

연구보고서 2015-35

2015년 소셜 빅데이터 기반 보건복지 이슈 동향 분석



송태민 ·진달래

【책임연구자】

송태민 한국보건사회연구원 연구위원

【주요저서】

보건복지 빅데이터 효율적 관리방안 연구

한국보건사회연구원, 2014(공저)

빅데이터 분석 방법론

한나래아카데미, 2013(공저)

【공동연구진】

진달래 한국보건사회연구원 연구원

연구보고서 2015-35

2015년 소셜 빅데이터 기반 보건복지 이슈 동향 분석

발 행 일 2015년 12월 31일

저 자 송 태 민

발 행 인 김 상 호

발 행 처 한국보건사회연구원

주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1층~5층)

전 화 대표전화: 044)287-8000

홈페이지 <http://www.kihasa.re.kr>

등 록 1994년 7월 1일 (제8-142호)

인 쇄 처 (주)한디자인코퍼레이션

정 가 6,000원

© 한국보건사회연구원 2015
ISBN 978-89-6827-296-7 93510

발간사 <<

최근 스마트폰, 스마트TV, RFID, 센서 등의 급속한 보급과 모바일 인터넷과 소셜미디어의 확산으로 데이터량이 기하급수적으로 증가하고 데이터의 생산, 유통, 소비 체계에 큰 변화를 주면서 데이터가 경제적 자산이 될 수 있는 빅데이터 시대를 맞이하게 되었다. 세계 각국의 정부와 기업들은 빅데이터가 향후 국가와 기업의 성패를 가름할 새로운 경제적 가치의 원천이 될 것으로 기대하고 있으며, The Economist, Gartner, McKinsey 등은 빅데이터를 활용한 시장변동 예측과 신사업 발굴 등 경제적 가치창출 사례 및 효과를 제시하고 있다. 특히, 빅데이터는 미래 국가 경쟁력에도 큰 영향을 미칠 것으로 기대하여 국가별로는 안전을 위협하는 글로벌 요인이나 테러, 재난재해, 질병, 위기 등에 선제적으로 대응하기 위해 우선적으로 도입하고 있다.

한국은 최근 정부3.0과 창조경제의 추진과 실현을 위하여 현 정부의 주요 정책과제를 지원하기 위하여 다양한 분야에 빅데이터의 활용가치가 강조되고 있다. 빅데이터는 데이터의 형식이 다양하고, 방대할 뿐만 아니라 그 생성 속도가 매우 빨라 기존의 데이터를 처리하던 방식이 아닌 새로운 관리 및 분석방법을 요구한다. 또한, 트위터, 페이스북 등 소셜미디어에 남긴 정치, 경제, 사회, 문화에 대한 메시지는 그 시대의 감성과 정서를 파악할 수 있는 원천으로 등장함에 따라, 대중매체에 의해 수립된 정책의제는 이제 소셜미디어로부터 파악할 수 있으며, 개인이 주고받은 수많은 댓글과 소셜 로그정보는 공공정책을 위한 공공재로서 진화중에 있다.

이와 같이 많은 국가와 기업에서는 SNS를 통하여 생산되는 소셜 빅데이터의 활용과 분석을 통하여 새로운 경제적 효과와 일자리 창출은 물론 사회적 문제의 해결을 위하여 적극적으로 노력하고 있다.

이에 본 연구는 소셜 빅데이터를 기반으로 우리나라 보건복지 주요 이슈에 대한 동향을 분석하기 위한 것으로 2015년의 주요 이슈로는 메르스, 담배, 통일, 청소년 음란물 유통의 4개의 주제에 대해 소셜 빅데이터를 수집하여 동향분석과 수요예측 및 위험예측 모형을 제시하였다. 이와 같은 연구는 보건복지 주요 이슈를 적시에 분석하고 수요를 예측할 수 있다는 점에서 정책적·분석방법론적으로 의의가 있으며, 실제로도 내용을 보다 빠르게 효과적으로 파악하여 사회조사가 지닌 한계를 보완할 수 있는 새로운 조사방법으로서의 소셜 빅데이터 가치를 확인하였다는 점에서 조사방법론적 의의를 가진다고 할 수 있다. 본 연구에 귀중한 조언을 아끼지 않으신 많은 전문가 분들과 주요 이슈에 대한 논문 집필에 참여해주신 교수님들께 감사드린다. 끝으로 본 보고서에 수록된 모든 내용은 우리 연구원의 공식적인 견해는 아니며 어디까지나 참여한 연구진의 의견임을 밝힌다.

2015년 12월

한국보건사회연구원장

김 상 호

목 차

Abstract	1
요 약	3
제1장 소셜 빅데이터 분석과 활용 방안	9
1. 서론	11
2. 빅데이터 개요	12
3. 소셜 빅데이터 분석 방법	15
4. 소셜 빅데이터 수집 및 분류 방법	18
참고문헌	23
제2장 소셜 빅데이터 분석 기반 메르스 감정 위험 예측	25
1. 서론	27
2. 이론적 배경	29
3. 연구 방법	35
4. 분석 결과	40
5. 분석 결과 요약	54
참고문헌	57
제3장 소셜 빅데이터를 활용한 담배 위험 예측	61
1. 서론	63
2. 연구 방법	66
3. 연구 결과	72



4. 결론	84
참고문헌	87
제4장 소셜 빅데이터를 활용한 통일인식 동향 분석 및 예측	91
1. 서론	93
2. 연구 방법	95
3. 연구 결과	101
4. 결론	118
참고문헌	125
제5장 소셜 빅데이터를 활용한 한국의 섹스팅(Sexting) 위험 예측 ...	127
1. 서론	129
2. 이론적 배경	131
3. 연구 방법	133
4. 분석 방법	137
5. 연구 결과	138
6. 결론 및 고찰	154
참고문헌	159

표 목차

〈표 1- 1〉 '빅데이터 개인정보보호 가이드라인'의 주요내용	15
〈표 1- 2〉 우울 온톨로지 분류에 따른 영역수준	20
〈표 2- 1〉 메르스 부정(불안) 감정 예측	43
〈표 2- 2〉 메르스 관련 버즈 현황	46
〈표 2- 3〉 예방요인, 대처/치료요인, 증상요인에 대한 메르스 감정 예측	47
〈표 2- 4〉 메르스의 감정에 영향을 미치는 요인	49
〈표 2- 5〉 메르스 관련 증상요인의 예측 모형에 대한 이익도표	53
〈표 3- 1〉 담배 감정 키워드 연관성 예측	74
〈표 3- 2〉 담배 관련 버즈 현황	76
〈표 3- 3〉 담배 관련 연도별 감정변화	77
〈표 3- 4〉 정책요인에 대한 담배 위험 예측	78
〈표 3- 5〉 담배의 위험에 영향을 미치는 정책 및 도구 요인	79
〈표 3- 6〉 정책요인의 예측 모형에 대한 이익도표	81
〈표 3- 7〉 질병요인의 예측 모형에 대한 이익도표	83
〈표 4- 1〉 통일인식의 감정 키워드 연관성 예	103
〈표 4- 2〉 통일 관련 문서(버즈) 현황	106
〈표 4- 3〉 연도별 통일 관련 국민인식	107
〈표 4- 4〉 미국대사 피습사건 전후 국민 통일인식 변화	107
〈표 4- 5〉 통일관련 안보이슈의 국가별 버즈 현황	108
〈표 4- 6〉 통일관련 안보이슈의 기관별 버즈 현황	109
〈표 4- 7〉 안보와 이슈 요인에 대한 통일인식의 연관성 예측	111
〈표 4- 8〉 주변국가(4국)의 통일인식의 연관성 예측	112
〈표 4- 9〉 통일인식에 영향을 미치는 안보이슈 요인	114
〈표 4-10〉 통일인식 관련 안보이슈 요인의 예측 모형에 대한 이익도표	116
〈표 4-11〉 통일인식 관련 국가 요인의 예측 모형에 대한 이익도표(주변 4국)	118
〈표 5- 1〉 섹스팅의 위험 감정 키워드 연관성 예측	141

〈표 5- 2〉 섹스팅 관련 버즈 현황	143
〈표 5- 3〉 유형과 내용 요인에 대한 섹스팅 위험 예측	145
〈표 5- 4〉 섹스팅에 영향을 요인	147
〈표 5- 5〉 도움요인의 섹스팅 위험 예측 모형에 대한 이익도표	150
〈표 5- 6〉 내용 요인의 섹스팅 위험 예측 모형에 대한 이익도표	151
〈표 5- 7〉 유형 요인의 섹스팅 위험 예측 모형에 대한 이익도표	153

그림 목차

[그림 1-1] 빅데이터의 특성과 정부 3.0 추진 전략	13
[그림 1-2] 소셜 빅데이터 분석 절차 및 방법(메르스 버즈 분석사례)	16
[그림 1-3] 빅데이터 분석기반의 위기청소년 예측 및 적시대응 기술 개발 연계사례	17
[그림 1-4] 청소년 우울관리 온톨로지	19
[그림 1-5] 메르스 분류(범주화, 유목화)체계	21
[그림 2-1] 세 과정 사이에서의 정보확산	34
[그림 2-2] 메르스 관련 문서(버즈)량의 일별 추이	41
[그림 2-3] 메르스에 대한 일자별 감정(상위 50개) 변화	42
[그림 2-4] 메르스 감정의 연관규칙에 대한 병렬좌표와 그래프 시각화	44
[그림 2-5] 지역별 메르스 위험(불안) 감정	44
[그림 2-6] 메르스 관련 예방 요인의 예측모형	51
[그림 2-7] 메르스 관련 증상 요인의 예측모형	52
[그림 3-1] 담배관련 시간별 및 요일별 버즈 현황	72
[그림 3-2] 연도별 담배 감정 변화	73
[그림 3-3] 담배 감정의 연관규칙에 대한 병렬좌표 시각화	75
[그림 3-4] 정책요인의 예측모형	80
[그림 3-5] 질병요인의 예측모형	83
[그림 4-1] 통일 관련 문서(버즈)량 일별 추이	102

[그림 4-2] 통일 연도별 감정 변화	102
[그림 4-3] 통일인식 감정의 연관규칙에 대한 병렬좌표 시각화	104
[그림 4-4] 통일 감정 응집구조 분석	110
[그림 4-5] 통일인식 관련 안보이슈 요인의 예측모형	115
[그림 4-6] 통일인식 관련 주요 주변 국가 예측모형	117
[그림 5-1] 섹스팅 관련 시간별 및 요일별 버즈 현황	139
[그림 5-2] 연도별 섹스팅 감정 변화	140
[그림 5-3] 지역별 섹스팅 감정(일반, 위험)	142
[그림 5-4] 섹스팅의 내용유형 및 폐해도움 간 외부 근접중심성	145
[그림 5-5] 도움요인의 섹스팅 위험 예측모형	149
[그림 5-6] 내용 요인의 섹스팅 위험 예측모형	151
[그림 5-7] 유형 요인의 섹스팅 위험 예측 모형	153

Abstract <<

Social big data trend analysis based on health and welfare issues in 2015

Big data consists of various forms of data in large volumes, which are rapidly created. Therefore, it requires a new management and analysis methodology. Moreover, as social media platforms emerge as the source of information about the feelings and sentiments of the current times with messages on politics, economics, society, and culture, policy agendas set up in public spheres can be identified from social media. Countless comments between individuals and society evolve as a log of information, which continues to evolve as an asset for public policy¹⁾. As such, many nations and businesses actively strive for new economic effects, job creation, as well as solving social problems through the utilization and analysis of social big data created through social networking services (SNS). This study proposes study methodologies and utilization strategies of social big data that can create value and predict the future by gathering and analyzing social big data from various fields.

1) Song YJ (2012) Age of Big Data! The Evolution of SNS and Public Policy. National Information Society Agency.

1. 연구의 배경 및 목적

빅데이터는 미래 국가 경쟁력에도 큰 영향을 미칠 것으로 기대하여 국가별로는 안전을 위협하는 글로벌 요인이나 테러, 재난재해, 질병, 위기 등에 선제적으로 대응하기 위해 우선적으로 도입하고 있다. 빅데이터는 데이터의 형식이 다양하고, 방대할 뿐만 아니라 그 생성 속도가 매우 빨라 기존의 데이터를 처리하던 방식이 아닌 새로운 관리 및 분석방법을 요구한다. 또한, 트위터, 페이스북 등 소셜미디어에 남긴 정치, 경제, 사회, 문화에 대한 메시지는 그 시대의 감성과 정서를 파악할 수 있는 원천으로 등장함에 따라, 대중매체에 의해 수립된 정책의제는 이제 소셜미디어로부터 파악할 수 있으며, 개인이 주고받은 수많은 댓글과 소셜 로그정보는 공공정책을 위한 공공재로서 진화중에 있다. 이와 같이 많은 국가와 기업에서는 SNS를 통하여 생산되는 소셜 빅데이터의 활용과 분석을 통하여 새로운 경제적 효과와 일자리 창출은 물론 사회적 문제의 해결을 위하여 적극적으로 노력하고 있다.

기존에 실시하던 횡단적 조사나 종단적 조사 등을 대상으로 한 연구는 정해진 변인들에 대한 개인과 집단의 관계를 보는 데에는 유용하나 사이버 상에서 언급된 개인별 온라인 문서에서 논의된 관련 정보 상호 간의 연관관계를 밝히고 원인을 파악하는 데는 한계가 있다. 이에 반해 소셜 빅데이터의 분석은 훨씬 방대한 량의 데이터를 활용하여 다양한 참여자의 생각과 의견을 확인할 수 있기 때문에 사회적 문제의 예측과 현상에 대한 복잡한 연관관계를 보다 정확하게 밝혀낼 수 있다. 본 연구에서는

다양한 분야의 소셜 빅데이터를 수집·분석하여 가치를 창출하고 미래를 예측할 수 있는 소셜 빅데이터 연구방법과 활용방안을 제시하였다. 그리고 2015년의 주요 이슈(메르스, 담배, 통일, 청소년 음란물 유통)에 대한 소셜 빅데이터를 수집하여 동향분석과 수요예측 및 위험예측 모형을 제시하였다.

2. 주요 연구결과

제1장 ‘서론’에서는 소셜 빅데이터 분석 절차 및 분류 방법에 대해 설명하였다. 소셜 빅데이터의 분석은 해당 주제와 관련한 문서를 분석 모터를 통해 수집대상과 수집범위를 설정한 후, 대상채널(뉴스·블로그·카페·게시판·SNS 등)에서 크롤러 등 수집엔진(로봇)을 이용하여 수집한다. 이때 불용어를 지정하여 수집의 오류를 방지하고 관련 연관 키워드 그룹을 지정한다. 수집한 비정형 데이터를 텍스트 마이닝, 오피니언 마이닝을 통하여 분류하고 정제하는 절차가 필요하다. 정제된 비정형 데이터 분석은 버즈 분석, 키워드분석, 감성분석, 계정분석 등으로 진행한다. 소셜 빅데이터의 수집 및 분류는 해당 토픽에 대한 이론적 배경 등을 분석하여 온톨로지를 개발한 후, 온톨로지의 키워드를 수집하여 분류하는 Top-down 방법과 해당 토픽을 웹크롤로 수집한 후 범용 사전이나 사용자 사전으로 분류하는 Bottom-up 방법이 있다.

제2장 ‘소셜 빅데이터 기반 메르스 감정 위험예측’에서는 우리나라 온라인 뉴스사이트, 블로그, 카페, SNS, 게시판 등에서 수집한 소셜 빅데이터를 바탕으로 우리나라에서 발생한 메르스 정보 확산 위험요인을 예측하였다. 메르스 관련 버즈는 2015년 5월 28일 급속히 증가하여 5월 30일 감소하였다가 6월 1일 이후 메르스 추가 환자 발생과 사망자 발생 보

도 후, 급속히 증가한 것으로 나타났다. 메르스에 대한 부정적 감정(불안)의 표현 단어는 ‘무시, 한심, 판단’, ‘거부, 비난, 무능, 불구’, ‘답답, 공포, 스트레스’, ‘무책임, 비판, 실패’, ‘괴담, 협박’ 키워드와 강하게 연결되어 있는 것으로 나타나, 이는 정부의 초기 대응 미흡에 대한 국민의 실망감과 SNS를 통한 메르스 괴담의 급속한 전파로 인한 불안감이 표출된 것으로 보인다.

제3장 ‘소셜 빅데이터를 활용한 담배 위험 예측’에서는 우리나라 온라인 뉴스사이트, 블로그, 카페, SNS, 게시판 등에서 수집한 소셜 빅데이터를 바탕으로 우리나라 국민의 담배에 대한 위험 예측모형과 연관규칙을 파악하였다. 담뱃값 인상 이후 위험군은 5.6% 감소하고, 일반군은 6.1% 증가한 것으로 나타났다. 온라인 버즈에서 담뱃값 인상, 금연관련법이 동시에 언급되면 일반군이 될 확률이 증가하며, 담뱃값 인상만 언급되어도 위험군을 감소시키는 것으로 나타났다.

제4장 ‘소셜 빅데이터를 활용한 통일인식 동향 분석 및 예측’에서는 우리나라 온라인 뉴스사이트, 블로그, 카페, SNS, 게시판 등에서 수집한 소셜 빅데이터를 바탕으로 우리나라 국민의 통일의 인식에 대한 동향을 분석하고 통일인식의 예측모형과 연관규칙을 파악하였다. 통일의 필요성에 대한 국민인식은 통일대박 감정을 포함할 경우 찬성은 2011년 55.0%, 2012년 62.2%, 2013년 57.8%, 2014년 77.1%, 2015년 56.2%로 나타났다. 안보·이슈 요인이 통일인식에 미치는 영향은 통일대박의 영향력이 가장 큰 것으로 나타나 온라인 문서 중, 통일대박이 있을 경우 통일에 대한 찬성의 인식이 이전의 68.8%에서 85.8%로 증가하였고, 특히 통일대박이 있고, 핵무기와 휴전선이 없는 경우 통일에 대해 찬성하는 확률이 가장 높은 것으로 나타났다.

제5장 ‘소셜 빅데이터를 활용한 한국의 섹스팅 위험 예측’에서는 우리

나라 온라인 뉴스사이트, 블로그, 카페, SNS, 게시판 등에서 수집한 소셜 빅데이터를 바탕으로 우리나라 섹스팅에 대한 위험요인을 예측하였다. 우리나라 섹스팅에 대한 긍정적 감정(위험)을 나타내는 버즈는 38.3%로 나타났다. 섹스팅의 영향은 윤리의식, 대인관계, 성욕, 건강, 공부, 비용 순으로 위험한 것으로 나타났으며, 유통방식은 수요보다 공유의 위험이 더 큰 것으로 나타났다.

3. 결론 및 시사점

소셜 빅데이터는 다양한 분야에 활용할 수 있다.

첫째, 조사를 통한 기존의 정보수집 체계의 한계를 보완할 수 있는 새로운 자료수집 방법으로 활용할 수 있다. 국민의 통일에 대한 인식조사, 정부의 금연정책(가격정책, 비가격정책 등) 실시 이후 흡연실태 조사, 스마트폰 및 인터넷 중독 실태 조사 등 다양한 분야의 조사에 활용할 수 있다. 둘째, 보건복지 정책 수요를 예측(저출산 정책 수요 예측 등)할 수 있다. 새정부 출범이후 건강보험보장성 강화에 대한 국민의 요구가 커지고 인구고령화와 저출산이 사회적 문제로 대두됨에 따라 대상자별·분야별로 다양한 보건복지정책이 요구됨에 따라 오프라인 보건복지 욕구 조사와 더불어 소셜 미디어에 남긴 다양한 정책 의제를 분석하여 수요를 파악해야 한다. 셋째, 사회적 위기상황에 대한 모니터링과 예측으로 위험에 대한 사전 대응체계를 구축할 수 있다. 따라서 청소년의 자살과 사이버폭력 대응체계 구축, 질병에 대한 위험 예측, 식품안전 모니터링 등에 활용할 수 있다. 넷째, 새로운 기술에 대한 동향을 파악할 수 있다. 빅데이터, 사물인터넷, 인공지능 등 새로운 기술에 대한 수요자와 공급자가 요구하는 기술 동향을 파악할 수 있다. 끝으로 정부와 공공기관이 보유·관리하

고 있는 빅데이터는 통합방안보다는 각각의 빅데이터의 집단별 특성을 분석하여 위험(또는 수요) 집단 간 연계를 통한 예측(위험예측 또는 질병 예측 등) 서비스가 제공되어야 할 것이다. 즉, 빅데이터 분석을 통한 개인별 맞춤형 서비스는 프라이버시를 침해할 수 있기 때문에 위험(또는 수요)집단별 맞춤형 서비스가 제공되어야 할 것이다. 또한, 빅 데이터를 분석하여 인과성을 발견하고 미래를 예측하기 위해서는 정부차원의 데이터 사이언티스터 양성을 위한 노력이 필요할 것으로 본다.

*주요용어: 소셜빅데이터, 보건복지, 데이터마이닝, 다변량분석, 시각화, R

제 1 장

소셜 빅데이터 분석과 활용 방안

1. 서론
 2. 빅데이터 개요
 3. 소셜 빅데이터 분석 방법
 4. 소셜 빅데이터 수집 및 분류 방법
- 참고문헌

1

소셜 빅데이터 분석과 << 활용 방안

1. 서론

최근 스마트폰, 스마트TV, RFID, 센서 등의 급속한 보급과 모바일 인터넷과 소셜미디어의 확산으로 데이터량이 기하급수적으로 증가하고 데이터의 생산, 유통, 소비 체계에 큰 변화를 주면서 데이터가 경제적 자산이 될 수 있는 빅데이터 시대를 맞이하게 되었다(송태민, 2012). 세계 각국의 정부와 기업들은 빅데이터가 향후 국가와 기업의 성패를 가름할 새로운 경제적 가치의 원천이 될 것으로 기대하고 있으며, The Economist, Gartner, McKinsey 등은 빅데이터를 활용한 시장변동 예측과 신사업 발굴 등 경제적 가치창출 사례 및 효과를 제시하고 있다. 특히, 빅데이터는 미래 국가 경쟁력에도 큰 영향을 미칠 것으로 기대하여 국가별로는 안전을 위협하는 글로벌 요인이나 테러, 재난재해, 질병, 위기 등에 선제적으로 대응하기 위해 우선적으로 도입하고 있다.

빅데이터는 데이터의 형식이 다양하고, 방대할 뿐만 아니라 그 생성 속도가 매우 빨라 기존의 데이터를 처리하던 방식이 아닌 새로운 관리 및 분석방법을 요구한다. 또한, 트위터, 페이스북 등 소셜미디어에 남긴 정치, 경제, 사회, 문화에 대한 메시지는 그 시대의 감성과 정서를 파악할 수 있는 원천으로 등장함에 따라, 대중매체에 의해 수립된 정책의제는 이제 소셜미디어로부터 파악할 수 있으며, 개인이 주고받은 수많은 댓글과 소셜 로그정보는 공공정책을 위한 공공재로서 진화중에 있다(송태민, 2012). 이와 같이 많은 국가와 기업에서는 SNS를 통하여 생산되는 소셜

빅데이터의 활용과 분석을 통하여 새로운 경제적 효과와 일자리 창출은 물론 사회적 문제의 해결을 위하여 적극적으로 노력하고 있다.

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)에서 논의된 관련 정보 상호간의 연관관계를 밝히고 원인을 파악하는 데는 한계가 있다(송주영, 송태민, 2014). 이에 반해 소셜 빅데이터의 분석은 훨씬 방대한 량의 데이터를 활용하여 다양한 참여자의 생각과 의견을 확인할 수 있기 때문에 사회적 문제의 예측과 현상에 대한 복잡한 연관관계를 보다 정확하게 밝혀낼 수 있다. 본고는 다양한 분야의 소셜 빅데이터를 수집·분석하여 가치를 창출하고 미래를 예측할 수 있는 소셜 빅데이터 연구방법과 활용방안을 제시코자 한다.

2. 빅데이터 개요

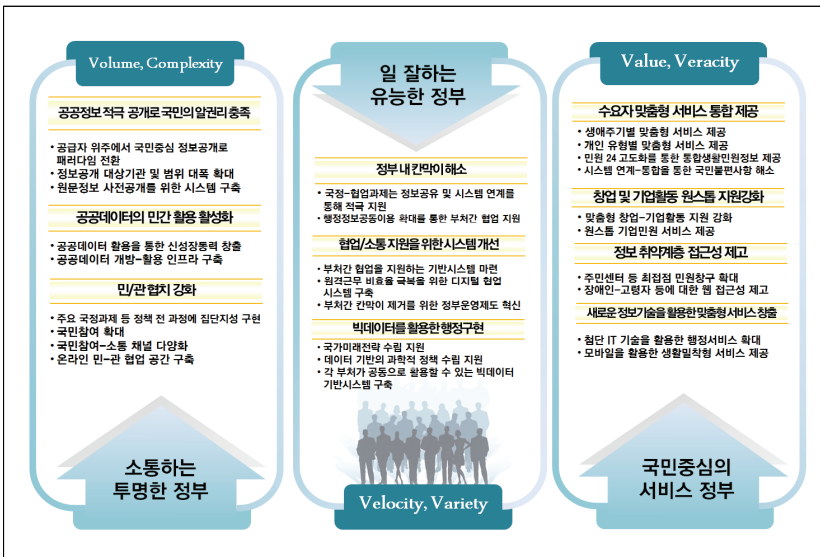
가. 빅데이터 정의

빅데이터(Big Data)는 Wikipedia(2015. 8. 5.)에서 '기존 데이터베이스 관리도구로 데이터를 수집·저장·관리·분석의 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술'로 정의하고 있다. Gartner(2012)는 더 나은 의사결정, 시사점 발견 및 프로세스 최적화를 위해 사용되는 새로운 형태의 정보처리가 필요한 대용량, 초고속 및 다양성의 특성을 가진 정보자산으로 정의하고 있으며, McKinsey(2011)는 일반적인 데이터베이스 소프트웨어 도구가 수집, 저장, 관리, 분석하기 어려운 대규모의 데이터로

정의하고 있다. 이와 같은 정의를 살펴볼 때 빅데이터란 엄청나게 많은 데이터로 양적인 의미를 벗어나 데이터 분석과 활용을 포괄하는 개념으로 사용되고 있다(송태민, 2012).

우리나라는 정부3.0의 효과적인 추진과 생애주기별 맞춤형 서비스 및 국민 행복 실현을 위하여 정부 차원의 빅데이터 추진방안이 마련되었다. 빅데이터의 주요 특성은 일반적으로 3V(Volume, Variety, Velocity)를 기본으로 2V(Value, Veracity)나 1C(Complexity)의 특성을 추가하여 설명하고 있다(그림 1-1 참조). 특히, 보건복지 분야에서는 국민의 생명과 직결되는 정보를 다루고 있어 빅데이터에 있어 가치(Value)와 신뢰성(Veracity)은 매우 중요하다고 할 수 있다.

[그림 1-1] 빅데이터의 특성과 정부 3.0 추진 전략



나. 빅데이터 개인정보보호 방안

현재 어느 나라를 막론하고 빅데이터의 활용에 있어 가장 큰 과제는 개인의 사생활 비밀보호 및 개인정보보호이다(송태민 외, 2014). 개인정보 보호에 중점을 두면 빅데이터의 활용을 저해하게 될 우려가 있다. 개인정보보호법의 목적이 ‘개인정보의 수집·유출·오용·남용으로부터 사생활의 비밀 등을 보호함으로써 국민의 권리와 이익을 증진하고’로 되어있지만 개인정보와 비개인정보를 명확히 구분하기가 어렵고 비즈니스에 있어 자동적으로 수집되는 데이터가 비개인정보라고 할지라도 프라이버시를 침해할 가능성이 있다. 특히 소셜미디어에 공개된 개인정보는 위변조와 오남용이 쉽고 상업적 이용을 위한 정보수집 등에 노출이 될 수 있기 때문에 프라이버시 침해 등의 문제가 발생할 가능성이 매우 높다. 방송통신위원회는 2013년 12월 18일 ‘빅데이터 개인정보보호 토론회’와 2014년 3월 19일 ‘온라인 개인정보보호 세미나’를 통해 의견을 수렴하고 2014년 12월 23일 ‘빅데이터 개인정보보호 가이드라인’을 발표하였다. 가이드라인의 주요 내용은 <표 1-1>과 같이 빅데이터 수집 시부터 개인식별 정보에 대한 철저한 비식별화 조치와 개인의 사상·신념, 정치적 견해 등 민감 정보에 대한 조합·분석 등 처리 금지 등에 관한 것이다. 빅데이터로부터 개인을 보호하기 위해 가장 중요한 것은 특정 개인을 식별하지 못하도록 하는 익명화와 정보접근 및 정보처리에 대한 통제다. 그러나 정보접근 및 정보처리에 대한 통제를 강하게 하면 정보활용을 활성화 할 수 없기 때문에 빅데이터의 ‘활용과 보호의 균형’에 대한 효과적인 정책이 우선적으로 마련되어야 할 것이다(송태민, 2013).

〈표 1-1〉 ‘빅데이터 개인정보보호 가이드라인’의 주요내용

구분	내용
비식별화 조치	수집 시부터 개인식별 정보에 대한 철저한 비식별화 조치 (제3조·제4조·제5조·제10조) 개인정보가 포함된 공개된 정보 및 이용내역정보는 비식별화 조치를 취한 후 수집·저장·조합·분석 및 제3자 제공 등 가능
공개로 통한 투명성 확보	빅데이터 처리 사실·목적 등의 공개를 통한 투명성 확보 (제4조·제5조·제9조) - 개인정보 취급방침을 통해 비식별화 조치후 빅데이터 처리 사실·목적·수집 출처 및 정보 활용 거부권 행사 방법 등을 이용자에게 투명하게 공개 • (개인정보 취급방침) 비식별화 조치후 빅데이터 처리 사실·목적 등을 이용자 등에게 공개하고 ‘정보 활용 거부 페이지 링크’를 제공하여 이용자가 거부권을 행사 할수 있도록 조치 • (수집 출처 고지) 이용자 이외의 자료부터 수집한 개인정보 처리 시 ‘수집·출처·목적, 개인 정보 처리 정지 요구권’을 이용자에게 고지
재식별시, 비식별화 조치	개인정보 재식별시, 즉시 파기 및 비식별화 조치(제3조·제6조) 빅데이터 처리 과정 및 생성정보에 개인정보가 재식별 될 경우, 즉시 파기하거나 추가적인 비식별화 조치토록 함
민감정보의 처리 금지	민감정보 및 통신비밀의 수집·이용·분석 등 처리 금지 (제7조·제8조) 특정 개인의 사상·신념, 정치적 견해 등 민감정보의 생성을 목적으로 정보의 수집·이용·저장·조합·분석 등 처리 금지 이메일, 문자메시지 등 통신 내용의 수집·이용·저장·조합·분석 등 처리 금지
기술적·관리적 보호조치	수집된 정보의 저장·관리 시 ‘기술적·관리적 보호조치’ 시행 (제3조·제2항) 비식별화 조치가 취해진 정보를 저장·관리하고 있는 정보 처리시스템에 대한 기술적·관리적 보호조치 적용 ※ (보호조치) 침입차단시스템 등 접근 통제장치 설치, 접속 기록에 대한 위·변조 방지조치 백신 소프트웨어 설치·운영 등 악성프로그램에 의한 침해 방지 조치

3. 소셜 빅데이터 분석 방법

소셜 빅데이터 분석 절차 및 방법은 [그림 1-2]와 같다. 첫째, 해당 주

제와 관련한 문서(메르스)를 분석 모델링을 통해 수집대상과 수집범위를 설정한 후, 대상채널(뉴스·블로그·카페·게시판·SNS 등)에서 크롤러 등 수집엔진(로봇)을 이용하여 수집한다. 이때 불용어(메르스벤츠, 메르스테스벤츠)를 지정하여 수집의 오류를 방지하고 메르스 관련 연관 키워드 그룹(메르스 바이러스, 중동 호흡기 증후군, 메르스 코로나 바이러스, 매르스)을 지정한다. 둘째, 수집된 메르스 원데이터(raw data)는 텍스트 형태의 비정형 데이터로 연구자가 수집된 원상태로 분석하기에는 어려움이 있다. 따라서 수집한 비정형 데이터를 텍스트 마이닝, 오피니언 마이닝을 통하여 분류하고 정제하는 절차가 필요하다. 정제된 비정형 데이터 분석은 버즈 분석, 키워드분석, 감성분석, 계정분석 등으로 진행한다.

[그림 1-2] 소셜 빅데이터 분석 절차 및 방법(메르스 버즈 분석사례)



셋째, 비정형 빅데이터를 정형 빅데이터로 변환해야 한다. 메르스 관련 주제 분석 사례를 살펴보면, 메르스 버즈 각각의 문서는 ID로 코드화하여야 하고, 버즈 내에서 발생하는 키워드는 모두 코드화하여야 한다. 넷째, 사회현상과 연계하여 분석하기 위해서는 정형화된 빅데이터를 오피라인

통계(조사) 자료와 연계해야 한다. 오프라인 통계(조사) 자료는 대부분 정부나 공공기관에서 유료 또는 무료로 제공하기 때문에, 연계대상 자료와 함께 연계 가능한 식별자(일별·월별·연별·지역별)를 확인한 후 오프라인 자료를 수집하여 연계(link)할 수 있다. 다섯째, 오프라인 통계(조사) 자료와 연계된 정형화된 빅데이터의 분석은 요인 간의 인과관계나 시간별 변화궤적을 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 데이터마ining 분석이나 시각화를 실시할 수 있다.

빅데이터 연계방법[big data linkage(matching)]으로는 정확매칭(exact matching)과 통계적매칭(statistical matching)이 있다. 정확매칭은 고유식별 정보가 존재할 때 사용하며, 통계적매칭은 고유식별 정보가 존재하지 않기 때문에 유사한 개체를 찾아 상호 데이터를 결합시킬 때 사용한다. 소셜 빅데이터와 공공 빅데이터의 연계는 시간 변수와 지역 변수 등을 고유식별 정보로 하여 상호 매칭하는 정확매칭 방법을 활용할 수 있다(그림 1-3 참조).

[그림 1-3] 빅데이터 분석기반의 위기청소년 예측 및 적시대응 기술 개발 연계사례

소셜빅데이터	1388정형빅데이터	최종연계빅데이터
일자	일자	일자
성적	가출	성적
성	가정폭력	성
우울	학업중단	우울
경제	학교폭력	경제
질병	성	질병
외모	흡연을주	외모
자살위험	자살	자살위험
	인터넷중독	가출
	은둔형	가정폭력
	비행	학업중단
		학교폭력
		성
		흡연을주
		자살
		인터넷중독
		은둔형
		비행

4. 소셜 빅데이터 수집 및 분류 방법

소셜 빅데이터의 수집 및 분류는 해당 토픽에 대한 이론적 배경 등을 분석하여 온톨로지(ontology)를 개발한 후, 온톨로지의 키워드를 수집하여 분류하는 Top-down 방법과 해당 토픽을 웹크롤로 수집한 후 범용 사전이나 사용자 사전으로 분류(유목화 또는 범주화)하는 Bottom-up 방법이 있다.

가. Top-down 방법²⁾

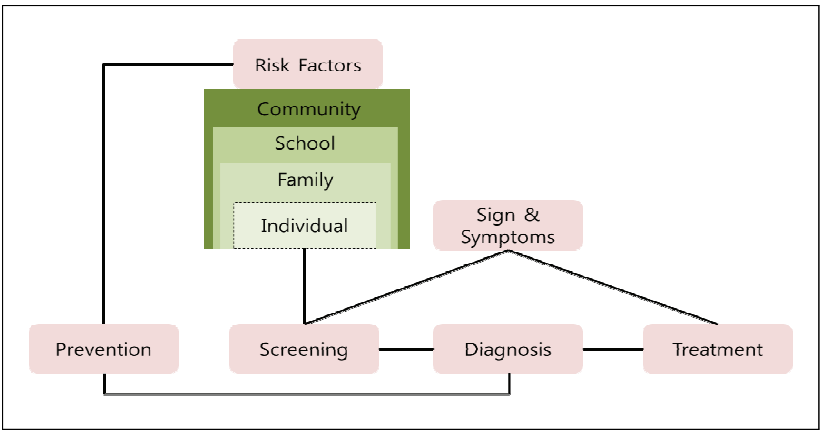
소셜 미디어에서 표현되는 언어들은 주로 사람들이 일상 대화에서 쓰이는 구어체 문장으로 이루어진 비정형 데이터이기 때문에(노진석, 2012), 이를 보다 효과적으로 수집 및 분석하기 위한 분석틀이 필요하다. 분석틀을 내용은 관련 주제가 어떤 개념 영역들로 구성되어 있는지와 각 개념 간 관계에 대한 정의가 필요하기 때문에, 이를 반영한 온톨로지(ontology)가 개발될 필요가 있다. 온톨로지(ontology)는 관심 주제의 공유된 개념(shared concepts)을 형식화하고(formalizing) 표현하기 위한(representing), 컴퓨터가 해석 가능한 지식 모델(computer-interpretable knowledge model)이다(Kim HY 외, 2013). 수집되는 소셜 빅데이터 자료는 비정형적으로 다양하게 표현됨으로써 온톨로지를 구성하는 개념을 설명하는 용어와 그 유의어를 정의하여 기술하여 용어체계를 마련하는 것이 필요하다. 본 연구는 온라인 상의 청소년 우울 주제

2) 본 절의 내용은 '송태민 외(2015). 빅데이터 분석 기반의 위기청소년 예측 및 적시 대응 기술 개발'의 일환으로 우울 빅데이터 수집을 위해 서울대학교 간호대학 박현애교수 연구팀과 공동으로 수행되었으며, '정혜실(2015). 청소년 우울관련 소셜 빅데이터 수집과 분석을 위한 온톨로지 개발 및 평가. 서울대학교 대학원 석사학위논문' 발표예정인 연구결과를 참고하였음 밝힌다.

