http://dx.doi.org/10.5392/JKCA.2016.16.08.001

소셜 빅데이터 기반 사회적 이슈 리스크 유형 분류

Social Issue Risk Type Classification based on Social Bigdata

오효정*, 안승권**, 김 용***
전북대학교 대학원 기록관리학과*, ㈜바른교육**, 전북대학교 문헌정보학과***

Hyo-Jung Oh(ohj@jbnu.ac.kr)*, Seung-Kwon An(ceo@trueedu.co.kr)**,
Yong Kim(yk9118@jbnu.ac.kr)***

요약

소셜미디어의 정치사회적인 활용도가 높아짐에 따라 소셜빅데이터 기반 온라인 동향분석 및 모니터링 기술에 대한 수요 역시 급증하고 있다. 본 논문에서는 이러한 요구에 부합, 특히 여론 형성의 악영향을 끼치는 부정적 이슈 탐지를 위해 사회적으로 파장이 큰 이슈 중 공공여론이 부정적으로 형성될 이슈를 '리스크'로 정의하고 세부 유형을 분류한다. 리스크 유형 정의를 위해 뉴스 문서집합을 대상으로 전수조사를 실시하였으며, 이슈 분야 즉 도메인별 특성을 파악하여 세부 유형을 정의한다. 또한 뉴스와 같은 공적미디어를 통해 정의된 리스크 유형이 개인화된 소셜 미디어에 나타난 리스크 유형과 어떤 차이가 있는지를 알아보기 위해 교차분석을 수행한다. 조사 결과에 따라 6개의 도메인별로 58개의 세부 유형을 정의하고 기계학습 방법을 통해 자동 분류 학습 모델을 구축한다. 실험 결과를 통해 소셜 미디어에 나타난 사회적 이슈 리스크를 자동으로 탐지, 분류가 가능함을 보인다.

■ 중심어 : | 소셜빅데이터 | 리스크 | 유형분석 | 모니터링 | 기계학습 |

Abstract

In accordance with the increased political and social utilization of social media, demands on online trend analysis and monitoring technologies based on social bigdata are also increasing rapidly. In this paper, we define 'risk' as issues which have probability of turn to negative public opinion among big social issues and classify their types in details. To define risk types, we conduct a complete survey on news documents and analyzed characteristics according to issue domains. We also investigate cross-medias analysis to find out how different public media and personalized social media. At the result, we define 58 risk types for 6 domains and developed automatic classification model based on machine learning algorithm. Based on empirical experiments, we prove the possibility of automatic detection for social issue risk in social media.

■ keyword : | Social Bigdata | Risk | Type Analysis | Monitoring | Machine Learning |

I. 서 론

최근 기존의 단순 웹 문서 뿐 아니라 소통과 여론 형성의 장으로 빠르게 확산되고 있는 소셜미디어 상의 다양한 이슈를 실시간으로 탐지하고 그에 대한 국민 여론을 수렴하고자 하는 요구가 대두되고 있다[1][2]. 한편,

* 이 논문은 2015년도 전북대학교 신임교수 연구비 지원에 의하여 연구되었음
* 이 논문은 2016년도 한국연구재단 연구비 지원에 의한 결과의 일부임 (과제번호: 2016R1A2B100800)
접수일자 : 2016년 02월 04일
수정일자 : 2016년 03월 29일
심사완료일 : 2016년 05월 02일
교신저자 : 김용, e-mail : yk9118@jbnu.ac.kr

여론 수렴 방식 역시 기존의 전화여론조사, 탐방조사 등 편향되고 지엽적인 여론 수렴에서 탈피, 투명하고 공정한 웹 문화 정착 및 활용을 위한 온라인 여론 수렴 기술에 대한 필요성이 증가함에 따라[3] 기업·정부 등에서 정치/경제/사회문화적 이슈들에 대한 온라인 동향 분석 및 이슈 예측 기술의 수요가 급증하고 있으며[4], 특히 여론 형성의 악영향을 끼치는 부정적 이슈 탐지에 대한 요구가 급증하고 있다.

소셜미디어를 통해 여론의 동향을 파악하고 이를 선제적 대응 매체로 활용한 연구들이 국내외적으로 매우 활발히 진행 중인데, 특히 개인, 기업 차원을 넘어서 국가 차원에서의 활용 방안도 다양하게 모색되고 있다. 싱가포르의 RAHS(Risk Assessment Horizontal Scanning)[5] 프로젝트, 영국의 호라이즌 스캐닝 센터(The Foresight Horizon Scanning Center)[6] 등이 그 예로, 다양한 사회 문제에 대한 여론 수렴 및 대응책을 마련하는 연구를 진행 중에 있다. 그 밖에도 범국가 차원에서 EU는 금융위기 극복과 사회의 복잡성을 이해하기 위한 FutureICT와 불확실한 미래 탐구를 위한 iKnow 프로젝트를 추진, 과거 단순한 데이터 분석 위주에서 최근 SNS와 결합한 미래 예측에 관한 연구를 활발히 수행 중이다[7].

