$)에서 단어 X가 없을 때 보다 있을 때 단어 Y가 발생할 비율을 나타낸다. 연관분석 과정은 연구자가 지정한 최소 지지도를 만족시키는 빈발항목집합 (frequent itemset)을 생성한 후, 이들에 대해 최저 신뢰도 기준을 마련하고 향상도가 1인 이상인 것을 규칙으로 채택한다 (Park, 2013). 본 연구의 연관분석은 선험적 규칙 (apriori principle) 알고리즘을 사용하였으며, 담배감정에 사용된 연관분석의 측도는 지지도 0.001, 신뢰도 0.01을 기준으로 시뮬레이션 하였다. 본 연구의 의사결정나무 형성을 위한 분석 알고리즘은 CHAID (Chi-squared Automatic Interaction Detection)를 사용하였다. 정지규칙 (stopping rule)으로 관찰치가 충분하여 상위노드 (부모마디)의 최소케이스 수는 100으로 하위노드 (자식마디)의 최소 케이스 수는 50으로 설정하였고, 나무깊이는 3수준으로 정하였다. 본 연구의 기술 분석, 다중응답 분석, 의사결정나무분석은 SPSS v. 22.0을 사용하였고, 연관분석과 시각화는 R을 사용하였다.

4. 연구 결과

4.1. 담배 관련 버즈 현황

담배와 관련된 버즈는 연도별로 비슷하게 8시부터 증가하여 11시 이후 감소하며, 다시 12시 이후 증가하여 17시 이후 감소하고, 20시 이후 증가하여 23시 이후 급감하는 추세를 보이고 있는 것으로 나타났다. 요일별로 평일에는 수요일, 목요일, 화요일, 월요일, 금요일 순으로 높은 추이를 보이는 반면, 주말에는 감소하는 것으로 나타났다 (Figure 4.1).

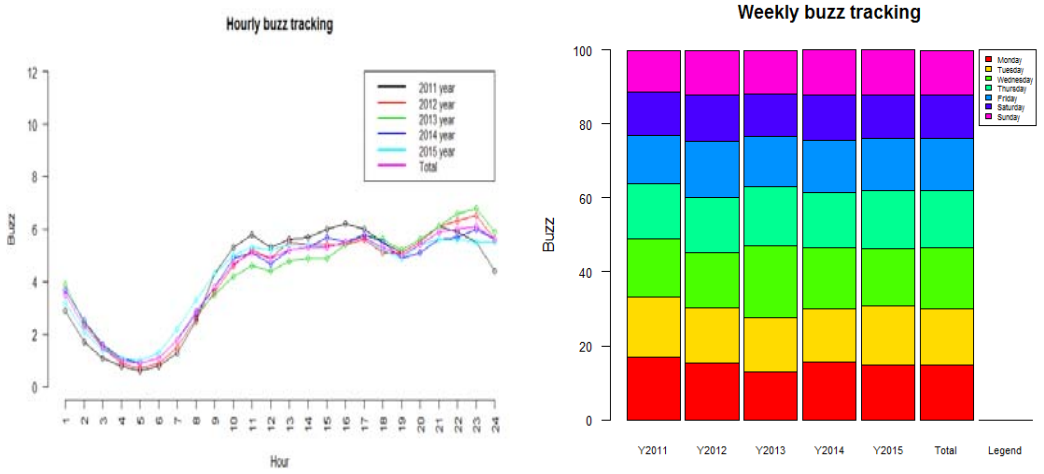
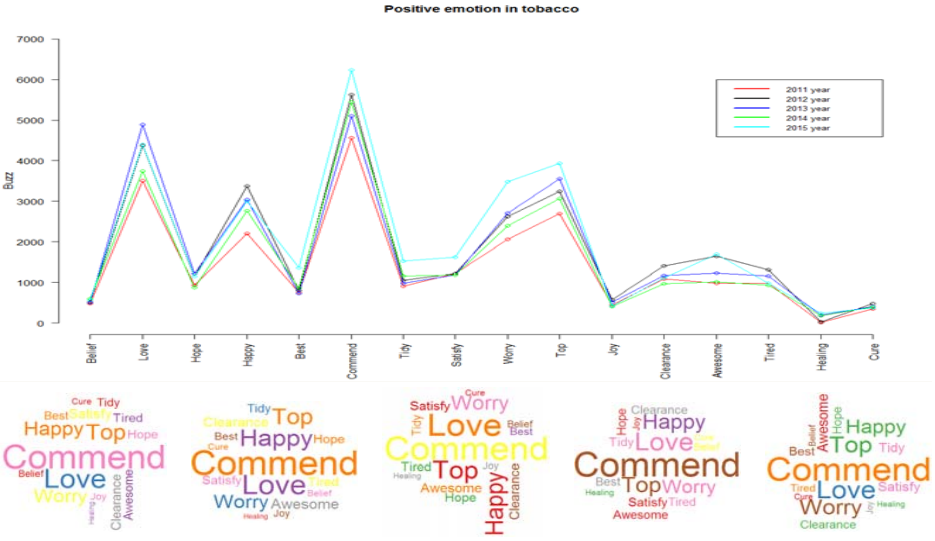


Figure 4.1 Buzz tracking in tobacco

Figure 4.2와 같이 연도별 담배에 대한 위험군의 변화는 2011년 대비 평균 1.8배씩 증가하였으며, 주된 위험군 변수는 '추천, 사랑, 최고, 행복, 고민 등'의 순으로 집중된 것으로 나타났다.