에 대해 수집된 빅데이터 자료를 식별하고 활용하기 위한 분석틀로서, 우울 관리 관련 주제를 분류하고, 우울 관리 온톨로지와 용어체계를 개발하는 것이다.

청소년 우울증을 중심으로 살펴본 정신건강관리 주제분류는 위험요인, 증상 및 징후, 스크리닝, 진단, 치료 및 예방으로 총 6개 영역이 도출되었고 이들 영역의 관계를 그림으로 나타내면 [그림 1-4]와 같다.

[그림 1-4] 청소년 우울관리 온톨로지



온톨로지 개발은 우울 관리 주제를 설명하는 분류틀에 해당하는 용어에 대하여, ‘대분류-중분류-소분류’의 각 영역 수준별로 용어를 추출하여 영역 수준별로 제시해야 한다. 따라서 <표 1-2>와 같이 각 용어별로 인터넷 검색과 선행 문헌 검색 등의 방법을 이용하여 동의어와 유사어를 정의해야 한다.

〈표 1-2〉 우울 온톨로지 분류에 따른 영역수준

대분류	중분류	소분류1	소분류2	소분류3	동익어·유익어	수준
위험요인						대분류(위험요인)
	대상자 특성요인				personal factor	대분류(위험요인) > 중분류(대상자특성 요인)
		인구사회학적 요인				대분류(위험요인) > 중분류(대상자특성 요인) > 소분류1(인구사회학적 요인)
			인구학적 특성		인구통계학적 특성	대분류(위험요인) > 중분류(대상자특성 요인) > 소분류1(인구사회학적 요인) > 소분류2 (인구학적 특성)
				성별	성, gender	대분류(위험요인) > 중분류(대상자특성 요인) > 소분류1(인구사회학적 요인) > 소분류2 (인구학적 특성) > 소분류3(성별)
				연령	나이, 발달단계, stage of puberty	대분류(위험요인) > 중분류(대상자특성 요인) > 소분류1(인구사회학적 요인) > 소분류2 (인구학적 특성) > 소분류3(연령)
				재학 여부	학교다님, 학교중퇴	대분류(위험요인) > 중분류(대상자특성 요인) > 소분류1(인구사회학적 요인) > 소분류2 (인구학적 특성) > 소분류3(재학여부)
...	...	...	...	...	...	이 외 분류를 포함할 수 있음.
증상 및 징후					emotional change	대분류(증상 및 징후)
	감정변화					대분류(증상 및 징후) > 중분류(감정변화)
		불안			걱정, 공포, 불안정, 조마조마, 뒤숭숭	대분류(증상 및 징후) > 중분류(감정변화) > 소분류1(불안)
		우울			공허함, 외로움, 소외감, 침울, 고독	대분류(증상 및 징후) > 중분류(감정변화) > 소분류1(우울)
		슬픔			불행, 울음, 비통	대분류(증상 및 징후) > 중분류(감정변화) > 소분류1(슬픔)
...	...	...	...	...	...	...

나. Bottom-up 방법

소셜 빅데이터를 수집 분류하기 위해서는 ‘21 세기 세종계획’과 같은 범용사전이 있지만 대부분 분석 목적에 맞게 사용자가 설계한 사전을 사용한다. 예를 들면, 메르스 정보확산 위험 예측을 위하여 소셜 빅데이터를 수집한다고 하면, 웹크롤의 수집조건으로 메르스 토픽(topic)은 모든 관련 문서를 수집하기 위해 ‘메르스’를 사용하며, 토픽과 같은 의미로 사

용되는 토픽 유사어로는 ‘메르스 바이러스, 중동 호흡기 증후군, 메르스 코로나 바이러스, 매르스’ 용어를 사용해야 한다. 그리고, 불용어는 ‘메르스벤츠, 메르스데스벤츠’를 사용해야 한다. 수집 가능한 채널[(메르스 키워드의 수집 가능채널은 105개의 온라인 뉴스사이트, 4개의 블로그, 2개의 카페, 1개의 SNS(트위트), 8개의 게시판 등 총 120개의 온라인 채널]에서 수집된 메르스 온라인 문서는 범용사전이나 사용자 사전을 이용하여 [그림 1-5]와 같이 유목화(범주화)한 후, 해당 키워드의 출현 유무를 확인하여 정형화 빅데이터로 변환해야 한다.

[그림 1-5] 메르스 분류(범주화, 유목화)체계

[illegible]

수집된 소셜 빅데이터의 분류 및 변환(정형 빅데이터 변환)이 완료된 후, 분류된 키워드에 대해 감성분석을 실시하여 요인을 추출(변수 축약)

해야 한다. 감성분석은 사용자가 감성어 사전을 개발하여 해당 문서의 감성을 분석하는 방법³⁾과 요인분석과 주제분석을 통한 감성분석 방법이 있다. 메르스 정보확산 위험을 예측하기 위해서는 해당 문서에 대해 ‘안심, 불안’를 정의하는 감성분석을 실시해야 한다. 따라서 메르스 감정 키워드는 온라인 문서 수집 이후, 주제분석을 통하여 총 163개의 긍정감정 키워드(다행, 해결, 행복, 든든, 완벽, 안정, 안전, 깨끗, 기대, 감동, 격려, 극복, 긍정, 기대감, 기쁨, 도움, 미소, 믿음, 따뜻, 선호, 성공, 소망, 소중, 희망 등)와 229개의 부정감정 키워드(답답, 거짓말, 비상, 판단, 불안, 스트레스, 괴담, 냉소, 공포, 혼란, 엄벌, 공포증, 위험, 우려, 문제, 긴급, 부담, 악화, 난리, 비판, 무책임, 갈등, 감소, 갑갑, 강제 등)으로 분류하고 문서 상의 긍정과 부정 키워드를 각각 합산한 후, 감성분석(Opinion Mining)을 실시하여야 한다. 긍정은 메르스에 대해 안심하는 감정이고, 부정은 메르스에 대해 불안한 감정이며, 보통은 긍정과 부정이 동일한 감정을 나타낸다.

3) 우울관련 감정은 감성어 사전을 개발하여 긍정(‘스트레스 받지 않다, 우울증 퇴치하다, 행복 넘치다’ 등), 보통(‘한국인 우울하다, 청소년 우울하다’ 등), 부정(‘친구자살하다, 스트레스 심각하다, 왕따 심각하다’ 등)으로 구분하여 감성분석을 실시함

참고문헌 <<

- 노진석(2012). 빅데이터와 소셜 분석: 빅데이터의 바다에서 '의미'를 찾다. http://www.imaso.co.kr/?doc=bbs/gnuboard.php&bo_table=article&wr_id=40725
- 송태민(2012). 보건복지 빅데이터 효율적 활용방안, 보건복지포럼, 통권 제193호, pp.68~76.
- 송영조(2012). 빅데이터 시대! SNS의 진화와 공공정책. 한국정보화진흥원
- 송주영·송태민(2014). 소셜 빅데이터를 활용한 북한 관련 위협인식 요인 예측. 국제문제연구, 가을. pp.209~243.
- 송태민 외(2014). 보건복지 빅데이터 효율적 관리방안 연구. 한국보건사회연구원.
- 송태민(2013. 9). 우리나라 보건복지 빅데이터 동향 및 활용방안, 과학기술정책, 192, 과학기술정책연구원.
- Kim HY·Park HA·Min YH·Jeon E(2013). Development of an obesity management ontology based on the nursing process for the mobile-device domain. J Med Internet Res, 15(6), e130. doi: 10.2196/jmir.2512
- Gartner(2012)(www.gartner.com/newsroom/id/2124315, 2015. 8. 5. 인출).
- McKinsey Global Institute(2011). Big data: The next frontier for innovation, competition, and productivity, 2015. 8. 5. 인출

제 2 장

소셜 빅데이터 분석 기반 메르스 감정 위험 예측

1. 서론
2. 이론적 배경
3. 연구 방법
4. 분석 결과
5. 분석 결과 요약

참고문헌

2

소셜 빅데이터 분석 기반 << 메르스 감정 위험 예측>>

1. 서론

보건복지부「중앙메르스관리대책본부」는 7월 4일 이후(30일째) 신규 확진환자는 없으며, 8월 4일 현재 총 186명의 메르스 확진 환자가 발생하여 이중 36명은 사망하고 12명은 치료중인 것으로 보도하였다(보건복지부·질병관리본부, 2015. 8. 4. 보도자료). 최초의 메르스 감염자는 농작물 재배관련 일에 종사하던 68세 남성으로, 5월 4일 카타르를 경유하여 인천공항에 입국한 뒤, 입국 7일 후인 5월 11일 발열 및 기침 등의 증상이 발생하여 여러 병원을 방문하던 도중 5월 18일 서울 국립중앙의료원에 입원하였고, 5월 19일에 검체 의뢰를 통해 5월 20일에 확진 판정을 받았다(보건복지부·질병관리본부, 2015. 5. 20. 보도자료). 이후 최초 메르스 감염자에게 2차적으로 감염된 확진환자 6명이 추가적으로 확인됨에 따라 5월 26일 메르스 확진 환자는 7명으로 늘어났으며, 6월 2일 최초 메르스 감염자 남성과 같은 병동이던 57세 여성과 71세 남성이 사망함에 따라 메르스 확산 방지를 위한 국가적 보건역량을 총 동원키로 하고, 5월 31일 ‘민관합동대책반’을 구성한데 이어, 6월 2일 메르스 확산 방지 강화 대책을 발표하였다(보건복지부·질병관리본부, 2015. 6. 2. 보도자료).

한국 정부와 세계보건기구(WHO)는 급변 한국 메르스 코로나 바이러스(MERS-CoV) 전개 양상이 사우디아라비아 등을 통해 알려진 전개와

4) 본 연구의 일부 내용은 해외 학술지에 게재하기 위하여 ‘송주영 교수(펜실베이니아주립대학), 송태민 박사(한국보건사회연구원) 서동철 교수(이화여자대학교), 진달래 연구원(한국보건사회연구원), 김정선 박사(SK텔레콤 스마트인사이트)’에서 공동 수행한 것임을 밝힘.

다소간의 차이를 보이고 있는데 대한 국제사회의 우려를 고려하여, ‘국제 보건규칙(International Health Regulation, IHR)’에 의거 한-WHO 합동평가단(Joint Mission)을 구성하여 그 결과를 발표하였다(WHO, 2015). 첫째, 메르스는 한국의 대다수 의료인들에게 기대하지 못했던 낮은 질병이었으며, 어떤 병원에서는 침실이 많은 다인용 응급실에 환자들을 넘치게 수용하기만 했다는 측면에서 문제가 발생되었다. 둘째, ‘의료 쇼핑’이라고 해서 환자들이 많은 의료시설을 이곳저곳 다녀보는 습관이 있었고, 병원에 입원한 환자에게 친구와 가족 구성원들이 무분별하게 문병을 오는 사태가 질병을 퍼트리는 데에 영향력을 행사했을 가능성이 제기하였다. 셋째, WHO는 보건시설로 인한 감염을 막을 수 있도록 보건시설 위생유지를 조언했고, 또한 메르스 초기 증세가 불분명해서 메르스 초기 진단이 어려운 특성이 있기 때문에 보건의로 인력은 환자의 메르스 감염여부와 상관없이 모든 환자를 진찰 시에 항상 의료적 표준주의 지침을 충분히 숙고할 것을 당부하였다. 넷째, 현재 지역감염의 증거는 없으며, 접촉자 추적향상을 위한 광범위한 노력, (잠복기간 동안) 확진자 및 접촉자에 대한 적절한 격리, 검역, 감시 및 여행 제한을 포함한 메르스 발병 억제를 위한 지속적인 공중 보건 조치로 확진자 발생이 감소추세로 접어드는 것으로 보인다는 결과를 발표하였다.

2015년 7월 4일 이후 3주 이상 신규 확진환자가 발생하지 않자, 국무총리는 7월 28일 메르스와 관련하여 ‘7월 27일로 격리자가 모두 해제되는 등 여러 상황을 종합해 볼 때 국민께서 이제는 안심해도 좋다는 것이 의료계와 정부의 판단’이라며 사실상 메르스 종식을 선언하였다.

한편 모바일 인터넷과 소셜 미디어의 확산으로 데이터 량이 증가하여 데이터의 생산, 유통 소비체계에 큰 변화가 일어나면서 데이터가 경제적 자산이 될 수 있는 빅데이터 시대를 맞이하게 되었다. 세계 각국의 정부와 기업들이 빅데이터가 공공과 민간에 미치는 파급효과를 전망함에 따

라 SNS를 통해 생산되는 소셜 빅데이터의 활용과 분석을 통하여 사회적 문제의 해결과 정부의 정책을 효과적으로 추진할 수 있을 것으로 예측하고 있다. 빅데이터는 미래 국가 경쟁력에도 큰 영향을 미칠 것으로 예측하고 국가별로는 안전을 위협하는 글로벌 요인이나 테러, 재난재해, 질병, 위기 등에 선제적으로 대응하기 위해 우선적으로 도입하고 있다.

소셜 빅데이터의 분석은 사용자가 남긴 문서의 의미를 분석하는 것으로 자연어 처리기술인 주제분석(Text Mining)과 감성분석 기술인 오피니언 마이닝(Opinion Mining)을 실시한 후, 네트워크 분석(Network Analysis)과 통계분석(Statistical Analysis)을 실시해야 한다. 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)에서 논의된 관련 정보 상호간의 연관관계를 밝히고 원인을 파악하는 데는 한계가 있다(송주영, 송태민, 2014). 이에 반해 소셜 빅데이터의 분석은 훨씬 방대한 량의 데이터를 활용하여 다양한 참여자의 생각과 의견을 확인할 수 있기 때문에 사회적 문제의 예측과 현상에 대한 복잡한 연관관계를 보다 정확하게 밝혀낼 수 있다. 본 연구는 우리나라 온라인 뉴스사이트, 블로그, 카페, SNS, 게시판 등에서 수집한 소셜 빅데이터를 바탕으로 우리나라에서 발생한 메르스 정보 확산 위험요인을 예측코자 한다.

2. 이론적 배경

가. 메르스 관련 이론적 배경

메르스(Middle East Respiratory Syndrome, MERS)는 2012년 중동 지역의 국가에 살거나 여행하는 사람들에게서 발견되는 전염병으로

(WHO, 2013), 신종 코로나바이러스가 전염병의 원인이라고 규정되었기 때문에 국제 위원회(the International Committee on Taxonomy of Viruses)에서 ‘중동 지역 호흡기 증후군-코로나 바이러스(MERS-CoV)’이라는 명칭으로 공식적으로 명명하였다(Groot, R.J. 외, 2013). 메르스를 최초로 발견한 사람은 이집트의 질병학자 알리 모하메드 자키(Ali Mohamed zaki)로서, 메르스를 일으키는 코로나 바이러스는 HKU4와 HKU5와 같은 박쥐에게서 자생하는 코로나바이러스가 다른 동물에게도 퍼지게 되면서 나타나게 된 것으로 파악하였다(Zaki, A.M. 외 2012). 이후 박쥐의 코로나바이러스는 중동지역의 단봉낙타에게 옮겨졌고, 낙타가 사람에게 메르스 바이러스를 옮기는 주된 전파수단이 되었다(Alagaili, A.N. 외, 2014). 메르스는 사람과 사람의 접촉으로 감염되는 경우도 존재하는데, 이 경우에는 메르스 감염환자와의 가까운 접촉 또는 비말접촉으로 주로 감염되며 이러한 감염은 병원과 같은 보건시설에서 자주 일어난 것으로 보고하고 있다(Assiri, A. 외, 2013). 메르스의 전파경로는 2012년 4월부터 사우디아라비아에서 주로 발생하다가, 2012년 9월에 중동 걸프지역에서 영국 런던으로 메르스 감염 환자가 이동하면서 영국 내에도 메르스 감염이 일어났다(Bermingham, A. 외 2012). 2013년의 경우 사우디아라비아에서 레바논, 요르단, 아랍에미레이트연합 등에 전파되었고, 2014년에는 터키, 카타르, 오스트리아에도 메르스가 전파가 이루어져 메르스 감염 환자가 나타났다. 2012년 4월부터 2015년 7월 21일까지 국내외 메르스 감염사례는 1,392명, 사망사례는 538명으로 보고하고 있다(European Centre for Disease Prevention and Control, 2015a).

메르스 바이러스가 사우디아라비아를 비롯한 여러 국가에서 확산됨에 따라 WHO에서는 감염예방을 위한 지침과 메르스의 전염경로, 메르스 증상판별법 및 메르스 환자에 대한 치료법 등을 담은 매뉴얼을 2013년 7

월에 발간하였다(WHO, 2013). 메르스 증상으로는 고열, 기침, 호흡곤란 등의 증상이 있으며, 어떤 사람들의 경우 설사와 메스꺼움, 구토와 같은 위장 관련 증상을 경험하기도 한다(European Centre for Disease Prevention and Control, 2015b). 폐렴이나 신부전증이 있는 사람의 경우 메르스에 전염되었을 때 조금 더 심각한 합병증을 겪을 수 있으며, 공존증(Comorbidities, 1차적 질환을 포함해서 하나 이상의 질환이 몸 안에 존재하는 의학적 상황)을 지닌 사람들의 경우 메르스 감염에 더 취약하거나 증상이 심각해지는 경향이 있다. 단봉낙타에게서만 사례 수의 90% 이상이 넘는 280마리에서 메르스 바이러스 양성반응이 나타나 낙타가 주요한 전염원인으로 밝혀짐에(Hemida, M.G. 외 2013) 따라, 최근에는 아라비아 사막지역을 다녀 왔거나 아라비아 지역 여행자와 가까운 접촉을 한 사람, 그리고 낙타와 근거리에서 접촉하거나 멸균되지 않은 낙타유 또는 낙타고기를 섭취한 사람을 중심으로 메르스가 전염되고 있다고 보고되고 있다(European Centre for Disease Prevention and Control, 2015a). 메르스를 예방할 수 있는 예방백신은 아직까진 없으며, 미국 국립보건원에서 예방백신으로 기능할 만한 것을 개발 중이다. 예방백신 대신 메르스 감염을 예방하는 방법으로는 20초 동안 손을 비누로 깨끗하게 씻고, 재채기를 할 시에 티슈로 코와 입을 가리고 재채기를 하는 것이 있다. 또한, 다른 사람과 컵을 같이 쓰는 것을 피하고, 문의 손잡이나 살림도구를 깨끗하게 소독하는 방법 등이 있다(European Centre for Disease Prevention and Control, 2015c).

나. SNS상 전염병 확산 이론

최근 SNS 환경의 규모와 영향력이 점차 증대됨에 따라(Jong-Hwan

Kong, 2014), 개인의 커뮤니케이션 수단으로 SNS는 긴급 상황과 위기 대응에서 결정적인 요소로 진화하고 있다(Ryu, Hyeon Suk, 2013). 전 세계 SNS 사용자가 2013년 현재 17억명으로 2017년에는 25억명이 SNS를 이용할 것으로 전망하고 있다(eMarketer, 2013). SNS는 2008년 미국 대통령 대선과 2010년 ‘아랍의 봄’ 민주화 운동 등의 긍정적인 영향력(Adrien Guille 외, 2014)이 있는 반면, 광우병사태, 사스(SARS), 조류독감 등 사람들의 관심을 집중적으로 받는 사안은 미디어의 반복된 조명으로 사람들의 공포심을 자극하게 되고, 정부 혹은 관련 조직에 대한 불신 등과 겹쳐 위험이 확산될 수 있다(Kim, Young Wook, 2013).

정보확산(Information Diffusion)은 사회구성원들 사이에서 시간의 흐름에 따라 특정 채널을 통해 커뮤니케이션 되는 과정으로(Rogers Everett M., 1983) SNS상의 정보 확산 능력이 점차 커지면서 이에 대한 역기능 또한 증가하고 있다(Jong-Hwan Kong, 2014). 신종플루, 조류독감, 중증급성호흡기증후군과 같은 감염성이 강한 질환의 발생은 전 세계적으로 질병에 대한 두려움을 키울 뿐 아니라(Liang Mao, 2014), 이용자의 접근성이 높고 이동성이 강한 SNS는 이러한 질병에 대한 유언비어의 확산 채널 역할을 한다(Hong, Ju-hyeon, Yun, Hye-jin 2014). 온라인 공간은 다양한 정보를 소비하는 곳일 뿐만 아니라 새로운 정보를 확산시키고 생산하고 더 나아가 현실 세계에서 직접적인 행동을 이끄는 원천이 되고 있으며(Park, Min-Gyeong, Lee, Gun-Ho, 2011) 이렇게 형성된 특정 정보는 낙인의 물결 효과를 거쳐서 낙인화(Stigmatization)의 충격을 양산하고 있다. 낙인화는 위험의 사회 확산에 따라 형성이 되며, 위험과 관련된 특정 사람, 상품, 장소, 기술 등에 붙여지는 부정적인 이미지, 감정적인 반응, 사회행동 차원의 효과를 의미한다. 위험의 확산에 따른 낙인화 과정은 특정 대상에서 위험 사안이 발생, 정보 확산, 공중

인식과 표식화 과정, 특정 대상의 정체성 형성, 낙인의 물결 효과, 낙인화 영향과 충격의 6가지로 단계로 이어진다(Kim, Young Wook. 2014).

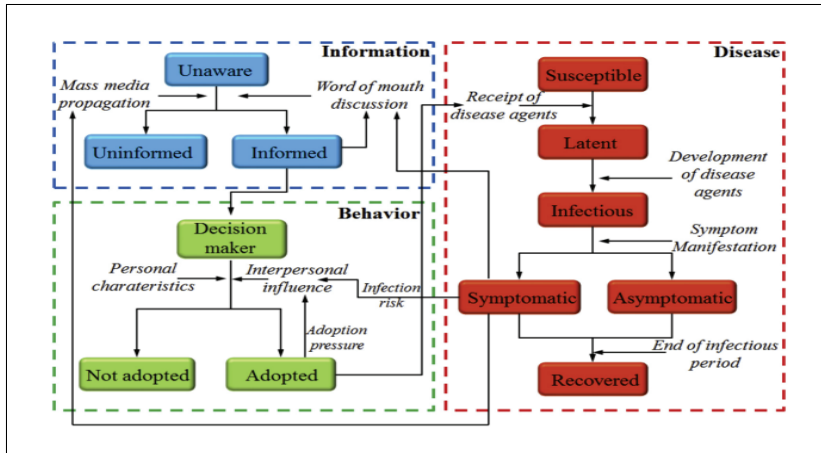
확산되는 루머가 자극적일수록 이용자들이 활발하게 댓글을 올리면서 상호작용을 할 뿐 아니라 국민들의 불안감을 유발하고, 정부 정책에 대한 불신을 초래하는 등 개인적 및 사회적으로 손실이 크게 다가온다(Hong, Ju-hyeon, 2014).

SNS를 통한 정보의 확산 연구는 데이터 분석을 통하여 사회문제에 대한 궁극적인 예측 결과를 형성 할 수 있을 뿐만 아니라 더 큰 이해관계를 얻을 수 있다(Dinyakant Agrawal, 2011). 질병의 확산과 같은 정보 확산의 과정은 네트워크를 통하여 발생하며(Mostafa Salehj 외 2015), 사회 확산(Social Amplification)은 질병과 같은 위험 정보가 인터넷 등의 채널을 통해 집중되었다가 빠르게 증폭됨으로써 사회적인 충격을 일으키는 경로와 과정을 추적하는 것으로, 일반적인 전염병은 질병 전파, 질병에 관한 정보의 흐름, 질병에 대한 예방 행동 수칙들을 확산 한다(Liang Mao, 2014).

[그림 2-1]과 같이 빨간 점선 안의 감염상태에 따른 프로세스는 ‘미보균자(Susceptible)’인 개인의 경우 감염인자가 있는 이웃과의 접촉했을 경우 감염성 질환에 감염될 수 있 수 있지만 감염인자가 신체 내부에서 방출되지 않은 ‘잠복(Latent)’상태로 있다. 질병인자가 신체 내부에 있을 때에는 다른 사람에게 전파되지 않으며, 잠복기의 다음 단계는 ‘전염(Infectious)’상태로 한 개인이 다른 사람에게 감염성 질환을 전파 할 수 있고 확산경로는 늘어난다. 감염 기간동안 ‘증상(Symptomatic)’ 또는 ‘무증상(Asymptomatic)’ 상태를 가질 수 있고, 감염 관련 정보들은 각각 대중매체를 통하여 전파, 주변을 통한 입소문 전파, 소셜 네트워크의 대인관계에 의하여 퍼져나가며 감염기간이 지나면 ‘회복(Recovered)’ 단

체를 거치게 되며 이때에는 감염에 면역성이 생긴 것으로 간주한다.

[그림 2-1] 세 과정 사이에서의 정보확산



주: Liang Mao (2014). Modeling Triple-diffusions of infractions disease, information, and preventive behaviors through a metropolitan social networks: an agent- based simulation. *Applied Geography*, 50, pp.31~49.

[그림 2-1]의 파란색 점선 안의 정보확산 프로세스는 질병이 처음 발병했을 때에는 개개인은 인식하지 못하지만(Unaware) 이 때 입소문과 대중매체를 통하여 정보를 습득하게 된다. 전자(uninformed)의 경우는 로컬 네트워크를 통하여 정보가 순환시켜 지지만, 후자(informed)의 경우는 전 세계적으로 정보가 확산될 뿐만 아니라 예방차원에서의 의사결정 행동을 불러일으킨다. [그림 2-1]과 같이 녹색 점선 안에 프로세스는 예방 행동이 확산되는 과정으로 자신의 개인적인 특성과 질병의 확산과정에서 증상(Symptomatic)을 통하여 ‘감염위험’으로 인지된 정보는 SNS의 대인관계 의한 영향력에 따라서 정보를 수용하는데 영향을 미치고, 질병의 확산은 개인의 인지된 위험과 자극을 통하여 예방 행동을 채택할 수 있도록 하며, 예방 행동의 채택을 저해하는 질병확산은 사람과 질

병 시스템간의 부정적인 연결고리를 형성 할 수 있다(Liang Mao, 2014).

3. 연구 방법

가. 연구대상⁵⁾ 및 분석방법

본 연구는 105개의 온라인 뉴스사이트, 8개의 게시판, 1개의 SNS(트위터), 4개의 블로그 등 총 120개의 온라인 채널을 통해 수집 가능한 텍스트 기반의 웹문서(버즈)를 소셜 빅데이터로 정의하였다. 메르스 토픽(topic)은 모든 관련 문서를 수집하기 위해 ‘메르스’를 사용하였으며, 토픽과 같은 의미로 사용되는 토픽 유사어로는 ‘메르스 바이러스, 중동 호흡기 증후군, 메르스 코로나 바이러스, 매르스’ 용어를 사용하였고, 불용어는 ‘메르스벤츠, 메르스테스벤츠’로 하였다. 소셜 빅데이터의 수집은 우리나라에 메르스의 발생이 처음으로 알려진 시점인 2015년 5월 19일부터 6월 2일 동안 해당 채널에서 매 시간 단위로 수집하였으며⁶⁾, 수집된 총 666,510건⁷⁾의 텍스트(Text) 문서를 본 연구의 분석에 포함하였다. 메르스 위험을 설명하는 가장 효율적인 예측모형을 구축하기 위해 데이터마이닝의 연관규칙과 의사결정나무 분석, 그리고 시각화 분석을 사용하였다. 연관규칙의 분석 알고리즘은 선형적 규칙(apriori principle)을 사용하였고, 의사결정나무 형성을 위한 분석 알고리즘은 훈련표본과 검정표본의 정분류율이 높게 나타난 Exhaustive CHAID(Chi-squared

5) 본 고의 연구대상은 1부 1장의 ‘소셜 빅데이터 분석 및 활용 방안’의 분석 사례의 대상과 달리 5월 19일~6월 2일(15일간) 온라인 문서를 대상으로 하였다.

6) 본 연구를 위한 소셜 빅데이터의 수집 및 토픽 분류는 ‘(주)SK텔레콤 스마트인사이트’에서 수행함.

7) 블로그 5,056건(0.8%), 카페 7,133건(1.1%), SNS 618,471건(92.8%), 게시판 12,693건(1.9%), 뉴스 23,157건(3.5%).

Automatic Interaction Detection) 알고리즘을 사용하였다. 기술분석, 다중응답분석, 로지스틱회귀분석, 의사결정나무 분석은 SPSS 22.0을 사용하였고, 연관분석과 시각화는 R version 3.1.3을 사용하였다.

나. 연구도구

메르스 관련하여 수집·분류된 문서는 주제분석(text mining)과정을 거쳐 다음과 같이 정형화 데이터로 코드화하여 사용하였다.

1) 메르스 관련 감정

메르스 감정 키워드는 온라인 문서 수집 이후, 주제분석을 통하여 총 163개의 긍정감정 키워드(다행, 해결, 행복, 든든, 완벽, 안정, 안전, 깨끗, 기대, 감동, 격려, 극복, 긍정, 기대감, 기쁨, 도움, 미소, 믿음, 따뜻, 선호, 성공, 소망, 소중, 희망 등)와 229개의 부정감정 키워드(답답, 거짓말, 비상, 판단, 불안, 스트레스, 괴담, 냉소, 공포, 혼란, 엄벌, 공포증, 위험, 우려, 문제, 긴급, 부담, 악화, 난리, 비판, 무책임, 갈등, 감소, 갑갑, 강제 등)으로 분류하고 문서 상의 긍정과 부정 키워드를 각각 합산한 후, 감성분석(Opinion Mining)을 실시하였다. 긍정은 메르스에 대해 안심하는 감정이고, 부정은 메르스에 대해 불안한 감정이며, 보통은 긍정과 부정이 동일한 감정을 나타낸다.

2) 메르스 감염대상

메르스 감염대상은 주제분석 과정을 거쳐 ‘일반인(국민, 국민들, 사람,

사람들, 성인, 시민 등), 남성(남편, 신랑, 아빠, 오빠, 아버지, 기러기아빠, 계부 등), 여성(부인, 아내, 어머니, 언니, 엄마 등), 노인(노약자, 노인, 할머니, 할아버지, 노인들, 조부모, 조부 등), 아이학생(대딩, 대학생, 아기, 아들, 아이, 아이들, 애들, 어린이 등), 가족(가족들, 가족, 부모, 패밀리, 부부 등), 싱글(개인, 본인, 나), 외국인(외국인, 미군병사, 중국인, 미국인, 중동인 등), 여행객(여행객들, 여행객, 관광객 등), 증상자(보균자, 증상자, 첫감염자, 감염환자, 메르스환자 등), 의료계(의료진들, 의료진, 교수, 간호사, 의사 등), 군인(파병, 장병, 파병자, 사병, 병사), 직장인(직원, 업주, 해외건설근로자, 건설근로자 등)의 13개 요인으로 대상이인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

3) 메르스 관련 국가

메르스 관련 국가는 주제분석 과정을 거쳐 '아시아(중국, 홍콩, 대만, 대한민국, 한국, 아시아 등), 중동(바레인, 이라크, 이란, 사우디아라비아, 중동국가, 레바논 등), 아프리카(수단, Sudan, 지부티, Djibouti, 서아프리카, 기니 등), 유럽(스페인, Spain, 영국, UnitedKingdom, 프랑스, 독일 등), 미국(미국, 뉴욕, 하와이 등)'의 5개 요인으로 국가요인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

4) 메르스 관련 기관

메르스 관련 기관은 주제분석 과정을 거쳐 '정부(감염병리과장, 복지장관, 보건장관, 보건복지위원장, 국무총리, 장관 등), 민간기관(영양사협회, 역사학연구회, 국경없는의사회 등), 정당(새정치민주연합, 새누리당 등),

국제기구(세계보건기구, WHO, CDC, 미국보건당국, 유럽질병통제청 등), 중국국가기관(중국보건당국 등), 병원(전남대병원, 국가격리병원, 국가지정격리치료병원, 고대구로병원, 서울대학교병원, 서울대병원 등), 항공사(이스타항공, 도하공항, 김해공항, 제주공항, 홍콩공항, 인천국제공항, 인천공항 등), 학교(대학교, 고등학교, 초등학교, 제주대학교)의 8개 요인으로 기관요인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

5) 메르스 감염경로

메르스 감염경로는 주제분석 과정을 거쳐 '1차감염, 2차감염, 3차감염, 감염경로, 감염원, 낙타(낙타, 낙타시장, 낙타접촉, 낙타체험, 낙타체험프로그램, 낙타타기, 낙타고기, 낙타요리, 생낙타유), 공기(공기, 공기감염, 공기전염, 공기호흡, 호흡기감염), 기타동물(당나귀, 염소, 동물, 동물들, 가금류, 박쥐), 접촉(밀접, 접촉, 밀접접촉, 재채기, 인체감염, 비말감염)의 5개 요인으로 감염경로 요인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

6) 메르스 관련 바이러스

메르스 관련 바이러스는 주제분석 과정을 거쳐 '코로나바이러스(베타코로나바이러스, 베타코로나, 메르스코로나, 메르스코로나바이러스, 코로나바이러스, 코로나바이러스과, Coronavirus), 사스(사스, SARS, SARScoronavirus, 중증급성호흡기증후군, 중증급성호흡기질환, 중증호흡기곤란, 급성호흡곤란증후군, 다기관부전증), 신종플루(H1N1, 신종인플루엔자, 신종플루), 조류인플루엔자(조류독감, 조류인플루엔자, avi-

aninfluenza), 에볼라(ebola, ebolahemorrhagicfever, 에볼라, 에볼라바이러스, 에볼라출혈열), 기타바이러스(바이러스, 신종감염병, 신종바이러스, 전염병, 감염병, 병원, 중증열성혈소판감소증후군, 인수공통전염병, 생탄저균, Lassafever, 라사열)의 7개 요인으로 바이러스 요인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

7) 메르스 관련 증상

메르스 관련 증상은 주제분석 과정을 거쳐 '전파(전염, 전염력, 전염성, 전파력, 발병, 발생, 감염, 감염력, 감염율, 확산), 의심증상(양성, 메르스 감염, 음성, 의심증상, 의심증세, 잠복기간, 잠복기, 최대잠복기), 열(오한, 38도, 발열, 고열, 고열증세, 미열, 미열증세), 호흡기(숨가쁨, 숨가쁨, 호흡곤란, 호흡곤란증세, 호흡기, 호흡기이상, 호흡기증상, 감기, 목감기, 인후통, 가래), 소화기증상(구토, 설사, 식욕부진, 위장장애, 복통), 신장질환(급성신부전, 신부전증세, 급성신부전증, 신장기능, 심낭액저류, 콩팥내종양, 콩팥종양, 심부전), 사망(사망, 취사율, 치사율, 목숨, 생명), 기타증상(혼수상태, 사구체신염, 폐렴, 폐렴증세, 폐감염, 합병증, 혈소판감소, 복막염, 패혈증, 채장염, 두통, 섬망, 흥통, 혈액, 혈전증, 죽상경화증, 간질, 감각이상, 결석, 경련, 과다출혈, 근육통, 수지진전, 중증질환)의 8개 요인으로 증상 요인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

8) 메르스 관련 대처

메르스 관련 대치는 주제분석 과정을 거쳐 '초기대응(신고, 초기, 초기대응, 초기발견, 진단, 초기증상, 병원진찰, 진료, 검사결과, 발견), 치료

(1차치료, 2차치료, 약물, 엑스레이, 항체검사, 치료, 집중치료, 인공호흡기, 치료백신, 치료약, 치료제, 해결방법, 해열제, 백신, ZMapp, 브린시도피어, 지맵, TKM에볼라, 파비피라비르, 입원, 산소공급), 격리(확진, 가택격리, 격리, 격리대상, 격리조치, 격리종료, 자가격리, 자택격리), 감염가능검사(검사, 격리검사, 발열감시, 발열감지, 발열검사, 양성반응, 양성판정, 유전자검사, 음성반응, 음성판정, 채혈, 판정, 감염가능성, 감염여부, 감염증세, 전파가능성), 정복대응(위급상황, 위기대응, 특별검역, 폐쇄, 지원, 검역절차, 검역체계, 교육비, 국가지정입원치료, 긴급복지지원제도, 긴급비상회의, 긴급현안, 대응단계, 대응책, 대응현황, 대처상황, 발표, 방역, 방역체계, 역학조사, 종합대응방안, 종합대책, 초강경대책)의 5개 요인으로 대처 요인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화하였다.

