한국환경정책·평가연구원

빅데이터와 환경

03. 환경 분야 빅데이터 분석과 응용

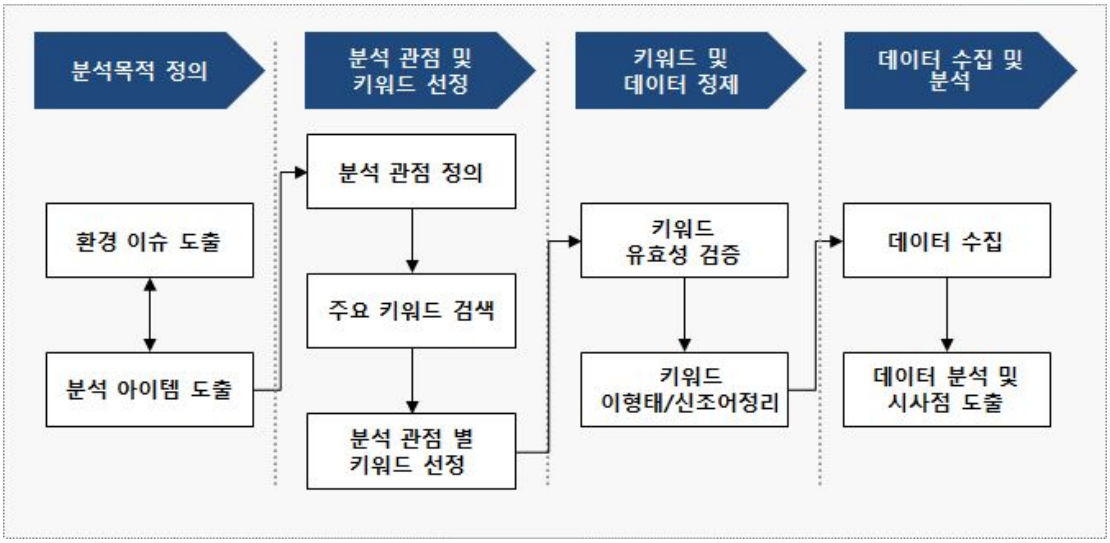
1. 소셜 빅데이터 분석 절차

지난 강의에서 살펴본 바와 같이 소셜 빅데이터 분석은 목적에 따라 다양한 분석 기법이 사용되며, 대표적으로 빈도 분석, 감성 분석, 연관어 분석, 이슈 반응 분석, 네트워크 분석 등이 있다. 이번 강의에서는 이러한 소셜 빅데이터 기법을 직접 활용하여 환경 분야의 소셜 빅데이터 분석하는 과정과 결과를 함께 살펴보도록 한다.

일반적으로 소셜 빅데이터 분석은 분석 목적 및 관점의 정의, 키워드 선정, 키워드 검증, 데이터 수집 및 분석, 결과 도출의 과정을 거쳐 수행된다. 연구의 출발점에서는 가장 먼저 분석 목적과 관점이 명확히 정립되어야 하며, 특히 분석 관점은 연구 주제와 빅데이터 분석의 일관성 확보를 위해 반드시 사전적으로 논의되어야 한다.

분석 목적과 관점에 따라 연구 주제를 선정하는 작업을 마치면, 이에 따라 데이터 추출을 위한 주요 키워드를 선정한다. 키워드는 분석 대상이 되는 데이터를 수집하는 데 직접적으로 영향을 미치기 때문에 객관적이고 전문적인 검증을 거쳐 선정되어야 한다. 검증된 키워드를 바탕으로 전체 소셜 빅데이터 중 분석 대상이 되는 데이터를 수집한 후, 상업성 광고글, 정치 관련 글 등 연구 주제에서 벗어나는 노이즈를 제거하는 정제 과정을 거친다. 분석 엔진마다 데이터 검색 메커니즘이 상이하기 때문에 분석 신뢰도를 높이기 위해서는 키워드를 정제하는 작업이 매우 중요하다. 정제된 데이터가 확보되면 앞서 언급한 다양한 분석 기법을 활용하여 연구 주제에 적합한 분석을 수행하게 된다.