2011 year 2012 year 2013 year 2014 year 2015 year

Figure 4.2 Positive emotion in tobacco

Table 4.1 Descriptive Statistics

Division	Items	*N* (%)		Division	Items	N (%)	
Emotion	Negative	110,401	(58.6)	Channel	Blog	147,235	(13.5)
	Usually	16,206	(8.6)		Cafe	268,463	(24.6)
	Positive	61,660	(32.8)		SNS	577,125	(52.9)
	Total	188,267			Board	53,243	(4.9)
Policy	Raising cigarette price	58,267	(50.0)		News	45,892	(4.2)
	FCTC	454	(0.4)		Total	1,091,958	
	Non-smoking laws	13,528	(11.6)	Tools	Non-smoking gum	4,260	(7.5)
	Smoking regulations	27,828	(23.9)		Non-smoking drug	4,778	(8.4)
	Non-smoking ads	5,626	(4.8)		Electronic cigarette	38,600	(68.1)
	Non-smoking business	10,887	(9.3)		Non-smoking patches	5,200	(9.2)
	Total	116,590			Supplements	3,819	(6.7)
Therapy	Non-smoking clinic	10,015	(25.6)		Total	56,657	
	Non-smoking helpline	360	(0.9)	Place	PC rooms	3,932	(2.1)
	Hospital	28,062	(71.7)		House	12,414	(6.7)
	Non-smoking class	694	(1.8)		Non-smoking building	1,609	(0.9)
	Total	39,131			Apartment	14,915	(8.1)
Harmful effect	Secondhand smoke	88,855	(48.6)		Public places	34,775	(18.9)
	Alcohol	16,217	(8.9)		Smoking area	14,029	(7.6)
	Addiction	13,333	(7.3)		Office	19,666	(10.7)
	Memory	12,396	(6.8)		Bar	26,000	(14.1)
	Cigarette butt	29,651	(16.2)		Restaurant	28,659	(15.5)
	Gambling drugs	8,469	(4.6)		School	28,314	(15.4)
	Divorce	3,333	(1.8)		Total	184,313	
	Mental health	3,907	(2.1)	Organization	Blue House	32,311	(48.6)
	Violence	6,611	(3.6)		Congress	5,803	(8.7)
	Total	182,772			Ministry of Health & welfare	7,894	(11.9)
Hazardous substances	Nicotine	18,496	(46.8)		Ministry of Gender Equality & Family	1,502	(2.3)
	Carcinogen	5,651	(14.3)		Ministry of Strategy & Finance	2,971	(4.5)
	Hazardous	3,142	(8.0)		Local Government	4,963	(7.5)
	Carbon monoxide	2,894	(7.3)		Public institutions	4,021	(6.1)
	Tar	4,707	(11.9)		WHO	2,227	(3.4)
	Chemicals	1,875	(4.7)		Non-smoking groups	1,079	(1.6)
	Waste	2,745	(6.9)		Tobacco companies	3,673	(5.5)
	Total	39,510			Total	66,444	

Table 4.1과 같이 담배의 위험군을 나타내는 버즈는 32.8%, 잠재군은 8.6%, 일반군은 58.6%로 나타났다. 담배와 관련된 정책은 담뱃값인상 (50.0%), 흡연규제 (23.9%), 금연사업 (9.3%) 등의 순으

로 나타났다. 담배에 대한 치료는 병원 (71.7%), 금연클리닉 (25.6%), 금연교실 (1.8%), 금연상담전화 (0.9%)의 순으로 나타났다. 담배와 관련된 폐해로는 간접흡연 (48.6%), 담배꽁초 (16.2%), 알코올 (8.9%), 중독 (7.3%) 등의 순으로 나타났다. 담배에 대한 유해물질로는 니코틴 (46.8%), 발암물질 (14.3%), 타르 (11.9%), 유해물질 (8.0%) 등의 순으로 나타났다. 담배와 관련된 채널로는 SNS (52.9%), 카페 (24.6%), 블로그 (13.5%) 등의 순으로 나타났다. 담배에 대한 금연도구로는 전자담배 (68.1%), 금연패치 (9.2%), 금연약 (8.4%) 등의 순으로 나타났다. 담배에 대한 장소로는 공공장소 (18.9%), 식당 (15.5%), 학교 (15.4%) 등의 순으로 나타났다. 담배에 대한 관련기관으로는 청와대 (48.6%), 보건복지부 (11.9%), 국회 (8.7%) 등의 순으로 나타났다.