9) 메르스 관련 예방

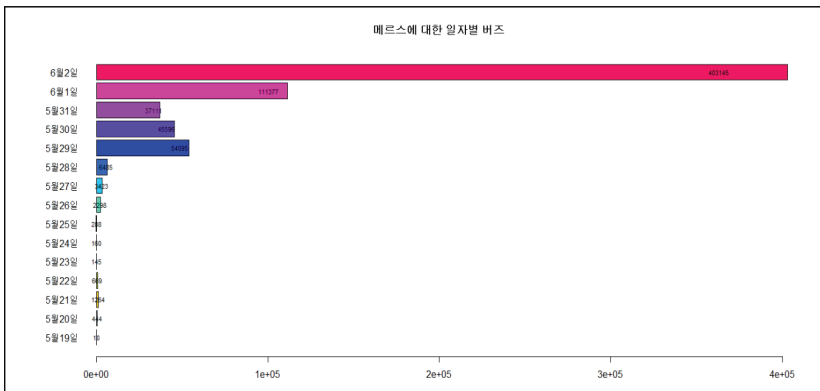
메르스 관련 예방은 주제분석 과정을 거쳐 '예방수칙(감염예방, 감염예방수칙, 예방, 예방법, 예방수칙), 외출(쇼핑, 야외활동, 소풍, 나들이), 위생(개인위생, 비누, 소독, 손세정제, 손소독젤, 손씻기, 위생, 청결), 면역강화(면역력, 항바이러스, 항체, 건강관리, 유산균), 마스크(마스크, 손수건, 입막음, 티슈)'의 5개 요인으로 예방 요인이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

4. 분석 결과

[그림 2-2]와 같이 메르스와 관련된 버즈는 2015년 5월 28일 '내국

인 메르스 의심자 1명 중국으로 출국’ 보도(2015. 5. 28. 보도자료) 후 급속히 증가하여 5월 30일 ‘유엔비어 관련 당부사항’과 5월 31일 ‘문형표 장관, 메르스 확산 방지 위해 민관 합동 총력 대응 선언’보도 이후 감소하였다가, 6월 1일 이후 메르스 추가 환자 발생과 사망자 발생 보도 후 급속히 증가하였다.

[그림 2-2] 메르스 관련 문서(버즈)량의 일별 추이



[그림 2-3]과 같이 메르스에 대한 단계별⁸⁾ 긍정적인 감정(안심) 표현 단어는 경계단계(3단계)까지 ‘가능성, 안전, 해결’ 키워드에 집중되었으며, 심각1단계(4단계) 이후, ‘다행, 가능성, 기대, 해결’ 키워드에 집중된 것으로 나타났다. 메르스에 대한 부정적인 감정(불안) 표현 단어는 경계 단계까지 ‘의심, 우려, 문제’ 키워드에 집중되었으며, 심각1단계(4단계) 이후, ‘괴담, 의심, 불안, 실패’ 키워드에 집중된 것으로 나타났다.

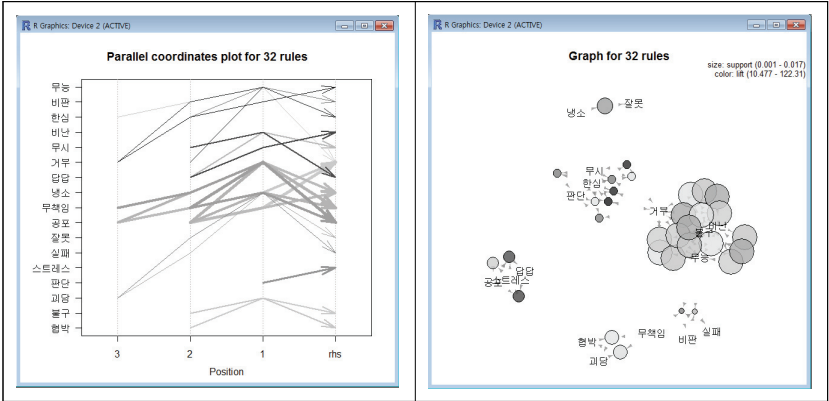
8) 메르스가 국내에 알려진 5월 19일~25일을 관심단계, 내국인 메르스 환자 중국 출국이 알려진 5월 26일~28일을 주의단계, 메르스 확진환자 추가발생과 메르스 괴담이 급속히 전파된 5월 29일~30일을 경계단계, 메르스 확산 방지를 위한 민관 합동 총력 대응을 선언한 5월 31일을 심각1단계, 메르스 추가환자와 사망자가 발생한 6월 1일~2일을 심각2 단계의 5단계로 구분함

〈표 2-1〉과 같이 메르스 부정 감정의 연관성 예측에서 가장 신뢰도가 높은 연관규칙으로는 ‘{거부, 비판, 무시} ⇒ {한심}’이며 네 변인의 연관성은 지지도 0.003, 신뢰도는 0.985, 향상도는 122.31로 나타나, 온라인 문서(버즈)에서 ‘거부, 비판, 무시’가 언급되면 정부의 대처방안에 대해 한심하다는 부정적 감정으로 생각할 확률이 98.5%이며, ‘거부, 비판, 무시’가 언급되지 않은 버즈 보다 메르스에 대해 한심하다는 부정적 감정일 확률이 약 122.3배 높아지는 것으로 나타났다. [그림 2-4]와 같이 메르스에 대한 부정적 표현 단어는 ‘무시, 한심, 판단’, ‘거부, 비난, 무능, 불구’, ‘답답, 공포, 스트레스’, ‘무책임, 비판, 실패’, ‘괴담, 헐박’ 키워드와 강하게 연결되어 있는 것으로 나타났다.

〈표 2-1〉 메르스 부정(불안) 감정 예측

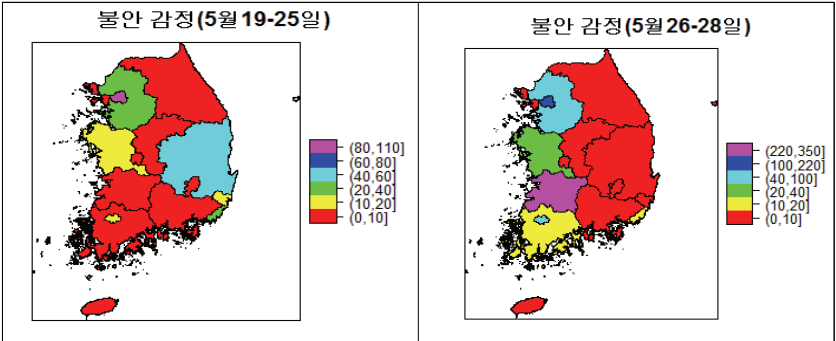
순위	규칙	지지도	신뢰도	향상도
1	{거부, 판단, 무시} ⇒ {한심}	0.002811300	0.9852632	122.31002
2	{거부, 판단, 한심} ⇒ {무시}	0.002811300	0.9936306	112.83129
3	{거부, 한심} ⇒ {무시}	0.002811300	0.9811321	111.41203
4	{공포, 답답} ⇒ {스트레스}	0.006109172	0.9835590	92.40071
5	{공포, 스트레스} ⇒ {답답}	0.006109172	0.9960823	88.95859
6	{거부, 무시, 한심} ⇒ {판단}	0.002811300	1.0000000	60.31558
7	{무시, 한심} ⇒ {판단}	0.002811300	0.9957447	60.05892
8	{거부, 한심} ⇒ {판단}	0.002829322	0.9874214	59.55689
9	{무시, 실패, 무책임} ⇒ {비판}	0.001129326	0.9947090	57.08039
10	{냉소} ⇒ {잘못}	0.009479128	0.9825654	44.08858
11	{거부, 무능, 불구} ⇒ {비난}	0.017228226	1.0000000	38.12022
12	{무능, 불구} ⇒ {비난}	0.017228226	0.9989551	38.08039
13	{거부, 무능} ⇒ {비난}	0.017246247	0.9954924	37.94839
14	{거부, 불구} ⇒ {비난}	0.017288296	0.9941278	37.89637
15	{무능, 비판, 무책임} ⇒ {실패}	0.001129326	0.9842932	30.41698
16	{거부, 무능, 비난} ⇒ {불구}	0.017228226	0.9989551	24.97328
17	{거부, 무능} ⇒ {불구}	0.017228226	0.9944521	24.86071
18	{무능, 비난} ⇒ {불구}	0.017228226	0.9927310	24.81768
19	{거부, 비난} ⇒ {불구}	0.017288296	0.9862920	24.65671
20	{거부, 불구, 비난} ⇒ {무능}	0.017228226	0.9965254	20.43012

[그림 2-4] 메르스 감정의 연관규칙에 대한 병렬좌표표와 그래프 시각화

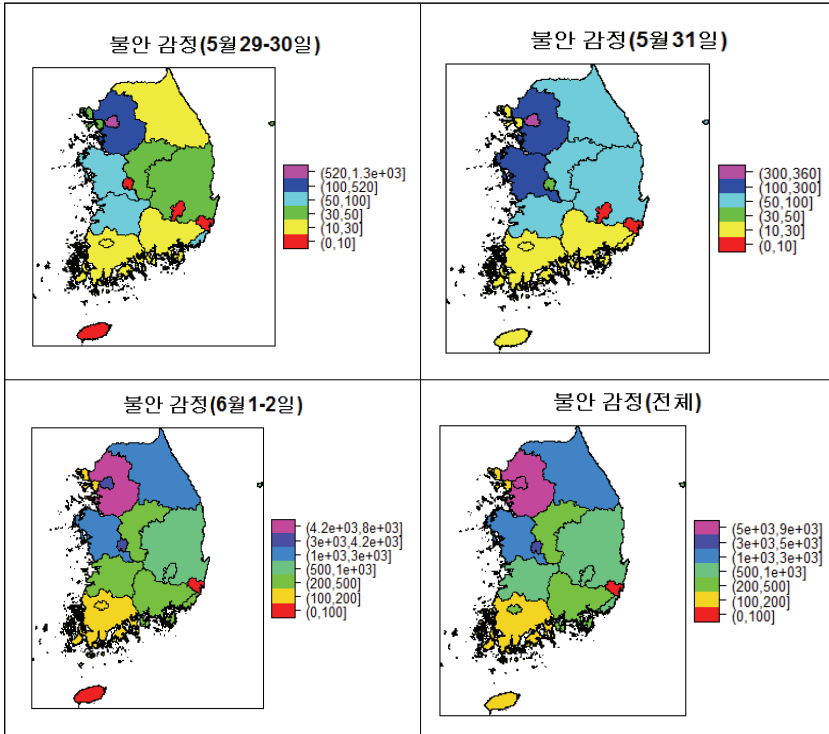


[그림 2-5]와 같이 지역별 메르스에 대한 부정적(불안) 감정⁹⁾은 심각1 단계(4단계)까지 서울, 경기, 충 남, 부산 등의 순으로 높은 것으로 나타났고, 심각2단계(5단계)부터는 경기, 대전, 서울, 강원 등의 순 으로 높은 것으로 나타났다.

[그림 2-5] 지역별 메르스 위험(불안) 감정



9) 총 버즈 666,510건 중 지역을 식별할 수 있는 버즈 76,316건(11.45%)에 대한 지역별 메르스에 대한 부정적(불안) 감정의 빈도를 나타냄



〈표 2-2〉와 같이 메르스와 관련하여 긍정적인 감정(안심)을 나타내는 온라인 문서(버즈)는 22.3%, 보통의 감정을 나타내는 버즈는 6.5%, 부정적인 감정(불안)은 71.2%로 나타났다. 메르스 관련 국가는 아시아(71.1%), 아메리카(16.5%), 중동(10.1%) 등의 순으로 나타났다. 메르스 관련 기관은 정부(68.9%), 병원(23.9%), 학교(1.8%), 정당(1.8%) 등의 순으로 나타났다. 메르스 관련 감염은 접촉(35.3%), 낙타(34.6%), 감염 경로(14.4%) 등의 순으로 나타났다. 메르스 관련 증상으로는 전파(49.7%), 사망(21.0%), 의심증상(13.6%) 등의 순으로 나타났다. 메르스 관련 대처/치료로는 정부대응(31.6%), 격리(28.9%), 감염가능검사(18.5%) 등의 순으로 나타났다. 메르스 관련 예방으로는 예방수칙(37.4%), 마스크

크(35.4%), 위생(20.3%) 등의 순으로 나타났다. 메르스 관련 대상으로는 증상자(52.1%), 일반인(21.4%), 의료인(8.6%) 등의 순으로 나타났다. 메르스 관련 바이러스로는 기타바이러스(37.6%), 사스(31.3%), 신종플루(12.6%) 등의 순으로 나타났다.

〈표 2-2〉 메르스 관련 버즈 현황

구분	항목	N(%)	구분	항목	N(%)
감정	긍정(안심)	51,998(22.3)	대처/치료	초기대응	37,147(13.6)
	보통	15,176(6.5)		치료	20,295(7.4)
	부정(불안)	166,471(71.2)		격리	78,985(28.9)
	계	233,645		감염가능검사	50,472(18.5)
국가	아시아	83,970(71.1)	예방	정부대응	86,289(31.6)
	중동	11,945(10.1)		계	273,188
	아프리카	840(0.7)		예방수칙	28.8(37.4)
	유럽	1,967(1.7)		외출자제	501(0.7)
	아메리카	19,450(16.5)	대상	위생	15,606(20.3)
기관	계	118,172		면약강화	11,306(14.7)
	정부	172,872(68.9)		마스크	27,164(35.4)
	민간기관	479(0.2)		계	76,691
	정당	4,531(1.8)	대상	일반인	88,928(21.4)
	국제기구	3,934(1.6)		남성	6,902(1.7)
	중국국가기관	1,073(0.4)		여성	4,638(1.1)
	병원	59,897(23.9)		노인	1,440(0.3)
	공항	3,625(1.4)		아이학생	12,305(3.0)
	학교	4,622(1.8)		가족	15,894(3.8)
	계	251,033		싱글	13,669(3.3)
감염	감염경로	13,137(14.4)		외국인	1,962(0.5)
	낙타	31,436(34.6)		여행객	3,425(0.8)
	공기	10,863(11.9)		증상자	216,660(52.1)
	기타동물	3,394(3.7)		의료인	35,666(8.6)
	접촉	32,119(35.3)		군인	844(0.2)
	계	90,949		직장인	13,895(3.3)
증상	전파	119,859(49.7)		계	416,228
	의심증상	32,753(13.6)	바이러스	코로나바이러스	2,126(3.0)
	열	12,133(5.0)		사스	21,955(31.3)
	호흡기증상	18,795(7.8)		신종플루	8,849(12.6)
	소화기증상	1,848(0.8)		조류인플루엔자	2,526(3.6)
	신장질환	1,429(0.6)		에볼라	8,307(11.8)
	사망	50,716(21.0)		기타바이러스	26,347(37.6)
	기타증상	3,576(1.5)		계	70,110
	계	241,109			

〈표 2-3〉과 같이 예방요인에 대한 메르스 감정의 연관성 예측에서 가장 신뢰도가 높은 연관규칙으로는 ‘{위생, 면역강화}=>{안심}’이며 세 변인의 연관성은 지지도 0.003, 신뢰도는 0.882, 향상도는 11.298로 나타나, 온라인 문서(버즈)에서 위생, 면역강화가 언급되면 메르스를 긍정적(안심)으로 생각할 확률이 88.2%이며, 위생, 면역강화가 언급되지 않은 버즈 보다 메르스에 대한 감정이 긍정적일 확률이 11.3배 높아지는 것으로 나타났다. 대처/치료요인에 대한 메르스 감정의 연관성 예측에서 가장 신뢰도가 높은 연관규칙으로는 ‘{초기대응, 격리, 감염검사}=>{불안}’이며 네 변인의 연관성은 지지도 0.02, 신뢰도는 0.904, 향상도는 3.619로 나타나, 온라인 문서(버즈)에서 초기대응, 격리, 감염검사가 언급되면 메르스를 부정적(불안)으로 생각할 확률이 90.4%이며, 초기대응, 격리, 감염검사가 언급되지 않은 버즈 보다 메르스에 대한 감정이 부정적일 확률이 3.6배 높아지는 것으로 나타났다. 증상요인에 대한 메르스 감정의 연관성 예측에서 가장 신뢰도가 높은 연관규칙으로는 ‘{전파, 열, 사망}=>{불안}’이며 세 변인의 연관성은 지지도 0.002, 신뢰도는 0.5676, 향상도는 2.273로 나타나, 온라인 문서(버즈)에서 전파, 열, 사망이 언급되면 메르스를 부정적(불안)으로 생각할 확률이 56.8%이며, 전파, 열, 사망이 언급되지 않은 버즈 보다 메르스에 대한 감정이 부정적일 확률이 2.27배 높아지는 것으로 나타났다.

〈표 2-3〉 예방요인, 대처/치료요인, 증상요인에 대한 메르스 감정 예측

구분	규칙	지지도	신뢰도	향상도
예방 요인	{위생, 면역강화} => {안심}	0.002955695	0.88143177	11.2981862
	{위생} => {안심}	0.007423745	0.31705754	4.0640414
	{면역강화} => {안심}	0.003318780	0.19564833	2.5078189
	{위생, 마스크} => {안심}	0.001947458	0.15676329	2.0093907
	{예방수칙, 위생, 마스크} => {불안}	0.001756913	0.35345608	1.4151535
	{예방수칙, 위생} => {불안}	0.002049482	0.34460141	1.3797015
	{예방수칙} => {안심}	0.003357789	0.10120286	1.2972175

48 2015년 소셜 빅데이터 기반 보건복지 이슈 동향 분석

구분	규칙	지지도	신뢰도	향상도
	{마스크 => {안심}}	0.003230259	0.07925931	1.0159453
	{> => {안심}}	0.078015334	0.07801533	1.0000000
	{> => {불안}}	0.249765195	0.24976519	1.0000000
	{예방수칙, 마스크} => {불안}	0.002074988	0.23742489	0.9505924
	{예방수칙} => {불안}	0.005713343	0.17219861	0.6894420
	{위생, 마스크} => {불안}	0.002106495	0.16956522	0.6788985
	{면역강화} => {불안}	0.002604612	0.15354679	0.6147646
	{마스크} => {불안}	0.005434277	0.13333824	0.5338544
	{위생} => {불안}	0.003053218	0.13039856	0.5220846
	{초기대응, 격리, 감염검사} => {불안}	0.015567658	0.9041478	3.619991
	{초기대응, 격리} => {불안}	0.018457337	0.8382393	3.356109
	{감염검사, 정부대응} => {불안}	0.013081574	0.8202258	3.283987
	{초기대응, 감염검사} => {불안}	0.015969753	0.7587141	3.037709
	{초기대응, 격리, 정부대응} => {불안}	0.004430541	0.7129406	2.854443
	{초기대응, 정부대응} => {불안}	0.008370467	0.7023795	2.812159
	{초기대응, 격리, 감염검사, 정부대응} => {불안}	0.002298540	0.6913357	2.767943
대체/치료 요인	{초기대응, 치료, 격리, 감염검사, 정부대응} => {불안}	0.001354818	0.6866920	2.749350
	{초기대응, 치료, 감염검사, 정부대응} => {불안}	0.001422334	0.6742532	2.699548
	{치료, 격리, 감염검사, 정부대응} => {불안}	0.001923452	0.6719078	2.690158
	{초기대응, 치료, 격리, 감염검사} => {불안}	0.001578371	0.6666667	2.669174
	{초기대응, 감염검사, 정부대응} => {불안}	0.002445575	0.6631408	2.655057
	{치료, 감염검사, 정부대응} => {불안}	0.002040480	0.6605148	2.644543
	{초기대응, 치료, 격리, 정부대응} => {불안}	0.001750911	0.6570946	2.630849
	{초기대응, 치료, 감염검사} => {불안}	0.001663891	0.6531213	2.614941
	{치료, 격리, 정부대응} => {불안}	0.002651123	0.6453616	2.583873
	{전파, 호흡기증상, 사망} => {안전}	0.001930954	0.4030692	5.166538
	{전파, 열, 사망} => {불안}	0.002082489	0.5676892	2.272891
	{전파, 열, 호흡기증상, 사망} => {불안}	0.001771916	0.5472660	2.191122
	{전파, 의심증상, 열, 사망} => {불안}	0.001543863	0.5455992	2.184448
	{전파, 의심증상, 호흡기증상, 사망} => {불안}	0.001518357	0.5403097	2.163270
	{전파, 열} => {불안}	0.004037449	0.5391705	2.158710
	{전파, 의심증상, 열} => {불안}	0.002715638	0.5375705	2.152304
증상 요인	{전파, 열, 호흡기증상} => {불안}	0.002600111	0.5333949	2.135585
	{전파, 의심증상, 열, 호흡기증상, 사망} => {불안}	0.001360820	0.5307197	2.124875
	{전파, 의심증상, 호흡기증상} => {불안}	0.002070487	0.5257143	2.104834
	{전파, 기타증상} => {불안}	0.001330813	0.5153980	2.063530
	{전파, 의심증상, 열, 호흡기증상} => {불안}	0.001756913	0.5113537	2.047338
	{전파, 열, 기타증상} => {불안}	0.001053248	0.5090645	2.038172
	{전파, 의심증상, 사망} => {불안}	0.002187514	0.4753831	1.903320
	{전파, 호흡기증상} => {불안}	0.003501823	0.4589971	1.837714
	{전파, 의심증상} => {불안}	0.008245938	0.4570474	1.829910
	{의심증상, 열} => {불안}	0.003776387	0.4384254	1.755350
	{전파, 호흡기증상, 사망} => {불안}	0.002055483	0.4290636	1.717868

메르스의 감정에 영향을 미치는 요인은 다음과 같다. <표 2-4>와 같이 메르스와 관련한 예방수칙, 위생, 면역강화는 정적의 영향을 미치는 것으로 나타나 예방수칙, 위생, 면역강화와 관련한 예방요인이 온라인 상에 많이 언급될수록 메르스에 대한 부정적인 감정(불안)이 감소하는 것으로 나타났으나, 마스크와 외출자제는 부적인 영향을 미치는 것으로 나타나 부정적인 감정(불안)을 증가시키는 것으로 나타났다. 메르스와 관련한 치료와 정부대응은 정적의 영향을 미치는 것으로 나타나 치료와 정부대응과 관련한 대처요인이 온라인 상에 많이 언급될수록 메르스에 대한 부정적인 감정(불안)이 감소하는 것으로 나타났으나, 초기대응, 격리, 감염가능 검사는 부적인 영향을 미치는 것으로 나타나 부정적인 감정(불안)을 증가시키는 것으로 나타났다. 메르스와 관련한 호흡기증상, 신장질환, 기타질환은 정적의 영향을 미치는 것으로 나타나 호흡기증상, 신장질환과 관련한 증상요인이 온라인 상에 많이 언급될수록 메르스에 대한 부정적인 감정(불안)이 감소하는 것으로 나타났으나, 전파, 의심증상, 열, 사망은 부적인 영향을 미치는 것으로 나타나 부정적인 감정(불안)을 증가시키는 것으로 나타났다. 메르스와 관련한 채널요인은 SNS만 부적인 영향을 미치는 것으로 나타나 SNS로 확산되는 온라인 문서가 부정적인 감정(불안)을 증가시키는 것으로 나타났다.

<표 2-4> 메르스의 감정에 영향을 미치는 요인¹⁾

변수		긍정				보통			
		b [†]	S.E. [‡]	OR [§]	P	b [†]	S.E. [‡]	OR [§]	P
예방	예방수칙	.082	.032	1.086	.011	-.327	.058	.721	.000
	외출자제	-.514	.152	.598	.001	.443	.187	1.557	.018
	위생	2.273	.035	9.707	0.000	1.292	.057	3.640	.000
	면역강화	.387	.040	1.472	.000	-.619	.092	.538	.000
	마스크	-.599	.039	.549	.000	.473	.050	1.605	.000

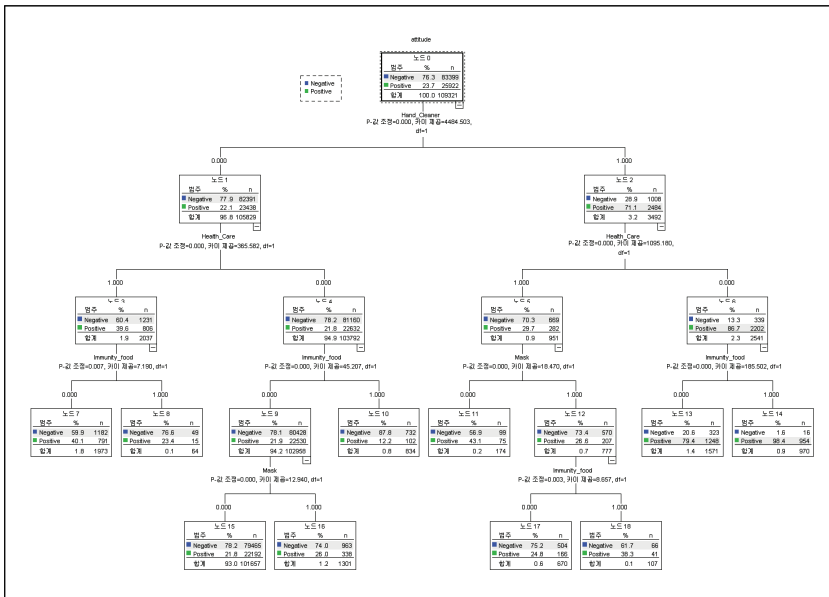
변수		긍정				보통			
		b [†]	S.E. [‡]	OR [§]	P	b [†]	S.E. [‡]	OR [§]	P
대처/ 치료	초기대응	-1.407	.030	.245	0.000	-.368	.032	.692	.000
	치료	1.566	.030	4.786	0.000	1.007	.044	2.736	.000
	격리	-1.192	.020	.304	0.000	-.772	.028	.462	.000
	감염가능검사	-1.583	.027	.205	0.000	.349	.024	1.418	.000
	정부대응	.682	.012	1.978	0.000	.040	.022	1.040	.070
증상	전파	-.481	.015	.618	.000	-.261	.023	.770	.000
	의심증상	-.410	.027	.664	.000	-.429	.043	.651	.000
	열	-.195	.043	.823	.000	.352	.058	1.422	.000
	호흡기증상	1.936	.029	6.929	0.000	1.581	.042	4.859	.000
	소화기증상	.115	.121	1.122	.340	-.347	.176	.707	.048
	신장질환	.267	.078	1.307	.001	.192	.101	1.211	.057
	사망	-.550	.027	.577	.000	-.298	.041	.742	.000
	기타질환	.244	.068	1.276	.000	.379	.087	1.461	.000
채널	블로그	.597	.047	1.817	.000	.863	.067	2.370	.000
	카페	.623	.052	1.865	.000	.446	.090	1.561	.000
	SNS	-.227	.018	.797	.000	-.614	.026	.541	.000
	게시판	.210	.040	1.234	.000	.263	.065	1.301	.000
	뉴스	.047	.024	1.048	.052	.624	.032	1.867	.000

주: 1) 기본범주: 부정, † Standardized coefficients, ‡ Standard error, § Adjusted odds ratio

[그림 2-6]과 같이 메르스 관련 예방요인이 메르스의 감정 예측 모형에 미치는 영향은 ‘Hand_Claner’의 영향력이 가장 큰 것으로 나타났다. ‘Hand_Claner’가 있을 경우 메르스의 부정(불안 감정)은 이전의 76.3%에서 28.9%로 크게 감소한 반면, 긍정(안심 감정)은 이전의 23.7%에서 71.1%로 증가하였다. ‘Hand_Claner’가 있고 ‘Hand_Care’가 없는 경우 메르스의 부정은 이전의 28.9%에서 13.3%로 증가한 반면, 긍정적 감정은 이전의 71.1%에서 86.7%로 감소하였다. ‘Hand_Claner’가 없을 경우 메르스의 부정(불안 감정)은 이전의 76.3%에서 77.9%로 증가한 반면, 긍정(안심 감정)은 이전의 23.7%에서 22.1%로 감소하였다. ‘Hand_Claner’가 없고 ‘Hand_Care’가 없는 경우 메르스의 부정은 이

전의 77.9%에서 78.2%로 증가한 반면, 긍정적 감정은 이전의 22.1%에서 21.8%로 감소하였다.

[그림 2-6] 메르스 관련 예방 요인의 예측모형

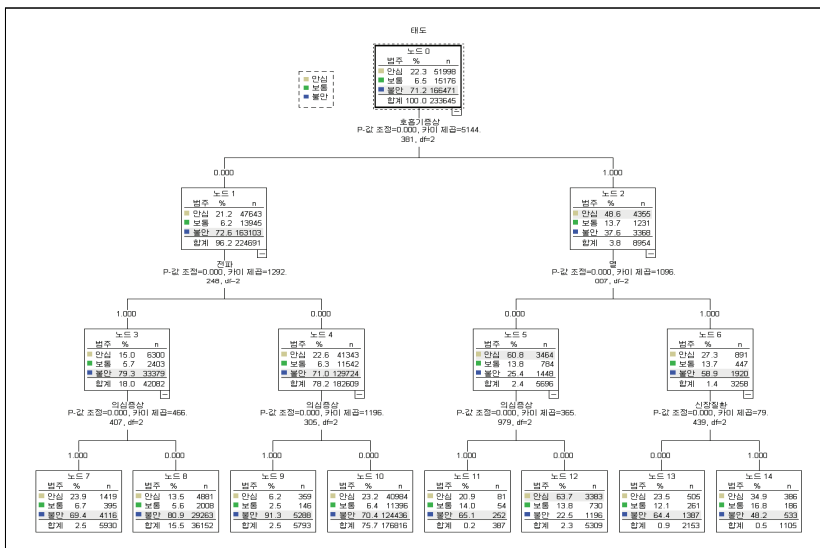


[그림 2-7]과 같이 메르스 관련 증상요인이 메르스의 감정 예측 모형에 미치는 영향은 ‘호흡기 증상’의 영향력이 가장 큰 것으로 나타났다. ‘호흡기 증상’이 있을 경우 메르스의 부정(불안 감정)은 이전의 71.2%에서 37.6%로 크게 감소¹⁰⁾한 반면, 보통 감정은 이전의 6.5%에서 13.7%, 긍정(안심 감정)은 이전의 22.3%에서 48.6%로 증가하였다. ‘호흡기 증상’이 있고 ‘열’이 있는 경우 메르스의 부정(불안)은 이전의 37.6%에서 58.9%

10) ‘호흡기 증상’이 있을 경우 부정(불안)의 감정이 감소한 것은 ‘호흡기 증상’ 보다는 ‘발열’에 대한 불안한 감정이 많이 전파되어 평소 ‘호흡기 증상’만 있는 사람은 안심을 하지만 ‘열’이 발생할 경우 불안한 감정을 증가시키는 것으로 판단된다.

로 증가한 반면, 긍정적 감정은 이전의 48.6%에서 27.3%로 증가하였다. ‘호흡기 증상’이 없을 경우 메르스의 부정(불안 감정)은 이전의 71.2%에서 72.6%로 증가한 반면, 긍정(안심 감정)은 이전의 22.3%에서 21.2%로 감소하였다. ‘호흡기 증상’이 없고 ‘전파’가 있는 경우 메르스의 부정(불안 감정)은 이전의 72.6%에서 79.3%로 증가한 반면, 긍정적 감정은 이전의 21.2%에서 15.0%로 감소하였다. ‘호흡기 증상’이 없고 ‘전파’가 없고 ‘의심증상’이 있는 경우 메르스의 부정(불안 감정)은 이전의 71.0%에서 91.3%로 증가한 반면, 긍정적 감정은 이전의 22.6%에서 6.2%로 크게 감소하였다.

[그림 2-7] 메르스 관련 증상 요인의 예측모형



〈표 2-5〉의 메르스의 증상요인의 예측 모형에 대한 이익도표와 같이 메르스의 긍정감정에 가장 영향력이 높은 경우는 ‘호흡기 증상’이 있고 ‘열’이 없고 ‘의심증상’이 없는 조합으로 나타났다. 즉, 12번 노드의 지수(index)가 286.3%로 뿌리마디와 비교했을 때 12번 노드의 조건을 가진

집단이 메르스를 긍정적으로 느끼는 확률이 2.86배로 나타남. 메르스 부정적인 감정에 가장 영향력이 높은 경우는 ‘호흡기증상’이 없고 ‘전파’가 없고 ‘의심증상’이 있는 조합으로 나타났다. 즉, 9번 노드의 지수가 128.1%로 뿌리마디와 비교했을 때 9번 노드의 조건을 가진 집단이 메르스에 대한 부정적인 확률이 1.28배로 나타났다.

〈표 2-5〉 메르스 관련 증상요인의 예측 모형에 대한 이익도표

구분	노드	이익자수				누적자수			
		노드(n)	노드(%)	이익(%)	자수(%)	노드(n)	노드(%)	이익(%)	자수(%)
안심	12	5309	2.3	6.5	286.3	5309	2.3	6.5	286.3
	14	1105	.5	.7	157.0	6414	2.7	7.2	264.0
	7	5930	2.5	2.7	107.5	12344	5.3	10.0	188.8
	13	2153	.9	1.0	105.4	14497	6.2	10.9	176.5
	10	176816	75.7	78.8	104.2	191313	81.9	89.8	109.6
	11	387	.2	.2	94.0	191700	82.0	89.9	109.6
	8	36152	15.5	9.4	60.7	227852	97.5	99.3	101.8
	9	5793	2.5	.7	27.8	233645	100.0	100.0	100.0
보통	14	1105	.5	1.2	259.1	1105	.5	1.2	259.1
	11	387	.2	.4	214.8	1492	.6	1.6	247.7
	12	5309	2.3	4.8	211.7	6801	2.9	6.4	219.6
	13	2153	.9	1.7	186.6	8954	3.8	8.1	211.7
	7	5930	2.5	2.6	102.6	14884	6.4	10.7	168.2
	10	176816	75.7	75.1	99.2	191700	82.0	85.8	104.6
	8	36152	15.5	13.2	85.5	227852	97.5	99.0	101.6
	9	5793	2.5	1.0	38.8	233645	100.0	100.0	100.0
불안	9	5793	2.5	3.2	128.1	5793	2.5	3.2	128.1
	8	36152	15.5	17.6	113.6	41945	18.0	20.8	115.6
	10	176816	75.7	74.7	98.8	218761	93.6	95.5	102.0
	7	5930	2.5	2.5	97.4	224691	96.2	98.0	101.9
	11	387	.2	.2	91.4	225078	96.3	98.1	101.9
	13	2153	.9	.8	90.4	227231	97.3	99.0	101.8
	14	1105	.5	.3	67.7	228336	97.7	99.3	101.6
	12	5309	2.3	.7	31.6	233645	100.0	100.0	100.0

5. 분석 결과 요약

메르스 관련 버즈는 2015년 5월 28일 급속히 증가하여 5월 30일 감소하였다가 6월 1일 이후 메르스 추가 환자 발생과 사망자 발생 보도 후, 급속히 증가한 것으로 나타났다. 메르스에 대한 부정적 감정(불안)의 표현 단어는 ‘무시, 한심, 판단’, ‘거부, 비난, 무능, 불구’, ‘답답, 공포, 스트레스’, ‘무책임, 비판, 실패’, ‘괴담, 협박’ 키워드와 강하게 연결되어 있는 것으로 나타나, 이는 정부의 초기 대응 미흡에 대한 국민의 실망감과 SNS를 통한 메르스 괴담의 급속한 전파로 인한 불안감이 표출된 것으로 보인다.

메르스와 관련하여 긍정적인 감정(안심)을 나타내는 온라인 문서(버즈)는 22.3%, 부정적인 감정(불안)은 71.2%로 나타나, 메르스에 대한 부정적 감정이 약 3.2배 높은 것으로 나타났다.

메르스 사태에서는 온라인 문서 중 트위터 등 SNS를 통해 정보가 많이 유통된 것으로 나타났다. 메르스를 키워드로 추출한 데이터 중 SNS가 차지하는 비율이 92.8%(61만8471건)로 나타나, 담뱃값 논란 당시 SNS 비율(52.9%) 보다 높은 비율로 나타났다. 또한, SNS 게시물들은 메르스에 대한 불안을 심화시킨 것으로 나타났다. 블로그, 카페 등을 통해 메르스 정보를 접한 사람은 안심 등 긍정적인 마음이 약 1.8배 증가했지만, SNS를 통해 메르스 정보를 접한 사람은 메르스에 대해 안심하는 비율이 20% 가량 감소한 것으로 나타났다¹¹⁾.

온라인 문서(버즈)에서 {위생, 면역강화}가 동시에 언급되면 메르스를 긍정적으로 생각할 감정(안심)이 증가하며, {초기대응, 격리, 감염검사}가 동시에 언급되면 메르스를 부정적으로 생각하는 감정(불안)이 증가하고,

11) 동아일보(2015. 6. 11.) ‘기침환자는 차분한데 .. 건강한 사람이 더 불안에 떨어’ <http://news.donga.com/3/all/20150611/71758969/1>.