국내에서는 주로 기업과 기관을 중심으로 상품 브랜드 및 기관에 대한 여론 동향을 파악하고 의사결정을 지원하기 위한 소셜미디어 분석이 활발하게 이루어지고 있다. 특히 2012년 총선을 기점으로 가장 큰 화두였던 대선에 이르기까지 소셜미디어 상의 민심 동향을 보여 줄 수 있는 선거 관련 서비스 사례들이 두드러지고 있다[8][9]. 주로 관련 정치인 트위터 실시간 업데이트, 정치 팟캐스트, 선거구 분석, 일반인들의 정치 관련 소셜미디어 동향 분석을 위주로 서비스를 구성하고 있다. 그러나 이러한 서비스들은 단기간 내에 화제가 된 사건을 중심으로 단편적인 정보를 분석하여 제공하는 것으로 대량의 소셜미디어를 포괄적으로 분석하여 시계열에 기반한 모니터링 서비스를 제시하는 리스닝 플랫폼 서비스는 아직 시작 단계에 머무르고 있다[10].

종합해보면 대용량 소셜미디어로부터 다양한 사용자들의 목소리(buzz)를 수렴, 그 의미를 파악하여 종합적으로 분석하는 기술에 대한 필요가 증대되고 있다. 그 첫걸음으로 본 논문에서는 사회적으로 파장이 큰 이슈 중 공공여론이 부정적으로 형성될 이슈를 '리스크'로 정의, 그 유형을 분류하고자 한다. 특히 본 논문에서 제안하는 방법은 기존에 분석하고자 하는 이슈를 미리 지정하는 하향식(Top-down) 방식이 아닌 기계학습을 통해 실제 소셜미디어에 나타난 리스크를 상향식(bottom-up) 방식으로 선정한다는 점에 차이가 있다.

II. 연구 방법

1. 사회적 이슈 리스크 개념 정의

앞서 기술한 대로 본 논문은 소셜 빅데이터로부터 사회적으로 파장이 큰 이슈 중 공공여론이 부정적으로 형성될 이슈를 '리스크'로 정의, 그 유형을 분류하는데 주안점을 둔다. 여기서 분석 대상으로 한 소셜 빅데이터는 뉴스와 같은 공공미디어 뿐 아니라 블로그, 게시판, 트위터 등 개인미디어를 포함한다. 또한 소셜웹에서의 '이슈'란 시간/지역별 특성을 반영한 정치·경제·사회·문화적으로 중요한 주제 또는 사건을 의미한다.

본 논문에서 다루고자 한 이슈 '리스크'란 소셜웹 미디어에서 발생한 다양한 이슈 중에서 이슈 대상 주체(target)에게 위협이 될만한 혹은 잠재적으로 위험을 내포하고 있는 사건을 의미한다. 그 중에서도 특히 공공의 공통관심 이슈로 보건, 복지 관련 정책이나 국민안전사고 등 사회적 파장이 큰 공공분야 이슈로 조사 대상을 한정, 이를 '사회적' 이슈 리스크로 선별하고 세부 유형을 분류하고자 한다.

사회적 이슈 리스크의 대표적인 예로는 최근 국방부에 불거진 "한국형 전투기(KF-X) 방산 비리"라던가 "누리사업 예산 중지" 등을 들 수 있다. 이들 사건은 이슈가 미치는 사회적 파장이 크고 특정 이슈 대상에 부정적 여론을 형성할 수 있는 사건이다. 본 연구의 궁극적인 목적은 이러한 사회적 이슈 리스크를 다양한 소셜미디어로부터 자동으로 탐지, 분류하는 것이다.

2. 연구 방법 및 범위

본 논문에서는 다음과 같은 과정을 통해 사회적 이슈 리스크 유형을 분류하고자 한다. 먼저 특정 기간에 보도된 뉴스 문서집합을 대상으로 전수조사를 통해 리스크 후보 문서를 선별하였다. 선별된 문서들에 나타난 이슈 분야 즉 도메인별 특성을 파악하여 세부 유형을 정의하였으며, 이 중에서 현재 언어처리 및 텍스트 마이닝 기법을 통해 기계적으로 구별이 가능한 유형을 분류하였다. 또한 리스크의 위험 대상이 되는 주체가 명확한 사건으로 한정하였다. 마지막으로 뉴스 미디어를 통해 정의된 리스크 유형이 개인화된 소셜 미디어에 나타난 리스크 유형과 어떤 차이가 있는지를 교차 분석하였다. 위 과정을 간략화 하면 다음과 같다.