[그림 1] 소셜 빅데이터 분석 절차



이번 강의에서 살펴볼 환경 분야 소셜 빅데이터 분석은 환경 전반 및 세부 분야별로 국민의 인식이 어떻게 나타나는지 분석하는 것을 목적으로 설정하였다. 그 다음으로 환경 관련 문서를 추출하기 위해 세부 환경분야별 검색 키워드를 도출하고 검증하는 절차를 수행하였다. 소셜 빅데이터 분석을 수행하는 절차 중에서 검색 키워드를 선정하는 것은 향후 도출되는 분석 결과에 직접적으로 영향을 미치며, 결과의 신뢰성을 확보하기 위해 가장 중요한 작업이라 할 수 있다. 이번 강의에서 살펴볼 소셜 빅데이터 사례(이미숙 외, 2014)는 최종적으로 13개 분야, 252개의 검색 키워드를 바탕으로 분석한 결과이다. 여기서 13개 환경 분야는 환경 전반 외에 기후, 대기, 생물다양성, 소음, 수질, 쓰레기(폐기물), 인적재해, 토양, 해양, 유해물질(보건), 빛공해 등으로 구분된다.

아래 [표 1]은 세부 환경 분야별로 구성한 검색 키워드 집합을 나타낸다. 소셜 빅데이터 분석에 있어서 검색 키워드를 어떻게 설정하느냐에 따라 분석 결과는 달라질 수 있음에 유의해야 한다. 또한 분석 신뢰도를 높이기 위해서는 상업성 광고글, 정치 관련 글 등 연구 주제에서 벗어나는 문서들을 배제할 수 있도록 데이터를 충분히 정제하는 과정이 반드시 필요하다. [표 1]에 제시된 검색 키워드마다 상당수의 배제 키워드를 설정하여 데이터를 정제하였음을 기억할 필요가 있다.

[표 1] 세부 환경 분야별 검색 키워드 설정

분야	검색 키워드
환경일반	환경
기후	기후, 온실가스, 온난화, 이산화탄소, 탄소, 폭설, 폭우, 홍수, 가뭄, 태풍, 혹서, 혹한, 해수면, 열섬, 이상고온, 집중호우, 해일, 쓰나미, 북극빙하, 남극빙하, 히말라야빙하, 그린란드빙하, 사막화, 녹색성장
대기	대기오염, 대기질, 미세먼지, 황사, 공해, 배기가스, 스모그, 공기, 매연, SOx, NOx, 이산화질소, 이산화황, 질소산화물, 황산화물, VOC, PM_{10}, $PM_{2.5}$, 대기환경, 맑은하늘, 산성비, 가시거리, 오존, 다이옥신, 대기배출, 특정대기유해물질, 휘발성유기화학물질, 월경성이동, 디젤입자, 실내공기, 배가스, 청정연료, 소각로, 벤젠, 대기오염측정망, 배경농도, 사업장총량제, 수도권대기, 분진
생물 다양성	생물다양성, 자연보전, 자연환경, 자연보호, 야생동물, 종다양성, 생태계교란종, 외래종, 멸종위기, 야생식물, 야생조류, 산림보호, 산림파괴, 자연훼손, 산림훼손, 야생동식물, 자연경관, 국립공원, 밀렵, 생태관광, 철새, 생물자원, 유전자원, 생물산업, 수렵, 고유종, CITES, 자연공원, 보전지역, 생태계서비스, 해양생물, 나고야의정서, 유전자변형생물체, GMO, 마을숲, 습지, 평화생태공원, 생태계다양성, 유전다양성, 생태발자국, 생태용량
소음	소음, 방음
수질	녹조, 수돗물, 수질, 폐수, BOD, COD, 4대강, 하천오염, 생태하천, 맑은물, 깨끗한물, 수생태, 하천생태, 물환경, 수변구역, 총인, 물고기폐사, 총질소, 관거, 분뇨, 정화조, 절수, 물놀이, 빗물관리, 침수, 1급수, 치어방류, 하천복원, 강살리기, 습지, 물고기떼죽음, 하천악취, 무단방류, 먹는물, 물부족, 물절약, 물낭비, 수도요금, 빗물모으기, 물재이용, 물재사용, 물재생, 해수담수화, 수처리
쓰레기/	쓰레기, 폐기물, 분리수거, 재활용, 종량제, 소각, 매립, 리싸이클, 리사이클, 음폐수,