{전파, 열, 사망}이 동시에 언급되면 메르스를 부정적으로 생각하는 감정(불안)이 증가하는 것으로 나타났다.

온라인상에 ‘마스크, 외출자제, 초기대응, 격리, 감염검사, 전파, 의심 증상, 열, 사망’이 언급되면 메르스를 부정적으로 생각하는 감정(불안)이 증가하는 것으로 나타났다.

SNS상에서 확산되는 메르스 관련 온라인 문서는 메르스를 부정적으로 생각하는 감정(불안)을 증가시키는 것으로 나타났다. 메르스에 긍정적인 감정에 가장 영향력이 높은 경우는 ‘호흡기 증상’이 있고 ‘열’이 없고, 의심 증상이 없는 조합으로 나타났으며, 부정적인 감정에 가장 영향력이 높은 경우는 ‘호흡기증상’이 없고, ‘전파’가 없고 ‘의심증상’이 있는 조합으로 나타났다.

참고문헌 <<

- 송주영, 송태민(2014). 소셜 빅데이터를 활용한 북한 관련 위협인식 요인 예측. 국제문제연구, 가을. pp.209~243.
- Adrien Guille, Hakim Hacid, C. Favre, Djamel Abdlkader Zighed(2013). *Information Diffusion in Online Social Networks: A Survey*. Association for Computing Machinery, 42(2). pp.17~28
- Alagaili, A.N., Briese, T., Mishra, N., Kapoor, V., Sameroff, S.C., de Wit, E., Munster V.J., Hensley, L.E., Zalmout, I.S., Kapoor, A., Epstein, J.H., Karesh, W.B., Daszak, P., Mohammed, O.B., Lipkin, W.I.(2014). *Middle East Respiratory Syndrome Coronavirus Infection in Dromedary Camels in Saudi Arabia.*, mBio, 5(2); e000884~14.
- Assiri, A., McGeer, A., Peri, T.M., Price, C.S., Rabeeah A.A., Cummings, D.A., Alabdullatif, Z.N., Assad, M., Almulhim, A., Makhdoom, H., Madani, H., Alhakeem, R., Al-Tawfig, J.A., Cotten, M., Watson, S.J., Kellam, P., Zumla, A., Memish, Z.A.(2013). Hospital Outbreak of Middle East Respiratory Syndrome Coronavirus, *The New England Journal of Medicine*, 369(5), pp.407~416.
- Bermingham, A., Chand, M.A., Brown, C.S., Asrons, E., Tong, C., Langrish, C., Hoschler, K., Brown, K., Galiano, M., Myers, R., Pebody, R.G., Green, H.K., Boddington N.L., Gopal, R., Price, N., Newsholme, W., Drosten, C., Fouchier, R.A., Zambon, M.(2012). *Severe Respiratory Illness Caused by a Novel Coronavirus, in a Patient Transferred to the United Kingdom from the Middle East*, Euro Surveillance, 17(4), pp.1~5.
- Centers for Disease Control and Prevention(2015b). *Middle East*

- Respiratory Syndrome(MERS): Symptoms & Complications.*
Http://www.cdc.gov/coronavirus/mers/about/symptoms. html.
Centers for Disease Control and Prevention(2015a). *People Who May Be at Increased Risk for MERS.*
Http://www.cdc.gov/coronavirus/mers/risk.html.
Centers for Disease Control and Prevention(2015c). *Middle East Respiratory Syndrome(MERS): Prevention & Treatment.*
Http://www.cdc.gov/coronavirus/mers/about/prevention.html
Dinyakant Agrawal, Caren Budak, Amr El Abbadi(2011). *Information diffusion in Social Networks: Observing and influencing Societal Interests.* in Proceeding of International Conference on Very Large Data Bases. pp.1~5.
European Centre for Disease Prevention and Control(2015a), *Epidemiological Update: Middle East Respiratory Syndrome Coronavirus(MERS-CoV).*
eMarkerter(2013). *Social Networking Reaches Nearly One in Four Around the World: By 2014, the Ranking of Regions by Social Networkd Users will Reflect Regional Shares of the Global Population.* 2013/06/18 [검색일: 2015.06.04., http://emarketer.com/Article/Social-Networking-Reaches-Nearly-One-Four-Around-World/1009976]
Groot, R.J., Baker, S.C., Baric, R.S., Brown, C.S., Drosten, C., Enjuances, L., Fouchier, R.A., Galiano, M., Gorbalenya, A.E., Memish, Z., Perlman, S., Poon, L.L., Snijer, E.J., Stephens, G.M., Woo, P.C., Zaki, A.M., Zambon, M., Ziebuhr, J.(2013). Middle East Respiratory Syndrome Coronavirus(MERS-CoV); Announcement of the Coronavirus Study Group, *Journal of Virology*, 87(14), pp.7790~779.

- Hemida, M.G., Perera, R.A., Wang, P., Alhammadi, M.A., Siu, L.Y., Li, M., Poon, L.L., Saif, L., Alnaeem, A., Peiris, M.(2013). Middle East Respiratory Syndrome(MERS) Coronavirus Seroprevalence in Domestic Livestock in Saudi Arabia, 2010 to 2013, *Euro Surveillance*, 18(50), pp.1~7.
- Hong, Ju-hyeon, Yun, Hye-jin(2014). *The Diffusion of Rumor Via Twitter: The diffusion and the user interactivity in the KOREA U.S. FTA Case*, Korean Association for Communication and Information Studies 66. pp.59~84.
- Hong, Ju-hyeon(2014). *A Crisis of Confidence and the Media: Newspapaer and Broadcast* 10, pp.15~20.
- Jong-Hwan Kong, Ik-Kun Kim, Myung-Mook Han (2014). Propagation Models for Structural Parameters in Online Social Networks, *Journal of Internet Computing and Services* 15(1) pp.125~134
- Kim, Young Wook (2014). *Risk Communication*. Communicationbooks
- Liang Mao (2014). Modeling Triple-diffusions of Infectious Disease, Information, and Preventive Behaviors through a Metropolitan Social Networks: ans Agent-based Simulation. *Applied Geography*. 50, pp.31~39.
- Mostafa Salehj, Payam Siyari, Matteo Magnani, Danilo Montesi(2015). Multidimensional Epidemic Thresholds in Diffusion Process over Interdependent Networks. *Multiplex Networks: Structure, Dynamics and Application*. 72. pp.59~67.
- Park, Min-Gyeong, Lee, Gun-Ho (2011). *Analysis of Online Opinion Leader's Discourse Patterns: Regarding Opinion aobut Sejong City posted on Agora, the discussion borad of the Portal Daum*. Korean Association for Communication and Information Studies, 48(1), pp.114~149.

- Ryu, Hyeon Suk (2013). *A Study on Risk Perception and Communication via Social Media*. The Korea Institute of Public Administration. Research Report 25(3).
- Rogers Everett M. (1983). *Diffusion of Innovations*, Third Edition. The American Center Library. p.5.
- WHO(2013). Middle East Respiratory Syndrome Coronavirus Joint Kingdom of Saudi Arabia/WHO mission. *Media Centre News Releases*. 2013-06-10.
- WHO(2013). WHO Guidelines for Investigation of Cases of Human Infection with Middle East Respiratory Syndrome Coronavirus (MERS-CoV).
- WHO(2015). Recommends Continuation of Strong Disease Control Measures to bring MERS-CoV Outbreak in Republic of Korea to an end , *Media Centre*, 2015-06-23.
- Zaki, A.M., Boheemen, S.V., Bestebroer, T.M., Osterhaus, A.D., Fouchier R.A(2012). Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia, *The New England Journal of Medicine*, 367(19), pp.1814~1820.

제 3 장

소셜 빅데이터를 활용한 담배 위험 예측

1. 서론
 2. 연구 방법
 3. 연구 결과
 4. 결론
- 참고문헌

3

소셜 빅데이터를 활용한 << 담배 위험 예측12)

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 Health Data, 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 건강보험진료비는 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, 만성

12) 본 연구는 '송태민(KIHASA), 송주영(PSU, 교신저자), 천미경(KIHASA). 소셜 빅데이터를 활용한 담배 위험 예측. 한국데이터정보과학회지. 2015, 제26권 5호'에 게재된 논문을 밝힌다.

질환 등 다양한 질병과 관련있는것으로 알려져 있다 (Carter 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 제도로 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 금

연클리닉, 금연상담전화, 온라인 상담 등 1:1 맞춤형 금연상담서비스도 대폭 강화할 계획이다(Ministry of Health and Welfare, 2014 Press release).

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 학문 연구에서도 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. 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. 연구 방법

가. 연구대상

본 연구는 국내의 SNS, 온라인 뉴스 사이트 등 인터넷을 통해 수집된 소셜 빅데이터를 대상으로 하였다. 본 분석에서는 200개의 온라인 뉴스 사이트, 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)¹³⁾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. 담배 토픽은 모든 관련

13) 토픽은 소셜 분석 및 모니터링의 '대상이 되는 주제어'를 의미하며, 문서 내에 관련 토픽이 포함된 문서를 수집함.

문서를 수집하기 위해 ‘담배’를 사용하였으며, 토픽과 같은 의미로 사용되는 토픽 유사어로는 ‘흡연, 담뱃값, 담배 피, 담배 추천, 담배가격, 혼녀 생정담배, 중딩담배, 고딩담배, 중고딩 담배, 청소년 담배’ 용어를 사용하였다. 본 연구를 위한 소셜 빅데이터의 수집¹⁴⁾은 크롤러 (Crawler)를 사용하였고, 이후 주제분석을 통해 분류된 명사형 어휘를 유목화 (categorization)하여 분석요인으로 설정하였다.

나. 연구도구

담배와 관련하여 수집된 문서는 주제분석¹⁵⁾의 과정을 거쳐 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 가

14) 본 연구를 위한 소셜 빅데이터의 수집 및 토픽 분류는 ‘(주)SK텔레콤 스마트인사이트’에서 수행함.

15) 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 문서 수집 후 원시자료 (raw data)에서 나타난 상위 2,000개의 키워드들을 대상으로 유목화를 하여 사용자 사전을 구축하였음.

지는 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 횡수로 표현한 문서는 잠재군으로 분류하였다. 일반군은 담배를 혐오적으로 생각하는 감정이고, 위험군은 담배를 애호적으로 생각하는 감정이며, 잠재군은 담배를 보통으로 생각하는 감정을 나타낸다.

2) 담배와 관련된 정책

담배와 관련된 정책의 정의는 주제 분석 과정을 거쳐 ‘담뱃값인상, FCTC(담배규제기본협약 등), 금연관련법(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, 동

맥경화, 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, 중

독, 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, 금연 건물, 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.

다. 분석 방법

본 연구에서 우리나라 담배의 위험을 설명하는 가장 효율적인 예측모

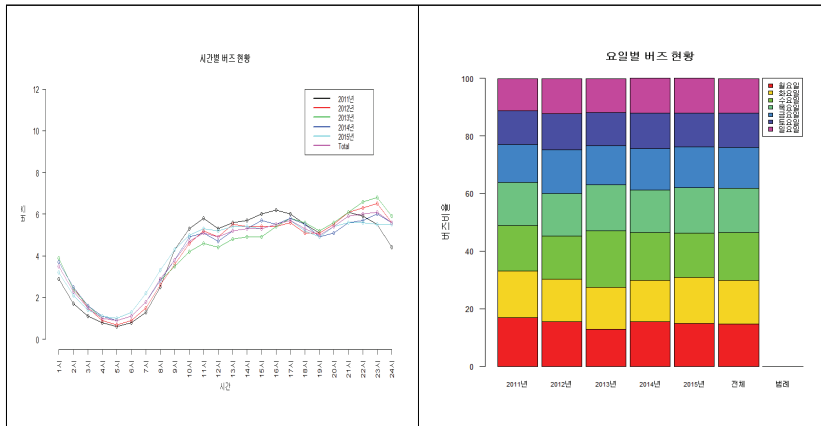
형을 구축하기 위해 특별한 통계적 가정이 필요하지 않은 데이터마이닝의 연관분석 (association analysis)과 의사결정나무 (decision tree) 방법을 사용하였다. 소셜 빅데이터 분석에서 연관분석은 하나의 온라인 문서 (transaction)에 포함된 둘 이상의 단어들에 대한 상호관련성을 발견하는 것으로 동시에 발생한 어떤 단어들의 집합에 대해 조건과 연관규칙을 찾는 분석방법이다. 전체 문서에서 연관규칙의 평가 측도는 지지도 (support), 신뢰도 (confidence), 향상도 (lift)로 나타낼 수 있다. 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, 2010). 본 연구의 연관분석은 선형적 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, 다중응답분석, 의사결정나무분석은 SPSS v. 22.0을 사용하였고, 연관분석과 시각화는 R version 3.1.3을 사용하였다.

3. 연구 결과

가. 담배 관련 문서(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. 담배와 관련된 버즈는 평일에는 수요일, 목요일, 화요일, 월요일, 금요일 순으로 높은 추이를 보이는 반면, 주말에는 감소하는 것으로 나타났다.

[그림 3-1] 담배관련 시간별 및 요일별 버즈 현황

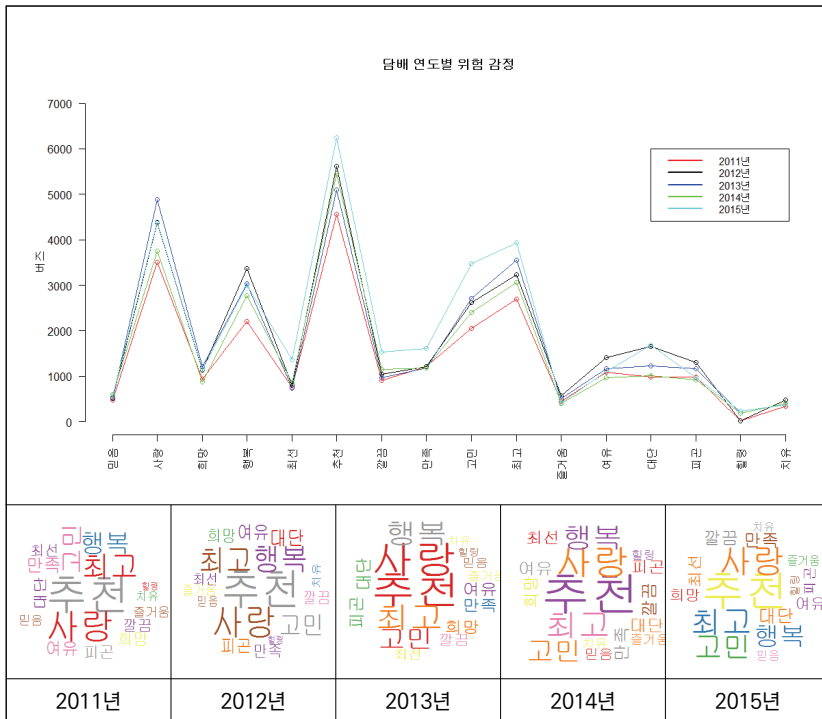


[그림 3-2]와 같이 연도별 담배에 대해 긍정적인 감정(위험) 변화는 2011년 대비 평균 1.8배씩 증가하였으며, 위험 감정의 표현 단어는 추천, 사랑, 최고, 행복, 고민 등의 순으로 집중된 것으로 나타났다. 담배에 대해 부정적인 감정(일반) 변화는 2011년 대비 평균 1.54배씩 하였으며, 일반 감정의 표현 단어는 문제, 스트레스, 걱정, 위험, 부담 등의 순으로

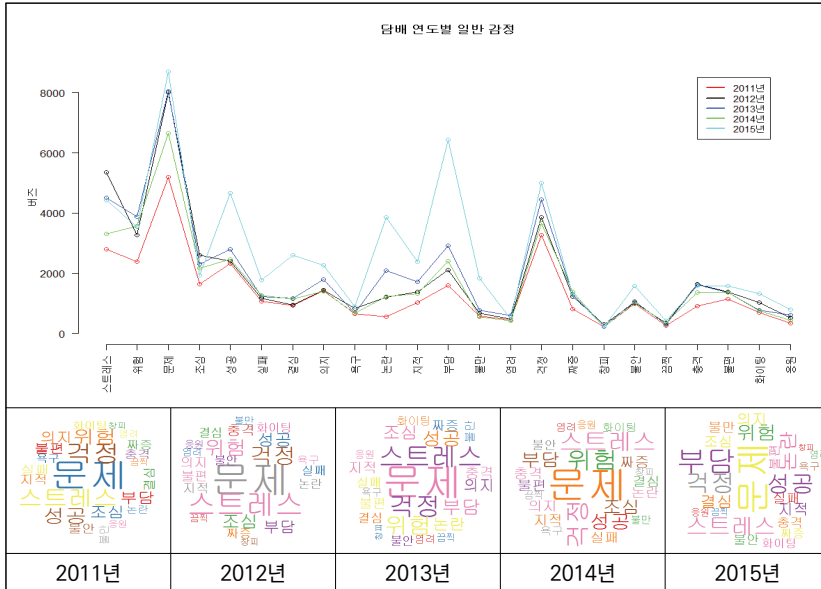
집중된 것으로 나타났다. [그림 3-2]과 같이 연도별 담배와 관련한 질병의 문서는 심혈관질환, 폐암, 사망, 감기, 고혈압, 당뇨병 등의 순으로 집중된 것으로 나타났다. 담배와 관련한 정책의 문서는 담배값인상, 흡연구제, 금연관련법, 금연상비, 금연광고, FCTC 순으로 언급된 것으로 나타났다. 담배와 관련한 금연도구의 문서는 전자담배, 금연패치, 금연약, 금연껌, 금연보조제 순으로 언급된 것으로 나타났다.

[그림 3-2]와 같이 지역별 담배에 대한 감정은 위험은 서울, 경기, 부산, 제주, 인천 등의 순으로 높은 것으로 나타났고, 잠재는 서울, 경기, 부산, 인천, 제주 등의 순으로 높은 것으로 나타났다.

[그림 3-2] 연도별 담배 감정 변화



74 2015년 소셜 빅데이터 기반 보건복지 이슈 동향 분석



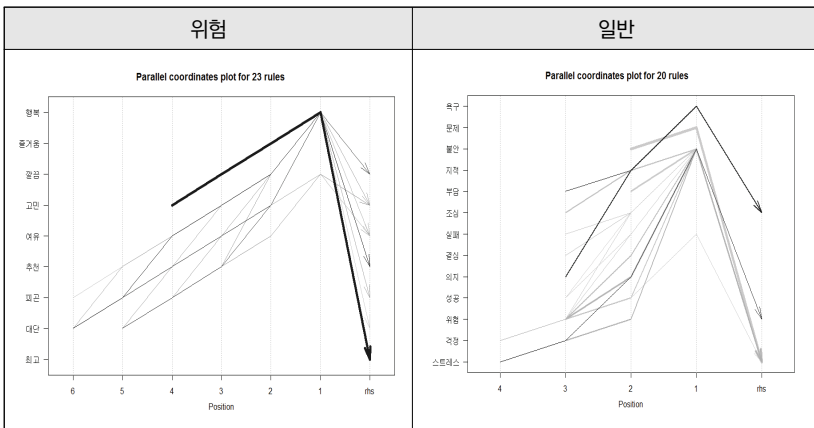
〈표 3-1〉과 같이 담배에 대한 감정 키워드의 연관성 예측에서 위험 감정의 경우 행복, 추천, 고민, 최고, 대단, 피곤, 깔끔, 여유의 감정에 위험 감정이 강하게 연결되어 있는 것으로 나타났으며, 일반 감정의 경우 스트레스, 문제, 조심, 지적, 위험의 감정에 일반 감정이 강하게 연결되어 있는 것으로 나타났다.

〈표 3-1〉 담배 감정 키워드 연관성 예측

규칙		지지도	신뢰도	향상도
위험	{행복, 추천, 고민, 최고, 대단, 피곤}=>{갈굼}	0.001110136	0.9722222	16.394312
	{고민, 즐거움, 대단, 피곤}=>{여유}	0.001416745	0.9852941	16.221423
	{행복, 추천, 갈굼고민, 최고, 피곤}=>{대단}	0.001110136	0.9905660	14.277767
	{행복, 추천, 갈굼, 고민, 피곤}=>{대단}	0.001120709	0.9724771	14.017037
	{행복, 추천, 갈굼, 최고, 대단}=>{고민}	0.001152427	0.9909091	7.064914
	{행복, 추천, 갈굼, 대단, 피곤}=>{고민}	0.001120709	0.9906542	7.063097
	{추천, 갈굼, 최고, 대단, 피곤}=>{고민}	0.001120709	0.9906542	7.063097
	{행복, 추천, 갈굼, 최고, 대단, 피곤}=>{고민}	0.001110136	0.9905660	7.062469
	{행복, 갈굼, 최고, 대단, 피곤}=>{고민}	0.001131282	0.9816514	6.998909
	{행복, 갈굼, 대단, 피곤}=>{고민}	0.001163000	0.9734513	6.940445

	규칙	지지도	신뢰도	향상도
일반	{스트레스, 문제, 조심, 지적}=>{위험}	0.001163219	0.9080460	7.419833
	{실패, 결심, 욕구}=>{성공}	0.001413532	0.7804878	7.227547
	{스트레스, 실패, 결심}=>{성공}	0.001516602	0.7518248	6.962119
	{스트레스, 조심, 지적}=>{위험}	0.001214754	0.7894737	6.450954
	{스트레스, 위험, 조심, 지적}=>{문제}	0.001163219	0.9575758	3.553563
	{위험, 조심, 지적}=>{문제}	0.001641758	0.9330544	3.462564
	{스트레스, 위험, 지적}=>{문제}	0.002061400	0.8562691	3.177614
	{위험, 의지, 지적}=>{문제}	0.001119046	0.8491620	3.151240
	{위험, 지적, 걱정}=>{문제}	0.001523964	0.8448980	3.135416
	{성공, 지적, 걱정}=>{문제}	0.001096959	0.8370787	3.106398

[그림 3-3] 담배 감정의 연관규칙에 대한 병렬좌표 시각화



〈표 3-2〉와 같이 담배와 관련하여 긍정적 감정(위험군)을 나타내는 온라인 문서(버즈)는 32.8%, 보통의 감정(잠재군)을 나타내는 버즈는 8.6%, 부정의 감정(일반군)은 58.6%로 나타났다. 담배와 관련한 정책은 담배값 인상(50.0%), 흡연구제(23.9%), 금연사업(9.3%) 등의 순으로 나타났다. 담배와 관련한 도움·치료는 병원(71.7%), 금연클리닉(25.6%), 금연교실(1.8%), 금연상담전화(0.9%)의 순으로 나타났다. 담배와 관련한 폐해로는 간접흡연(48.6%), 담배꽂초(16.2%), 알코올(8.9%), 중독(7.3%) 등의

순으로 나타났다. 담배와 관련한 유해물질로는 니코틴(48.6%), 발암물질(14.3%), 타르(11.9%), 유해물질(8.0%) 등의 순으로 나타났다. 담배와 관련한 채널로는 SNS(52.9%), 카페(24.6%), 블로그(13.5%) 등의 순으로 나타났다. 담배와 관련한 금연도구로는 전자담배(50.0%), 금연보조제(33.1%), 금연패치(6.0%) 등의 순으로 나타났다. 담배와 관련한 장소로는 공공장소(19.9%), 식당(15.5%), 학교(15.4%) 등의 순으로 나타났다. 담배와 관련한 기관으로는 청와대(48.6%), 보건복지부(11.9%), 국회(8.9%) 등의 순으로 나타났다. <표 3-3>과 같이 담배와 관련한 연도별 긍정적 감정(위험)을 나타내는 버즈는 청소년의 경우 2011년(30.9%), 2012년(29.2%), 2013년(27.2%), 2014년(26.8%), 2015년(25.3%)로 나타났으며, 성인의 경우 2011년(39.9%), 2012년(36.4%), 2013년(34.0%), 2014년(36.2%), 2015년(29.1%)로 나타났다. 따라서 2015년 담배값 인상은 청소년 보다 성인에게 더 많은 영향을 미친 것으로 나타났다.

<표 3-2> 담배 관련 버즈 현황

구분	항목	N(%)	구분	항목	N(%)
감정	일반	110,401(58.6)	채널	블로그	147,235(13.5)
	잠재	16,206(8.6)		카페	268,463(24.6)
	위험	61,660(32.8)		SNS	577,125(52.9)
	계	188,267		게시판	53,243(4.9)
정책	담배값인상	58,267(50.0)		뉴스	45,892(4.2)
	FCTC	454(0.4)		계	1,091,958
	금연관련법	13,528(11.6)	도구	금연껌	4,260(7.5)
	흡연규제	27,828(23.9)		금연약	4,778(8.4)
	금연광고	5,626(4.8)		전자담배	38,600(68.1)
	금연사업	10,887(9.3)		금연패치	5,200(9.2)
도움	계	116,590		금연보조제	3,819(6.7)
	금연클리닉	10,015(25.6)		계	56,657
	금연상담전화	360(0.9)	장소	PC방	3,932(2.1)
	병원	28,062(71.7)		가정	12,414(6.7)
	금연교실	694(1.8)		금연건물	1,609(0.9)
	계	39,131		아파트	14,915(8.1)

구분	항목	N(%)	구분	항목	N(%)
폐해	간접흡연	88,855(48.6)	기관	공공장소	34,775(18.9)
	알코올	16,217(8.9)		흡연구역	14,029(7.6)
	중독	13,333(7.3)		직장	19,666(10.7)
	기억력	12,396(6.8)		술집	26,000(14.1)
	담배꽁초	29,651(16.2)		식당	28,659(15.5)
	도박마약	8,469(4.6)		학교	28,314(15.4)
	이혼	3,333(1.8)		계	184,313
	정신건강	3,907(2.1)		청와대	32,311(48.6)
	폭력	6,611(3.6)		국회	5,803(8.7)
유해 물질	계	182,772	기관	보건복지부	7,894(11.9)
	니코틴	18,496(48.6)		여성가족부	1,502(2.3)
	발암물질	5,651(14.3)		기획재정부	2,971(4.5)
	유해물질	3,142(8.0)		지방자치단체	4,963(7.5)
	일산화탄소	2,894(7.3)		공공기관	4,021(6.1)
	타르	4,707(11.9)		세계보건기구	2,227(3.4)
	화학물질	1,875(4.7)		금연단체	1,079(1.6)
	노폐물	2,745(6.9)		담배회사	3,673(5.5)
	계	39,510		계	66,444

〈표 3-3〉 담배 관련 연도별 감정변화

연도	청소년				성인				전체			
	위험	잠재	일반	계	위험	잠재	일반	계	위험	잠재	일반	계
2011	2,283 (30.9)	1,017 (13.8)	4,085 (55.3)	7,385	6,873 (39.9)	1,747 (10.2)	8,589 (49.9)	17,209	9,156 (37.2)	2,764 (11.2)	12,674 (51.5)	24,594
2012	2,767 (29.2)	1,174 (12.4)	5,542 (58.4)	9,483	10,537 (36.4)	1,938 (6.7)	16,508 (57.0)	28,983	13,304 (34.6)	3,112 (8.1)	22,050 (57.3)	38,466
2013	2,478 (27.9)	1,178 (13.3)	5,228 (58.8)	8,884	11,037 (34.0)	1,956 (6.0)	19,432 (59.9)	32,425	13,515 (32.7)	3,134 (7.6)	24,660 (59.7)	41,309
2014	2,145 (26.8)	1,050 (13.1)	4,799 (60.0)	7,994	9,106 (36.2)	1,906 (7.6)	14,125 (56.2)	25,137	11,251 (34.0)	2,956 (8.9)	18,924 (57.1)	33,131
2015	2,160 (25.3)	1,068 (12.5)	5,324 (62.3)	8,552	12,274 (29.1)	3,172 (7.5)	26,769 (63.4)	42,215	14,434 (28.4)	4,240 (8.4)	32,093 (63.2)	50,767
계	11,833 (28.0)	5,487 (13.0)	24,978 (59.1)	42,298	49,827 (34.1)	10,719 (7.3)	85,423 (58.5)	145,969	61,660 (32.8)	16,206 (8.6)	110,401 (58.6)	188,267

나. 담배 위험 관련 연관성 분석

〈표 3-4〉와 같이 정책요인에 대한 담배 위험에 연관성 예측에서 가장

신뢰도가 높은 연관규칙으로는 {담배값인상, 금연관련법} ⇒ {일반}이며 세 변인의 연관성은 지지도 0.002, 신뢰도는 0.539, 향상도는 5.338로 나타났다. 이는 온라인 문서에서 담배값인상, 금연관련법이 언급되면 담배를 부정적(일반)으로 생각할 확률이 53.9%이며, 담배값인상, 금연관련법이 언급되지 않은 문서 보다 담배에 대한 감정이 일반일 확률이 5.34배 높아지는 것을 나타낸다. 특히, {담배값인상} ⇒ {위험} 두 변인의 연관성은 지지도 0.002, 신뢰도는 0.04, 향상도는 0.72로 나타나 담배값인상은 담배에 대한 긍정적(위험) 감정에 감소시키는 것으로 나타났다. 반면, {담배값인상} ⇒ {잠재} 두 변인의 연관성은 지지도 0.001, 신뢰도는 0.02, 향상도는 1.49로 나타나 담배값인상은 담배에 대한 보통(잠재) 감정을 증가시키는 것으로 나타났다.

〈표 3-4〉 정책요인에 대한 담배 위험 예측

규칙	지지도	신뢰도	향상도
{담배값인상, 금연관련법} ⇒ {일반}	0.001776625	0.53978854	5.3389590
{담배값인상, 흡연규제} ⇒ {일반}	0.001778457	0.49465104	4.8925115
{금연사업} ⇒ {일반}	0.004685162	0.46991825	4.6478836
{금연관련법, 흡연규제} ⇒ {일반}	0.001862709	0.46395985	4.5889500
{금연관련법} ⇒ {일반}	0.005130234	0.41410408	4.0958349
{금연광고} ⇒ {일반}	0.001716183	0.33309634	3.2946007
{흡연규제} ⇒ {일반}	0.007551572	0.29632025	2.9308545
{담배값인상} ⇒ {일반}	0.010004048	0.18748176	1.8543511
{금연사업} ⇒ {위험}	0.001788530	0.17938826	3.1768480
∅ ⇒ {일반}	0.101103705	0.10110371	1.0000000
{흡연규제} ⇒ {위험}	0.001625520	0.06378468	1.1295846
∅ ⇒ {위험}	0.056467373	0.05646737	1.0000000
{담배값인상} ⇒ {위험}	0.002182318	0.04089794	0.7242755
{담배값인상} ⇒ {잠재}	0.001182280	0.02215662	1.4929102
∅ ⇒ {잠재}	0.014841230	0.01484123	1.0000000

다. 담배의 위험에 영향을 미치는 요인

〈표 3-5〉와 같이 금연과 관련한 모든 정책 요인은 담배의 위험에 부적인 영향을 미치는 것으로 나타나, FCTC, 담배값 인상, 금연관련법, 흡연규제, 금연광고, 금연사업과 관련한 정책이 온라인 상에 많이 언급될수록 담배에 대한 긍정적 감정(위험)은 감소하는 것으로 나타났다. 금연과 관련한 도구 요인의 영향은 금연약, 금연패치, 금연껌은 부적인 영향을 미치는 것으로 나타나, 금연약, 금연패치, 금연껌과 관련한 금연도구가 온라인 상에 많이 언급될수록 담배에 대한 긍정적 감정(위험)은 감소하는 것으로 나타났으나, 전자담배와 금연보조제는 정적인 영향을 미치는 것으로 나타나 전자담배와 금연보조제와 관련한 금연도구가 많이 언급될수록 담배에 대한 긍정적 감정(위험)은 증가하는 것으로 나타났다.

〈표 3-5〉 담배의 위험에 영향을 미치는 정책 및 도구 요인¹⁾

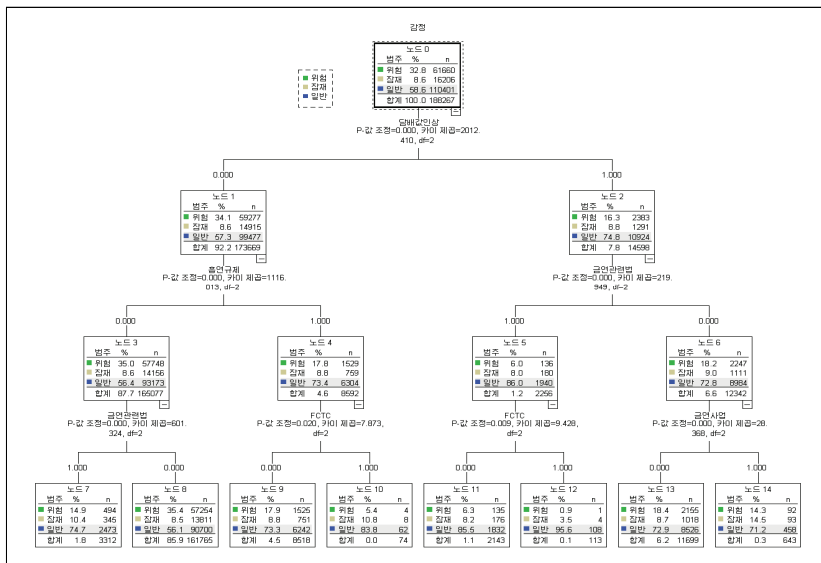
변수		위험				잠재			
		b	S.E. [†]	OR [§]	P	b [†]	S.E. [†]	OR [§]	P
정책	담배값인상	-.854	.024	.426	.000	-.207	.031	.813	.000
	FCTC	-1.328	.269	.265	.000	-.451	.215	.637	.036
	금연관련법	-.845	.037	.430	.000	-.153	.044	.858	.001
	흡연규제	-.742	.027	.476	.000	-.191	.036	.826	.000
	금연광고	-.275	.049	.760	.000	.076	.065	1.079	.240
	금연사업	-.242	.028	.785	.000	.410	.035	1.507	.000
도구	금연껌	-.357	.051	.700	.000	.068	.069	1.071	.324
	금연약	-1.556	.060	.211	.000	-.176	.058	.839	.003
	전자담배	.206	.019	1.229	.000	.155	.032	1.167	.000
	금연패치	-1.091	.051	.336	.000	-.414	.065	.661	.000
	금연보조제	.374	.060	1.454	.000	.688	.081	1.990	.000

주: 1) 기본범주: 일반, † Standardized coefficients, ‡ Standard error, § Adjusted odds ratio

라. 담배 관련 위험 예측모형

본 연구에서는 담배 관련 위험을 예측하기 위하여 담배와 관련한 정책 요인과 금연도구요인에 대해 데이터마이닝 분석을 실시하였다. 담배 관련 정책요인이 담배의 위험예측 모형에 미치는 영향은 [그림 3-4]와 같다.

[그림 3-4] 정책요인의 예측모형



나무구조의 최상위에 있는 네모는 루트노드로서, 예측변수(독립변수)가 투입되지 않은 종속변수(위험, 잠재, 일반)의 빈도를 나타낸다. 루트노드에서 담배의 위험은 32.8%(61,660건), 잠재는 8.5%(16,206건), 일반은 58.6%(110,401건)으로 나타났다. 루트노드의 하단의 가장 상위에 위치하는 요인은 담배의 위험예측에 가장 영향력이 높은(관련성이 깊은) 정책요인으로 '담배값인상요인'의 영향력이 가장 큰 것으로 나타났다. '담배값인상요인'이 있을 경우 담배의 위험은 이전의 32.8%에서 16.3%로

크게 감소한 반면, 잠재는 이전의 8.5%에서 8.8%, 일반은 이전의 58.6%에서 74.8%로 증가하였다. ‘담배값인상요인’이 있고 ‘금연관련법요인’이 있는 경우 담배의 위험은 이전의 16.3%에서 6.0%, 잠재는 이전의 8.8%에서 8.0%로 감소한 반면, 일반은 이전의 74.8%에서 88.0%로 증가하였다. <표 3-6>의 담배와 관련한 정책요인의 위험예측 모형에 대한 이익도표와 같이 담배의 위험에 가장 영향력이 높은 경우는 ‘담배값인상요인’이 없고 ‘흡연구제요인’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 위험이 높을 확률이 1.08배로 나타났다. 담배의 잠재에 가장 영향력이 높은 경우는 ‘담배값인상요인’이 있고 ‘금연관련법요인’이 없으며 ‘금연사업요인’이 있는 조합으로 나타났다. 즉, 14번 노드의 지수(index)가 168.0%로 뿌리마디와 비교했을 때 14번 노드의 조건을 가진 집단이 담배에 대한 보통의 감정(잠재)이 높을 확률이 1.68배로 나타났다. 담배의 일반에 가장 영향력이 높은 경우는 ‘담배값인상요인’이 있고 ‘금연관련법요인’이 있으며 ‘FCTC요인’이 있는 조합으로 나타났다. 즉, 12번 노드의 지수(index)가 163.0%로 뿌리마디와 비교했을 때 12번 노드의 조건을 가진 집단이 담배에 대한 부정의 감정(일반)이 높을 확률이 1.63배로 나타났다.