1) 뉴스 미디어 전수조사
2) 리스크 후보 대상 표본조사
3) 이슈 도메인별 세부 유형 분석
4) 소셜 미디어 교차분석

표 1. 수집 문서집합 통계

미디어	수집 문서 수	특성
뉴스	1,437,118	평균 23.5 개 문장으로 구성
트위터	1,002,240,097	140자 이하 (평균 123.2자)

본 논문에서는 한국전자통신연구원(이하 ETRI)에서 개발한 소셜 이슈 분석 플랫폼인 WISDOM[8]에서 수집한 원시데이터를 활용하였다. 분석 대상은 중 19대 총선과 18대 대선 등 정치·사회적 이슈가 많이 발생한 2012년도[8][9] 문서집합을 조사 대상으로 삼았으며 수집된 데이터 통계는 [표 1]과 같다.

Ⅲ. 사회적 이슈 리스크 유형 정의

1. 사회적 이슈 리스크 대상 수집 및 특성 조사

1.1 전수조사

후보 리스크를 선정하기 위해 먼저 수집된 문서집합을 월별로 구별, 정치·사회적 이슈를 다루고 있는 문서만을 선별하였다. 특히 뉴스 미디어에서 리스크의 후보가 될만한 부정적 사건이나 사고를 보도한 기사를 선별하고 이후 관련 리스크에 해당하는 트윗 멘션을 검색하는 방식으로 조사 대상문서집합을 선정하였다. 부정적인 사건/사고를 판단하는 기준은 뉴스 문장의 구문적 특성과 논조에 따라 판단하였는데, 예를 들면 “물의를 일으키다 / 심각한 타격을 얻었다 / 구설수에 오르다” 등과 같이 직접적으로 문제가 있음을 밝힌 경우나, “파장이 일고 있다 / 우려가 있다” 등의 어구로 부정적인 여론이 형성됨을 알리는 어구가 포함된 뉴스를 대상으로 하였다.

표 2. 전수대상 뉴스 문서집합 통계

월별	전체 뉴스 문서 수	조사 대상 문서 수	리스크후보 문서수
201201	123,414	20,899	1,866
201202	121,605	18,276	1,523
201203	131,927	34,665	2,311
201204	121,650	21,838	1,790
201205	134,272	24,780	1,652
201206	106,979	20,965	1,839
201207	126,658	34,391	2,023
201208	113,747	33,991	1,789
201209	101,001	22,789	1,753
201210	120,349	27,365	2,243
201211	116,602	34,150	2,422
201212	118,914	20,418	1,823
전체	1,437,118	314,526	23,034
평균	119,760	45,406 (37.9%)	1,920 (1.6%)

[표 2]는 [표 1]의 수집된 뉴스 데이터의 월별 통계와 리스크 후보를 선정하기 위해 전수조사를 수행한 대상 문서 통계를 나타내는 표이다. [표 2]를 살펴보면 평균적으로 한 달에 생성되는 기사 중 정치·사회적으로 부정적인 영향을 미친 사건, 사고를 다룬 기사가 전체 37.9%에 해당하는 것을 알 수 있다. 이 중에서 처음 사건이 보도된 이후 해당 사건에 관해 중복으로 다룬 기사를 제외한 리스크 후보 대상 문서 수는 월별 평균 1,920건으로 전체 기사 중 1.6%에 해당하는 것으로 분석되었다. 이는 하나의 사건당 평균 23.65개의 후속보도가 다루어졌음을 의미한다.

1.2 리스크 후보 대상 표본 조사

다음으로는 선별된 리스크 후보 대상 문서의 내용에 따라 이슈 분야 즉 도메인(domain)별로 나누어 분석하

였다. 도메인은 수집된 뉴스 웹 사이트들에서 섹션 혹은 카테고리 등으로 구분하여 제공하는 하위범주명을 종합하여 작성하였다.

표 3. 1월, 8월 뉴스 문서의 도메인별 분포도

도메인	1월		8월	
	문서 수	비율	문서 수	비율
공공/자연재해	105	5.6%	132	7.4%
공공/질병	16	0.9%	28	1.6%
공공/사건	411	22.0%	382	21.4%
공공/기타	259	13.9%	209	11.7%
공공기관	197	10.6%	164	9.2%
정책	22	1.2%	22	1.2%
인물	259	13.9%	311	17.4%
먹거리	47	2.5%	58	3.2%
자동차	11	0.6%	12	0.7%
스마트기기	13	0.7%	31	1.7%
가전류	2	0.1%	8	0.4%
유통(쇼핑)	22	1.2%	15	0.8%
서비스업	157	8.4%	123	6.9%
제조업	68	3.6%	59	3.3%
IT업체	64	3.4%	95	5.3%
기타	213	11.4%	140	7.8%
전체	1,866	100%	1,789	100%