Table 4.2와 같이 담배와 관련된 연도별 위험군의 변화는 2011년 37.2%, 2012년 34.6%, 2013년 32.7%, 2014년 34.0%, 2015년 28.4%로 나타났으며, 2015년 담뱃값 인상 이후 위험군이 5.6% 감소하고, 일반군이 6.1% 증가한 것으로 나타났다.

Table 4.2 Emotion changes in tobacco (unit: n (%))

Year	Positive		Usually		Negative		Total
2011	9,156	(37.2)	2,764	(11.2)	12,674	(51.5)	24,594
2012	13,304	(34.6)	3,112	(8.1)	22,050	(57.3)	38,466
2013	13,515	(32.7)	3,134	(7.6)	24,660	(59.7)	41,309
2014	11,251	(34.0)	2,956	(8.9)	18,924	(57.1)	33,131
2015	14,434	(28.4)	4,240	(8.4)	32,093	(63.2)	50,767
Total	61,660	(32.8)	16,206	(8.6)	110,401	(58.6)	188,267

4.2. 담배 위험 관련 연관성 분석

Table 4.3과 같이 정책요인에 대한 담배 위험 연관성 예측에서 가장 신뢰도가 높은 연관규칙으로는 담뱃값인상, 금연관련법 => 일반군이며 세 변인의 연관성은 지지도 0.002, 신뢰도는 0.540, 향상도는 5.339로 나타났다. 이는 온라인 문서에서 담뱃값인상, 금연관련법이 언급되면 담배를 혐오적으로 생각하는 일반군이 될 확률이 54.0%이며, 담뱃값인상, 금연관련법이 언급되지 않은 문서 보다 담배에 대한 감정이 혐오적일 확률이 5.34배 높아지는 것을 나타낸다. 특히, 담뱃값인상 => 위험군 두 변인의 연관성은 지지도 0.002, 신뢰도는 0.041, 향상도는 0.724로 나타나 담뱃값인상은 위험군을 감소시키는 것으로 나타났다. 반면, 담뱃값인상 => 일반군의 향상도 (1.854)가 담뱃값인상 => 잠재의 향상도 (1.493)보다 높게 나타나 온라인 문서에 담뱃값인상이 언급될 경우 잠재군 보다 일반군의 확률이 더 높은 것으로 나타났다.

Table 4.3 Association analysis in tobacco policy

Rule	Support	Confidence	LIft
{Rasing cigarette price, Non-smoking laws} =>{Negative)}	0.001776625	0.53978854	5.3389590
{Rasing cigarette price, Smoking regulations} => {Negative)}	0.001778457	0.49465104	4.8925115
{Non-smoking business} => {Negative)}	0.004685162	0.46991825	4.6478836
{Non-smoking laws, Smoking regulations} => {Negative)}	0.001862709	0.46395985	4.5889500
{Non-smoking laws} => {Negative)}	0.005130234	0.41410408	4.0958349
{Non-smoking ads} => {Negative)}	0.001716183	0.33309634	3.2946007
{Smoking regulations} => {Negative)}	0.007551572	0.29632025	2.9308545
{Rasing cigarette price} => {Negative)}	0.010004048	0.18748176	1.8543511
{Non-smoking business} => {Positive)}	0.001788530	0.17938826	3.1768480
{} => {Negative)}	0.101103705	0.10110371	1.0000000
{Smoking regulations} => {Positive)}	0.001625520	0.06378468	1.1295846
{} => {Positive)}	0.056467373	0.05646737	1.0000000
{Rasing cigarette price} => {Positive)}	0.002182318	0.04089794	0.7242755
{Rasing cigarette price} => {Usually)}	0.001182280	0.02215662	1.4929102
{} => {Usually)}	0.014841230	0.01484123	1.0000000

4.3. 담배의 위험에 영향을 미치는 요인

금연정책의 중요한 요인인 흡연규제와 금연도구가 담배의 위험에 미치는 요인은 Table 4.4와 같이 금연과 관련한 모든 정책 요인은 담배의 위험군에 부정적인 영향을 미치는 것으로 나타나, FCTC, 담뱃값 인상, 금연관련법, 흡연규제, 금연광고, 금연사업과 관련한 정책이 온라인상에 많이 언급될수록 위험군은 감소하는 것으로 나타났다. 금연과 관련한 도구 요인의 영향은 금연약, 금연패치, 금연껌은 부적인 영향을 미치는 것으로 나타나, 금연약, 금연패치, 금연껌과 관련한 금연도구가 온라인상에 많이 언급될수록 위험군은 감소하는 것으로 나타났으나, 전자담배와 금연보조제는 정적인 영향을 미치는 것으로 나타나 전자담배와 보조제와 관련한 금연도구가 많이 언급될수록 담배에 대한 위험군은 증가하는 것으로 나타났다.