폐기물	무단투기, 폐비닐, 해양투기, 폐자원, 폐열, 업사이클, 업싸이클, 폐금속, 물질흐름분석, 용출시험, 자원순환, 폐전기전자, 폐자동차, 다이옥신
인적 재해	화학사고, 기름유출, 오염사고, 불산사고, 페놀사고, 원전사고, 원자력사고, 후쿠시마사고, 방사능누출, 누출사고, 태안사고, 구미사고, 사고대비물질, 원유유출, 화학물질유출, 화학테러
악취	악취
토양	토양오염, 토양정화, 토양환경, 지하수, 토양생태, 폐광산, 가축폐사, 가축매몰지, 중금속오염, 유류오염, 토양복원, 오염부지, 토양미생물, 토양질, 토양독성
해양	갯벌, 연안, 해수면, 적조, 해일, 쓰나미, 해양쓰레기, 해양투기, 기름유출, 해양생태, 해양환경, 바다경관, 간척, 새만금, 부영양화, 해안선침식, 해안침식, 해양수질, 이상파랑, 이안류, 물고기집단폐사, 해안오염, 해양오염
유해 물질/ 보건	환경호르몬, 새집증후군, 발암물질, 환경보건, 중금속오염, 가습기살균제, 유해화학물질, 아토피, 환경성질환, 내분비계장애물질, 유해중금속, 다이옥신, 석면, 프탈레이트, 비스페놀A, 화평법, 화관법, 라돈, 고엽제, DDT, 폼알데하이드, 포름알데히드, 잔류성유기오염물질, POPs, 카드뮴, 수은, 불산, 염산, 페놀, 유독물, 위해성
빛공해	빛공해

이러한 검색 키워드를 바탕으로 환경 관련 문서를 수집한 결과, 본 연구에서는 2012년 11월 1일부터 2014년 4월 30일까지 1년 6개월 동안 뉴스, 블로그, 트위터 채널에서 한국어로 발현된 문서 7억 5,538만 268건 중에서 데이터 정제 과정을 통해 547만 6,652건의 환경 관련 문서가 수집되었다. 물론 분석 기간과 분석 대상 문서를 어떻게 설정하느냐에 따라 소셜 빅데이터 분석 결과는 달라질 수 있다.

분석기간 동안 발현된 전체 문서 중에서 환경 관련 키워드를 포함한 문서가 차지하는 비중은 0.73% 정도에 해당한다. 환경 관련 문서 중에서 트위터가 411만 6,820건(75.2%)으로 가장 많은 비중을 차지하며, 뉴스 76만 5,204건(14.0%), 블로그 59만 4,628건(10.8%) 순으로 나타났다. 또한 환경 관련 문서 중에서 긍정 또는 부정의 감성이 발현된 건수는 총 35만 5,303건이며, 부정 감성(19만 7,696건, 55.6%)이 긍정 감성(15만 7,607건, 44.4%)에 비해 상대적으로 많이 발현되었다.[1] 다음 절에서는 환경 전반 및 세부 환경 분야별로 소셜 빅데이터의 특성을 분석한 결과를 보여주며, 이를 통해 소셜 빅데이터 분석 결과를 해석하는 방법을 확인할 수 있다.

2. 환경 분야 소셜 빅데이터 분석

(1) 환경 전반에 대한 분석 결과

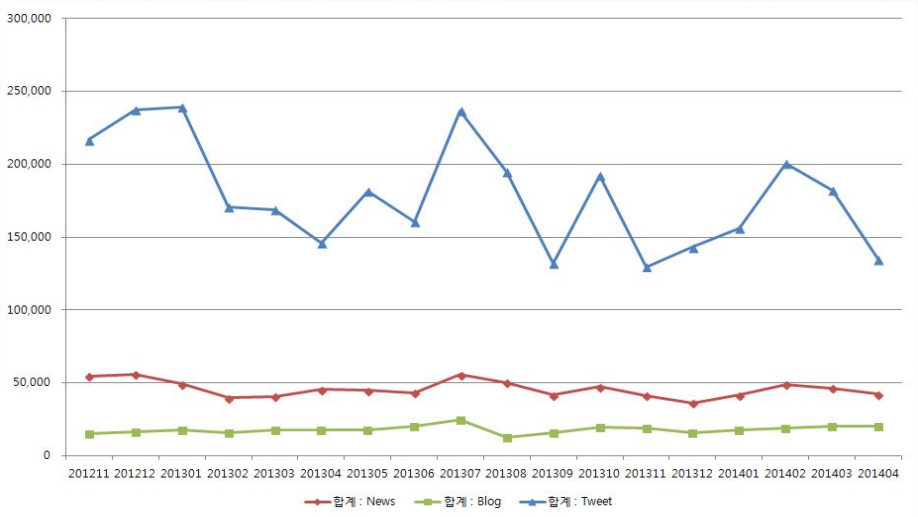
먼저 환경 전반의 빈도 분석 결과[그림 2]를 살펴보면, 뉴스 채널에서는 증감을