[표 3]은 1월과 8월에 보도된 리스크 후보 뉴스 집합의 도메인별 분포도를 나타내는 것으로, 1월 뉴스 기사 중 가장 많이 나타난 분야는 공공분야의 '사건' 으로 전체 22.0%를 차지하였으며 그 외 '인물', '공공기관', '서비스업' 분야가 각각 13.9%, 10.6%, 8.4% 순으로 기사가 생성되었음을 알 수 있다. 이는 8월에도 유사한 양상을 보이는데 1월과 마찬가지로 가장 많은 건수의 기사가 보도된 분야는 공공분야의 '사건' 으로 전체 21.40%를 차지하였다. 특히 '인물' 분야의 뉴스가 17.4%로 1월보다 3.5% 정도 상승했는데 이는 2012년 5월 총선 이후와 12월 대선을 앞두고 정치인에 관한 사건, 사고 관련 기사가 증가하였기 때문으로 풀이된다. 그 밖에 '공공기관', '서비스업' 분야의 기사가 역시 각각 9.2%와 6.9%를 차지한 것으로 1월과 유사한 양상을 보였다.

1.3 리스트 후보 대상 도메인 선정

[표 3]을 기반으로 선별된 문서들에 나타난 이슈의 분야별 즉 도메인별 특성을 분석하였다. 가장 많은 뉴스를 차지한 공공분야의 '사건' 도메인을 살펴본 결과, 대부분의 기사가 범죄 및 도난, 자동차 사고 등으로 특정 대상에게 잠재적 위험요소로 작용할 '이슈'가 아닌 일상생활에서 발생한 다양한 사고 소식을 다룬 내용으로 파악되었다. 공공분야의 '자연재해' 도메인 역시 날씨에 따른 한파나, 홍수, 장마와 같은 재해에 따른 피해 상황을 보도하는 뉴스가 대부분이었다. 이와 같이 공공분야의 뉴스들은 국정 운영을 하는 정부 차원에서 모니터링해야 할 필요성은 있으나 이슈 리스크의 위험 대상 주체가 명확히 특정되지 않아 본 연구의 대상에서 제외하였다. 또한 특정 도메인이 명확하지 않은 '기타' 도메인도 리스크 유형 분류 대상에서 제외하였다.

위와 같은 기준으로 본 논문에서는 리스크 유형 분류 대상으로 11개 세부 도메인을 선정하였다. 이는 뉴스 집합 전체 문서(1,866건) 중 리스크 위험 대상이 명확한 사건, 사고 문서 1,075건 중 862건에 해당하는 것으로, 후보 도메인 중 80.2%를 포괄하는 범위로 선정하였다. [그림 1]은 선정된 대상 도메인별 세부 분포도를 도식화한 것으로 가장 많은 비율을 차지하는 도메인은 인물(30%)과 공공기관(23%)로 나타났다.

2. 사회적 이슈 리스크 유형 모델링

2.1 도메인별 리스크 유형 조사

[그림 1]의 대상 도메인별 리스크 유형 유형을 세부적으로 살펴보니 '인물'의 경우 주로 구속이나 체포 등의 법적조치 당한 경우, 혹은 공천비리, 뇌물 비리와 같이 위법 행위에 연루된 경우, 개인 구설수 등으로 비난 여론이 거세진 경우 등에 대한 뉴스가 주를 이루었다. 구체적인 예로는 2012년 1월 2일 보도된 "주진우 기자와 나경원 의원 쌍방 맞고소 남발" 사건이나 같은 달 26일 보도된 "최시중 종편 돈봉투 살포" 등과 같은 뉴스 등이다.

한편, '공공기관'에 대한 리스크 유형도 이와 유사한 경향을 보였는데, 주로 해당 기관에서 벌어진 부정부패, 비리 연루 사건이나 위법행위가 적발된 경우에 대한 보도가 다수를 이루었다. 구체적인 예로는 2012년 1월 14일 보도된 '선관위 홈페이지 또 디도스 공격', 같은 달 17일 '목포지청 운영비 황령 5명 기소' 등의 기사를 들 수 있다.