<표 3-6> 정책요인의 예측 모형에 대한 이익도표

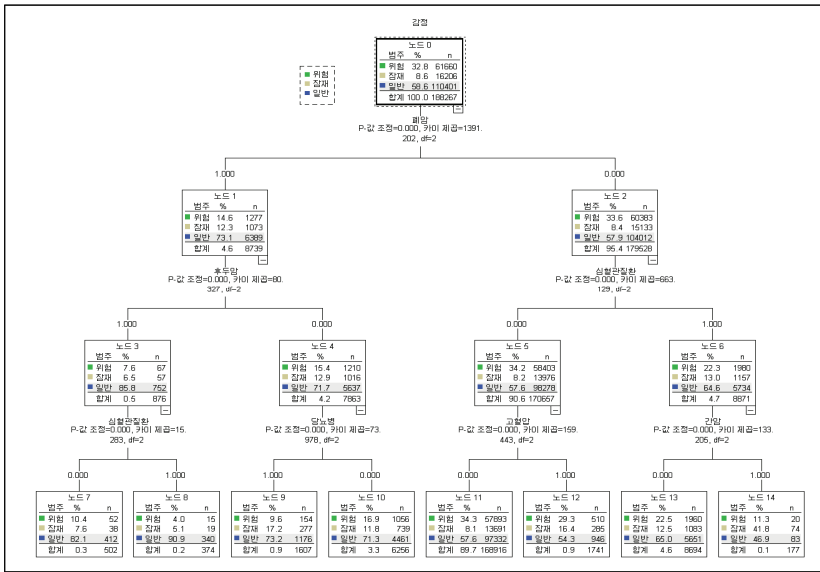
구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
위험	8	161765	85.9	92.9	108.1	161765	85.9	92.9	108.1
	13	11699	6.2	3.5	56.2	173464	92.1	96.3	104.6
	9	8518	4.5	2.5	54.7	181982	96.7	98.8	102.2
	7	3312	1.8	.8	45.5	185294	98.4	99.6	101.2
	14	643	.3	.1	43.7	185937	98.8	99.8	101.0
	11	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

구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
잠재	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
일반	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

담배의 위험예측에 가장 영향력이 높은 질병요인으로 ‘폐암’의 영향력이 가장 큰 것으로 나타났다. ‘폐암’이 있을 경우 담배의 위험은 이전의 32.8%에서 14.6%로 크게 감소한 반면, 잠재는 이전의 8.6%에서 12.3%, 일반은 이전의 58.6%에서 73.1%로 증가하였다. ‘폐암’이 있고 ‘후두암’이 있는 경우 담배의 위험은 이전의 14.6%에서 7.6%, 잠재는 이전의 12.3%에서 6.5%로 감소한 반면, 일반은 이전의 73.1%에서 85.8%로 증가하였다. <표 3-7>의 담배와 관련한 질병요인의 위험예측 모형에 대한 이익도표와 같이 담배의 위험에 가장 영향력이 높은 경우는 ‘폐암’이 없고 ‘심혈관질환’이 없으며 ‘고혈압’이 없는 조합으로 나타났다. 즉, 11번 노드의 지수(index)가 104.6%로 뿌리마디와 비교했을 때 11번 노드의 조건을 가진 집단이 담배에 대한 위험이 높을 확률이 1.05배로 나타났다. 담배의 잠재에 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 집단이 담

배에 대한 부정의 감정(일반)이 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 집단이 담배에 대한 부정의 감정(일반)이 높을 확률이 1.55배로 나타났다.

[그림 3-5] 질병요인의 예측모형



〈표 3-7〉 질병요인의 예측 모형에 대한 이익도표

구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
위험	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

구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
잠재	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
일반	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

4. 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, 흡연규

제, 금연광고, 금연사업과 관련된 정책이 온라인상에 많이 언급될수록 위험군이 감소하는 것으로 나타났다. 금연약, 금연패치, 금연껌과 같은 도구가 온라인상에 많이 언급될수록 위험군은 감소하는 것으로 나타났다. 다섯째, 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.

- 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 mMrality : Beyond Established Causes. *N Engl J Med*. 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 General*. GA: US Department of Health and Human Services, Atlanta.
- Centers for Disease Control and Prevention. (2014). Cigarette Prices and Smoking Prevalence after a Tobacco Tax Increase-Turkey, 2008 and 2012. *MMWR Morb Mortal Wkly Rep*. 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.
- 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. *Journal 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 VI*, 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. (2010). Standardization for Basic Association Measures in Association Rule Mining. *Journal of the Korean Data & Information Science Society*, 21, 891~899.
- 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. *PLoS Med*, 11, e1001631.

제 4 장

소셜 빅데이터를 활용한 통일인식 동향 분석 및 예측

1. 서론
 2. 연구 방법
 3. 연구 결과
 4. 결론
- 참고문헌

4

소셜 빅데이터를 활용한 << 통일인식 동향 분석 및 예측>>¹⁶⁾

1. 서론

현 정부는 2013년 출범 이후 튼튼한 안보를 바탕으로 남북한 신뢰를 형성함으로써 남북관계를 발전시키고 한반도에 평화를 정착시키며 통일 기반을 구축하려는 한반도 신뢰 프로세스(통일교육원, 2013: p.3)를 통한 남북관계 정상화를 주요 국정과제로 설정한데 이어, 2014년 신년 기자회견에서 박근혜 대통령이 발언한 ‘통일은 대박이다’라는 한마디로 평화통일의 기반 구축에 적극적으로 나설 것임을 선언하였다. 과거 통일논의는 통일비용과 대북정책을 중심으로 이루어지면서 통일의 부정적 측면이 부각되었으나, 통일대박론은 편익중심의 통일논의를 선언한 것으로(통일연구원, 2014: p.7), 통일이 되면 천문학적인 통일비용과 사회적 혼란이 야기될 것이라는 부정적 인식을 극복하고 통일을 기회와 희망으로 보는 긍정적 통일담론을 확산시키기 위한 것이다. 통일대박론은 통일 논의의 위축에서 벗어나 통일문제에 대한 국민적 합의를 이루는 계기가 되었다는 긍정적 주장(김창수, 2014: p.120.)과 함께 현 정부는 통일대박론을 시작으로 대통령 직속으로 통일준비위원회¹⁷⁾를 출범하여 통일준비에 박차를 가하게 되었다. 한반도 통일은 남북한의 문제이자 동북아 주변

16) 본 연구의 일부내용은 “The 11th International Conference on Multimedia Information Technology and Applications (MITA 2015). ‘Predicting Koreans’ Perceptions About Reunification Using Social Big Data. Tae Min Song, Juyoung Song(교신저자), Dal Lae Jin’와 ‘송태민(2015). 소셜 빅데이터를 활용한 통일인식 동향 분석 및 예측. 북한경제리뷰. KDI’에 발표된 논문임을 밝힌다.

17) 대통령을 위원장으로 하여 총 50명의 위원으로 구성된 통일준비위원회 출범(위원장, 민간위원 30명, 국회의원 2명, 정부위원 11명, 국책연구기관장 6명으로 구성)

국의 미래를 좌우할 국제적 사안으로(김규륜, 2013), 한반도 평화체제와 본격적인 통일과정에서 국제사회의 지지는 필수불가결한 요소이며(차문석, 2013) 국내 차원에서는 통일에 대한 국민적 공감대 형성과 통일을 맞이할 수 있는 역량 구축이 필요하다. 따라서 통일을 위해서 국제차원의 한반도 통일에 대한 국제적 공감대 형성과 미·중·일·러 주변 4국의 협조와(이규창, 2014) 남북 간의 신뢰 구축이 필요할 것이다.

평화로운 통일한국을 실현하기 위해서는 주변국들의 반응과 함께 우리 국민들의 통일에 대한 인식과 태도를 분석하는 것이 필요하다. 통일연구원은 2012년 7월 19세 이상 성인 남녀 1,000명 대상의 전화조사 결과 63.1%가 통일의 필요성에 공감한다고 응답하였다(김규륜·김형기, 2012: p.14). 서울대 통일평화연구원은 2007년부터 매년 7월부터 8월 사이 정기적으로 전국의 성인 1,200명을 대상으로 면대면 설문조사를 실시하여 통일이 필요하다는 인식은 2014년 55.8%로 나타났다. 아산정책연구원은 2014년 3월 만 19세 이상 성인남녀 1,000명을 대상으로 통일인식 여론조사를 실시한 결과 80.5%가 통일이 필요하다고 보고 있다¹⁸⁾. 이와 같이 그동안 우리사회는 대북정책의 주요한 기초자료로 사용하기 위하여 통일·북한 관련 연구소나 대학과 정부 산하기관 등에서 일반 국민들을 대상으로 통일의식조사를 정기적으로 시행해 왔다(강동완·박정란, 2014:p 2.).

한편, 모바일 인터넷과 소셜 미디어의 확산으로 데이터량이 증가하여 데이터의 생산, 유통 소비체계에 큰 변화가 일어나면서 데이터가 경제적 자산이 될 수 있는 빅데이터 시대가 도래되었다. 많은 국가에서는 빅데이터가 공공과 민간에 미치는 파급효과를 전망함에 따라 빅데이터의 활용은 정부의 정책을 효과적으로 추진하기 위한 새로운 성장동력이 될 것으로 예측하고 있다(김정선 외, 2014). 세계 각국의 정부와 기업들은 SNS

18) 아산정책연구원 2014년 4월 8일 보도자료.

를 통해 생산되는 소셜 빅데이터의 활용과 분석을 통하여 사회적 문제의 해결은 물론 새로운 경제적 효과와 일자리 창출을 위하여 적극적으로 노력하고 있으며, 우리나라는 정부 3.0과 창조경제의 추진과 실현을 위하여 다양한 분야에 빅데이터의 효율적 활용을 적극적으로 모색하고 있다. 소셜 빅데이터의 분석은 사용자가 남긴 문서의 의미를 분석하는 것으로 자연어 처리기술인 주제분석(text mining)과 감성분석 기술인 오피니언 마이닝(opinion mining)을 실시한 후, 네트워크 분석(network analysis)과 통계분석(statistical analysis)을 실시해야 한다. 우리 국민의 통일에 대한 다양한 인식을 살펴보기 위하여 그동안 실시하던 설문조사는 정해진 변인에 대한 개인과 집단의 관계를 보는 데는 유용하나, 사이버 상에 언급된 개인별 담론(buzz)가 사회적 현상들과 얼마나 어떻게 연관되어 있는지 밝히고 원인을 파악하는 데는 한계가 있다(송주영·송태민, 2014). 이에 반해 소셜 빅데이터의 분석은 훨씬 방대한 량의 데이터를 활용하여 다양한 참여자의 생각과 의견을 확인할 수 있기 때문에 기존의 오프라인 조사와 함께 활용하면 통일 인식의 예측을 보다 정확하게 할 수 있다. 본 연구는 우리나라 온라인 뉴스사이트, 블로그, 카페, SNS, 게시판 등에서 수집한 소셜 빅데이터를 바탕으로 우리나라 국민의 통일의 인식에 대한 동향을 분석하고 통일인식의 예측모형과 연관규칙을 파악한다.

2. 연구 방법

가. 연구대상

본 연구는 국내의 온라인 뉴스 사이트, 블로그, 카페, 소셜 네트워크 서비스, 게시판 등 인터넷을 통해 수집된 소셜 빅데이터를 대상으로 하였다. 본 분석에서는 160개의 온라인 뉴스사이트, 4개의 블로그(네이버, 네

이트, 다음, 티스토리), 2개의 카페(네이버, 다음), 1개의 SNS(트위터), 9개의 게시판(네이버지식인, 네이버지식, 네이버톡, 다음지식인, 다음아고라, SLR클럽 등) 등 총 176개의 온라인 채널을 통해 수집 가능한 텍스트 기반의 온라인 문서(버즈)를 소셜 빅데이터로 정의하였다. 통일 관련 토픽(topic)¹⁹⁾의 수집은 2011년 ~ 2015년의 매년 1/4분기(1월 ~ 3월) 기간 동안(총 15개월)²⁰⁾ 해당 채널에서 요일, 주말, 휴일을 고려하지 않고 매 시간단위로 수집하였으며, 수집된 총 411,135건(2011년: 10,211건, 2012년: 53,884건, 2013년: 83,268건, 2014년: 166,952건, 2015년: 96,820건)의 텍스트(Text) 문서²¹⁾를 본 연구의 분석에 포함시켰다. 통일 토픽은 모든 관련 문서를 수집하기 위해 ‘통일’을 사용하였으며, 토픽과 같은 의미로 사용되는 토픽 유사어로는 ‘남북통일, 한반도통일’ 용어를 사용하였다. 그리고 불용어는 ‘통일신라, 통일교 등’을 사용하였다. 본 연구를 위한 소셜 빅데이터의 수집²²⁾은 크롤러(Crawler)를 사용하였고, 이후 주제분석을 통해 분류된 명사형 어휘를 유목화(categorization)하여 분석요인으로 설정하였다.

나. 연구도구

통일과 관련하여 수집된 문서는 주제분석²³⁾의 과정을 거쳐 다음과 같

19) 토픽은 소셜 분석 및 모니터링의 ‘대상이 되는 주제어’를 의미하며, 문서 내에 관련 토픽이 포함된 문서를 수집함.

20) 본 연구의 연구대상은 2014년 1월 2일 통일 대박 발언 이후, 통일담론이 확산됨에 따라 연도별 통일인식 비교를 위해 1/4분기를 분석시기로 결정함.

21) 수집된 문서는 SNS(82.3%, 338,501건), 뉴스(6.0%, 24,693건), 블로그(5.4%, 22,220건), 카페(4.8%, 19,785건), 게시판(1.4%, 5,939건)의 순으로 나타남.

22) 본 연구를 위한 소셜 빅데이터의 수집 및 토픽 분류는 ‘(주)SK텔레콤 스마트인사이트’에서 수행함.

23) 주제분석에 사용되는 사전은 ‘21세기 세종계획’과 같은 범용사전도 있지만 대부분 분석의 목적에 맞게 사용자가 설계한 사전을 사용하게 된다. 본 연구의 통일관련 주제분석

이 정형화 데이터로 코드화하여 사용하였다.

1) 통일 관련 감정

본 연구의 통일 감정 키워드는 문서 수집 이후, 주제 분석을 통하여 총 56개(대박, 문제, 평화, 필요, 통일대박, 애국심, 중요, 노력, 반대, 쪽박, 관심, 희망, 신뢰, 우려, 불가능, 잘못, 사랑, 성공, 부정, 최고, 재앙, 위협, 행복, 공감, 빨갱이, 평화적, 천박, 위험, 염원, 포기, 비난, 고통, 강력, 걱정, 갈등, 분열, 혼란, 위대, 환영, 경박, 압박, 비아냥, 조롱, 든든, 불신, 환장, 친절, 한심, 다행, 응원, 긴장, 아픔, 천박한, 소중, 부담, 충격) 키워드로 분류하였다. 본 연구에서는 56개의 통일 감정 키워드(변수)가 가지는 통일 인식의 정도를 판단하기 위해 2차 요인분석을 통하여 14개의 요인(38개 변수)으로 축약을 실시한 후, 감성분석을 실시하였다. 요인분석에서 결정된 2개의 요인에 대한 주제어의 의미를 파악하여 ‘찬성, 보통, 반대’로 감성분석을 실시하였다. 일반적으로 감성분석은 긍정과 부정의 감성어 사전으로 분석해야 하나, 본 연구에서는 요인분석의 결과로 분류된 주제어의 의미를 파악하여 감성분석을 실시하였다. 따라서 본 연구에서 찬성(13개: 관심, 필요, 중요, 노력, 신뢰, 평화(평화적), 든든, 다행, 공감, 행복, 사랑, 위대, 대박), 반대(18개: 쪽박, 천박, 경박, 빨갱이, 환장, 친절, 한심, 문제, 위협, 압박, 포기, 재앙, 혼란, 분열, 갈등, 아픔, 고통, 불신)로 분류하였다. 그리고 찬성과 반대의 감정을 동일한 횟수로 표현한 문서는 보통의 감정으로 분류하였다.

은 SKT에서 관련 문서 수집 후 원시자료(raw data)에서 나타난 상위 2,000개의 키워드들을 대상으로 유목화를 하여 사용자 사전을 구축하였다.

2) 통일에 대한 정책

통일에 대한 정책의 정의는 주제 분석의 과정을 거쳐 '진보, 보수, 공산주의, 사회주의, 대북정책(대북정책, 정책, 통일정책), 창조경제(창조경제, 박근혜정권), 민주주의(민주정부, 민주주의), 햇볕정책'의 8개 정책으로 정책이 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

3) 통일에 대한 안보

통일에 대한 안보의 정의는 주제 분석 과정을 거쳐 '핵무기(미사일, 핵무기, 도발, 전쟁), 정상회담(회담, 정상회담, 비핵화), 휴전선(휴전선, DMZ, 판문점), 남북공동선언(104선언, 남북공동선언), 간첩(고정간첩, 간첩)'의 5개 안보로 안보가 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

4) 통일에 대한 이슈

통일에 대한 이슈의 정의는 주제 분석을 거쳐 '천안함(연평도, 천안함), 민영화(민영화, 특검), 이산가족상봉(이산가족상봉, 이산가족), 선거(대선, 부정선거, 지방선거, 선거), 통일대박(통일대박론, 대박, 대박론)'의 5개 이슈로 이슈가 있는 경우는 '1', 없는 경우는 '0'으로 코드화 하였다.

5) 통일에 대한 주변국가

통일에 대한 주변국가의 정의는 주제 분석 과정을 거쳐 '중국, 독일(독일, 동독, 서독), 일본, 미국, 러시아(소련, 러시아), 유럽(유럽, 영국, 스위

스, 네덜란드)’의 6개로 해당 국가가 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

6) 통일에 대한 관련기관

통일에 대한 관련기관의 정의는 요인분석과 주제 분석의 과정을 거쳐 ‘범민련(범청학련, 범민련, 조국통일범민족연합, 한총련), 시민단체(실천연대, 시민단체, 전교조), 정부(외교부, 국방부, 통일부, 정부, 법무부, 국가보훈처, 해수부), 청와대(청와대, 박근혜정부), 국정원, 통일준비위원회(통일준비위원회, 통일준비위), 조국통일평화위원회, 자유통일포럼, 국회(국회, 외교통일위원회)’의 9개 관련기관으로 해당 관련기관이 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

7) 통일에 대한 정당

통일에 대한 정당의 정의는 주제 분석의 과정을 거쳐 ‘새정치민주연합(민주당, 새정치, 새천년민주, 새정치민주연합), 새누리당(새누리, 보수당, 새누리당), 통합진보당(통진당, 통합진보당), 자유선진당, 노동당’의 5개 정당으로 해당 정당이 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

8) 통일에 대한 통일방식

통일에 대한 통일방식의 정의는 주제 분석의 과정을 거쳐 ‘흡수통일(흡수, 흡수통일), 자유통일(자유통일, 자유북진통일, 북진통일), 평화통일, 선진통일, 공존통일’의 5개 통일방식으로 해당 통일방식이 있는 경우는

‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

다. 분석방법

본 연구에서는 대국민 통일인식을 설명하는 가장 효율적인 예측모형을 구축하기 위해 특별한 통계적 가정이 필요하지 않은 데이터마이닝(data mining)의 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 척도는 지지도(support), 신뢰도(confidence), 향상도(lift)로 나타낼 수 있다. 연관분석 과정은 연구자가 지정한 최소 지지도를 만족시키는 빈발항목집합(frequent item set)을 생성한 후, 이들에 대해 최저 신뢰도 기준을 마련하고 향상도가 1인 이상인 것을 규칙으로 채택한다(박희창, 2010). 본 연구의 연관분석은 선험적 규칙(apriori principle)²⁴⁾ 알고리즘을 사용하였다. 본 연구의 통일감정에 사용된 연관분석의 척도는 지지도 0.001, 신뢰도 0.15를 기준으로 시뮬레이션하였다. 데이터마이닝의 의사결정나무 분석은 방대한 자료 속에서 종속변인을 가장 잘 설명하는 예측모형을 자동적으로 산출해 줌으로써 각기 다른 속성을 가진 통일 인식에 대한 요인을 쉽게 파악할 수 있다. 본 연구의 의사결정나무 형성을 위한 분석 알고리즘은 CHAID(Chi-squared Automatic Interaction Detection)를 사용하였다. CHAID(Kass, 1980)는 이산형인 종속변수의 분리기준으로

24) 한 항목집합이 빈발하다면 이 항목집합의 모든 부분집합은 역시 빈발항목집합이며, 한 항목집합이 비빈발하다면 이 항목집합을 포함하는 모든 집합은 비빈발항목 집합이다(이정진, 2011, p123).

카이제곱(χ^2 -검정)을 사용하며, 모든 가능한 조합을 탐색하여 최적분리를 찾는다. 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, 다중응답분석, 로지스틱 회귀분석, 의사결정나무분석은 SPSS v. 22.0을 사용하였고, 연관분석과 시각화 분석은 R version 3.1.3, 소셜 네트워크 분석은 NetMiner²⁵⁾를 사용하였다.

3. 연구 결과

가. 통일 관련 문서(버즈) 현황

통일과 관련된 이슈발생 시에 커뮤니케이션이 급증하는 양상을 보이고 있으며, 특히 2014년 첫 신년 기자회견에서 언급된 박근혜 대통령의 통일대박론(통일은 대박이다) 강조 이후 문서량이 급증한 것으로 나타났다(그림 4-1 참조).

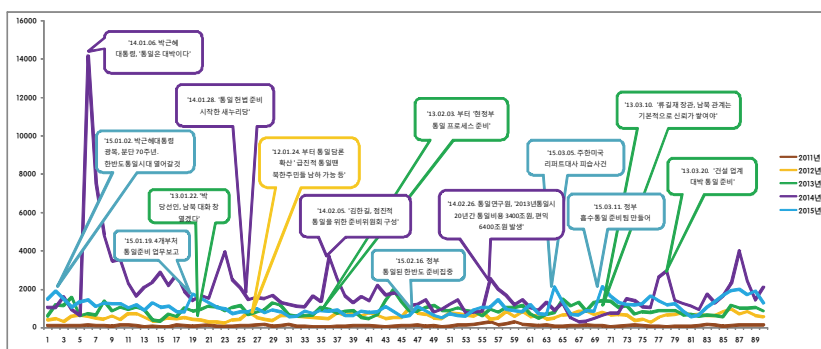
[그림 4-2]와 같이 연도별 통일찬성의 감정변화는 대박을 제외하고 2011년 대비 평균 2.23배 증가(평화 2.42배, 필요 1.68배, 중요 1.54배, 노력 1.95배, 관심 1.75배 등) 하였으며, 찬성 감정의 표현 단어는 평화, 필요, 중요, 노력, 관심 등의 순으로 집중된 것으로 나타났다. 특히 대박은 2011년 56건에서 2014년 47,480건²⁶⁾으로 크게 증가한 것으로 나타났다. 연도별 통일반대의 감정변화는 쪽박, 천박, 경박을 제외하고 2011년 대비 평균 3.25배 증가(문제 1.71배, 위협 1.64배, 갈등 1.56배, 포기

25) NetMiner v4.2.0.140122 Seoul: Cyram Inc.

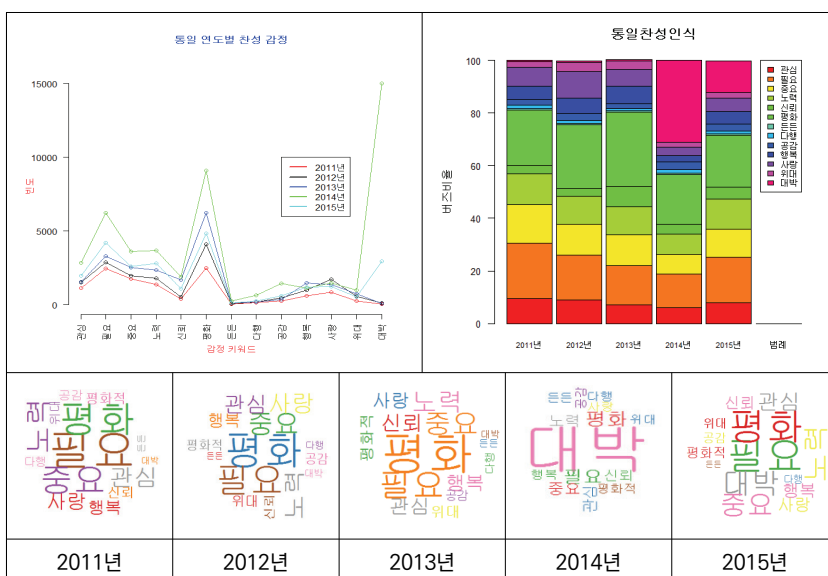
26) [그림 4-2]의 통일 찬성인식에서 '대박'은 15,000건으로 표기함.

1.58배, 분열 2.29배 등) 하였으며, 반대 감정의 표현 단어는 문제, 위협, 갈등, 포기, 분열 등의 순으로 집중된 것으로 나타났다. 특히, 쪽박, 천박, 절박은 2011년 30건에서 2014년 5,254건으로 증가한 것으로 나타났다.

〔그림 4-1〕 통일 관련 문서(버즈)량 일별 추이

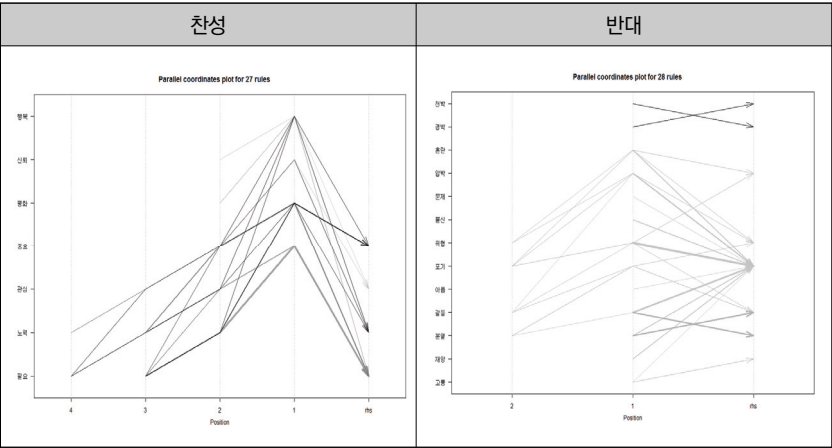


[그림 4-2] 통일 연도별 감정 변화



	규칙	지지도	신뢰도	향상도
반대	{관심, 중요, 평화}=>{필요}	0.001753669	0.6986434	15.07262
	{경박}=>{천박}	0.001041013	0.5638999	163.729283
	{천박}=>{경박}	0.001041013	0.3022599	163.729283
	{분열}=>{갈등}	0.003877044	0.3904949	27.637678
	{문제, 분열}=>{갈등}	0.001145601	0.3732171	26.414828
	{문제, 압박}=>{위협}	0.001087226	0.3284350	20.269003
	{문제, 포기}=>{위협}	0.001578545	0.3042663	18.777459
	{포기, 갈등}=>{문제}	0.001028852	0.7050000	12.094817
	{위협, 압박}=>{문제}	0.001087226	0.6908810	11.852595
	{위협, 포기}=>{문제}	0.001578545	0.6781609	11.634372
	{위협, 갈등}=>{문제}	0.001493416	0.6623517	11.363152

[그림 4-3] 통일인식 감정의 연관규칙에 대한 병렬좌표 시각화



〈표 4-2〉와 같이 통일과 관련하여 통일대박 감정을 포함하는 경우 통일을 찬성하는 문서는 68.8%, 중립의 문서는 10.2%, 반대하는 문서는 21.0%로 나타났다. 통일관련 정책의 문서는 대북정책(37.7%), 민주주의(18.6%), 보수(11.2%), 진보(10.3%) 등의 순으로 나타났다. 통일관련 안보의 문서는 핵무기(52.7%), 정상회담(22.0%), 휴전선(12.3%) 등의 순

으로 나타났다. 통일관련 이슈의 문서는 통일대박(62.2%), 선거(11.2%), 천안함(7.3%), 이산가족상봉(6.8%) 등의 순으로 나타났다. 통일관련 국가의 문서는 중국(24.6%), 미국(23.6%), 일본(16.7%) 등의 순으로 나타났다. 통일관련 기관의 문서는 정부(46.1%), 국회(11.6%), 청와대(11.3%), 국정원(7.1%) 등의 순으로 나타났다. 통일관련 정당의 문서는 새누리당(43.5%), 새정치민주연합(35.3%), 통합진보당(13.2%) 등의 순으로 나타났다. 통일방식 관련 문서는 평화통일(44.5%), 자유통일(28.2%), 흡수통일(25.7%) 등의 순으로 나타났다.

나. 통일 관련 국민 인식

〈표 4-3〉과 같이 통일의 필요성에 대한 국민의 통일인식은 조사기관별로 차이를 보이고 있으나, 2011년부터 2014년 까지 통일대박을 제거한 소셜 빅데이터의 분석 결과는 서울대 통일평화연구원의 여론조사 결과와 비슷한 추이를 보이는 것으로 나타났다. 통일대박을 포함한 연도별 통일 관련 찬성 인식은 2011년 55.0%, 2012년 62.2%, 2013년 57.8%, 2014년 77.1%, 2015년 59.8%로 나타났다. 특히, 아산정책연구원의 2014년 통일인식 여론조사에서 80.5%(통일은 가능한 빨리되어야+점진적으로 이루어져야)가 통일이 필요하다고 보고 있어, 통일 대박을 포함한 소셜 빅데이터 분석결과(77.1%)와 비슷한 것으로 나타났다.

〈표 4-4〉와 같이 리퍼트 미국 대사 피습사건(2015년 3월 5일) 전의 통일에 대한 찬성은 55.0%에서 피습 후 1주일간²⁷⁾ 50.9%로 낮아졌다가 일주일 후부터 61.9%로 상승한 것으로 나타났다.

27) URL을 통해 확산되는 온라인 문서(자살 등)와 3주 정도의 생명주기를 가지며, 발생 후 첫 주에 급속히 전파되는 경향을 보인다(National Information Society Agency, 2012).

〈표 4-2〉 통일 관련 문서(버즈) 현황

구분	항목	N(%)	구분	항목	N(%)
연도	2011년	10,211(2.5)	감정 [통일대박 포함]	찬성	52,562(58.0) [90,993(68.8)]
	2012년	53,884(13.1)		중립	8,817(9.7) [13,496(10.2)]
	2013년	83,268(20.3)		반대	29,287(32.3) [27,692(21.0)]
	2014년	166,952(40.6)		계	90,999 [132,181]
	2015년	96,820(23.5)			
	계	411,135			
채널	카페	19,785(4.8)	국가	중국	21,270(24.6)
	SNS	338,501(82.3)		독일	13,049(15.1)
	블로그	22,220(5.4)		미국	20,474(23.6)
	게시판	5,939(1.4)		일본	14,492(16.7)
	뉴스	24,693(6.0)		러시아	10,040(11.6)
	계	411,138		유럽	7,250(8.4)
정책	진보	4,504(10.3)	기관	계	86,575
	보수	4,925(11.2)		법민련	3,673(4.2)
	공산주의	2,472(5.6)		시민단체	6,402(7.3)
	사회주의	2,475(5.6)		정부	40,174(46.1)
	대북정책	16,535(37.7)		청와대	9,859(11.3)
	창조경제	2,760(6.3)		국정원	6,203(7.1)
	민주주의	8,128(18.6)		통일준비위원회	5,521(6.3)
	햇볕정책	2,007(4.6)		조국평화통일위원회	2,148(2.5)
안보	계	43,806	정당	자유통일포럼	3,082(3.5)
	핵무기	22,126(52.7)		국회	10,161(11.6)
	정상회담	9,249(22.0)		계	87,223
	휴전선	5,177(12.3)		새정치민주연합	9,308(35.3)
	남북공동선언	2,547(6.1)		새누리	11,487(43.5)
	간첩	2,882(6.9)		통합진보당	3,473(13.2)
	계	41,981		자유선진당	187(0.7)
				노동당	1,933(7.3)
이슈	천안함	6,444(7.3)	통일방식	계	26,388
	민영화	1,873(2.1)		흡수통일	12,688(25.7)
	자유통일	13,927(28.2)		선진통일	236(0.5)
	이산가족상봉	6,050(6.8)		통일대박	55,176(62.2)
	평화통일	21,940(44.5)		공존통일	540(1.1)
	선거	9,978(11.2)		통일비용	9,184(10.4)
	계	88,705		계	49,331

〈표 4-3〉 연도별 통일 관련 국민인식

연도	찬성				중립				반대			
	소셜 ¹⁾		서울대 ²⁾	KBS ³⁾	소셜 ¹⁾		서울대 ²⁾	KBS ³⁾	소셜 ¹⁾		서울대 ²⁾	KBS ³⁾
	통일 대박 제거	통일 대박 포함			통일 대박 제거	통일 대박 포함			통일 대박 제거	통일 대박 포함		
2011	55.0	55.0	53.7	74.4	18.3	18.3	25.0	19.9	26.7	26.6	21.3	5.7
2012	62.0	62.2	57.0	68.4	10.7	10.7	21.6	24.6	27.3	27.1	21.4	7.0
2013	57.7	57.8	54.8	69.1	8.7	8.7	21.5	21.2	33.6	33.6	23.7	9.7
2014	58.2	77.1	55.8	-	9.0	10.3	22.5	-	32.7	12.6	21.7	-
2015	56.2	59.8			8.3	8.5			35.5	31.7		

주: 1) 본 연구의 소셜 빅데이터의 감성분석 결과

2) 서울대학교 통일평화연구원의 매년 7월부터 8월 사이(3주간) 1,200명 대상 대면조사 결과

3) KBS 방송문화연구소(2013). 2013년 국민통일 의식조사

〈표 4-4〉 미국대사 피습사건 전후 국민 통일인식 변화

(단위: N, %)

	찬성	중립	반대	계
2015. 1. 1. ~ 3. 4.	7,291(55.0)	1,081(8.1)	4,892(36.9)	13,264
2015. 3. 5. ~ 3. 12.	892(50.9)	141(8.0)	720(41.1)	1,753
2015. 3. 13. ~ 3. 31.	2,739(61.9)	396(8.9)	1,293(29.2)	4,428
계	10,922(56.2)	1,618(8.3)	6,905(35.5)	19,445

다. 통일 관련 안보이슈의 동향

〈표 4-5〉와 같이 통일관련 안보이슈에 대한 주변 국가의 문서는 미국, 중국, 일본, 러시아 등 대부분의 국가에서 핵무기, 정상회담, 천안함, 통일대박 등의 순으로 문서가 많이 언급된 것으로 나타났다.

〈표 4-5〉 통일관련 안보·이슈의 국가별 버즈 현황

(단위: N, %)

속성	핵무기	정상회담	휴전선	남북 공동선언	천안함	통일대박	통일비용	합계
중국	5973 (36.6)	3395 (20.8)	1222 (7.5)	164 (1.0)	1651 (10.1)	2160 (13.2)	1759 (10.8)	16324 (22.8)
독일	2421 (25.0)	1869 (19.3)	961 (9.9)	62 (0.6)	590 (6.1)	2214 (22.9)	1551 (16.0)	9668 (13.5)
미국	7103 (40.0)	3597 (20.2)	1277 (7.2)	209 (1.2)	1953 (11.0)	1896 (10.7)	1728 (9.7)	17763 (24.8)
일본	4579 (39.7)	2134 (18.5)	931 (8.1)	129 (1.1)	975 (8.4)	1420 (12.3)	1377 (11.9)	11545 (16.1)
러시아	3354 (37.1)	1814 (20.1)	868 (9.6)	99 (1.1)	705 (7.8)	1213 (13.4)	992 (11.0)	9045 (12.6)
유럽	2355 (32.6)	1476 (20.4)	659 (9.1)	70 (1.0)	486 (6.7)	1350 (18.7)	832 (11.5)	7228 (10.1)
계	25785 (36.0)	14285 (20.0)	5918 (8.3)	733 (1.0)	6360 (8.9)	10253 (14.3)	8239 (11.5)	71573 (100.0)

통일과 관련한 기관별 안보·이슈에 대한 버즈는 〈표 4-6〉와 같이 나타났다. 청와대는 통일대박(28.2%), 핵무기(23.9%), 정상회담(23.6%) 등의 순으로 나타났고, 정부는 핵무기(28.1%), 정상회담(20.5%), 통일대박(19.6%) 등의 순으로 나타났고, 국회는 핵무기(30.9%), 정상회담(22.2%), 통일대박(14.1%) 등의 순으로 나타났고, 국정원은 통일대박(35.9%), 핵무기(27.1%), 정상회담(13.2%) 등의 순으로 나타났고, 시민단체는 남북공동선언(50.7%), 핵무기(20.6%), 통일대박(10.6%) 등의 순으로 나타났다.