Table 4.4 Logistic regression analysis in tobacco risk factors*

Variables		Positive				Usually			
		b†	S.E.‡	OR§	P	b†	S.E.	OR§	P
Policy	Rasing cigarette price	-.854	.024	.426	.000	-.207	.031	.813	.000
	FCTC	-1.328	.269	.265	.000	-.451	.215	.637	.036
	Non-smoking laws	-.845	.037	.430	.000	-.153	.044	.858	.001
	Smoking regulations	-.742	.027	.476	.000	-.191	.036	.826	.000
	Non-smoking ads	-.275	.049	.760	.000	.076	.065	1.079	.240
	Non-smoking business	-.242	.028	.785	.000	.410	.035	1.507	.000
Tools	Non-smoking gum	-.357	.051	.700	.000	.068	.069	1.071	.324
	Non-smoking drug	-1.556	.060	.211	.000	-.176	.058	.839	.003
	Electronic cigarettes	.206	.019	1.229	.000	.155	.032	1.167	.000
	Non-smoking patches	-1.091	.051	.336	.000	-.414	.065	.661	.000
	Supplements	.374	.060	1.454	.000	.688	.081	1.990	.000

* base category: Negative, † Standardized coefficients, ‡ Standard error, § Adjusted odds ratio

4.4. 담배 관련 위험 예측모형

본 연구에서는 담배 관련 위험을 예측하기 위하여 담배와 관련된 정책요인과 금연도구요인에 대해 데이터마이닝 분석을 실시하였다. 담배와 관련된 정책요인이 담배의 위험예측 모형에 미치는 영향은 Figure 4.3과 같다. 나무구조의 최상위에 있는 네모는 루트노드로서, 예측변수 (독립변수)가 투입되지 않은 종속변수 (위험군, 잠재군, 일반군)의 빈도를 나타낸다. 루트노드에서 위험군은 32.8% (61,660건), 잠재군은 8.5% (16,206건), 일반군은 58.6% (110,401건)으로 나타났다. 루트노드 하단의 가장 상위에 위치하는 요인은 담배의 위험예측에 가장 영향력이 높은 (관련성이 깊은) 정책요인으로 '담뱃값인상요인'의 영향력이 가장 큰 것으로 나타났다. '담뱃값인상요인'이 있을 경우 담배의 위험군은 이전의 32.8%에서 16.3%로 크게 감소한 반면, 잠재군은 이전의 8.5%에서 8.8%, 일반군은 이전의 58.6%에서 74.8%로 증가하였다. '담뱃값인상요인'이 있고 '금연관련법요인'이 있는 경우 담배의 위험군은 이전의 16.3%에서 6.0%, 잠재군은 이전의 8.8%에서 8.0%로 감소한 반면, 일반군은 이전의 74.8%에서 88.0%로 증가하였다. Table 4.5의 담배와 관련한 정책요인의 위험예측 모형에 대한 이익도표와 같이 담배의 위험군에 가장 영향력이 높은 경우는 '담뱃값인상요인'이 없고 '흡연규제요인'이 없으며 '금연관련법요인'이 없는 조합으로 나타났다. 즉, 8번 노드의 지수 (index)가 108.1%로 뿌리마디와 비교했을 때 8번 노드의 조건을 가진 집단이 담배에 대한 위험군이 높을 확률이 1.08배로 나타났다. 담배의 잠재군에 가장 영향력이 높은 경우는 '담뱃값인상요인'이 있고 '금연관련법요인'이 없으며 '금연사업요인'이 있는 조합으로 나타났다. 즉, 14번 노드의 지수가 168.0%로 뿌리마디와 비교했을 때 14번 노드의 조건을 가진 집단에서 잠재군이 높을 확률이 1.68배로 나타났다. 담배의 일반군에 가장 영향력이 높은 경

우는 '담뱃값인상요인'이 있고 '금연관련법요인'이 있으며 'FCTC요인'이 있는 조합으로 나타났다. 즉, 12번 노드의 지수가 163.0%로 뿌리마디와 비교했을 때 12번 노드의 조건을 가진 집단에서 일반군이 높을 확률이 1.63배로 나타났다.