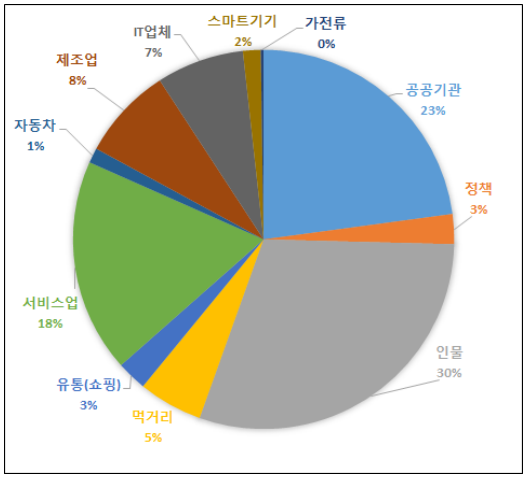

그림 1. 리스크 대상 도메인별 세부 분포

표 4. '인물'과 '공공기관'의 세부 리스크 유형 비교

인물		공공기관	
세부 리스크 유형	뉴스 건수	세부 리스크 유형	뉴스 건수
갈등	1	*갈등*	15
건강악화	2	경제위기	12
기밀유출	2	기밀유출	1
논란	3	낭비	1
비난여론	3	논란	3
비리	14	*비난여론*	7
사과	2	*비리*	11
사퇴	5	사회빈곤	1
사퇴요구	9	*소송*	3
소송	22	시위	2
시위	1	신용하락	8
악성루머	5	*악성루머*	4
위법행위	102	*위법행위*	45
의혹	1	의혹	2
조사	19	인재난	3
지지율하락	1	*조사*	8
징계	3	지지율하락	3
탈당	2	*징계*	1
퇴출	2	탈당	2
평가하락	2	투자손실	1
법정판결	15	*법정판결*	4
법적조치	16	부조리	1
경쟁	1	해킹	5
레임덕	1	감사	2
		실업증가	1
기타	22	기타	7.8%
합	256	합	197

[표 4]는 1월에 보도된 뉴스 중 '인물'과 '공공기관'에 해당하는 256건과 197건의 뉴스에 나타난 세부 리스크 유형을 분류한 표이다. 표에서 나타나듯이 '인물'과 '공공기관'에 나타난 세부 리스크 유형 중 상당 부분이 유사한 것을 볼 수 있다. 마찬가지로 '먹거리' 도메인에 나타난 리스크와 '서비스업' 분야에서 나타난 리스크 세부 유형 역시 '서비스불만', '가격상승/하락', '소송' 등, 상당수의 세부 유형을 공유함을 알 수 있다[표 5]. 이와 같이 서로 다른 도메인이지만 보도되는 뉴스의 유형이 유사한 도메인을 다시 대범주로 묶어본 결과 [표 6]과 같이 크게 4 범주로 구별됨을 알 수 있었다.

표 5. '먹거리'과 '서비스업'의 세부 리스크 유형 비교

먹거리		서비스업	
세부 리스크 유형	뉴스 건수	세부 리스크 유형	뉴스 건수
가격상승	4	*가격상승*	11
가격하락	5	*매출하락*	8
리콜	1	경쟁	8
불량식품	3	비난여론	32
소비자불만	3	*소비자불만*	45
소송	2	*소송*	8
위법행위	2		
유해물질	8		
질병	2		
판매금지	5		
판매중단	2	*판매중단*	29
기타	10	기타	52
합	256	합	197

표 6. 대범주별 리스크 도메인 분류

대범주	비율	세부 도메인	대상문서 수	비율
공공기관/정책/인물	55.5%	공공기관	197	22.9%
		정책	22	2.6%
		인물	259	30.0%
먹거리/유통/서비스업	26.2%	먹거리	47	5.5%
		유통(쇼핑)	22	2.6%
		서비스업	157	18.2%
자동차/제조업	9.2%	자동차	11	1.3%
		제조업	68	7.9%
스마트기기/IT/가전	9.2%	IT업체	64	7.4%
		스마트기기	13	1.5%
		가전류	2	0.2%
합	100%		862	100%

[표 6]를 기반으로 도메인별의 세부 유형을 대범주 단위로 비교, 공통된 리스크 유형을 정의하였다. 그 결과 대범주 [공공기관/정책/인물]은 각각 '공공기관/정책', '인물'로, [먹거리/유통/서비스업] 대범주와 [자동차/제조업] 대범주는 전체를 대표하여 각각 '먹거리'와 '자동차'로, 마지막으로 [스마트기기/가전/IT업체] 대범주는 '스마트기기'와 'IT업체'로 나누어 리스크 대상 도메

인을 통합, 분류하였다. 재정의된 도메인별로 수렴된 리스크 세부 후보 유형의 종류와 예는 [표 7]과 같다.

표 7. 도메인별 리스크 후보 유형 정의

대범주	세부 도메인	후보 유형 수	리스크 후보 유형 예
공공기관/정책/인물	공공기관	25	갈등, 기밀유출, 비난여론, 비리, 징계, 시위..
	인물	24	갈등, 건강악화, 법적조치..
먹거리/유통/서비스업	먹거리	11	유해식품, 가격상승/하락..
자동차/제조업	자동차	6	결함, 리콜,..
스마트기기/IT/가전	IT업체	20	정보유출,경쟁,..
	스마트기기	7	결함, 판매감소,..
합		93	

2.2 소셜미디어 교차 분석

다음으로는 뉴스 미디어를 통해 수렴된 리스크 후보 유형과 개인 소셜 미디어에 나타난 리스크 후보 유형과 어떤 차이가 있는지를 교차 분석하였다. 먼저 [표 7]에 선정된 1차 리스크 후보 유형에 해당하는 뉴스 기사와 연관된 트윗 멘션(mention)을 검색하는 방식으로 조사 대상문서집합을 선정하였다. 교차 분석 방법은 해당 리스크 뉴스와 관련된 트윗 멘션의 수와 리트윗(retweet) 수, 관련 트윗 생성 기간 등과의 상관관계를 비교하였다.