〈표 4-6〉 통일관련 안보이슈의 기관별 버즈 현황

(단위: N, %)

속성	핵무기	정상회담	휴전선	남북 공동선언	천안함	통일대박	통일비용	합계
범민련	131 (8.6)	39 (2.6)	20 (1.3)	1258 (82.8)	46 (3.0)	15 (1.0)	11 (.7)	1520 (3.1)
시민단체	506 (20.6)	106 (4.3)	52 (2.1)	1244 (50.7)	173 (7.0)	261 (10.6)	112 (4.6)	2454 (5.0)
정부	7435 (28.1)	5418 (20.5)	2162 (8.2)	368 (1.4)	3036 (11.5)	5191 (19.6)	2861 (10.8)	26471 (53.6)
청와대	2022 (23.9)	1999 (23.6)	581 (6.9)	78 (.9)	752 (8.9)	2392 (28.2)	652 (7.7)	8476 (17.2)
국정원	485 (27.1)	236 (13.2)	81 (4.5)	44 (2.5)	191 (10.7)	642 (35.9)	110 (6.1)	1789 (3.6)
통일준비 위원회	370 (15.0)	491 (19.9)	197 (8.0)	24 (1.0)	106 (4.3)	1044 (42.4)	231 (9.4)	2463 (5.0)
조국평화 통일위원회	445 (32.7)	329 (24.2)	282 (20.8)	10 (.7)	255 (18.8)	22 (1.6)	16 (1.2)	1359 (2.8)
자유통일 포럼	6 (2.4)	0 (0.0)	1 (.4)	0 (0.0)	116 (6.0)	128 (50.8)	1 (.4)	252 (.5)
국회	1422 (30.9)	1023 (22.2)	355 (7.7)	96 (2.1)	538 (1.7)	650 (14.1)	515 (11.2)	4599 (9.3)
계	12822 (26.0)	9641 (19.5)	3731 (7.6)	3122 (6.3)	5213 (10.6)	10345 (20.9)	4509 (9.1)	49383 (100.0)

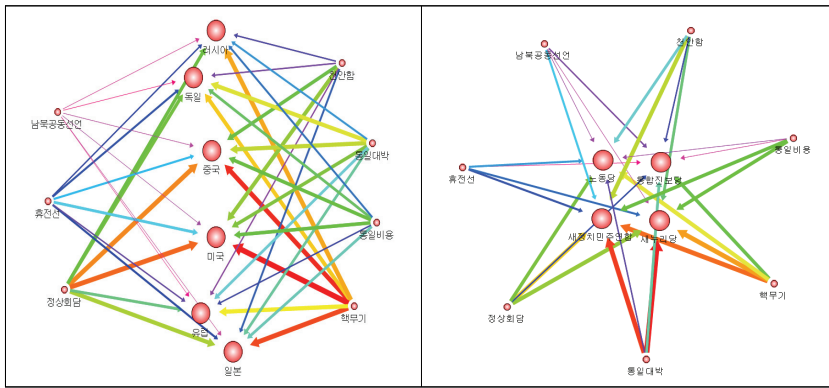
라. 통일 관련 소셜 네트워크 분석

[그림 4-4]와 같이 주변 국가와 안보이슈 간의 외부 근접중심성(out closeness Centrality)²⁸⁾을 살펴보면 핵무기와 천안함은 미국, 중국, 일본, 러시아 순으로 밀접하게 연결되어 있는 것으로 나타났으며, 정상회담

28) 근접중심성(Closeness Centrality)은 평균적으로 다른 노드들과의 거리가 짧은 노드의 중심성이 높은 경우로, 근접중심성이 높은 노드는 확률적으로 가장 빨리 다른 노드에 영향을 주거나 받을 수 있다.

은 미국, 중국, 일본, 독일의 순으로 밀접하게 연결된 것으로 나타났다. 정당과 안보·이슈 간의 외부 근접중심성을 살펴보면 통일대박은 새누리당, 새정치민주연합 등의 순으로 밀접하게 연결되어 있고, 천안함은 새정치민주연합, 새누리당 등의 순으로 밀접하게 연결되어 있는 것으로 나타났다.

[그림 4-4] 통일 감정 응집구조 분석



마. 통일 관련 인식에 대한 연관성 분석

〈표 4-7〉과 같이 안보와 이슈에 대한 통일인식의 연관성 예측에서 가장 신뢰도가 높은 연관규칙으로는 {정상회담, 선거} ⇒ {찬성}이며 세 변인의 연관성은 지지도 0.001, 신뢰도는 0.54, 향상도는 4.22로 나타났다. 이는 온라인 문서에서 정상회담과 선거가 언급되면 통일에 찬성할 확률이 53.9%이며, 정상회담과 선거가 언급되지 않은 문서 보다 통일에 대한 찬성의 확률이 4.22배 높아지는 것을 나타낸다. 특히, {정상회담, 통일대박} ⇒ {찬성} 세 변인의 연관성은 지지도 0.001, 신뢰도 0.47, 향상도는 3.71로 정상회담과 통일대박이 언급되지 않은 문서 보다 통일에 대한 찬성의 확률이 3.71배 높아지는 것으로 나타났다. {핵무기, 휴전선} ⇒

{반대} 세 변인의 연관성은 지지도 0.001, 신뢰도는 0.31, 향상도는 4.34로 나타났다. 이는 온라인 문서에서 핵무기와 휴전선이 언급되면 통일에 반대할 확률이 30.9%이며, 핵무기와 휴전선이 언급되지 않은 문서 보다 통일에 대한 반대의 확률이 4.34배 높아지는 것을 나타낸다.

〈표 4-7〉 안보와 이슈 요인에 대한 통일인식의 연관성 예측

규칙	지지도	신뢰도	향상도
{정상회담, 선거} => {찬성}	0.001366938	0.5393474	4.218755
{정상회담, 휴전선} => {찬성}	0.001357209	0.5224719	4.086756
{핵무기, 선거} => {찬성}	0.002045542	0.4973389	3.890166
{정상회담, 천안함} => {찬성}	0.001814476	0.4946950	3.869485
{핵무기, 정상회담, 천안함} => {찬성}	0.001216137	0.4935834	3.860791
{핵무기, 정상회담} => {찬성}	0.004181078	0.4926913	3.853813
{핵무기, 휴전선} => {찬성}	0.001770695	0.4840426	3.786163
{정상회담, 통일대박} => {찬성}	0.001804747	0.4747281	3.713305
{핵무기, 통일비용} => {찬성}	0.001442338	0.4640063	3.629440
{핵무기, 천안함} => {찬성}	0.002887108	0.4475867	3.501007
{정상회담} => {찬성}	0.009631802	0.4281544	3.349008
{핵무기, 통일대박} => {찬성}	0.001707456	0.4156306	3.251047
{휴전선} => {찬성}	0.005034806	0.3998455	3.127576
{천안함} => {찬성}	0.005153987	0.3288330	2.572120
{핵무기} => {찬성}	0.016687827	0.3100877	2.425494
{핵무기, 휴전선} => {반대}	0.001131007	0.3091755	4.340281
{통일비용, 통일대박} => {찬성}	0.002356873	0.3074239	2.404658
{핵무기, 천안함} => {반대}	0.001848528	0.2865762	4.023026
{선거} => {찬성}	0.006929547	0.2855282	2.233391
{핵무기, 정상회담} => {반대}	0.002400654	0.2828891	3.971265
{통일비용} => {찬성}	0.006287427	0.2814678	2.201630
{정상회담, 천안함} => {반대}	0.001014258	0.2765252	3.881928
{핵무기, 선거} => {반대}	0.001048310	0.2548788	3.578050
{정상회담} => {반대}	0.004424305	0.1966699	2.760900
{핵무기} => {반대}	0.010487963	0.1948838	2.735827
{천안함} => {반대}	0.002984399	0.1904097	2.673017
{남북공동선언} => {찬성}	0.001140736	0.1841382	1.440322
{휴전선} => {반대}	0.002242556	0.1780954	2.500147

〈표 4-8〉과 같이 주요 주변국가 대한 통일인식의 연관성 예측에서 가장 신뢰도가 높은 연관규칙으로는 {중국, 미국, 일본, 러시아} ⇒ {찬성}이며 다섯 변인의 연관성은 지지도 0.002, 신뢰도는 0.17, 향상도는 3.74로 나타났다. 이는 온라인 문서에서 중국, 미국, 일본, 러시아가 언급되면 통일에 찬성할 확률이 47.8%이며, 중국, 미국, 일본, 러시아가 언급되지 않은 문서 보다 통일에 대한 찬성의 확률이 3.74배 높아지는 것을 나타낸다. 특히, {중국, 미국, 일본} ⇒ {중립} 네 변인의 연관성은 지지도 0.001, 신뢰도 0.47, 향상도는 7.72로 중국, 미국, 일본이 언급되지 않은 문서보다 통일에 대한 중립의 확률이 7.72배 높아지는 것으로 나타났다²⁹⁾. {중국, 미국, 러시아} ⇒ {반대} 네 변인의 연관성은 지지도 0.003, 신뢰도는 0.27, 향상도는 3.78로 나타났다³⁰⁾. 이는 온라인 문서에서 중국, 미국, 러시아가 언급되면 통일에 반대할 확률이 26.9%이며, 중국, 미국, 러시아가 언급되지 않은 문서 보다 통일에 대한 반대의 확률이 3.78배 높아지는 것을 나타낸다.

〈표 4-8〉 주변국가(4국)의 통일인식의 연관성 예측

규칙	지지도	신뢰도	향상도
{중국,미국,일본,러시아}⇒{찬성}	0.00341248	0.477861	3.737811
{미국,일본,러시아}⇒{찬성}	0.003682462	0.473125	3.700766
{중국,일본,러시아}⇒{찬성}	0.003967038	0.4678715	3.659673
{중국,미국,러시아}⇒{찬성}	0.004480248	0.4622334	3.615572
{중국,미국,일본}⇒{찬성}	0.005548016	0.4519517	3.535149
{중국,일본}⇒{찬성}	0.007627609	0.4394619	3.437454

29) {중국, 미국, 일본} ⇒ {찬성}의 향상도는 3.53, {중국, 미국, 일본} ⇒ {반대}의 향상도는 3.41, {중국, 미국, 일본} ⇒ {중립}의 향상도는 7.72로 나타나, 중국, 미국, 일본의 3국이 문서에서 언급될 경우, 중립의 확률이 더 높은 것으로 나타났다.

30) {중국, 미국, 러시아} ⇒ {찬성}의 향상도는 3.62, {중국, 미국, 일본} ⇒ {반대}의 향상도는 3.79로 나타나 중국, 미국, 러시아의 3국이 언급될 경우, 찬성의 문서보다 반대의 문서가 조금 더 많은 것으로 나타났다.

규칙	지지도	신뢰도	향상도
{중국, 미국}=>{찬성}	0.009320471	0.426631	3.337092
{미국, 러시아}=>{찬성}	0.005336408	0.4263506	3.334898
{중국, 러시아}=>{찬성}	0.006036902	0.4239112	3.315817
{미국, 일본}=>{찬성}	0.007017109	0.4108516	3.213666
{일본, 러시아}=>{찬성}	0.004426737	0.4031008	3.153039
{러시아}=>{찬성}	0.008342698	0.3416335	2.672244
{일본}=>{찬성}	0.011942462	0.3388076	2.650141
{중국}=>{찬성}	0.017298328	0.3343677	2.615411
{미국}=>{찬성}	0.016571078	0.3327635	2.602864
{중국, 미국, 러시아}=>{반대}	0.002607397	0.2690088	3.77641
{미국, 일본, 러시아}=>{반대}	0.002016355	0.2590625	3.636782
{미국, 러시아}=>{반대}	0.00316682	0.253012	3.551844
{중국, 미국}=>{반대}	0.005399647	0.247161	3.469706
{미국, 일본}=>{반대}	0.004190807	0.2453717	3.444587
{중국, 일본, 러시아}=>{반대}	0.002065	0.2435456	3.418952
{중국, 미국, 일본}=>{반대}	0.002979535	0.2427184	3.40734
{중국, 러시아}=>{반대}	0.003417344	0.2399658	3.368699
{중국, 일본}=>{반대}	0.003830831	0.2207119	3.098407
{일본, 러시아}=>{반대}	0.002344712	0.2135105	2.997312
{미국}=>{반대}	0.009619641	0.1931718	2.711793
{중국}=>{반대}	0.009665854	0.1868359	2.622848
{러시아}=>{반대}	0.004387821	0.1796813	2.522409
{일본}=>{반대}	0.006263104	0.1776842	2.494374
{중국, 미국, 일본}=>{중립}	0.002033381	0.165643	7.723955
{중국, 미국}=>{중립}	0.003431938	0.157092	7.325221
{중국, 일본}=>{중립}	0.002629287	0.1514854	7.063788

바. 통일 인식에 영향을 미치는 안보이슈 요인

〈표 4-9〉와 같이 통일대박, 남북공동선언, 휴전선, 정상회담 순으로 통일의 찬성에 정적(+) 영향을 주는 것으로 나타났으나, 간첩, 핵무기, 통일비용, 천안함 순으로 통일의 찬성에 부적(-)인 영향을 주는 것으로 나타났다.

〈표 4-9〉 통일인식에 영향을 미치는 안보·이슈 요인¹⁾

변수	찬성				중립			
	b [†]	S.E. [‡]	OR [§]	P	b [†]	S.E. [‡]	OR [§]	P
핵무기	-.230	.023	.795	.000	.407	.032	1.502	.000
정상회담	.158	.032	1.171	.000	.601	.042	1.823	.000
휴전선	.159	.043	1.173	.000	.235	.060	1.265	.000
남북공동선언	.814	.105	2.257	.000	.194	.156	1.214	.214
간첩	-.525	.081	.592	.000	-.515	.128	.597	.000
천안함	-.051	.040	.951	.200	.198	.053	1.219	.000
선거	.076	.034	1.079	.025	.332	.046	1.393	.000
통일비용	-.099	.036	.906	.006	-.127	.051	.880	.012
통일대박	2.803	.027	16.489	0.000	2.725	.032	15.258	0.000

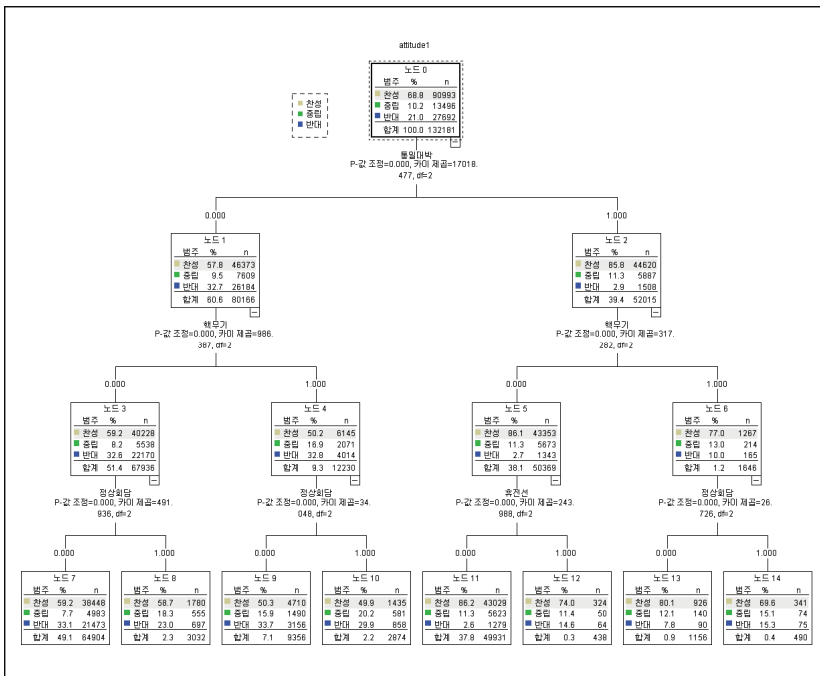
주: 1) 기본범주: 반대, † Standardized coefficients, ‡ Standard error, § Adjusted odds ratio

사. 통일 인식에 관련 예측 모형

[그림 4-5]과 같이 통일의 안보·이슈 요인이 통일인식에 미치는 영향은 ‘통일대박’의 영향력이 가장 큰 것으로 나타났다. ‘통일대박’이 있을 경우 통일에 대한 찬성의 인식이 이전의 68.8%에서 85.8%로 증가한 반면, 반대의 인식은 이전의 21.0%에서 2.9%로 감소하였다. ‘통일대박’이 있고 ‘핵무기’가 있는 경우 통일에 대한 찬성의 인식은 이전의 85.8%에서 77.0%로 감소한 반면, 반대의 인식은 이전의 2.8%에서 10.0%로 크게 증가하는 것으로 나타났다. 〈표 4-10〉의 통일인식의 안보·이슈요인의 예측 모형에 대한 이익도표와 같이 통일의 찬성에 가장 영향력이 높은 경우는 ‘통일대박’이 있고 ‘핵무기’가 없고 ‘휴전선’이 없는 조합으로 나타났다. 즉, 11번 노드의 지수(index)가 125.2%로 뿌리마디와 비교했을 때 11번 노드의 조건을 가진 집단이 통일을 찬성하는 확률이 1.25배로

나타났다. 통일의 반대에 가장 영향력이 높은 경우는 ‘통일대박’이 없고 ‘핵무기’가 있고 ‘정상회담’이 없는 조합으로 나타났다. 즉, 9번 노드의 지수가 161.1%로 뿌리마디와 비교했을 때 9번 노드의 조건을 가진 집단이 통일을 반대하는 확률이 1.61배로 나타났다.

[그림 4-5] 통일인식 관련 안보이슈 요인의 예측모형



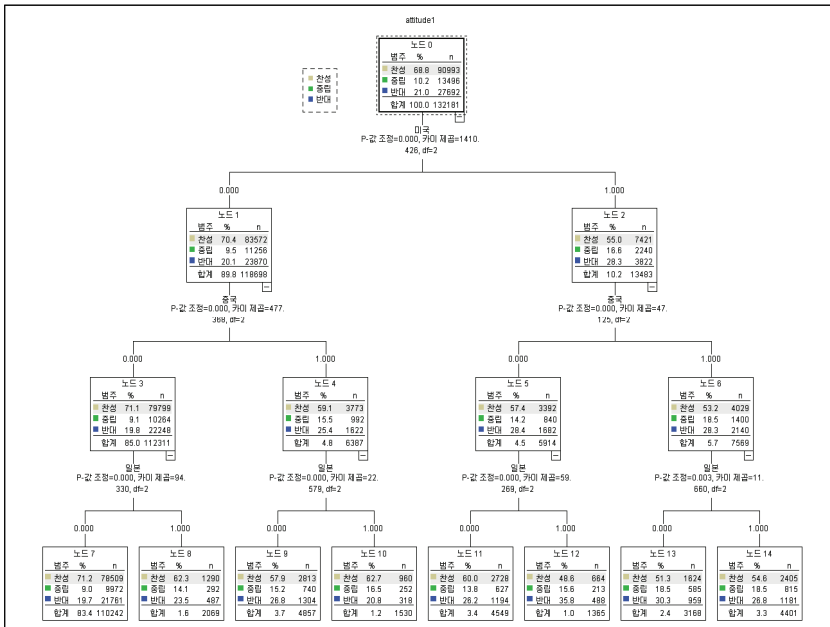
〈표 4-10〉 통일인식 관련 안보이슈 요인의 예측 모형에 대한 이익도표

구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
찬성	11	49931	37.8	47.3	125.2	49931	37.8	47.3	125.2
	13	1156	.9	1.0	116.4	51087	38.6	48.3	125.0
	12	438	.3	.4	107.5	51525	39.0	48.7	124.8
	14	490	.4	.4	101.1	52015	39.4	49.0	124.6
	7	64904	49.1	42.3	86.1	116919	88.5	91.3	103.2
	8	3032	2.3	2.0	85.3	119951	90.7	93.2	102.8
	9	9356	7.1	5.2	73.1	129307	97.8	98.4	100.6
	10	2874	2.2	1.6	72.5	132181	100.0	100.0	100.0
보통	10	2874	2.2	4.3	198.0	2874	2.2	4.3	198.0
	8	3032	2.3	4.1	179.3	5906	4.5	8.4	188.4
	9	9356	7.1	11.0	156.0	15262	11.5	19.5	168.5
	14	490	.4	.5	147.9	15752	11.9	20.0	167.9
	13	1156	.9	1.0	118.6	16908	12.8	21.0	164.5
	12	438	.3	.4	111.8	17346	13.1	21.4	163.2
	11	49931	37.8	41.7	110.3	67277	50.9	63.1	123.9
	7	64904	49.1	36.9	75.2	132181	100.0	100.0	100.0
반대	9	9356	7.1	11.4	161.0	9356	7.1	11.4	161.0
	7	64904	49.1	77.5	157.9	74260	56.2	88.9	158.3
	10	2874	2.2	3.1	142.5	77134	58.4	92.0	157.7
	8	3032	2.3	2.5	109.7	80166	60.6	94.6	155.9
	14	490	.4	.3	73.1	80656	61.0	94.8	155.4
	12	438	.3	.2	69.7	81094	61.4	95.1	154.9
	13	1156	.9	.3	37.2	82250	62.2	95.4	153.3
	11	49931	37.8	4.6	12.2	132181	100.0	100.0	100.0

〔그림 4-6〕와 같이 주변 4개국이 통일인식에 미치는 영향은 ‘미국’이 가장 높은 것으로 나타났다. ‘미국’이 있을 경우 통일에 대한 찬성의 인식이 이전의 68.8%에서 55.0% 감소한 반면, 반대의 인식은 이전의 21.0%에서 28.3%로 증가하였다. ‘미국’이 있고 ‘중국’이 있는 경우 통일에 대한 찬성의 인식은 이전의 55.0%에서 53.2%로 감소한 반면, 중립의 인식은 이전의 16.0%에서 18.5%로 증가하는 것으로 나타났다. 〈표 4-11〉의 통일인식의 주변 4개국 요인의 예측 모형에 대한 이익도표와 같이 통일

의 찬성에 가장 영향력이 높은 경우는 ‘미국’이 없고 ‘중국’이 없고 ‘일본’이 없는 조합으로 나타났다. 즉, 7번 노드의 지수가 103.5%로 뿌리마디와 비교했을 때 7번 노드의 조건을 가진 집단이 통일을 찬성하는 확률이 1.04배로 나타났다. 통일의 반대에 가장 영향력이 높은 경우는 ‘미국’이 있고 ‘중국’이 없고 ‘일본’이 있는 조합으로 나타났다. 즉, 12번 노드의 지수가 170.6%로 뿌리마디와 비교했을 때 12번 노드의 조건을 가진 집단이 통일을 반대하는 확률이 1.71배로 나타났다. 통일의 중립에 가장 영향력이 높은 경우는 ‘미국’이 있고 ‘중국’이 있고 ‘일본’이 있는 조합으로 나타났다. 즉, 14번 노드의 지수가 181.4%로 뿌리마디와 비교했을 때 14번 노드의 조건을 가진 집단이 통일을 반대하는 확률이 1.8배로 나타났다.

[그림 4-6] 통일인식 관련 주요 주변 국가 예측모형



〈표 4-11〉 통일인식 관련 국가 요인의 예측 모형에 대한 이익도표(주변 4국)

구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
찬성	7	110242	83.4	86.3	103.5	110242	83.4	86.3	103.5
	10	1530	1.2	1.1	91.1	111772	84.6	87.3	103.3
	8	2069	1.6	1.4	90.6	113841	86.1	88.8	103.1
	11	4549	3.4	3.0	87.1	118390	89.6	91.8	102.4
	9	4857	3.7	3.1	84.1	123247	93.2	94.8	101.7
	14	4401	3.3	2.6	79.4	127648	96.6	97.5	100.9
	13	3168	2.4	1.8	74.5	130816	99.0	99.3	100.3
	12	1365	1.0	.7	70.7	132181	100.0	100.0	100.0
보통	14	4401	3.3	6.0	181.4	4401	3.3	6.0	181.4
	13	3168	2.4	4.3	180.9	7569	5.7	10.4	181.2
	10	1530	1.2	1.9	161.3	9099	6.9	12.2	177.8
	12	1365	1.0	1.6	152.8	10464	7.9	13.8	174.6
	9	4857	3.7	5.5	149.2	15321	11.6	19.3	166.5
	8	2069	1.6	2.2	138.2	17390	13.2	21.5	163.2
	11	4549	3.4	4.6	135.0	21939	16.6	26.1	157.3
	7	110242	83.4	73.9	88.6	132181	100.0	100.0	100.0
반대	12	1365	1.0	1.8	170.6	1365	1.0	1.8	170.6
	13	3168	2.4	3.5	144.5	4533	3.4	5.2	152.4
	9	4857	3.7	4.7	128.2	9390	7.1	9.9	139.8
	14	4401	3.3	4.3	128.1	13791	10.4	14.2	136.1
	11	4549	3.4	4.3	125.3	18340	13.9	18.5	133.4
	8	2069	1.6	1.8	112.4	20409	15.4	20.3	131.3
	10	1530	1.2	1.1	99.2	21939	16.6	21.4	129.0
	7	110242	83.4	78.6	94.2	132181	100.0	100.0	100.0

4. 결론

본 연구는 국내의 온라인 뉴스 사이트, 블로그, 카페, SNS, 게시판 등 인터넷을 통해 수집된 소셜 빅데이터를 주제분석과 감성분석 기술로 분류하고 네트워크 분석과 데이터마이닝의 연관성 분석과 의사결정나무 분석 기법을 적용하여 분석함으로써 우리나라 국민의 통일인식에 대한 동향을 분석하고 통일인식의 연관규칙과 예측모형을 개발하고자 하였다.

본 연구의 결과를 요약하면 다음과 같다. 첫째, 통일과 관련된 이슈발생 시에 온라인 상에 통일관련 문서가 급증하는 양상을 보이고 있으며, 연도별 통일 찬성의 감정변화는 2011년 대비 평균 2.23배 증가하였으며, 찬성 감정의 표현단어는 평화, 필요, 중요, 노력, 관심 등의 순으로 집중된 것으로 나타났다. 통일 반대의 감정 변화는 2011년 대비 평균 3.25배 증가하였으며, 반대 감정의 표현 단어는 문제, 위협, 갈등, 포기, 분열 등의 순으로 집중된 것으로 나타났다.

둘째, 안보와 관련된 문서는 핵무기, 정상회담, 휴전선, 남북공동선언, 간첩 순으로 나타났으며, 이슈와 관련한 문서는 통일대박, 선거, 통일비용, 천안함, 이산가족상봉, 민영화 순으로 나타났다. 그리고 통일 방식과 관련한 문서는 평화통일, 자유통일, 흡수통일 등의 순으로 나타났다.

셋째, 통일의 필요성에 대한 국민인식은 통일대박 감정을 포함할 경우 찬성은 2011년 55.0%, 2012년 62.2%, 2013년 57.8%, 2014년 77.1%, 2015년 56.2%로 나타났다.

넷째, 통일관련 네트워크 분석에서 안보·이슈 간의 외부 근접중심성은 핵무기와 천안함은 미국, 중국, 러시아 순으로 밀접하게 연결되어 있으며, 천안함은 새정치민주연합, 새누리당 등의 순으로 밀접하게 연결되어 있는 것으로 나타났다.

다섯째, 안보와 이슈에 대한 통일 인식의 연관성예측에서 {정상회담, 선거} => {찬성}의 신뢰도가 0.54로 온라인 문서 중 정상회담과 선거가 동시에 언급되면 언급되지 않는 문서보다 통일에 대한 찬성의 확률이 4.22배가 높아지는 것으로 나타났다. {핵무기, 휴전선} => {반대}의 신뢰도는 0.31로 온라인 문서 중 핵무기와 휴전선이 동시에 언급되면 언급되지 않은 문서 보다 통일에 대해 반대할 확률이 4.34배 높아지는 것으로 나타났다. 주변 국가의 통일인식의 연관성 예측에서 {중국, 미국, 일본, 러시아}

=> {찬성}의 신뢰도가 0.17로 온라인 문서에서 중국, 미국, 일본, 러시아가 동시에 언급되면 언급되지 않은 문서 보다 통일에 대한 찬성의 확률이 3.74배 높아지며, {중국, 미국, 일본} => {중립}은 신뢰도가 0.47로 온라인 문서에서 중국, 미국, 일본이 동시에 언급되면 통일에 대한 중립의 확률이 7.72배 높아지는 것으로 나타났다. 그리고, {중국, 미국, 러시아} => {반대}의 신뢰도는 0.27로 온라인 문서에서 중국, 미국, 러시아가 동시에 언급되면 통일에 대한 반대의 확률이 3.78배 높아지는 것으로 나타났다.

여섯째, 통일 인식에 영향을 미치는 안보·이슈요인에 대한 다중로지스틱 회귀분석 결과 통일대박, 남북공동선언, 휴전선, 정상회담 순으로 통일의 찬성에 정적(+) 영향을 미치는 것으로 나타났으나, 간첩, 핵무기, 통일비용, 천안함 순으로 통일의 찬성에 부적(-)인 영향을 주는 것으로 나타났다.

마지막으로 안보·이슈 요인이 통일인식에 미치는 영향은 통일대박의 영향력이 가장 큰 것으로 나타나 온라인 문서 중, 통일대박이 있을 경우 통일에 대한 찬성의 인식이 이전의 68.8%에서 85.8%로 증가하였고, 특히 통일대박이 있고, 핵무기와 휴전선이 없는 경우 통일에 대해 찬성하는 확률이 가장 높은 것으로 나타났다. 주변 4개국이 통일인식에 미치는 영향은 미국이 가장 높은 것으로 나타나, 온라인 문서 중 미국이 없고, 중국이 없고, 일본이 없는 조합이 통일에 대해 찬성할 확률이 가장 높은 반면, 미국이 있고, 중국이 없고, 일본이 있는 조합이 통일에 대해 반대할 확률이 가장 높은 것으로 나타났다.

본 연구를 근거로 우리나라의 통일관련 인식에 대한 예측과 관련하여 다음과 같은 정책적 함의를 도출할 수 있다.

첫째, 통일과 관련된 이슈발생 시에 온라인 상에 통일 관련 커뮤니케이션이 급증하며, 2014년 첫 신년 기자회견에서 언급된 통일대박론 강조

이후 온라인 문서량이 급증한 것으로 나타났다. 이는 통일이 되면 천문학적 통일비용과 사회적 혼란이 야기될 것이라는 부정적 인식을 극복하고 통일을 기대와 희망으로 보는 긍정적 담론이 확산된 이유로 보고 있으며, 통일대박론이 통일문제에 대한 국민적 합의를 이루는 계기가 되었다는 긍정적 주장(김창수, 2014)을 지지하는 것으로 나타났다.

둘째, 통일관련 국민인식은 소셜 빅데이터의 분석결과와 정기적인 여론조사 결과와 비슷한 추이를 보이는 것으로 나타났다. 이는 본 연구에서 제시한 통일관련 감정 키워드의 감성분석 방법으로 통일의 필요성에 대한 인식을 찬성, 보통, 반대로 분류하는 것에 대한 타당성이 어느 정도 확보된 것으로 볼 수 있다.

셋째, 리퍼트 대사 피습사건 후의 통일에 대한 인식이 발생 후 1주일 정도 영향을 받는 것으로 나타나, SNS를 통해 확산되는 부정적 이슈(질병, 자살, 테러 등)는 발생 후 첫 주에 급속히 전파되는 경향을 보이고 있다는 연구³¹⁾를 지지하는 것으로 나타났다.

넷째, 정상회담과 선거가 동시에 언급된 온라인 문서의 경우 통일에 대한 찬성의 확률이 높은 것은 정상회담과 선거가 통일에 대한 긍정적 담론을 확산시키며, 핵무기와 휴전선이 동시에 언급된 온라인 문서의 경우 통일에 대한 반대의 확률이 높은 것은 핵무기와 휴전선에 대한 북한의 위협이 고조될 때, 통일에 대한 부정적인 담론이 확산되기 때문인 것으로 본다.

다섯째, 중국, 미국, 일본, 러시아가 동시에 언급된 온라인 문서의 경우 통일에 대한 찬성의 확률이 높은 것은 북한의 핵 문제를 해결하고, 한반도의 비핵화를 실현하기 위해 한국·북한·미국·중국·러시아·일본이 참석하는 6자회담이 언급될 경우 통일에 대한 국민의 인식이 긍정적 감정으로 나타나는 것으로 보인다. 이는 한반도 통일은 남북한의 문제이자 동북

31) National INformation Society Agency(2012)

아 주변국의 미래를 좌우하며(김규륜, 2013), 한반도 평화체제와 통일과정에 국제사회의 지지가 필수불가결함을 보여주는 것이다(차문석, 2013).

여섯째, 간첩, 핵무기, 통일비용, 천안함 요인은 통일을 반대할 확률이 높은 것은 북한의 위협과 같은 통일에 대한 부정적 요소의 문서가 언급될 때 통일에 대한 반대의 담론이 확산되는 것으로 나타났다. 특히 통일비용에 대한 부정적 인식은 1990년대 후반 IMF 구제금융사태와 독일 통일의 후유증이 우리사회에 알려지면서 급속히 확산되었으며, 통일비용을 통해 실제로 얻어질 수 있는 이득은 계산하지 않고 투입된 비용만 계산하여 천문학적 수치만 제시함으로써 통일에 대한 두려움이 증가한 것으로 보고 있다(김규륜·김형기, 2012).

마지막으로 통일의 반대에 가장 영향력이 높은 경우는 통일대박이 없고, 핵무기가 있고, 정상회담이 없는 문서들로 나타났으며, 통일대박이 없고, 핵무기가 있더라도 정상회담이 있는 문서들은 통일의 인식이 보통으로 나타나 남북정상회담의 이슈가 통일을 긍정적으로 작용하는 핵심요인으로 보여진다.

본 연구는 개개인의 특성을 가지고 분석한 것이 아니고 그 구성원이 속한 전체 집단의 자료를 대상으로 분석하였기 때문에 이를 개인에게 적용하였을 경우 생태학적 오류(ecological fallacy)가 발생할 수 있다(Song et al., 2014). 또한, 본 연구에서 감성분석 결과 정의된 통일인식은 온라인 문서 내에서 발생된 감정 단어의 빈도로 정의되었기 때문에 기존의 조사 등을 통한 통일인식의 조작적 정의와 다를 수 있으며, 2011년 ~ 2015년 기간의 1/4분기(15개월간)의 제한된 소셜 빅데이터를 분석함으로써 전체적인 통일관련 인식의 예측에 한계가 있을 수 있다. 그럼에도 불구하고 본 연구는 소셜 빅데이터에서 통일관련 주요 이슈에 대한 실제적인 내

용을 빠르게 효과적으로 파악함으로써 기존의 통일관련 인식의 정보수집 체계의 한계를 보완할 수 있는 새로운 분석방법을 제시하였다는 점에서 정책적·분석방법론적 의의를 가진다고 할 수 있다(송주영·송태민, 2014).

끝으로 통일에 대한 찬반, 통일방법 등에 대한 국민의 통일의식조사와 더불어 소셜 미디어에서 수집된 빅데이터의 활용과 분석을 할 경우, 통일 인식의 예측은 더욱 신뢰성이 있을 것으로 본다.

참고문헌 <<

- 강동완, 박정란(2014). 북한주민의 통일의식 조사연구: 북한주민 100명 면접조사를 중심으로. 통일정책연구 제23권 2호, p 2.
- 김규륜(2013). 한반도 통일의 미래와 주변 4국의 기대, 통일연구원, 2013.12. pp.1~215.
- 김규륜, 김형기(2012). 통일 재원 마련 및 통일의지 결집 관련 국민의 의식, 통일연구원 정책연구시리즈, 12-01. p. 14
- 김정선, 권은주, 송태민(2014). 분석지의 확장을 위한 소셜 빅데이터 활용연구-국내 '빅데이터' 수요공급 예측-. 지식경영연구, 15(3), pp.173~192.
- 김창수(2014). 통일대박론과 통일준비위원회-의의, 한계, 방향, 북한연구학회 춘계학술발표논문집, p. 120.
- 박희창(2010). 연관 규칙 마이닝에서의 평가기준 표준화 방안. 한국데이터정보과학회지, 제 21권, 제5호, pp. 891-899.
- 송주영, 송태민(2014). 소셜 빅데이터를 활용한 북한 관련 위협인식 요인 예측. 국제문제연구, 가을. pp. 209-243.
- 이규창(2014). 통일기반 조성과 법제준비, 북한연구학회 춘계학술발표논문집, pp.15-34.
- 이정진(2011). R, SAS, MS-SQL을 활용한 데이터마이닝. 자유아카데미.
- 차문석(2013). 한반도 평화통일 동북아 냉전적 대립 해소, 통일한국 제358호, pp.11-13.
- 통일연구원(2014). 드레스덴 구상과 행복한 통일, 제 1차 KINU 통일포럼, p.7
- 통일교육원(2013). 통일 문제 이해. p 93.
- Kass, G.(1980). *An exploratory technique for investigating large quantities of categorical data*. Applied Statistics, Vol. 292, pp.119~127.
- National Information Society Agency(2012). *Implications for Suicide Prevention Policy of Youth Described in the Social Analysis*.

Seoul, Korea: Author.