담배 관련 질병요인이 담배의 위험예측 모형에 미치는 영향은 Figure 4.4와 같다. 담배의 위험예측에 가장 영향력이 높은 질병요인으로 '폐암'의 영향력이 가장 큰 것으로 나타났다. '폐암'이 있을 경우 위험군은 이전의 32.8%에서 14.6%로 크게 감소한 반면, 잠재군은 이전의 8.6%에서 12.3%, 일반군은 이전의 58.6%에서 73.1%로 증가하였다. '폐암'이 있고 '후두암'이 있는 경우 담배의 위험군은 이전의 14.6%에서 7.6%, 잠재군은 이전의 12.3%에서 6.5%로 감소한 반면, 일반군은 이전의 73.1%에서 85.8%로 증가하였다. Table 4.6의 담배와 관련한 질병요인의 위험예측 모형에 대한 이익도표와 같이 담배의 위험군에 가장 영향력이 높은 경우는 '폐암'이 없고 '심혈관질환'이 없으며 '고혈압'이 없는 조합으로 나타났다. 즉, 11번 노드의 지수가 104.6%로 뿌리마디와 비교했을 때 11번 노드의 조건을 가진 집단이 담배에 대한 위험이 높을 확률이 1.05배로 나타났다. 잠재군에 가장 영향력이 높은 경우는 '폐암'이 없고 '심혈관질환'이 있으며 '간암'이 있는 조합으로 나타났다. 즉, 14번 노드의 지수가 485.7%로 뿌리마디와 비교했을 때 14번 노드의 조건을 가진 집단에서 잠재군이 높을 확률이 4.86배로 나타났다. 담배의 일반에 가장 영향력이 높은 경우는 '폐암'이 있고 '후두암'이 있으며 '심혈관질환'이 있는 조합으로 나타났다. 즉, 8번 노드의 지수가 155.0%로 뿌리마디와 비교했을 때 8번 노드의 조건을 가진 집단에서 일반군이 높을 확률이 1.55배로 나타났다.

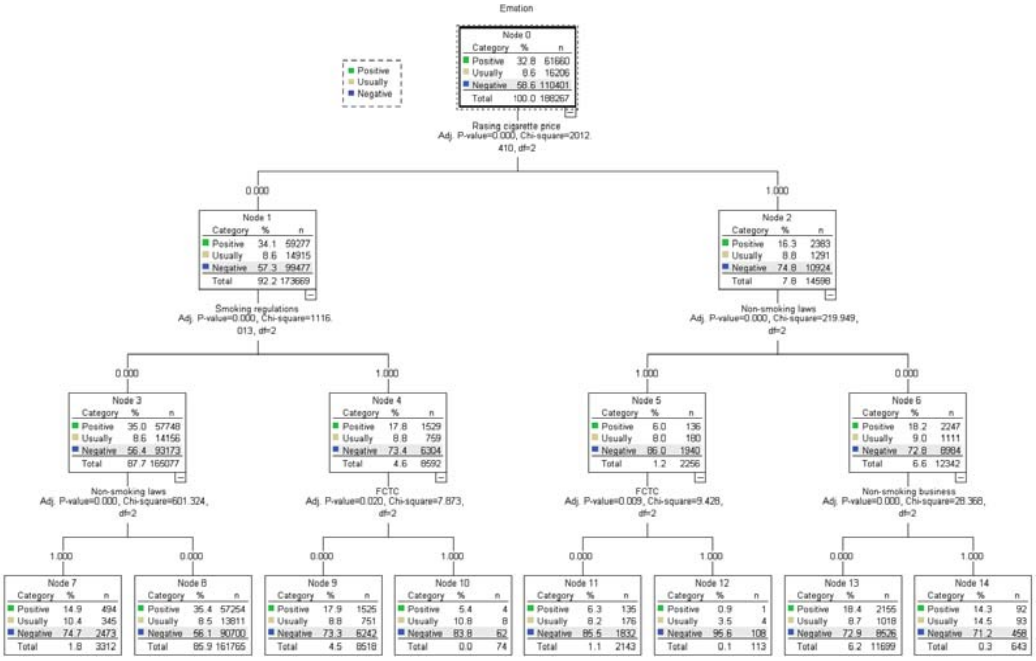


Figure 4.3 Prediction model in policy factors

Table 4.5 Profit chart of prediction model in policy factors

Part	Node	Profit Index				Cumulative Index			
		Node (n)	Node (%)	Profit (%)	Index (%)	Node (n)	Node (%)	Profit (%)	Index (%)
Positive	8	161765	85.9	92.9	108.1	161765	85.9	92.9	108.1
	13	11699	6.2	3.5	56.2	173464	92.1	96.3	104.6
	9	8518	4.5	2.5	54.7	181982	96.7	98.8	102.2
	7	3312	1.8	.8	45.5	185294	98.4	99.6	101.2
	14	643	.3	.1	43.7	185937	98.8	99.8	101.0
	11	2143	1.1	.2	19.2	188080	99.9	100.0	100.1
	10	74	.0	.0	16.5	188154	99.9	100.0	100.1
	12	113	.1	.0	2.7	188267	100.0	100.0	100.0
Usually	14	643	.3	.6	168.0	643	.3	.6	168.0
	10	74	.0	.0	125.6	717	.4	.6	163.6
	7	3312	1.8	2.1	121.0	4029	2.1	2.8	128.6
	9	8518	4.5	4.6	102.4	12547	6.7	7.4	110.8
	13	11699	6.2	6.3	101.1	24246	12.9	13.7	106.1
	8	161765	85.9	85.2	99.2	186011	98.8	98.9	100.1
	11	2143	1.1	1.1	95.4	188154	99.9	100.0	100.0
	12	113	.1	.0	41.1	188267	100.0	100.0	100.0
Negative	12	113	.1	.1	163.0	113	.1	.1	163.0
	11	2143	1.1	1.7	145.8	2256	1.2	1.8	146.6
	10	74	.0	.1	142.9	2330	1.2	1.8	146.5
	7	3312	1.8	2.2	127.3	5642	3.0	4.1	135.3
	9	8518	4.5	5.7	125.0	14160	7.5	9.7	129.1
	13	11699	6.2	7.7	124.3	25859	13.7	17.4	126.9
	14	643	.3	.4	121.5	26502	14.1	17.8	126.8
	8	161765	85.9	82.2	95.6	188267	100.0	100.0	100.0