1) 긍정 및 부정 감성의 발현 여부는 검색엔진의 학습을 통해 검토된다. 본 사례에서 활용한 분석엔진은 기쁨, 좋아함, 감동, 안심 등의 정서가 텍스트에 반영되어 있다고 판단될 경우 긍정 감성으로, 두려움, 슬픔, 싫어함, 실망, 화남 등의 정서가 텍스트에 반영되어 있다고 판단될 경우 부정 감성으로 판단한다.

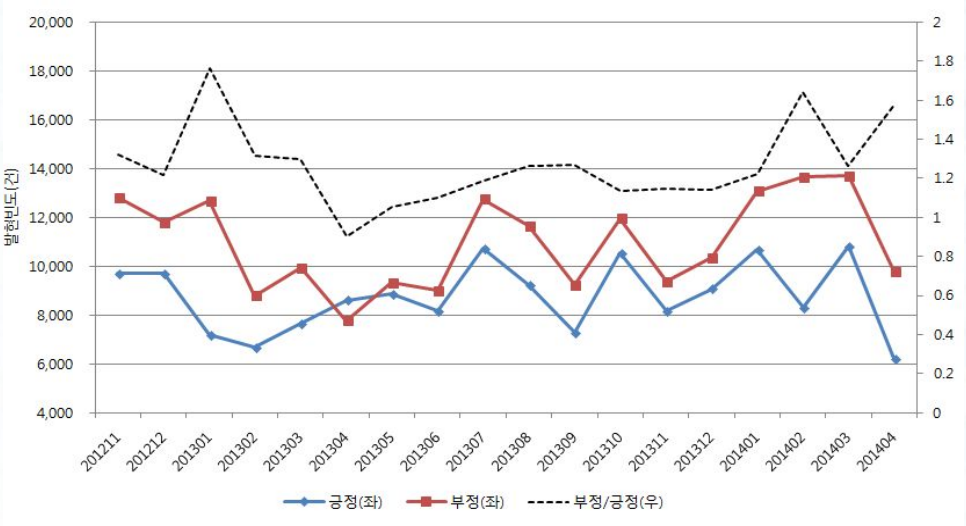
반복하다가 최근 다소 감소하였고, 블로그의 경우 전체적으로 발현빈도가 낮으나 일정한 수준을 유지하는 것으로 나타났다. 트위터는 뉴스와 블로그에 비해 상대적으로 발현빈도가 높으며 시기별 발현빈도의 증감이 크게 나타났다. 이로부터 환경에 대한 국민의 관심이 다소 감소하고 있다고 유추할 수 있다.

다음으로 환경 분야 전반에 대한 감성 분석 결과[그림 3], 전체적으로 긍정 감성보다 부정 감성의 비중이 더 큰 것으로 나타났으며, 대체로 긍정 감성과 부정 감성의 증감 방향이 일치하였다. 즉 긍정 감성이 증가하면 부정 감성도 같이 증가하며, 부정 감성이 감소하면 긍정 감성도 함께 감소하는 경향을 나타내고 있다. 긍정 대비 부정 감성의 비율을 살펴보면 2013년 1월과 2014년 2월에 급증하는 것으로 나타났는데, 해당 시기에 부정적 감성을 유발하는 원인이 있었을 것으로 추측된다. 전반적으로 환경 문제와 관련하여 우리나라 국민들은 부정적인 감정을 더 많이 표출한다고 볼 수 있다.