Song TM, Song J, An JY, Hayman LL, Woo JM(2014). *Psychological and Social Factors Affecting Internet Searches on Suicide in Korea: A Big Data Analysis of Google Search Trends*, Yonsei Med Journal, Vol. 55, No. 1, pp.254~263.

제 5 장

소셜 빅데이터를 활용한 한국의 섹스팅(Sexting) 위험 예측

1. 서론
2. 이론적 배경
3. 연구 방법
4. 분석 방법
5. 연구 결과
6. 결론 및 고찰

참고문헌

5

소셜 빅데이터를 활용한 한국의 <<섹스팅(Sexting) 위험 예측>>³²⁾

1. 서론

2014년 현재 10대 청소년의 99.7%가 스마트폰을 보유하고 있고(한국인터넷진흥원, 2014), 10대 청소년의 95.2%가 인터넷을 사용하며 고등학생의 78.1%가 SNS를 이용(미래창조과학부·한국인터넷진흥원, 2014)하는 것으로 나타나, 스마트폰을 이용한 인터넷 사용은 청소년의 필수적인 수단이 되고 있다. 이와 같이 청소년의 일상생활에서의 인터넷 및 스마트폰 이용이 증가함에 따라 긍정적인 효과와 더불어 인터넷 중독 등 역기능의 문제가 제기되고 있다. 2014년 청소년의 스마트폰 중독은 29.2%로 성인의 스마트폰 중독률 11.3%의 약 2.6배에 달하는 것으로 나타나고 있으며(미래창조과학부·한국정보화진흥원, 2014), 2014년 현재 중·고등학생의 52.6%가 스마트폰을 통하여 음란물(섹스팅)을 이용한 것으로(여성가족부, 2014년 청소년 유해환경 접촉 실태조사)나타나고 있고, 일부 청소년들은 자신의 성행동 장면을 촬영하여 실시간 인터넷 방송서비스(UCC 방송 등)나 웹하드 등 파일공유사이트를 통해 게시하는 등의 비행이 자행되고 있어 정부차원의 청소년에 대한 음란물 차단 대책이 요구되고 있다.

섹스팅(Sexting)은 성(sex)과 문자메시지 보내기(texting)의 합성어로만 18세 미만의 청소년들이 자신의 핸드폰상이나 인터넷상에서 만난 불

32) 본 연구는 해외 학술지에 게재하기 위하여 '송주영 교수(PSU), 송태민 박사(KIHASA)'가 작성한 논문임을 밝힌다.

특정한 이성에게 자신의 특정한 신체부위를 노출시킨 그림파일을 주고받는 것을 의미한다(Walker et al, 2011, p.8; Lounsbury et al., 2009, p1). 섹스팅의 문제점으로는 청소년의 신체노출 사진을 핸드폰에 가지고 있을 경우 아동 포르노그래피로 법적인 제재를 당할 수 있으며 섹스팅 상대방에게 정서적·신체적 상처를 입힐 수 있다는 것이다(Chalfen, 2009, p.263). 청소년기에 음란물에 대한 경험은 성에 대한 잘못된 현실 인식을 갖게 하며 지속적인 호기심으로 인해 더욱 음란물에 집착하게 만들 가능성이 있다(주지혁·김형일, 2010, p.18). 이와 같이 섹스팅 문제의 심각성에도 불구하고 국내에서는 섹스팅에 대한 과학적 연구가 부족한 실정에 있다.

한편 모바일 인터넷과 소셜 미디어의 확산으로 데이터량이 기하급수적으로 증가하여 데이터의 생산, 유통 소비체계에 큰 변화가 일어나면서 데이터가 경제적 자산이 될 수 있는 빅데이터 시대가 도래되었다. 세계 각국의 정부와 기업들은 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. 우리나라는 정부 3.0과 창조경제의 추진과 실현을 위하여 다양한 분야에 빅데이터를 효율적 활용을 적극적으로 모색하고 있다. 기존에 실시하던 횡단적 조사나 종단적 조사 등을 대상으로 한 연구는 정해진 변인들에 대한 개인과 집단의 관계를 보는 데는 유용하나, 사이버 상에서 언급된 개인별 담론(buzz)에서 논의된 관련 정보간의 연관 관계를 밝히고 원인을 파악하는 데는 한계가 있다(song et al., 2014). 이에 반해 소셜 빅데이터 분석은 훨씬 방대한 량의 데이터를 활용하여 다양한 참여자의 생각과 의견을 확인 할 수 있기 때문에 사회적 문제의 예측을 보다 정확하게 할 수가 있다. 본 연구는 우리나라 온라인 뉴스사이

트, 블로그, 카페, SNS, 게시판 등에서 수집한 소셜 빅데이터를 바탕으로 우리나라 섹스팅에 대한 위험요인을 예측하고자 한다.

2. 이론적 배경

청소년 시기는 일반적으로 성적인 호기심이 높은 시기이며, 성에 대한 사실과 환상의 구분에 취약함으로써 온라인의 성(性) 콘텐츠에 노출되기 쉬운 시기이다(신선미, 2013. p276). 음란물은 인간의 성적 행위를 노골적으로 묘사하여 음탕하고 난잡한 느낌을 주는 사진이나 잡지, 영상물 등을 통틀어 이르는 말로 주로 상업적 목적으로 성기와 성행위만을 강조해서 보여줌으로써 그것을 읽거나 보는 사람을 성적으로 흥분하게 만드는 글, 사진, 만화, 잡지 등을 말한다(김문녕, 2012, p52). 특히, 한국에서의 "아동·청소년이용음란물"은 아동·청소년 또는 아동·청소년으로 명백하게 인식될 수 있는 사람이나 표현물이 등장하여 '성교행위', '구강·항문 등 신체의 일부나 도구를 이용한 유사 성교 행위', '신체의 전부 또는 일부를 접촉·노출하는 행위로서 일반인의 성적 수치심이나 혐오감을 일으키는 행위', '자위행위' 중 어느 하나에 해당하는 행위를 하거나 그 밖의 성적 행위를 하는 내용을 표현하는 것으로서 필름·비디오물·게임물 또는 컴퓨터나 그 밖의 통신매체를 통한 화상·영상 등의 형태로 된 것을 말한다(아동·청소년의 성보호에 관한 법률, 제2조 5호)³³⁾. 미국의 경우, 미국연방법인 PROTECT법에서 아동 포르노그래피의 개념을 "성적으로 노골적인 행위(Sexual explicit conduct)에 관여하고 있는 미성년자를 묘사한 영상"으로 정하고 있으며, 여기서 성적으로 노골적인 행위라 함은 '동성 또는 이성간에 생식을 이용한 성교행위, 구강 및 항문과 생식기를 이용한

33) <http://www.law.go.kr/lsInfoP.do?lsiSeq=150720&efYd=20140929#0000>. 2015. 6. 16 검색

유사 성교 행위, 수간, 자위, 가학적 또는 피가학적인 학대행위, 특정인의 생식기 혹은 음부의 외설적인 전시행위' 이라고 구체적으로 명시되어 있다(U.S. Department of state, 2003). 일본에서는 아동포르노그래피의 개념을 '아동매춘, 아동포르노에 관련된 행위 등에 대한 처벌 등 법률 제 2조의 3'과 '도쿄도 청소년 보호 조례 개정안 제 3장 7조 2항'에서는 18세 미만으로 표현되고 있다고 인식되는 청소년에 대해서 성적인 감정을 자극하거나, 아동을 상대로 한 성교, 아동을 상대로 한 성교유사행위, 아동에 의한 성교, 아동에 의한 성교유사행위가 포함되는 것으로 규정하고 있다(東京都議會, 2010). 현대사회에서 음란물이 문제가 된 사회적 배경으로는 그동안 미디어 소비의 주체가 기성세대였던 반면에 청소년들에게도 다양한 미디어를 소비할 수 있는 선택권이 주어져서 인터넷 음란물 접촉과도 같은 문제행동이 가능해진 것이다(Gruber & Thau, 2003, p.441~443).

한편, 섹스팅(sexting)이란 'sex'와 'texting'의 합성어로 만 18세 미만의 청소년들이 자신의 핸드폰상이나 인터넷상에서 만난 불특정한 이성에게 핸드폰의 카메라와 같은 디지털장비를 사용하여 찍은 성적인 이미지를 다른 핸드폰과 같은 장비 혹은 인터넷에 전송하는 행위를 의미한다(Lounsbury et al., 2009, p1; Walker et al, 2011, p.8).

섹스팅은 Macquarie 온라인 사전³⁴⁾에 2010년에 처음으로 등재된 이후 연구자들에 의해 쓰여지고 있고(Walker et al, 2011, p.8), 연구자들 사이에도 청소년들끼리 신체노출 사진을 교환하는 활동에 대해서 "나체 또는 반나체 사진교환"이라는 애매한 용어 대신 섹스팅이라는 용어를 사용하기로 합의가 되었으며 성적인 의미가 담긴 SMS 문자를 주고받는 활동은 제외하고 사진을 교환하는 행위만을 부르는 용어로 의미가 정해

34) <https://www.macquariedictionary.com.au>

지고 있다(Lounsbury et al, 2010, p.2).

3. 연구 방법

가. 연구대상

본 연구는 국내의 SNS, 온라인 뉴스 사이트 등 인터넷을 통해 수집된 소셜 빅데이터를 대상으로 하였다. 본 분석에서는 146개의 온라인 뉴스 사이트, 9개의 게시판, 1개의 SNS(트위터) 등 총 156개의 온라인 채널을 통해 수집 가능한 텍스트 기반의 웹문서(버즈)를 소셜 빅데이터로 정의하였다. 섹스팅 관련 토픽(topic)³⁵⁾의 수집은 2011. 1. 1 ~ 2015. 3. 31(4년 3개월간) 해당 채널에서 요일, 주말, 휴일을 고려하지 않고 매 시간단위로 수집하였으며, 수집된 총 65,611건 중 청소년 추정 문서 13,774건(2011년: 1,086건, 2012년: 5,352건, 2013년: 3,983건, 2014년: 2,319건, 2015년: 1,034건)의 텍스트(Text) 문서를 본 연구의 분석에 포함시켰다. 섹스팅 토픽은 모든 관련 문서를 수집하기 위해 ‘섹스팅’을 사용하였으며, 토픽과 같은 의미로 사용되는 토픽 유사어로는 ‘음란물 유통, 성인물 유통, sexting, 음란 유통, 음란 유포, 음란물 업로드, 음란물 다운, 음란 공유, 음란 채팅, 포르노 유통, 포르노 유포, 아동 유통, 아동 유포, 아동 업로드, 아동 다운’ 용어를 사용하였다. 본 연구를 위한 소셜 빅데이터의 수집³⁶⁾은 크롤러(Crawler)를 사용하였고, 이후 주제분석을 통해 분류된 명사형 어휘를 유목화(categorization)하여 분석요인으

35) 토픽은 소셜 분석 및 모니터링의 ‘대상이 되는 주제어’를 의미하며, 문서 내에 관련 토픽이 포함된 문서를 수집함.

36) 본 연구를 위한 소셜 빅데이터의 수집 및 토픽 분류는 ‘(주)SK텔레콤 스마트인사이트’에서 수행함.

로 설정하였다.

나. 연구도구

섹스팅과 관련하여 수집된 문서는 주제분석³⁷⁾의 과정을 거쳐 다음과 같이 정형화 데이터로 코드화하여 사용하였다.

1) 섹스팅 관련 감정

본 연구의 섹스팅 감정 키워드는 문서 수집 이후, 주제 분석을 통하여 총 106개(중독, 갈등, 강제, 걱정, 고민, 고생, 고통, 골치, 공감, 공포, 긍정, 기쁨, 논란, 눈물, 단절, 따뜻, 문제, 반대, 반대, 서명, 불안, 불편, 비난, 사회악, 상처, 서명, 스트레스, 실패, 심각, 악영향, 어려움, 우려, 우울증, 인정, 자유, 잘못, 재미, 중독성, 즐거움, 집착, 최고, 최악, 포기, 피로, 한숨, 한심, 행복, 호기심, 후회, 흥미, 희망, 긴급, 엄중, 비방, 강화, 흥분, 충격, 기대, 요구, 강경, 모욕, 중요, 집중, 협박, 검토, 해결, 부담, 위험, 비판, 장난, 적나라, 야한, 비하한, 자극적인, 곤혹, 막장, 유혹, 침해, 욕설, 자극, 쓰레기, 은밀, 기대감, 거짓, 혼란, 힘들다, 부적절, 현혹, 호소, 선처, 조롱, 불쌍, 위협한, 수치심, 잔인, 잔혹, 왜곡, 방탕, 배신감, 악마, 빠친다, 퇴치, 혐오감, 퇴폐적, 마음고생, 충격적인, 복수) 키워드로 분류하였다. 본 연구에서는 106개의 음란물 유통 감정 키워드(변수)가 가지는 음란물 유통 감정 정도를 판단하기 위해 2차 요인분석을 통하여 11개의 요인(67개 변수)으로 축약을 실시한 후, 감성분석을 실시하였다. 일

37) 주제분석에 사용되는 사전은 '21세기 세종계획'과 같은 범용사전도 있지만 대부분 분석의 목적에 맞게 사용자가 설계한 사전을 사용하게 된다. 본 연구의 섹스팅 관련 주제분석은 SKT에서 관련 문서 수집 후 원시자료(raw data)에서 나타난 상위 2,000개의 키워드들을 대상으로 유목화를 하여 사용자 사전을 구축하였다.

반적으로 감성분석은 긍정과 부정의 감성어 사전으로 분석해야 하나, 본 연구에서는 요인분석의 결과로 분류된 주제어의 의미를 파악하여 감성분석을 실시하였다. 요인분석에서 결정된 11개의 요인에 대한 주제어의 의미를 파악하여 ‘일반군, 위험군’으로 감성분석을 실시하였다. 따라서 본 연구에서는 일반군(27개: 강경, 고통, 상처, 수치심, 비방, 엄중, 마음고생, 한숨, 골치, 불편, 퇴치, 피로, 조롱, 악영향, 최악, 쓰레기, 단절, 한심, 서명, 비하한, 모욕, 거짓, 배신감, 사회악, 혼란, 불쌍, 장난), 위험군(28개: 잔혹, 잔인, 공포, 고생, 최고, 중요, 자유, 위험, 요구, 인정, 자국, 침해, 기대, 해결, 긍정, 충격, 적나라, 중독성, 중독, 방탕, 퇴폐적, 유혹, 은밀, 혐오감, 집착, 야한, 흥분, 흥미)으로 분류하였다. 그리고 일반군과 위험군의 감정을 동일한 횟수로 표현한 문서는 잠재군으로 분류하였다. 그리고, 최종 위험군과 잠재군은 ‘위험’으로 일반군과 감정을 나타내지 않은 문서는 ‘일반’으로 분류하였다. 위험군은 섹스팅을 긍정적으로 생각하는 감정이고, 일반군은 섹스팅을 부정적으로 생각하는 감정을 나타낸다.

2) 섹스팅에 대한 제도

섹스팅에 대한 제도 정의는 요인분석과 주제 분석 과정을 거쳐 ‘가중 처벌, 정보통신망법, 벌금, 아동청소년보호법’의 4개 제도로 제도가 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

3) 섹스팅에 대한 기관

섹스팅에 대한 기관 정의는 주제 분석 과정을 거쳐 ‘방송통신위원회, 경찰청, 국회, 청와대, 정부, 사법기관, 시민단체, 국제기구’의 8개 기관으로 기관이 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

4) 섹스팅에 대한 폐해

섹스팅에 대한 폐해의 정의는 요인분석과 주제 분석 과정을 거쳐 ‘ 명예 훼손, 성범죄, 사기, 음주, 사회문제’의 5개 폐해로 폐해가 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

5) 섹스팅에 대한 영향

섹스팅에 대한 영향의 정의는 주제 분석 과정을 거쳐 ‘ 공부, 건강, 대인 관계, 비용, 윤리의식, 성욕’의 6개로 해당 영향이 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

6) 섹스팅에 대한 도움

섹스팅에 대한 도움의 정의는 요인분석과 주제 분석 과정을 거쳐 ‘ 예방 교육, 전문가상담, 진전생활유도, 통제, 사랑’의 5개 도움으로 해당 도움이 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

7) 섹스팅에 대한 유형

섹스팅에 대한 유형의 정의는 주제 분석과 요인분석 과정을 거쳐 ‘ 성인 음란물, 유해광고, 스미싱³⁸⁾, 아동음란물’의 4개 유형으로 해당 유형이 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

38) 스미싱은 문자메시지(SMS)와 피싱(Phishing)의 합성어로 휴대전화 문자메시지를 통해 발송되는 피싱공격을 의미한다.

8) 섹스팅에 대한 내용

섹스팅에 대한 내용의 정의는 요인분석과 주제 분석 과정을 거쳐 ‘nud, 성행위, 원조교재, 문란행위, 폭력’의 5개 내용으로 해당 내용이 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

9) 섹스팅에 대한 유통방식

섹스팅에 대한 유통방식의 정의는 주제 분석 과정을 거쳐 ‘수요, 공급, 공유’의 3개 유통방식으로 해당 내용이 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

10) 섹스팅에 대한 채널

섹스팅에 대한 채널의 정의는 요인분석과 주제 분석 과정을 거쳐 ‘SNS, 온라인커뮤니티, 파일공유채널’의 3개 채널로 해당 내용이 있는 경우는 ‘1’, 없는 경우는 ‘0’으로 코드화 하였다.

4. 분석 방법

본 연구에서는 우리나라의 섹스팅의 위험을 설명하는 가장 효율적인 예측모형을 구축하기 위해 특별한 통계적 가정이 필요하지 않은 데이터 마이닝(data mining)의 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. 본 연구의 연관분석은 선험적 규칙(apriori principle) 알고리즘을 사용하였다. 본 연구의 섹스팅의 위험예측에 사용된 연관분석의 측도는 지지도 0.02, 신뢰도 0.2를 기준으로 시뮬레이션하였다. 데이터마이닝의 의사결정나무 분석은 방대한 자료 속에서 종속변인을 가장 잘 설명하는 예측모형을 자동적으로 산출해 줌으로써 각기 다른 속성을 가진 섹스팅에 대한 요인을 쉽게 파악할 수 있다. 본 연구의 의사결정나무 형성을 위한 분석 알고리즘은 CHAID (Chi-squared Automatic Interaction Detection)를 사용하였다. CHAID(Kass, 1980)는 이산형인 종속변수의 분리기준으로 카이제곱(χ^2 -검정)을 사용하며, 모든 가능한 조합을 탐색하여 최적분리를 찾는다. 정지규칙(stopping rule)으로 관찰치가 충분하여 상위노드(부모마디)의 최소 케이스 수는 100으로 하위노드(자식마디)의 최소 케이스 수는 50으로 설정하였고, 나무깊이는 3수준으로 정하였다. 본 연구의 기술 분석, 다중 응답분석, 의사결정나무분석은 SPSS v. 22.0을 사용하였고, 연관분석과 시각화는 R version 3.1.3, 소셜 네트워크 분석은 NetMiner³⁹⁾를 사용하였다.

5. 연구 결과

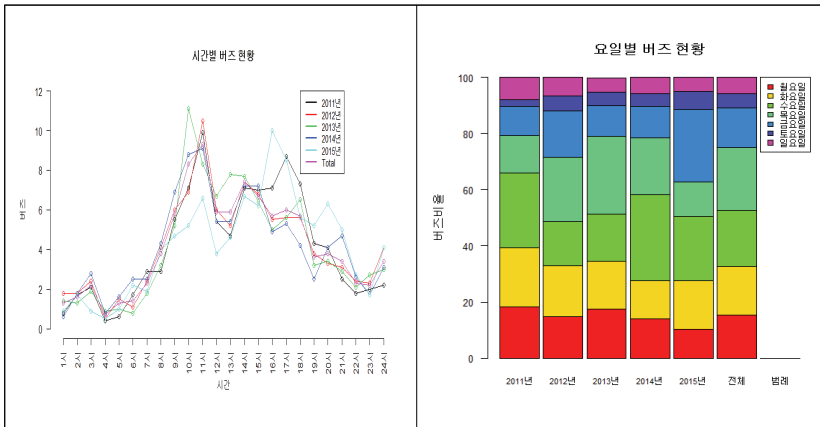
가. 섹스팅 관련 문서(버즈) 현황

섹스팅과 관련된 버즈는 년도별로 다르지만 10시부터 증가하여 11시 이후 급감하며, 다시 13시 이후 증가하여 15시 이후 감소하고, 23시 이후 증가하여 3시 이후 급감하는 패턴을 보이고 있는 것으로 나타났다. 섹

39) NetMiner v4.2.0.140122 Seoul: Cyram Inc.

스팅과 관련된 버즈는 목요일과 수요일에 가장 높은 추이를 보이는 반면, 주말에는 감소하는 것으로 나타났다.

[그림 5-1] 섹스팅 관련 시간별 및 요일별 버즈 현황

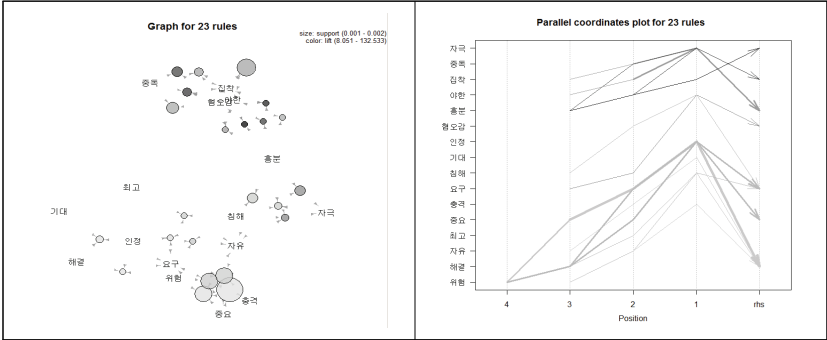


[그림 5-2]와 같이 연도별 섹스팅에 대해 긍정적인 감정(위험) 변화는 2011년 대비 평균 4.6배씩 증가하였으며, 위험 감정의 표현 단어는 요구, 충격, 인정, 자유, 중요, 침해 등의 순으로 집중된 것으로 나타났다. 섹스팅에 대해 부정적인 감정(일반) 변화는 2011년 대비 평균 2.5배씩 하였으며, 일반 감정의 표현 단어는 수치심, 상처, 고통, 모욕, 악영향, 장난 등의 순으로 집중된 것으로 나타났다.

〈표 5-1〉과 같이 섹스팅에 대한 위험 감정 키워드의 연관성 예측에서 {집착, 야한, 흥분} ⇒ {혐오감} 네 변인의 연관성은 지지도 0.001, 신뢰도는 0.933, 향상도는 132.53으로 나타났다. 이는 온라인 문서에서 집착, 야한, 흥분이 언급되면 혐오감 감정이 나타날 확률이 93.3%이며, 집착, 야한, 흥분이 언급되지 않은 문서 보다 혐오감 감정을 나타낼 확률이 132.5배 높아지는 것을 나타낸다. 따라서, 섹스팅의 위험 감정은 혐오감, 집착, 자극, 야한, 자유, 요구에 강하게 연결되어 있는 것으로 나타났다.

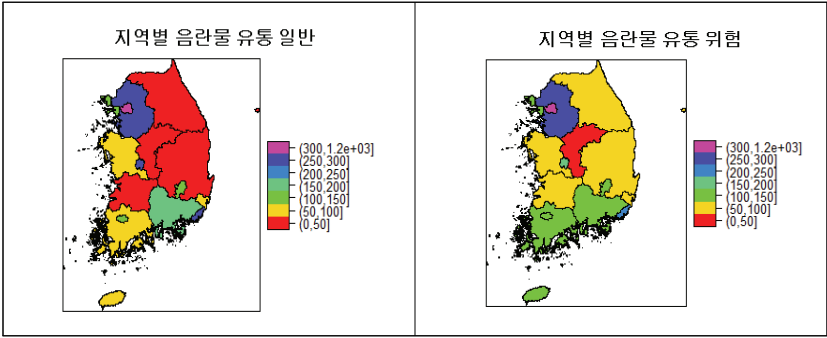
〈표 5-1〉 섹스팅의 위험 감정 키워드 연관성 예측

	지지도	신뢰도	향상도
{집착, 야한, 흥분} ⇒ {혐오감}	0.001016408	0.9333333	132.533333
{혐오감, 야한, 흥분} ⇒ {집착}	0.001016408	1.0000000	112.901639
{중독, 혐오감, 야한} ⇒ {집착}	0.001161609	0.8888889	100.357013
{혐오감, 흥분} ⇒ {집착}	0.001016408	0.8235294	92.977821
{중독, 혐오감} ⇒ {집착}	0.001234209	0.8095238	91.396565
{자유, 침해, 흥분} ⇒ {자극}	0.001089008	0.8823529	52.841432
{침해, 흥분} ⇒ {자극}	0.001234209	0.8500000	50.903913
{혐오감, 집착, 흥분} ⇒ {야한}	0.001016408	1.0000000	41.363363
{중독, 혐오감, 집착} ⇒ {야한}	0.001161609	0.9411765	38.930224
{혐오감, 집착} ⇒ {야한}	0.001597212	0.9166667	37.916416
{중독, 혐오감} ⇒ {야한}	0.001306810	0.8571429	35.454311
{혐오감, 흥분} ⇒ {야한}	0.001016408	0.8235294	34.063946
{위험, 인정, 침해} ⇒ {자유}	0.001016408	0.9333333	19.808526
{중요, 위험, 요구, 충격} ⇒ {자유}	0.001524612	0.9130435	19.377906
{자극, 침해, 흥분} ⇒ {자유}	0.001089008	0.8823529	18.726548
{침해, 흥분} ⇒ {자유}	0.001234209	0.8500000	18.039908
{자유, 위험, 요구, 충격} ⇒ {중요}	0.001524612	0.8076923	17.519927
{최고, 요구, 침해} ⇒ {자유}	0.001016408	0.8235294	17.478111
{위험, 인정, 침해} ⇒ {요구}	0.001016408	0.9333333	9.124012
{중요, 자유, 위험, 충격} ⇒ {요구}	0.001524612	0.9130435	8.925664



[그림 5-3]과 같이 지역별 섹스팅에 대한 감정은 일반은 서울, 경기, 대전, 부산, 경남 등의 순으로 높은 것으로 나타났고, 위험은 서울, 경기, 부산, 대전, 전남 등의 순으로 높은 것으로 나타났다.

[그림 5-3] 지역별 섹스팅 감정(일반, 위험)



<표 5-2>와 같이 섹스팅과 관련하여 긍정적 감정(위험)을 나타내는 버즈는 38.3%(2011년: 51.7%, 2012년 32.4%, 2013년 36.1%, 2014년 46.3%, 2015년 45.5%)로 나타났다. 섹스팅과 관련 폐해는 성범죄(71.2%), 명예훼손(9.5%), 사기(7.6%) 등의 순으로 나타났다. 섹스팅과

관련 유형은 성인음란물(71.3%), 아동음란물(16.1%), 유해광고(3.8%), 스미싱(3.8%) 순으로 나타났다. 섹스팅과 관련 내용으로는 성행위(52.7%), 누드(25.3%), 폭력(12.0%) 등의 순으로 나타났다. 섹스팅과 관련 도움으로는 전문가상담(33.8%), 통제(29.2%), 예방교육(17.7%) 등의 순으로 나타났다. 섹스팅과 관련 유통으로는 공급(58.2%), 수요(22.8%), 공유(19.0%)의 순으로 나타났다. 섹스팅과 관련 영향으로는 공부(61.7%), 건강(15.8%), 성욕(8.1%) 등의 순으로 나타났다. 섹스팅과 관련 제도로는 정보통신망법(61.2%), 가중처벌(13.6%), 벌금(13.3%), 아동청소년보호법(11.9%) 순으로 나타났다. 섹스팅과 관련 기관은 경찰청(49.6%), 방송통신위원회(15.0%), 정부(13.0%), 국회(9.2%) 등의 순으로 나타났다. 섹스팅과 관련 채널은 SNS(56.1%), 파일공유채널(33.3%), 온라인커뮤니티(10.6%) 순으로 나타났다.

〈표 5-2〉 섹스팅 관련 버즈 현황

구분	항목	N(%)	구분	항목	N(%)
감정	위험	5,277(38.3)	폐해	명예훼손	657(9.5)
	일반	8,497(61.7)		성범죄	4,931(71.2)
	계	13,774		사기	529(7.6)
채널	SNS	5,820(56.1)		음주	417(6.0)
	온라인커뮤니티	1,094(10.6)		사회문제	393(5.7)
	파일공유채널	3,454(33.3)		계	6,927
	계	10,368	영향	공부	1,432(61.7)
유형	성인음란물	7,440(71.3)		건강	367(15.8)
	유해광고	919(3.8)		대인관계	77(3.3)
	스미싱	398(3.8)		비용	133(5.7)
	아동음란물	1,676(16.1)		윤리의식	123(5.3)
	계	10,433		성욕	188(8.1)
내용	누드	1,893(25.3)		계	2,320
	성행위	3,943(52.7)	제도	가중처벌	2,105(13.6)
	원조교재	94(1.3)		정보통신망법	9,455(61.2)
	문란행위	659(8.8)		벌금	2,060(13.3)
	폭력	897(12.0)		아동청소년보호법	1,838(11.9)
	계	7,486		계	15,458

구분	항목	N(%)	구분	항목	N(%)
도움	예방교육	2,044(17.7)	기관	방송통신위원회	1,642(15.0)
	전문가상담	3,916(33.8)		경찰청	5,442(49.6)
	건강생활유도	708(6.1)		국회	1,006(9.2)
	통제	3,382(29.2)		청와대	528(4.8)
	사랑	1,530(13.2)		정부	1,427(13.0)
	계	11,580		사법기관	590(5.4)
유통	수요	2,681(22.8)		시민단체	244(2.2)
	공급	6,829(58.2)		국제기구	82(0.7)
	공유	2,232(19.0)		계	10,964
	계	11,742			

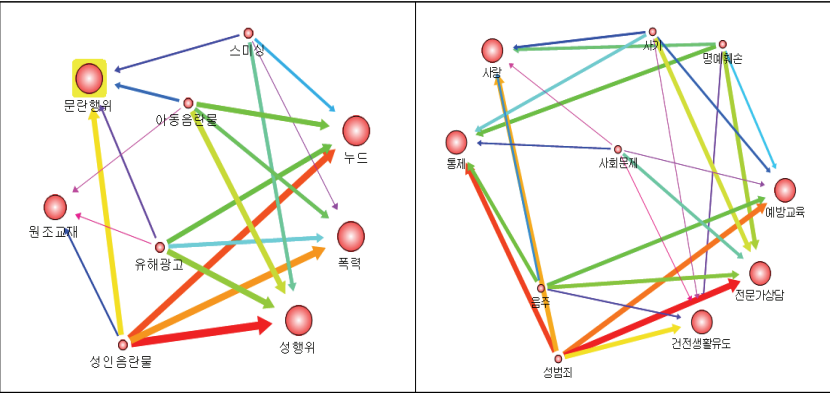
나. 섹스팅 관련 소셜 네트워크 분석

근접중심성은 평균적으로 다른 노드들과의 거리가 짧은 노드의 중심성이 높은 경우로, 근접중심성이 높은 노드는 확률적으로 가장 빨리 다른 노드에 영향을 주거나 받을 수 있다. 따라서 [그림 5-4]의 음란물 유통의 내용·유형 간의 외부 근접중심성(out closeness Centrality)을 살펴보면 성인음란물은 성행위, 누드, 폭력, 문란행위에 밀접하게 연결되어 있으며, 아동음란물은 성행위, 누드, 폭력과 접하게 연결되어 있는 것으로 나타났다. 그리고 스미싱은 성행위와 밀접하게 연결되어 있는 것으로 나타났다. 음란물 유통의 폐해·도움 간의 외부 근접중심성을 살펴보면 성범죄는 전문가상담, 통제, 예방교육과 밀접하게 연결되어 있으며, 명예훼손은 통제, 전문가상담과 밀접하게 연결되어 있는 것으로 나타났다.

〈표 5-3〉과 같이 섹스팅 위험에 대한 연관성 예측에서 가장 신뢰도가 높은 연관규칙으로는 {문란행위, 성인음란물} => {위험}이며 세 변인의 연관성은 지지도 0.031, 신뢰도는 0.765, 향상도는 1.996으로 나타났다. 이는 온라인 문서에서 문란행위, 성인음란물이 언급되면 섹스팅을 긍정적(위험)으로 생각할 확률이 76.5%이며, 문란행위, 성인음란물이 언급되

지 않은 문서 보다 섹스팅이 위험한 확률이 1.996배 높아지는 것을 나타낸다. 특히, {아동음란물} => {일반} 두 변인의 연관성은 지지도 0.074, 신뢰도는 0.609, 향상도는 0.987로 나타났다. 이는 온라인 문서에서 아동음란물이 언급되면 섹스팅이 부정적(일반)인 확률이 60.9%이며, 아동음란물이 언급되지 않은 문서 보다 섹스팅이 부정적 확률이 0.98배로 낮아지는 것을 나타낸다.

[그림 5-4] 섹스팅의 내용유형 및 폐해도움 간 외부 근접중심성



〈표 5-3〉 유형과 내용 요인에 대한 섹스팅 위험 예측

규칙	지지도	신뢰도	향상도
{문란행위, 성인음란물} => {위험}	0.03092784	0.7648115	1.9963073
{성행위, 폭력} => {위험}	0.02279657	0.7511962	1.9607686
{문란행위} => {위험}	0.03455786	0.7223065	1.8853610
{누드, 성행위, 성인음란물} => {위험}	0.04450414	0.7078522	1.8476324
{누드, 성행위} => {위험}	0.05227240	0.6792453	1.7729628
{폭력, 성인음란물} => {위험}	0.03071003	0.6438356	1.6805366
{누드, 성인음란물} => {위험}	0.06780892	0.6401645	1.6709543
{성행위, 성인음란물} => {위험}	0.13329461	0.6232179	1.6267204
{} => {일반}	0.61688689	0.6168869	1.0000000
{아동음란물} => {일반}	0.07412516	0.6091885	0.9875207
{성인음란물, 유해광고} => {위험}	0.02686220	0.6055646	1.5806419

규칙	지지도	신뢰도	향상도
{누드} => {위험}	0.08269203	0.6016904	1.5705295
{폭력} => {위험}	0.03818789	0.5863991	1.5306161
{성행위} => {위험}	0.16589226	0.5795080	1.5126290
{유해광고} => {위험}	0.03833309	0.5745375	1.4996551
{성행위, 아동음란물} => {위험}	0.02243357	0.5638686	1.4718071
{성인음란물, 아동음란물} => {일반}	0.03760709	0.5504782	0.8923487
{성인음란물} => {일반}	0.27152606	0.5026882	0.8148790
{성인음란물} => {위험}	0.26862204	0.4973118	1.2980809
{성인음란물, 아동음란물} => {위험}	0.03071003	0.4495218	1.1733396
{유해광고} => {일반}	0.02838682	0.4254625	0.6896928
{성행위} => {일반}	0.12037171	0.4204920	0.6816355
{폭력} => {일반}	0.02693480	0.4136009	0.6704647
{누드} => {일반}	0.05474082	0.3983096	0.6456768
{아동음란물} => {위험}	0.04755336	0.3908115	1.0200942
{ } => {위험}	0.38311311	0.3831131	1.0000000
{성행위, 성인음란물} => {일반}	0.08058661	0.3767821	0.6107798
{누드, 성인음란물} => {일반}	0.03811529	0.3598355	0.5833087
{누드, 성행위} => {일반}	0.02468419	0.3207547	0.5199571

다. 섹스팅의 위험에 미치는 요인

〈표 5-4〉와 같이 모든 도움요인은 섹스팅의 위험에 정적인 영향을 미치는 것으로 나타나, 건전생활유도, 사랑, 전문가상담, 통제, 예방교육 순으로 섹스팅의 위험에 도움이 되는 것으로 나타났다. 모든 영향요인은 섹스팅의 위험에 정적인 영향을 미치는 것으로 나타나, 윤리의식, 대인관계, 건강, 공부 등의 순으로 섹스팅의 위험에 영향을 주는 것으로 나타났다. 내용요인은 원조교제를 제외한 모든 요인이 섹스팅의 위험에 정적인 영향을 미치는 것으로 나타나, 문란행위, 성행위 등의 순으로 섹스팅의 위험에 영향을 주는 것으로 나타났다. 유형요인은 아동음란물을 제외한 모든 요인이 섹스팅의 위험에 정적인 영향을 미치는 것으로 나타나, 성인음란물, 스미싱, 유해광고의 순으로 섹스팅의 위험에 영향을 주는 것으로

나타났다. 폐해요인은 사회문제를 제외한 모든 요인이 섹스팅의 위험에 정적인 영향을 미치는 것으로 나타나, 음주, 성범죄 등의 순으로 섹스팅의 위험에 영향을 주는 것으로 나타났다. 모든 제도요인은 섹스팅의 위험에 정적인 영향을 미치는 것으로 나타나, 정보통신망법, 가중처벌 등의 순으로 섹스팅의 위험에 영향을 주는 것으로 나타났다. 유통요인은 공유가 섹스팅의 위험에 가장 큰 영향을 주는 것으로 나타났다.