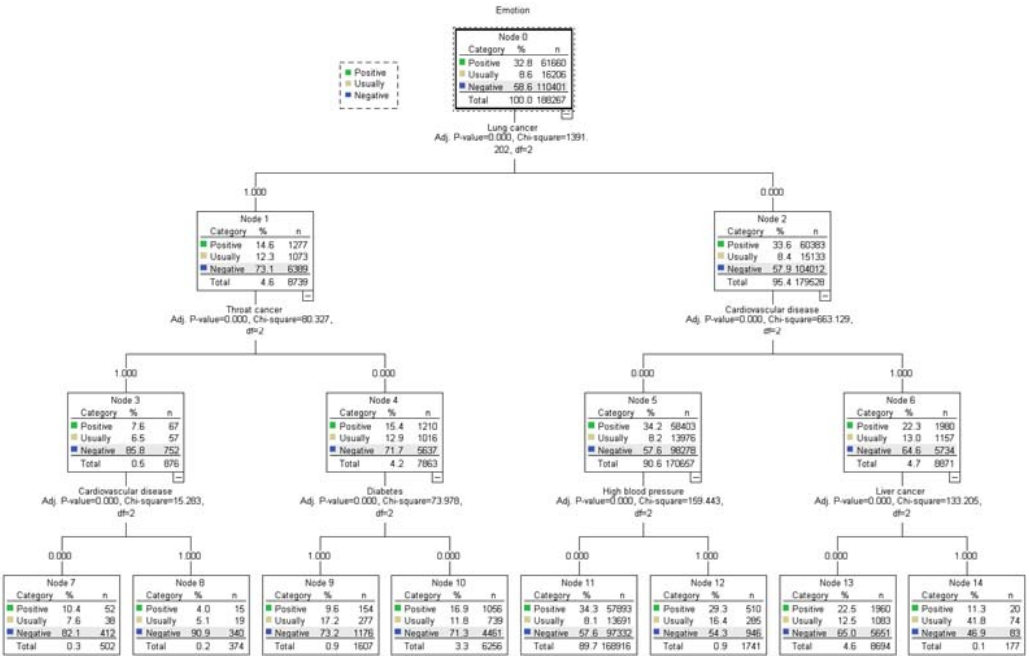


Figure 4.4 Prediction model in disease factors

Table 4.6 Profit chart of prediction model in disease factors

Part	Node	Profit Index				Cumulative Index			
		Node (n)	Node (%)	Profit (%)	Index (%)	Node (n)	Node (%)	Profit (%)	Index (%)
Positive	11	168916	89.7	93.9	104.6	168916	89.7	93.9	104.6
	12	1741	.9	.8	89.4	170657	90.6	94.7	104.5
	13	8694	4.6	3.2	68.8	179351	95.3	97.9	102.8
	10	6256	3.3	1.7	51.5	185607	98.6	99.6	101.0
	14	177	.1	.0	34.5	185784	98.7	99.6	101.0
	7	502	.3	.1	31.6	186286	98.9	99.7	100.8
	9	1607	.9	.2	29.3	187893	99.8	100.0	100.2
	8	374	.2	.0	12.2	188267	100.0	100.0	100.0
Usually	14	177	.1	.5	485.7	177	.1	.5	485.7
	9	1607	.9	1.7	200.2	1784	.9	2.2	228.6
	12	1741	.9	1.8	190.2	3525	1.9	3.9	209.6
	13	8694	4.6	6.7	144.7	12219	6.5	10.6	163.4
	10	6256	3.3	4.6	137.2	18475	9.8	15.2	154.6
	11	168916	89.7	84.5	94.2	187391	99.5	99.6	100.1
	7	502	.3	.2	87.9	187893	99.8	99.9	100.1
	8	374	.2	.1	59.0	188267	100.0	100.0	100.0
Negative	8	374	.2	.3	155.0	374	.2	.3	155.0
	7	502	.3	.4	140.0	876	.5	.7	146.4
	9	1607	.9	1.1	124.8	2483	1.3	1.7	132.4
	10	6256	3.3	4.0	121.6	8739	4.6	5.8	124.7
	13	8694	4.6	5.1	110.8	17433	9.3	10.9	117.8
	11	168916	89.7	88.2	98.3	186349	99.0	99.1	100.1
	12	1741	.9	.9	92.7	188090	99.9	99.9	100.0
	14	177	.1	.1	80.0	188267	100.0	100.0	100.0