[그림 2] 환경 전반에 대한 빈도 분석 결과



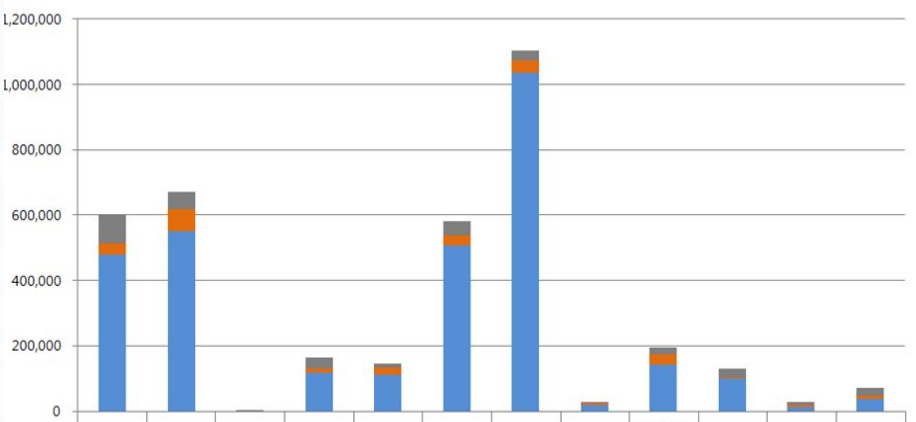
[그림 3] 환경 전반에 대한 감성 분석 결과



다음으로 세부 환경분야별 빈도 분석 수행 결과를 비교하면, 쓰레기(폐기물) 분야에서 약 110만 건의 문서가 발현되어 가장 높은 수치를 나타냈다[그림 4]. 대기와 기후 분야의 문서 수는 각각 67만 건, 60만 건 수준으로 비교적 높은 빈도를 나타냈으며, 수질분야의 경우는 약 58만 건의 자료가 확인되었다. 이 외에 유해물질/보건, 생물다양성, 소음, 인적재해 분야에 대해서도 일정 수준 이상의 문서가 수집되었으며, 해양, 악취, 토양, 빛공해 분야에서의 문서 발현량은 10만 건 이하로 비교적 낮은 비중을 나타냈다. 이번 강의에서는 기후, 대기, 수질, 쓰레기(폐기물) 등 비교적 관심도가 높은 것으로 확인된 세부 환경 분야의 분석 결과를 추가적으로 살펴본다.

또한 채널별로는 모든 세부 분야에서 트위터 문서량이 가장 많게 나타나, 환경문제에 대한 국민들의 인식을 분석할 때 트위터 채널이 유용하게 활용될 수 있다고 판단된다. 특히 쓰레기와 관련된 이슈는 다른 분야에 비해 트위터 채널에서의 언급량이 월등히 많았다. 대부분 트위터 다음으로 뉴스 채널에서 관련 언급량이 많았으며, 특히 기후 분야에서 뉴스 문서의 수가 상대적으로 많았다는 점이 특징적이다.

[그림 4] 세부 환경 분야별 빈도 분석 결과 비교



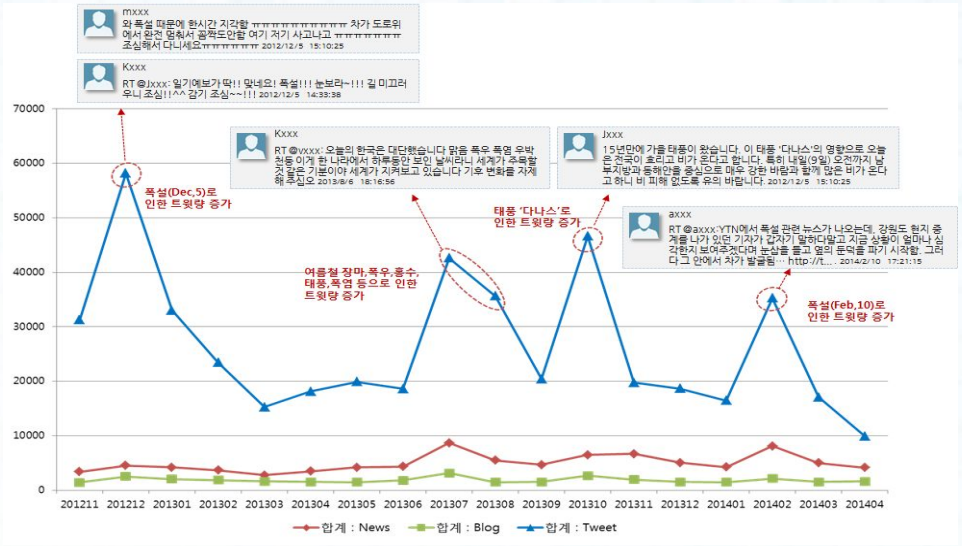
	기후	대기	빛공해	생물다양성	소음	수질	쓰레기	악취	유해물질/보건	인적재해	토양	해양
뉴스	88,992	52,879	107	31,393	12,510	42,339	33,581	4,108	22,038	27,294	9,403	25,364
블로그	33,484	66,056	45	14,694	22,299	30,474	35,376	3,620	32,185	4,593	4,469	7,800
트위터	480,089	551,389	1,108	116,829	112,255	508,169	1,035,814	20,168	142,335	99,040	13,122	37,817
총합	**602,565**	**670,324**	**1,260**	**162,916**	**147,064**	**580,982**	**1,104,771**	**27,896**	**196,558**	**130,927**	**26,994**	**70,981**