표 8. '인물' 도메인 리스크 후보 트위터 양상 분석

세부 리스크 유형	뉴스 건수	관련 트윗 수	관련 트윗 평균 수	트윗 생성 기간
갈등	1	221	221.0	2.3
건강악화	2	564	282.0	1.5
기밀유출	2	648	324.0	6
논란	3	123	41.0	1
비난여론	3	1,564	521.3	5
비리	14	7,756	554.0	2
사과	2	486	243.0	3
사퇴	5	4,708	941.5	5
사퇴요구	9	8,194	910.4	3
소송	22	7,154	325.2	5
시위	1	154	154.0	3.4
악성루머	5	660	132.0	3
위법행위	102	43,166	423.2	7.8
의혹	1	52	52.0	1
조사	19	2,337	123.0	3
지지율하락	1	199	199.0	3
징계	3	1,026	342.0	4
탈당	2	1,648	824.0	6
퇴출	2	484	242.0	5
평가하락	2	284	142.0	1.8
법정판결	15	13,835	922.3	8.2
법적조치	16	13,651	853.2	9
경쟁	1	151	151.0	2
레임덕	1	53	53.0	4
합	234	109,118	466.3	3.96

[표 8]는 [표 7]에 선정된 '인물' 도메인의 리스크 후보 유형별 트위터 양상을 분석한 결과이다. 전체 24개의 리스크 후보별로 총 234건의 리스크 뉴스와 관련된 트윗을 검색, 연관성을 판별하였다. 관련 멘션 수집은 트위터에서 제공하는 Open API인 스트림 API (http://api.twitter.com)를 사용하였으며[8], 수집된 트윗은 총 109,118개이다.

해당 리스크와 관련성이 있는지의 여부는 뉴스가 처음 보도된 시점부터 연속적으로 언급된 경우나 해당 뉴스를 리트윗(retweet) 한 경우를 기준으로 하였으며, 이후 3일 동안 관련 트윗이 수집되지 않은 경우를 리스크 전파 '종료' 시점으로 보았다. 분석 결과, '인물' 도메인의 리스크 뉴스가 트위터에서 회자되는 기간은 평균적으로 3.96일이고 466.3개의 트윗에서 언급되는 것으로 파악되었다. 그러나 이는 개별 리스크의 특성에 따라 편차가 매우 큰 것으로 분석되었다.

[표 8]을 자세히 살펴보면, '논란'이나 '의혹'의 경우 관련 트윗 생성 기간이 1일~2일 안팎으로 매우 짧은 반면, '위법행위' 이나 '법정판결', '법적조치' 등은 평균 일주일 넘게, 길게는 9일까지도 관련 트윗이 전파되는 양상을 보였다. 또한 같은 리스크 유형이라도 리스크 대상의 관심도에 따라 매우 다른 양상을 보였는데, 예를 들면 '법정판결' 유형 중 하나인 "한명숙, '피의사실 공표' 동아일보 10억 청구 '패소'판결" 사건의 경우, 관련 트윗이 1만5천 건은 넘었으며 트위터에서 언급된 기간 역시 11일이 넘는 것으로 분석된 반면, 같은 '법정판결' 유형이지만 "이국철 SLS그룹 회장, 허위공시 항소심서 집행유예 5년" 사건의 경우, 관련 트윗이 500여건, 언급된 기간 역시 2일로 매우 짧은 것을 알 수 있었다.

앞 절의 '인물' 도메인 교차 분석 방법과 같은 방법으로 [표 7]에 선정된 나머지 도메인의 리스크 유형의 트위터 양상을 분석한 결과, [표 8]의 결과와 같이 리스크 유형 및 개별 이슈의 사안에 따라 관련 트윗 수와 유지 기간이 차이가 난다는 결론을 얻을 수 있었다.

2.3 사회적 이슈 리스크 유형 정의

본 논문에서 분류 대상으로 하는 사회적 이슈 리스크는 뉴스와 같은 공적 매체 뿐 아니라 개인 미디어에서도 같이 회자가 되는 리스크를 대상으로 하는 점을 반

영, 뉴스와 트위터 미디어 모두에서 이슈가 되는 리스크 유형을 재선정 하였다. 선정 기준은 뉴스에서 선정된 유형 중 트위터에 2일 이상, 150건 이상 회자된 리스크를 대상으로 하였다. [표 9]는 최종 리스크 유형 선정 결과로, 총 6개의 도메인에 58개의 세부 유형으로 정의하였다.