〈표 5-4〉 섹스팅에 영향을 요인¹⁾

변수		위험			
		b [†]	S.E. [‡]	OR [§]	P
도움	예방교육	.568	.057	1.765	.000
	전문가상담	.835	.043	2.306	.000
	건전생활유도	1.326	.103	3.767	.000
	통제	.607	.045	1.834	.000
	사랑	.881	.062	2.414	.000
영향	공부	1.061	.060	2.891	.000
	건강	1.136	.123	3.114	.000
	대인관계	1.589	.316	4.901	.000
	비용	.800	.179	2.225	.000
	윤리의식	2.081	.268	8.016	.000
	성욕	.978	.173	2.659	.000
내용	누드	.739	.054	2.095	.000
	성행위	1.028	.041	2.797	.000
	원조교재	-.425	.242	.654	.079
	문란행위	1.595	.092	4.927	.000
	폭력	.664	.075	1.942	.000
유형	성인음란물	1.067	.037	2.906	.000
	유해광고	.689	.072	1.992	.000
	스미싱	.992	.112	2.695	.000
	아동음란물	.053	.056	1.054	.341
폐해	명예훼손	.412	.086	1.509	.000
	성범죄	1.162	.038	3.197	.000
	사기	.632	.096	1.881	.000
	음주	1.334	.120	3.795	.000
	사회문제	-.090	.110	.913	.411

변수		위험			
		b [†]	S.E. [‡]	OR [§]	P
제도	가중처벌	.528	.050	1.695	.000
	정보통신망법	.643	.042	1.903	.000
	벌금	.438	.051	1.550	.000
	아동청소년보호법	.439	.053	1.551	.000
유통	수요	.170	.044	1.186	.000
	공급	.006	.035	1.006	.876
	공유	.662	.047	1.939	.000

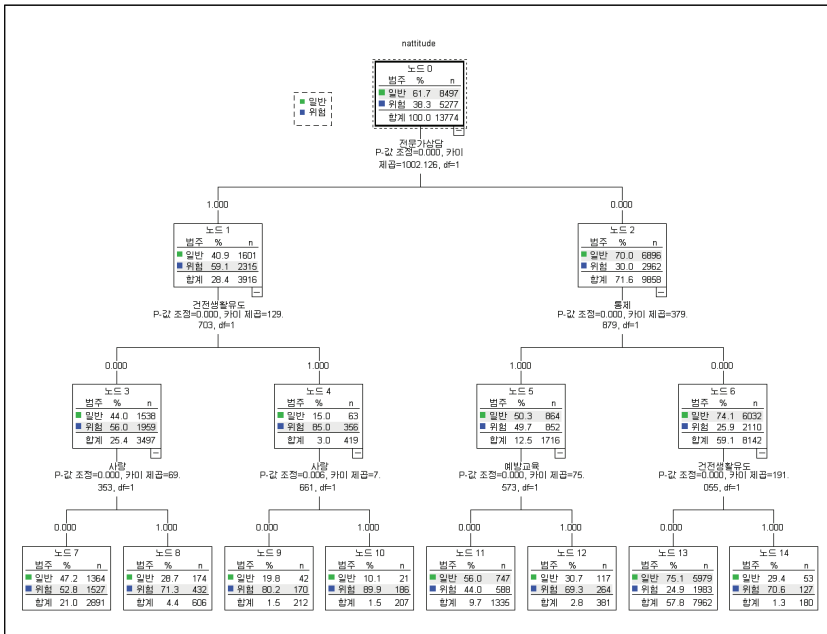
주: 1) 기본범주: 일반, † Standardized coefficients, ‡ Standard error, § Adjusted odds ratio

라. 섹스팅 관련 위험 예측모형

본 연구에서는 섹스팅 관련 위험을 예측하기 위하여 섹스팅의 도움요인, 내용요인, 유형요인에 대해 데이터마이닝 분석을 실시하였다. 섹스팅의 도움요인이 섹스팅의 위험예측 모형에 미치는 영향은 [그림 5-5]와 같다. 나무구조의 최상위에 있는 네모는 루트노드로서, 예측변수(독립변수)가 투입되지 않은 종속변수(위험, 일반)의 빈도를 나타낸다. 루트노드에서 섹스팅의 위험은 38.3%(5,277건), 일반은 61.7%(8,497건)으로 나타났다. 루트노드의 하단의 가장 상위에 위치하는 요인은 섹스팅 위험예측에 가장 영향력이 높은(관련성이 깊은)요인으로 ‘전문가상담요인’의 영향력이 가장 큰 것으로 나타났다. ‘전문가상담요인’이 있을 경우 섹스팅의 위험은 이전의 38.3%에서 58.1%로 증가한 반면, 일반은 이전의 61.7%에서 40.9%로 감소하였다. ‘전문가상담요인’이 있고 ‘건전생활유도요인’이 있는 경우 섹스팅의 위험은 이전의 58.1%에서 85.0%로 증가한 반면, 일반은 이전의 40.9%에서 15.0%로 감소하였다. <표 5-5>의 섹스팅 도움요인의 위험예측 모형에 대한 이익도표와 같이 섹스팅의 위험에 가장 영향력이 높은 경우는 ‘전문가상담요인’이 있고 ‘건전생활유도요인’이 있으며 ‘사랑요인’이 있는 조합으로 나타났다. 즉, 10번 노드의 지

수(index)가 234.5%로 뿌리마디와 비교했을 때 10번 노드의 조건을 가진 집단이 섹스팅 위험이 높을 확률이 2.34배로 나타났다. 일반인에게 가장 영향력이 높은 경우는 ‘전문가상담요인’이 없고 ‘통제요인’이 없고, ‘건전생활유도요인’이 없는 조합으로 나타났다. 즉 13번 노드의 지수가 121.7%로 뿌리마디와 비교했을 때 13번 노드의 조건을 가진 집단이 일반의 확률이 1.22배로 나타났다.

[그림 5-5] 도움요인의 섹스팅 위험 예측모형



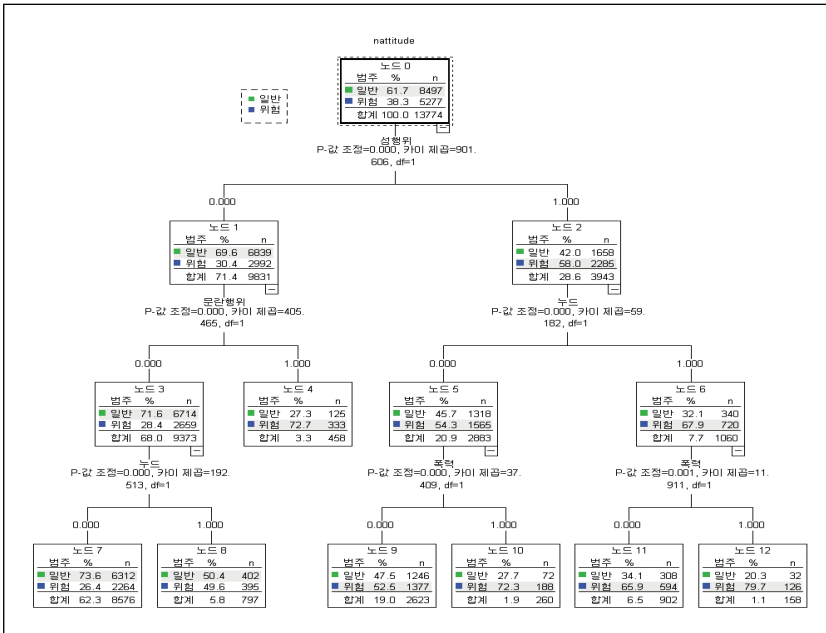
〈표 5-5〉 도움요인의 섹스팅 위험 예측 모형에 대한 이익도표

구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
위험	10	207	1.5	3.5	234.5	207	1.5	3.5	234.5
	9	212	1.5	3.2	209.3	419	3.0	6.7	221.8
	8	606	4.4	8.2	186.1	1025	7.4	14.9	200.7
	14	180	1.3	2.4	184.2	1205	8.7	17.3	198.2
	12	381	2.8	5.0	180.9	1586	11.5	22.3	194.0
	7	2891	21.0	28.9	137.9	4477	32.5	51.3	157.8
	11	1335	9.7	11.1	115.0	5812	42.2	62.4	147.9
	13	7962	57.8	37.6	65.0	13774	100.0	100.0	100.0
일반	13	7962	57.8	70.4	121.7	7962	57.8	70.4	121.7
	11	1335	9.7	8.8	90.7	9297	67.5	79.2	117.3
	7	2891	21.0	16.1	76.5	12188	88.5	95.2	107.6
	12	381	2.8	1.4	49.8	12569	91.3	96.6	105.8
	14	180	1.3	.6	47.7	12749	92.6	97.2	105.0
	8	606	4.4	2.0	46.5	13355	97.0	99.3	102.4
	9	212	1.5	.5	32.1	13567	98.5	99.8	101.3
	10	207	1.5	.2	16.4	13774	100.0	100.0	100.0

섹스팅의 내용요인이 섹스팅의 위험예측 모형에 미치는 영향은 [그림 5-6]과 같다. 섹스팅 위험예측에 가장 영향력이 높은 내용요인으로 ‘성행위요인’의 영향력이 가장 큰 것으로 나타났다. ‘성행위요인’이 높을 경우 섹스팅의 위험은 이전의 38.3%에서 58.0%로 증가한 반면, 일반인은 이전의 61.7%에서 42.0%로 감소하였다. ‘성행위요인’이 높고 ‘누드요인’이 높은 경우 섹스팅의 위험은 이전의 58.0%에서 67.9%로 증가한 반면, 일반인은 이전의 42.0%에서 32.1%로 감소하였다. 〈표 5-6〉의 섹스팅 내용요인의 위험예측 모형에 대한 이익도표와 같이 섹스팅의 위험에 가장 영향력이 높은 경우는 ‘성행위요인’이 있고 ‘누드요인’이 있고 ‘폭력요인’이 있는 조합으로 나타났다. 즉, 12번 노드의 지수가 208.2%로 뿌리마디와 비교했을 때 12번 노드의 조건을 가진 집단이 섹스팅 위험이 높을 확률이 2.08배로 나타났다. 일반인에게 가장 영향력이 높은 경우는 ‘성행

위요인'이 없고 '문란행위요인'이 없고 '누드요인'이 없는 조합으로 나타났다. 즉 7번 노드의 지수가 119.3%로 뿌리마디와 비교했을 때 7번 노드의 조건을 가진 집단이 일반의 확률이 1.19배로 나타났다.

[그림 5-6] 내용 요인의 섹스팅 위험 예측모형



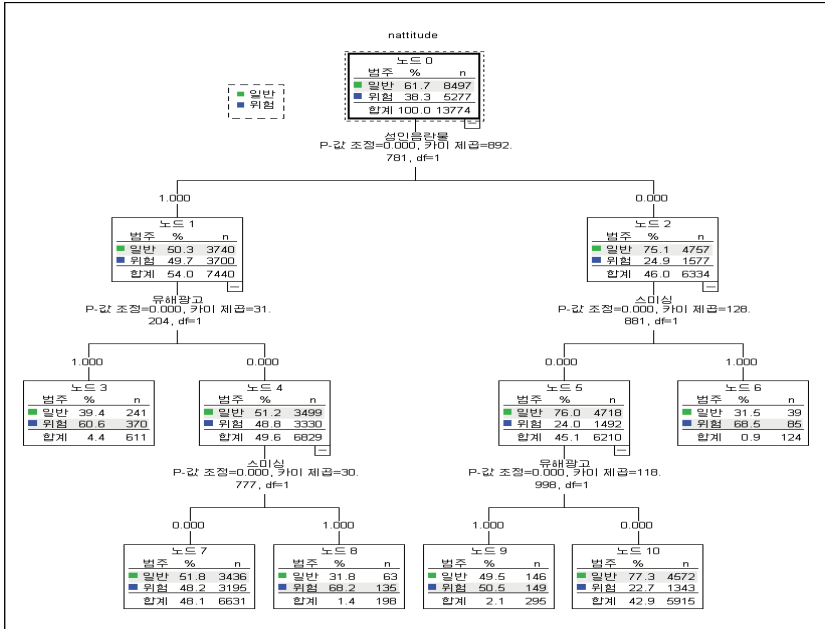
〈표 5-6〉 내용 요인의 섹스팅 위험 예측 모형에 대한 이익도표

구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
위험	12	158	1.1	2.4	208.2	158	1.1	2.4	208.2
	4	458	3.3	6.3	189.8	616	4.5	8.7	194.5
	10	260	1.9	3.6	188.7	876	6.4	12.3	192.8
	11	902	6.5	11.3	171.9	1778	12.9	23.5	182.2
	9	2623	19.0	26.1	137.0	4401	32.0	49.6	155.3
	8	797	5.8	7.5	129.4	5198	37.7	57.1	151.3
	7	8576	62.3	42.9	68.9	13774	100.0	100.0	100.0

구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
일반	7	8576	62.3	74.3	119.3	8576	62.3	74.3	119.3
	8	797	5.8	4.7	81.8	9373	68.0	79.0	116.1
	9	2623	19.0	14.7	77.0	11996	87.1	93.7	107.6
	11	902	6.5	3.6	55.4	12898	93.6	97.3	103.9
	10	260	1.9	.8	44.9	13158	95.5	98.2	102.7
	4	458	3.3	1.5	44.2	13616	98.9	99.6	100.8
	12	158	1.1	.4	32.8	13774	100.0	100.0	100.0

섹스팅의 유형요인이 섹스팅의 위험예측 모형에 미치는 영향은 [그림 5-7]과 같다. 섹스팅 위험예측에 가장 영향력이 높은 유형요인으로 ‘성인 음란물요인’의 영향력이 가장 큰 것으로 나타났다. ‘성인음란물요인’이 있을 경우 섹스팅의 위험은 이전의 38.3%에서 49.7%로 증가한 반면, 일반인은 이전의 61.7%에서 50.3%로 감소하였다. ‘성인음란물요인’이 있고 ‘유해광고요인’이 있는 경우 섹스팅의 위험은 이전의 49.7%에서 60.6%로 증가한 반면, 일반인은 이전의 50.3%에서 39.4%로 감소하였다. <표 5-7>의 섹스팅 내용요인의 위험예측 모형에 대한 이익도표와 같이 음섹스팅의 위험에 가장 영향력이 높은 경우는 성인음란물요인’이 없고 ‘스미싱요인’이 있는 조합으로 나타났다. 즉, 6번 노드의 지수가 178.9%로 뿌리마디와 비교했을 때 6번 노드의 조건을 가진 집단이 섹스팅 위험이 높을 확률이 약 1.78배로 나타났다. 일반인에게 가장 영향력이 높은 경우는 ‘성인음란물요인’이 없고 ‘스미싱요인’이 없고 ‘유해광고요인’이 없는 조합으로 나타났다. 낮은 조합으로 나타났다. 즉 10번 노드의 지수가 125.3%로 뿌리마디와 비교했을 때 10번 노드의 조건을 가진 집단이 일반의 확률이 1.25배로 나타났다.

[그림 5-7] 유형 요인의 섹스팅 위험 예측 모형



〈표 5-7〉 유형 요인의 섹스팅 위험 예측 모형에 대한 이익도표

구분	노드	이익지수				누적지수			
		노드(n)	노드(%)	이익(%)	지수(%)	노드(n)	노드(%)	이익(%)	지수(%)
위험	6	124	.9	1.6	178.9	124	.9	1.6	178.9
	8	198	1.4	2.6	178.0	322	2.3	4.2	178.3
	3	611	4.4	7.0	158.1	933	6.8	11.2	165.1
	9	295	2.1	2.8	131.8	1228	8.9	14.0	157.1
	7	6631	48.1	60.5	125.8	7859	57.1	74.5	130.7
	10	5915	42.9	25.5	59.3	13774	100.0	100.0	100.0
일반	10	5915	42.9	53.8	125.3	5915	42.9	53.8	125.3
	7	6631	48.1	40.4	84.0	12546	91.1	94.2	103.5
	9	295	2.1	1.7	80.2	12841	93.2	96.0	102.9
	3	611	4.4	2.8	63.9	13452	97.7	98.8	101.2
	8	198	1.4	.7	51.6	13650	99.1	99.5	100.4
	6	124	.9	.5	51.0	13774	100.0	100.0	100.0

6. 결론 및 고찰

본 연구는 국내의 온라인 뉴스 사이트, 블로그, 카페, SNS, 게시판 등 인터넷을 통해 수집된 소셜 빅데이터를 네트워크 분석과 데이터마이닝의 연관분석과 의사결정나무 분석기법을 적용하여 분석함으로써 우리나라 섹스팅의 위험요인을 예측하고자 하였다.

본 연구의 결과를 요약하면 다음과 같다. 첫째, 섹스팅과 관련된 버즈는 10시부터 증가하여 11시 이후 급감하며, 다시 13시 이후 증가하여 15시 이후 감소하고, 23시 이후 증가하여 3시 이후 급감하는 패턴을 보이고 있다. 섹스팅과 관련된 버즈는 목요일과 수요일이 가장 높은 반면, 주말에는 감소하는 것으로 나타났다. 둘째, 섹스팅에 대한 긍정적 감정(위험)의 표현 단어는 요구, 충격, 인정, 자유, 중요, 침해 등의 순으로 집중된 것으로 나타났다. 섹스팅에 대한 위험 감정 키워드의 연관성 예측에서 위험 감정은 ‘혐오감, 집착, 자극, 야한, 자유, 요구’에 강하게 연결되어 있는 것으로 나타났다. 셋째, 섹스팅에 대한 긍정적 감정(위험)을 나타내는 버즈는 38.3%로 나타났다. 넷째, 섹스팅의 영향은 윤리의식, 대인관계, 성욕, 건강, 공부, 비용 순으로 위험한 것으로 나타났으며, 유통방식은 수요보다 공유의 위험이 더 큰 것으로 나타났다. 다섯째, 섹스팅 관련 네트워크 분석에서 아동음란물은 ‘성행위, 누드, 폭력’과 밀접하게 연결되어 있으며, 스미싱은 성행위와 밀접하게 연결되어 있는 것으로 나타났다.

여섯째, 섹스팅과 내용이 섹스팅에 영향을 미치는 연관분석에서 온라인 문서 상에 ‘문란행위, 성인음란물’이 언급되면 섹스팅을 긍정적으로 생각(위험)하는 확률이 높은 것으로 나타났다. 일곱째, 섹스팅의 위험에 도움을 주는 요인으로는 ‘전문가상담’, ‘건강생활유도’, ‘사랑’ 요인이 있는 조합으로 나타났다. 섹스팅의 위험에 영향을 미치는 내용요인으로는 ‘성행위’와 ‘누드’ 요인이 있는 조합으로 나타났다. 섹스팅에 영향을 미치

는 유형요인으로는 ‘성인음란물’이 없고 ‘스미싱’ 요인이 있는 조합으로 나타났다. 본 연구를 근거로 우리나라의 섹스팅의 문제에 대해 다음과 같은 정책적 함의를 도출할 수 있다. 첫째, 섹스팅에 대한 온라인 문서가 23시에서 3시 사이에 집중하고 있어 늦은 시간에 청소년의 섹스팅을 방지할 수 있는 대책마련이 요구되고 있다. 현재 청소년의 게임 중독을 방지하기 위하여 여성가족부는 셧다운제, 문화체육관광부는 게임시간 선택제를 실시하고 있다. 셧다운제는 16세 미만의 청소년에게 오전 0시부터 오전 6시까지 심야 6시간 동안 인터넷 게임 제공을 제한 하는 것을 말하며, 게임시간 선택제는 청소년 게임중독을 막기 위해 보호자가 만 18세 미만 청소년의 게임 시간을 선택하여 제한할 수 있는 제도를 말한다. 섹스팅을 방지하기 위해서는 법령과 정책의 정합성 제고를 위한 일원화 방안이 검토되어야 하며, 특히 섹스팅이 심야시간에 발생한다는 것을 고려할 때 여성가족부에서 실시하는 셧다운제로 일원화하는 방안으로 검토되어야 할 것으로 본다. 둘째, 2011년부터 2015년 3월까지 한국의 섹스팅 위험은 38.3%로 나타났다. 이는 2014년 여성가족부의 조사결과 청소년의 섹스팅은 2011년 12.3%, 2012년 20.5%, 2014년 52.6%로 평균 28.5%를 어느 정도 지지하는 것으로 나타났다. 본 연구의 결과와 같이 2011년을 제외하고 섹스팅의 위험은 지속적으로 증가하는 것으로 나타난 것은 청소년들에게 스마트 기기의 보급의 확산과 관련이 있는 스마트폰 중독과 관련이 있는 것으로 보며, 정부차원에서 예방교육과 치료·상담을 통한 해소방안 마련과 함께 청소년의 유해정보 차단을 위한 다양한 어플리케이션이 개발되어야 할 것으로 본다. 셋째, 윤리의식, 대인관계, 성욕, 건강, 공부, 비용에 섹스팅의 위험이 있는 것은 기존의 연구(이창훈·김은영, 2009, p101)에서 친구사이에 주목을 받기 위해서 섹스팅을 하며 섹스팅을 하는 청소년들은 성적이 떨어지거나 수치심이나 도덕심을 고려치 않

고 자신의 정서적·심리적 쾌락을 추구하는 잠재적 sexter가 될 가능성이 높다는 연구를 지지하는 것이다. 넷째, 섹스팅의 유통방식이 수요보다 공유의 위험이 더 큰 것으로 나타나 청소년 등의 음란물 접근방법이 특정 인터넷 사이트를 거치지 않고도 일대일 파일 공유 네트워크를 인기가 있는 일대일 파일공유 네트워크로 음란물을 쉽게 전송받을 수 있기 때문에 정부차원에서 웹하드와 파일공유사이트에서 음란물 유통에 대한 지속적인 모니터링이 필요할 것으로 본다.

다섯째, 섹스팅의 내용요인의 위험예측에서 ‘성인물’이 없고 ‘스미싱’이 있는 문서인 경우 섹스팅 위험이 가장 높은 것으로 나타나, 일상적인 문자메시지에 성인음란물의 내용을 포함하여 유통시키는 섹스팅에 대한 지속적인 모니터링이 필요할 것으로 본다.

여섯째, 섹스팅의 위험에 ‘전문가상담, 건전생활유도, 사랑’이 가장 영향력이 높은 것으로 나타났다. 이는 청소년의 섹스팅의 대책을 위한 학교와 가정 차원에서의 접근이 필요한 것을 시사하는 것이다. 따라서 학교차원에서는 학생들에게 미디어 메시지를 비판적으로 읽어내고 이해하는 능력에 관한 교육과 함께 부모의 모니터링과 이해의 중요성을 학부모에게 이야기하고 성적으로 음란한 내용 대신 건전한 성과 이성관계에 대해서 배울 수 있도록 대안책을 마련해주는 것이 필요하며(Flood, 2009 p.143), 가정에서 부모의 규제적 노력이 청소년의 섹스팅을 방지한데 효과가 있을 것으로 본다. 끝으로 섹스팅을 모니터링 할 수 있는 정부차원의 대책이 마련되어야 할 것이다. 현재 사어버안전국(<http://cyberbureau.police.go.kr>)에서 음란물 근절을 위해 아동·청소년이용 음란물 전담조직을 설치하여 웹하드·P2P 사이트, 인터넷 음란물을 집중적으로 단속을 하고 있지만 실시간으로 이루어지는 섹스팅에 대한 모니터링은 어려운 실정이다. 따라서 섹스팅의 위험요소를 분석하여 섹스팅에 위험이 있는

온라인 문서를 모니터링하고 제제할 수 있는 어플리케이션이 개발되어야 할 것으로 본다.

본 연구는 개개인의 특성을 가지고 분석한 것이 아니고 그 구성원이 속한 전체 집단의 자료를 대상으로 분석하였기 때문에 이를 개인에게 적용하였을 경우 생태학적 오류(ecological fallacy)가 발생할 수 있다(Song et al., 2014). 또한, 본 연구에서 감성분석 결과 정의된 섹스팅의 위험은 온라인 문서 내에서 발생된 감정 단어에 대한 감성분석의 결과로 정의되었기 때문에 기존의 오프라인 조사 등을 통한 섹스팅 위험의 조작적 정의와 다를 수 있으며, 청소년으로 예상되는 온라인 문서(버즈)를 분석 대상으로 선정하였기 때문에 성인이 남긴 문서 중에 청소년의 키워드(10대 미만, 초등학생, 중등학생, 고등학생 등)가 있으면 대상으로 포함되어 분석 대상의 정확성에 한계가 있을 수 있다. 그럼에도 불구하고 본 연구는 소셜 빅데이터에서 섹스팅의 위험에 대한 실제적인 내용을 빠르게 효과적으로 파악함으로써 기존의 일부 학교를 대상으로 익명성을 전제로 한 횡단조사의 한계를 보완할 수 있는 새로운 분석방법을 제시하였다는 점에서 조사방법론적 의의를 가진다고 할 수 있다.

끝으로 섹스팅은 스마트폰의 문자메시지나 콘텐츠로 유통되기 때문에 기존의 표본추출을 통한 횡단조사 방법과 더불어 소셜 미디어에서 수집된 빅데이터의 활용과 분석을 할 경우, 섹스팅의 위험예측은 더욱 신뢰성이 있을 것으로 본다.

- Chalfen.R(2009). It's Only a Picture. *The Previous Cross-sectional Methodology Through Visual Studies*, 24(3); 258~269.
- Choi, JY & Chung, DH(2014). Teenagers with Smartphones Exposed to Sexual Content, *The Journal of the Korea Contents Association*, 14(4); 445~455.
- Eberstadt, M. & Layden, M.A.(2010). *The Social Costs of Pornography, A Statement of Findings and Recommendations*, The Witherspoon Institute.
- http://www.lifeissues.net/writers/may/may_17pornographycost.html
- Flood, M. (2009). Youth, Sex, and the Internet, *Counselling, Psychotherapy, and Health*, 5(1); 131~147.
- Greenfield, P.M.(2004). Inadvertent Exposure to Pornography on the Internet : Implications of Peer-to-peer File-sharing Networks for Child Development and Families, *Applied Development Psychology*, 25; 741~750.
- Gruber, E. & Thau, H.(2003). Sexually Related Content on Television and Adolescents of Color : Media Theory, Physiological Development, and Psychological Impact, *Journal of Negro Education*, 72(4); 438~456.
- Itzin, C.(1997). Pornography and the Organization of Intra familial and Extra familial Child Sexual Abuse: Developing a Conceptual Model. *Child Abuse Review*, 6; 94~106.
- Jaishankar.K(2009). Sexting: A New Form of Victimless Crime?, *International Journal of Cyber Criminology*, 3(1); 21~25.
- Joo, JH & Kim, HI(2013). Exploration of Relationship Among Korean Adolescents' Sexual Orientations, Exposure to Internet

- Pornography and Sexual Behaviors after Exposure : Focused on PLS Path Modeling Analysis. *The Journal of Digital Policy & Management*, 11(6); 11~21.
- Lounsbury, K., Mitchell, K.J. & Finkelhor, D.(2011). *The True Prevalence of 'Sexting'*, Crimes Against Children Research Center.
- Kim, JG(2012). The Predictive Factors on Adolescents' Exposure to Sexually Explicit Online Materials and Adolescents's Sexuality, *Social Science Studies*, 24(1); 1~33.
- Kim, M & Kawk, JB(2011). You Cybersex Addiction in the Digital Media Era. *Soon Chunhyang Journal of Humanities*, 29; 283~326.
- Kim, MN(2012). *A Study on Understanding of Adolescent Sex Offenses and Pornography Addiction*, Korean Association of Addiction Crime, 2(1); 47~71.
- Korea Internet and Security Agency(2014). *2014 Mobile Internet Use Survey*, 19.
- Lee, CH & Kim, EG(2009). A Study of Sexting Activities Among South Korean Youths. *Korean Institute of Criminology Research Series*, 09-17; 11~131.
- Lenhart, A.(2009). *Teens and Sexting: How and Why Minor Teens are Sending Sexually Suggestive Nude or Nearly Nude Images via Text Messaging*, Pew Internet & American Life Project.
- Malamuth, N. M. & Check, J.V.P.(1985). The Effects of Aggressive Pornography on Belief in Rape Myths: Individual Differences. *Journal of Research in Personality*, 19; 299~320.
- McLaughlin, J.H(2010). *Crime and Punishment: Teen Sexting in Context*, Florida Coastal School of Law.
- Ministry of Science, ICT and Future Planning & Korea Internet and

- Security Agency(2014). *2014 Internet use Survey*.
- Ministry of Science, ICT and Future Planning & National Information Society Agency(2014). *2014 Internet Addiction Survey*.
- Ministry of the Gender Equality & Family(2014). *2014 Youth Harmful Environment Contact Survey*.
- Mitchell, K.J., Finkelhor, D., Jones, L.M & Wolak, J.(2011). Prevalence and Characteristics of Young Sexting: A National Study, *Pediatrics*, 129(13); 13~20.
- Ringrose, L., Gill, R., Livingstone, R. & Harvey, L.(2012). A Qualitative Study of Children, Young People and d Tative Study of Children, *Young People and National Society for the Prevention of Cruelty to children*, London, UK.
- Sherman, S.J. & Fazio, R.H.(1983). Parallels between Attitudes and Traits as Predictors of Behavior, *Journal of Personality*, 51(3); 308~345.
- Shin, SM.(2013). Associations of Demographic and Psycho-Social Characteristics with Frequent Watching Pornography Material or the Adults-Only Internet Chatting, *The Korean Journal of Stress Research*, 21(4); 275~281.
- Shim, JW(2010). The Role of Timing to Exposure to Pronography in What Adolescent Boys and Girls Think About Sexual Issues, Media, *Gender & Culture*, 16; 75~105.
- Song TM, Song J, An JY, Hayman LL & Woo JM(2014). Psychological and Social Factors Affecting Internet Searches on Suicide in Korea: A Big Data Analysis of Google Search Trends, *Yonsei Med Journal*, 55(1); 254~263.
- Temple, J.R., Paul, J.A., Berg, P.V.D., Le, V.D., McElhany, A., Temple, B.W.(2012). Teen Sexting and Its Association With Sexual

- Behaviors, *Archives of Pediatrics and Adolescent Med*, 166(9); 828~833. doi:10.1001/Archpediatrics.2012.835
- Tokyo Congress(2010). *Tokyo Congress C Youth Protection Ordinance Revised*.
- U.S. Department of state(2003). Prosecutorial Remedies and Other Tools To end the Exploitation of Children Today Act of 2003. Title V, *Obscenity and Pornography*.
- Walker, S., Sanci, L. & Temple-Smith, M.(2011). Sexting and Young People, *Youth Studies Australia*. 30(4); 8~16.
- Ybarra, M.L. & Mitchell, K.J.(2005). Exposure to Internet Pornography among Children and Adolescents: A National Survey. *Cyber Psychology & Behavior*, 8(5); 473~486.

간행물회원제 안내

▶ 회원에 대한 특전

- 본 연구원이 발행하는 판매용 보고서는 물론 「보건복지포럼」, 「보건사회연구」도 무료로 받아보실 수 있으며 일반 서점에서 구입할 수 없는 비매용 간행물은 실비로 제공합니다.
- 가입기간 중 회비가 인상되는 경우라도 추가 부담이 없습니다.

▶ 회원종류

- 전체간행물회원 : 120,000원
- 보건분야 간행물회원 : 75,000원
- 사회분야 간행물회원 : 75,000원
- 정기간행물회원 : 35,000원

▶ 가입방법

- 홈페이지(www.kihasa.re.kr) - 발간자료 - 간행물구독안내

▶ 문의처

- (30147) 세종특별자치시 시청대로 370 세종국책연구단지 사회정책동 1F~5F
간행물 담당자 (Tel: 044-287-8157)

KIHASA 도서 판매처

- | | |
|---|---|
| ■ 한국경제서적(총판) 737-7498 | ■ 교보문고(광화문점) 1544-1900 |
| ■ 영풍문고(종로점) 399-5600 | ■ 서울문고(종로점) 2198-2307 |
| ■ Yes24 http://www.yes24.com | ■ 알라딘 http://www.aladdin.co.kr |

연구보고서 발간자료 목록

발간번호	보고서명	연구책임자
연구 2015-01	의료이용 합리화를 위한 실태분석과 제도 개선방안	김남순
연구 2015-02	보건의료인력의 연수교육 개선방안	오영호
연구 2015-03	의료패러다임 변화에 따른 미래 보건의료산업 정책과제	김대중
연구 2015-04	한국의 건강불평등 지표와 정책과제: 한국의 건강격차 모니터링	김동진
연구 2015-05	한국 의료 질 보고서: 의료서비스 질 향상에 대한 의료시스템의 성과와 과제	강희정
연구 2015-06	보건의료 공급체계 재설계를 통한 국민의료비 합리화 방안	정영호
연구 2015-07	호스피스완화의료 활성화 방안 -노인장기요양서비스 이용자를 중심으로	최정수
연구 2015-08	주요 소득보장정책의 효과성 평가	강신옥
연구 2015-09	돌봄보건의료 연합서비스(Joined-up Services)공급 모형에 관한 전망과 과제	박세경
연구 2015-10	가족형태 다변화에 따른 부양체계 변화전망과 공시간 부양 분담방안	김유경
연구 2015-11	공공 사회복지 전달체계의 변화와 정책적 함의	이현주
연구 2015-12	각국 공공부조제도 비교 연구: 스웨덴&프랑스&미국 편	임완섭
연구 2015-13	사회보장 역할분담 구조 변화와 정책적 대응방안 연구	정해식
연구 2015-14	시간제 일자리 확산이 소득불평등과 빈곤에 미치는 영향	김현경
연구 2015-15	사회보장재정 재구조화를 위한 중장기 전략연구	고제이
연구 2015-16	사회보장재정과 경제 선순환 국제비교연구	유근춘
연구 2015-17	공사적 연금 체계의 노후소득보장 효과 전망과 발전 방향	우해봉
연구 2015-18	사회보장 중장기 재정추계 모형 개발을 위한 연구: 장기재정전망과 재정평가	신화연
연구 2015-19	사회복지법인의 재정운용 실태와 제도개선 방안	고경환
연구 2015-20	지역단위 복지서비스 수요-공급 분석	정홍원
연구 2015-21-01	가족변화에 따른 결혼·출산·양육행태 변화와 정책과제	이삼식
연구 2015-21-02	임산출산 및 영아기 양육 인프라의 형평성과 정책과제	이소형
연구 2015-21-03	동아시아 국가의 가족정책 비교연구	신윤정
연구 2015-21-04	여성의 노동·출산·양육행태와 정책과제	박종서
연구 2015-21-05	저출산고령사회 동태적 분석을 위한 지역추적조사	오영희
연구 2015-21-06	은퇴전환기 중고령자의 일·여가현황과 여가 증진방안	강은나
연구 2015-21-07	노인 돌봄(케어)서비스의 제공주체 간 역할정립과 연계체계 구축 방안	선우덕
연구 2015-21-08	연령통합 지표 개발과 적용	정경희
연구 2015-21-09	인구구조 변화와 사회보장 재정의 사회경제적 파급 효과 연구	원종욱
연구 2015-21-10	저출산 극복을 위한 아동보호체계 국제비교 연구: 한중일 비교를 중심으로	류정희
연구 2015-21-11	소셜 빅데이터 기반 저출산 정책 수요 예측	송태민
연구 2015-22	사회정신건강연구센터 운영	이상영
연구 2015-23	건강영향평가 사업 운영	김정진
연구 2015-24	한국형 복지모형 구축: 생애주기별 소득·자산·소비 연계형 복지모형 구축	여유진
연구 2015-25	사회통합 실태진단 및 대응방안 II: 사회통합과 사회이동	여유진/정해식
연구 2015-26	정책결정자의 사회통합 인식에 관한 연구	김미근
연구 2015-27	아시아 각국의 복지제도 비교연구: 소득보장체계를 중심으로	노대명
연구 2015-28	지방자치단체 복지정책평가센터 운영	강혜규
연구 2015-29	보건복지통계정보 통합 관리 및 운영	오미애
연구 2015-30-1	국민건강과 안전을 위한 아동안전전략 구축방안	김미숙

발간번호	보고서명	연구책임자
연구 2015-30-2	국민건강과 안전을 위한 식품안전전략 구축방안	김정선
연구 2015-31	2015년 전국 출산력 및 가족보건·복지실태조사	이삼식
연구 2015-32	2015년 한국복지패널 기초분석: 한국복지패널로 본 한국의 복지실태	노대명
연구 2015-33	2013 한국의료패널 기초분석보고서(II)-만성질환, 임산출산, 보건의식행태-	이수형
연구 2015-34	2014년 빈곤통계연보	정은희
연구 2015-35	2015년 소셜 빅데이터 기반 보건복지 이슈 동향 분석	송태민
연구 2015-36	의료기술 혁신과 의료보장체계의 지속성을 위한 국제동향과 정책과제	박실비아
연구 2015-37	보건분야 국제공동연구사업: 주요국 보건의료산업 육성정책 변화와 동향연구	김대중