5. 결론

본 연구는 국내의 온라인 뉴스사이트, 블로그, 카페, SNS, 게시판 등 인터넷을 통해 수집된 소셜 빅데이터를 주제분석과 감성분석 기술로 분류하고 데이터마이닝의 연관성 분석과 의사결정나무 분석 방법을 적용하여 분석함으로써 우리나라 국민의 담배에 대한 위험요인을 예측하고자 하였다. 본 연구의 주요 분석결과는 다음과 같다. 첫째, 담배관련 버즈는 매일 8시부터 증가하여 11시 이후 감소하며, 20시 이후 증가하여 23시 이후 급감하고, 요일별로 수요일, 목요일, 화요일, 월요일, 금요일 순으로 높은 추이를 보이는 반면, 주말에는 감소하는 것으로 나타났다. 둘째, 담뱃값 인상 이후 위험군은 5.6% 감소하고, 일반군은 6.1% 증가한 것으로 나타났다. 셋째, 버즈에서 담뱃값 인상, 금연관련법이 동시에 언급되면 일반군이 될 확률이 증가하며, 담뱃값 인상만 언급되어도 위험군을 감소시키는 것으로 나타났다. 넷째, FCTC, 담뱃값 인상, 금연관련법, 흡연규제, 금연광고, 금연사업과 관련된 정책이 온라인상에 많이 언급될수록 위험군이 감소하는 것으로 나타났다. 금연약, 금연패치, 금연껌과 같은 도구가 온라인상에 많이 언급될수록 위험군은 감소하는 것으로 나타났으나, 전자담배와 보조제는 위험군을 증가시키는 것으로 나타났다. 다섯째, 담배 위험 예측 모형에서 온라인상에 '담뱃값인상'이 언급될 경우 일반군이 58.6%에서 74.8%로 증가하며, '폐암'이 언급될 경우 73.1%로 증가한 것으로 나타났다. 끝으로 금연정책의 효과에 대한 대국민 조사와 더불어 소셜 미디어에서 수집된 빅데이터의 활용과 분석을 병행할 경우, 정부의 금연정책에 대한 예측 및 평가의 신뢰성이 더욱 제고될 것으로 예상되며, 또한 국민들이 금연에 적극적으로 동참할 수 있도록 소셜 빅데이터 분석을 통하여 담배를 애호적으로 생각하는 위험군을 감소시킬 수 있는 SNS 홍보가 강화되어야 할 것이다.

References

Campaign for Tobacco-Free Kids. (2013). *Increasing the federal tobacco tax reduces tobacco use*, Washington DC.

Carter, B. D., Abnet, C. C., Feskanich, D., Freedman, N. D., Hartge, P., Lewis, C. E., Ockene, J. K., Prentice, R. L., Speizer, F. E., Thun, M. J. and Jacobs, E. J. (2015). Smoking and mortality : Beyond established causes. N *ew England Journal of Medicine*, 372, 631-640.

Center for Disease Control and Prevention. (2010). *How tobacco smoke cause disease : The biology and behavioral basis for smoking attributable disease: A report of the surgeon genera*, US Department of Health and Human Services, Atlanta, GA.

Centers for Disease Control and Prevention. (2014). Cigarette prices and smoking prevalence after a tobacco tax increase-Turkey, 2008 and 2012. *MMWR Morbidity and Mortality Weekly Report*, **63**, 457-461.

Chun, H. (2015). The comparison of coauthor networks of two statistical Journals of the Korean Statistical Society using social network analysis. *Journal of the Korean Data & Information Science Society*, **26**, 335-346.

Chun, H. and Leem. B. (2014). Face/non-face channel fit comparison of life insurance company and non-life insurance company using social network analysis. *Journal of the Korean Data & Information Science Society*, **25**, 1207-1219.

Hammond, D., Fong, G. T., McDonald, P. W., Brown, K. S. and Cameron, R. (2004). Graphic Canadian cigarette warning labels and adverse outcomes. *American Journal of Public Health*, **94**, 1442-1445.

Hong, Y. (2014). A study on the invigorating strategies for open government data. *Journal of the Korean Data & Information Science Society*, **25**, 769-777.

Ji, S., Jung, K., Jeon, C., Kim, H., Yun, Y. and Kim, I. (2014). Smoking attributable risk and medical care cost in 2012 in Korea. *Journal of Health Informatics and Statistics*, **39**, 25-41.

Jung, K., Yun, Y., Baek, S., Jee, S. and Kim, I. (2013). Smoking-attributable mortality among Korean adults, 2012. J *ournal of Health Informatics and Statistics*, **38**, 36-48.

Kang, E. and Lee, J. (2011) Factor related to willingness-to-quit smoking cigarette price among Korean adults. *Korean Journal of Health Education and Promotion*, **28**, 125-137.