표 9. 도메인별 사회적 이슈 리스크 유형 정의

도메인	유형 수	리스크 유형
공공기관	11	갈등, 경쟁, 법적조치, 법정판결, 부정여론, 소송, 시위, 위법행위, 정보유출, 제재, 조사
인물	12	갈등, 건강악화, 경쟁, 법적조치, 법정판결, 부정여론, 사퇴요구, 소송, 시위, 위법행위, 정보유출, 조사 , 징계
먹거리	9	가격상승/하락, 리콜, 부정여론, 불매운동, 위법행위, 유해식품, 제재, 판매감소
자동차	6	가격하락, 결함, 경쟁, 리콜, 부정여론, 판매감소
IT업체	12	갈등, 경영위기, 경쟁, 범정판결, 부정여론, 불매운동, 사업종료, 소송, 위법행위, 정보유출, 조사, 제재
스마트기기	8	가격하락, 결함, 경쟁, 부정여론, 정보유출, 제재, 판매감소
6	58	

Ⅳ. 기계학습 기반 리스크 유형 분류

서두에서 언급한 바와 같이 본 논문의 최종 목적은 다양한 소셜 미디어에 나타난 사회적 이슈 리스크를 자동으로 탐지, 분류하는 것이다. 이를 위해 본 논문에서는 미디어에 나타난 어휘적 특성을 자질로 활용하여 기계학습 기반 분류 모델을 적용하였다. 한국어 어휘 분석을 위해 ETRI에서 개발한 언어분석기[9][10]를 사용하였으며 기계학습 기법은 지지벡터머신(SVM: support vector machines) 알고리즘을 적용하였다[11]. 학습에 사용한 언어분석 결과는 형태소 분석, 개체명 인식 및 감성분석 결과로, 세부 어휘 자질은 [그림 2]와 같다.

원문	이수근, 탁재훈에 이어 불법 도박 혐의로 검찰 조사를 받은 것으로 ....비난을 받고 있다.
형태소 분석	+이수근/nc+,/s+ 탁재훈/nc+에/jc+ 잇/pv+어/ec+ 불법/nc+ 도박/nc+ 혐의/nc+로/jc+ 검찰/nc+ 조사/nc+를/jc+ 받/pv+은/etm+ 것/nb+으로/jc+ ...
개체명 인식	NE_PS_이수근 NE_PS_탁재훈 NE_CVL_CRIME_도박 NE_OGG_POLITICS_검찰....
감성 분석	비난/nc+받/pv : 부정/비난

그림 2. 리스크 대상 도메인별 세부 분포

[표 10]은 뉴스 미디어를 대상으로 리스크 유형 자동분류 구축을 위해 각각 학습 및 평가에 사용한 실험문서집합이 통계자료이다. [표 11]은 기계학습 이후 자동분류 정확도를 평가한 결과이다. 측정 결과, 전체 8,175개 문장 중 7,017개의 문장에 정확한 리스크 유형을 분류하여 85.83%의 정확률을 얻었다. 이는 전체 리스크 유형 58개를 대상으로 실험한 결과로, [표 4]부터 [표 6]에서 나타나듯이 리스크 유형에 따라 뉴스 미디어에 나타난 빈도의 편차가 매우 크다. 이 같은 특성을 반영하여 전체 58개 중 고빈도 리스크 유형 20개를 대상으로 실험한 결과, 전체 3,713개의 문장 중 3,419개의 리스크 정답 유형을 분류하여 92.08%의 높은 정확률을 얻었다. 이와 같이 고빈도 리스크 유형에 대한 정확도가 높다는 의미는 실제 사용자 체감 성능은 미치는 영향이 더 크다는 의미로 해석될 수 있다.

표 10. 리스크 유형 분류 실험집합 통계

문서 수	리스크 문장 수	학습셋	실험셋
3,000	24,385	16,210	8,175

표 11. 기계학습 기반 리스크 유형 분류 성능 평가 결과

평가기준	대상 문장 수	맞은 수	정확률 (Precision)
전체 58개 유형	8,175	7,017	85.83%
고빈도 20개 유형	3,713	3,419	92.08%

Ⅴ. 결론 및 향후 연구 계획

소셜미디어의 정치사회적인 활용도가 높아짐에 따라 소셜빅데이터 기반 온라인 동향분석 및 모니터링 기술에 대한 수요 역시 급증하고 있다. 본 논문에서는 이러한 요구에 부합, 특히 여론 형성의 악영향을 끼치는 부정적 이슈 탐지를 위해 사회적으로 파장이 큰 이슈 중 공공여론이 부정적으로 형성될 이슈를 '리스크'로 정의하고 세부 유형을 분류하였다.