Ministry of Health and Welfare. (2014). *Korea health statistics 2013: Korea national health and nutrition examination survey*, Ministry of Health and Welfare, Korea.

Ministry of Health and Welfare. (2014) press release. Government-wide, 「No smoking comprehensive plan」 retrieved September 11, 2014.

Organization for Economic Cooperation and Development. (2014). *Health data 2014*, Paris, OECD.

Park, H. C. (2013). Proposition of causal association rule thresholds. J *ournal of the Korean Data & Information Science Society*, **24**, 1189-1197.

Song, T. M. (2015). *Predicting tobacco risk factors by using social big data*, Health and Social Welfare Issue & Focus, Korea Institute for Health and Social affairs, Korea.

Song, T. M., Song, J., An, J. Y. and Jin, D. (2013). Multivariate analysis of factors for search on suicide using social big data. *Korean Journal of Health Education and Promotion*, **30**, 59-73.

Thun, M. J., Carter, B. D., Feskanich, D., Freedman, N. D., Prentice, R., Lopez, A. D., Hartge, P. and Gapstur, S. M. (2013). 50-year trends in smoking-related mortality in the United States. *New England Journal of Medicine*, **368**, 351-364.

World Health Organization. (2008). *Report on the global tobacco epidemic - The MPOWER package*, World Health Organization, Geneva.

Zheng, W., McLerran, D. F., Rolland, B. A., Fu, Z., Boffetta, P., He, J., Gupta, P. C., Ramadas, K., Tsugane, S., Irie, F., Tamakoshi, A., Gao, Y. T., Koh, W. P., Shu, X. O., Ozasa, K., Nishino, Y., Tsuji, I., Tanaka, H., Chen, C. J., Yuan, J. M., Ahn, Y. O., Yoo, K. Y., Ahsan, H., Pan, W. H., Qiao, Y. L., Gu, D., Pednekar, M. S., Sauvaget, C., Sawada, N., Sairenchi, T., Yang, G., Wang, R., Xiang, Y. B., Ohishi, W., Kakizaki, M., Watanabe, T., Oze, I., You, S. L., Sugawara, Y., Butler, L. M., Kim, D. H., Park, S. K., Parvez, F., Chuang, S. Y., Fan, J. H., Shen, C. Y., Chen, Y., Grant, E. J., Lee, J. E., Sinha, R., Matsuo, K., Thornquist, M., Inoue, M., Feng, Z., Kang, D. and Potter, J. D. (2014). Burden of total and cause-specific mortality related to tobacco smoking among adults aged 45 years in Asia: A pooled analysis of 21 cohorts. *Public Library of Science Medicine*, **11**, e1001631.

Journal of the Korean Data & Information Science Society
2015, **26**(5), 1047–1059

http://dx.doi.org/10.7465/jkdi.2015.26.5.1047
한국데이터정보과학회지

Predicting tobacco risk factors by using social big data[†]

Tae Min Song[1] · Juyoung Song[2] · Mi Kyung Cheon[3]

[13]Information and Statistics Department, Korea Institute for Health and Social Affairs
[2]Department of Criminal Justice, Pennsylvania State University

Received 5 August 2015, revised 1 September 2015, accepted 24 September 2015

Abstract

This study will predict risk factors associated with cigarettes in Korea by analyzing the social big data collected from the internet such as blogs, cafes, and SNSes in Korea, using data mining techniques. The key analysis results are as follows. First, when "raising cigarette price"is mentioned online, the negative group (i.e., the proportion of people holding negative views about smoking) increased from 58.6% to 74.8%, and when "lung cancer" is mentioned, it increased to 73.1%. Second, with regard to cigarettes in general, the positive group (i.e., the proportion of people holding positive views about smoking) decreased by 5.6% after the raising of cigarette prices, while the negative group increased by 6.1%. Third, when policies related to "FCTC, raising cigarette price, non-smoking laws, smoking regulations, non-smoking ads, and non-smoking business" are more frequently mentioned online, the positive group tended to decrease. Finally, when "non-smoking drugs, non-smoking patches, and non-smoking gums" are more frequently mentioned online, the positive group tended to decrease. However, when "electronic cigarettes and supplements" are more frequently mentioned online, the positive group increased.

Keywords: Association analysis, data mining, opinion mining, social big data, tobacco.

[†] Some of this research was published in Health and Social Welfare Issue & Focus : Song, T. M. (2015), Predicting tobacco risk factors by using social big data.

[1] Head, Research Center for Big Data, Information and Statistics Department, Korea Institute for Health and Social affairs, Sejong 339-007, Korea.

[2] Corresponding author: Assistant professor, Department of Criminal Justice, Pennsylvania State University, North Wales PA 19454, USA. E-mail: juyoung81@gmail.com

[3] Researcher, Information and Statistics Department, Korea Institute for Health and Social affairs, Sejong 339-007, Korea.