리스크 유형 정의를 위해 뉴스 문서집합을 대상으로 전수조사를 실시하였으며, 이슈 분야 즉 도메인별 특성을 파악하여 세부 유형을 정의하였다. 또한 공적미디어

인 뉴스를 통해 정의된 리스크 유형이 개인화된 소셜 미디어에 나타난 리스크 유형과 어떤 차이가 있는지를 알아보기 위해 교차분석을 수행하였다. 그 결과 6개의 도메인별로 58개의 세부 유형을 정의하였으며, 기계학습 방법을 통해 자동 분류 학습 모델을 구축하였다. 실험 결과를 통해 소셜 미디어에 나타난 사회적 이슈 리스크를 자동으로 탐지, 분류가 가능함을 보였다.

본 연구의 궁극적인 목표는 정부나 기업체에 미래 위험요소로 작용할 이슈를 미리 감지하여 이에 대한 선제적 대응이 가능하도록 하기 위함이다. 이를 위한 향후 연구 방향으로는 자동으로 탐지된 리스크의 위험 강도를 측정하고 향후 방향을 예측하여 리스크 대상에게 알림(alert) 기능을 구현하고자 한다.

참 고 문 헌

[1] 이지영, "'빅데이터 분석 = 소셜 분석'이 된 까닭," bloter.net, 2012.

[2] 이병엽, 임종태, 유재수, "빅 데이터를 이용한 소셜 미디어 분석 기법의 활용," 한국콘텐츠학회논문지, 제13권, 제2호, pp.211-219, 2013.

[3] 윤미영, 권정은, "빅데이터로 진화하는 세상," 한국정보화진흥원 IT & Future Strategy, 2012-제6호, 2012.

[4] 백인수, *데이터 강국을 위한 국가정보화사업 추진 방향*, 한국정보화진흥원, 2013.05.

[5] G. H. Kim, S. Trimi, and J. H. Chung, "Big-data applications in the government sector," Communications of the ACM, Vol.57, No.3, pp.78-85, 2014.

[6] W. J. Sutherland, M. J. Bailey, I. P. Bainbridge, T. Brereton, J. T. Dick, J. Drewitt, and P. M Gilder, "Future novel threats and opportunities facing UK biodiversity identified by horizon scanning," Journal of Applied Ecology, Vol.45, No.3, pp.821-833, 2008.

[7] 한국정보화진흥원, "이슈 스캐닝(Horizon Scanning) 기반 국가미래전략 수립방향," 한국정보화진흥원 IT & Future Strategy, 2013-제5호, 2013.

[8] 이충희, 허정, 오효정, 김현진, 류법모, 김현기, "소셜 빅데이터 이슈 탐지 및 예측분석 기술 동향," 전자통신동향분석, 제28권, 제1호, pp.62-71, 2013.

[9] 이상지, 장동혁, 박성운, 조원희, 이기철, "객체식별아이디 이포지션 기반의 LBSNS 앱이 19대 총선 후보 지지율의 변화에 미친 영향," 한국콘텐츠학회논문지, 제13권, 제8호, pp.171-179, 2013.

[10] 허정, 이충희, 오효정, 윤여찬 김현기, 조요한, 옥철영, "소셜 빅데이터 마이닝 기반 이슈 분석보고서 자동 생성," 정보처리학회논문지/소프트웨어 및 데이터 공학, 제3권, 제12호, pp.553-564, 2014(12).

[11] Y. J Choi, H. K Kim, and C. K. Lee, "Balanced Korean word spacing with structural SVM," Proc. of Empirical Methods in Natural Language Processing(EMNLP), pp.875-879, 2014.

저 자 소 개

오 효 정(Hyo-Jung Oh) 정회원

- 2000년 : 충남대학교 컴퓨터과학과(이학석사)
- 2008년 : 한국과학기술원 컴퓨터공학과(공학박사)
- 2000년 ~ 2015년 5월 : 한국전자통신연구원 지식마이닝연구실 책임연구원
- 2015년 5월 ~ 현재 : 전북대학교 기록관리학과 조교수

<관심분야> : 정보검색, 질의응답, 빅데이터정보처리

안 승 권(Seung-Kwon An) 정회원

- 2008년 : 중앙대학교 창업경영대학원(경영학석사)
- 2016년 : 중앙대학교 일반대학원 창업컨설팅(경영학 박사)
- 2006년 ~ 현재 : ㈜바른교육 대표이사

<관심분야> : 디지털콘텐츠, 이러닝, 창업교육

김 용(Yong Kim) 정회원

- 1995년 8월 : University of NorthTexas(MS in Information Science)
- 2006년 6월 : 연세대학교 문헌정보학과(문헌정보학 박사)
- 1996년 2월 ~ 2008년 8월 : ㈜KT 중앙연구소 책임연구원
- 2008년 9월 ~ 현재 : 전북대학교 문헌정보학과 부교수

<관심분야> : 정보검색, 디지털도서관, 데이터마이닝, e-Learning, 전자기록물, 전자기록관리시스템