(2) 기후 분야에 대한 분석 결과

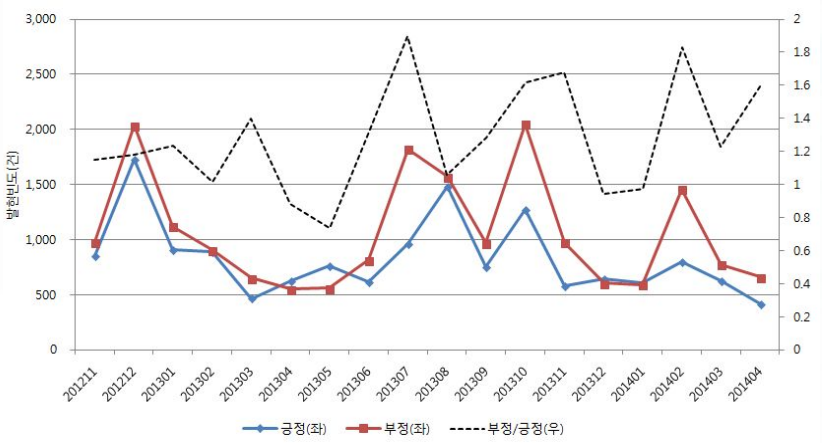
기후 분야는 키워드 발현 빈도가 상대적으로 높지만 전반적으로 발현 빈도가 감소하는 추세를 나타낸다[그림 5]. 기후 분야 문서의 대부분을 차지하는 트위터의 경우 특정 시기에 빈도가 큰 폭으로 증가하는 현상을 나타내며, 해당 시기의 원문을 검토하여 대략적인 원인을 파악할 수 있다. 예를 들어 2012년 12월에는 폭설로 인한 트윗량이 급증한 것이 원인이며, 2013년 7~8월은 여름철 장마, 폭우, 홍수, 태풍, 폭염 등과 관련된 문서량이 많은 것으로 확인되었다. 2013년 10월, 2014년 2월에도 각각 태풍과 폭설로 인한 트윗량이 급증한 것으로 확인되어 계절별 이상기후 현상에 대한 관심이 높은 것을 확인할 수 있다.

기후 분야에 대한 감성 분석 결과 역시 시기별로 큰 폭의 증감을 보이며, 특히 부정 감성의 추이는 전체 빈도 추이와 거의 일치하는 패턴을 보이고 있다[그림 6]. 전체적으로 부정 감성의 발현빈도가 긍정 감성의 발현빈도보다 조금 높게 나타났으며, 빈도가 급증한 시기에 긍정 감성 대비 부정 감성의 비율 역시 크게 증가한 것을 확인할 수 있다. 따라서 폭설, 태풍, 폭염 등 계절성 이상기후 현상이 부정적 감정으로 이어지는 경우가 많다고 볼 수 있다.

[그림 5] 기후 분야에 대한 빈도 분석 결과



[그림 6] 기후 분야에 대한 감성 분석 결과



기후 분야의 검색 키워드에 대한 상위 50개의 주요 연관어를 살펴보면 눈, 비, 바람, 오다, 내리다, 불다 등 날씨와 관련된 연관어가 높은 순위를 차지하고 있다. 그 외에 피해, 안전, 문제 등 기후변화로 인한 폐해를 나타내는 연관어도 상위에 나타나며 대비, 필요 등 기후변화 적응에 관한 연관어도 도출되었다[그림 7].

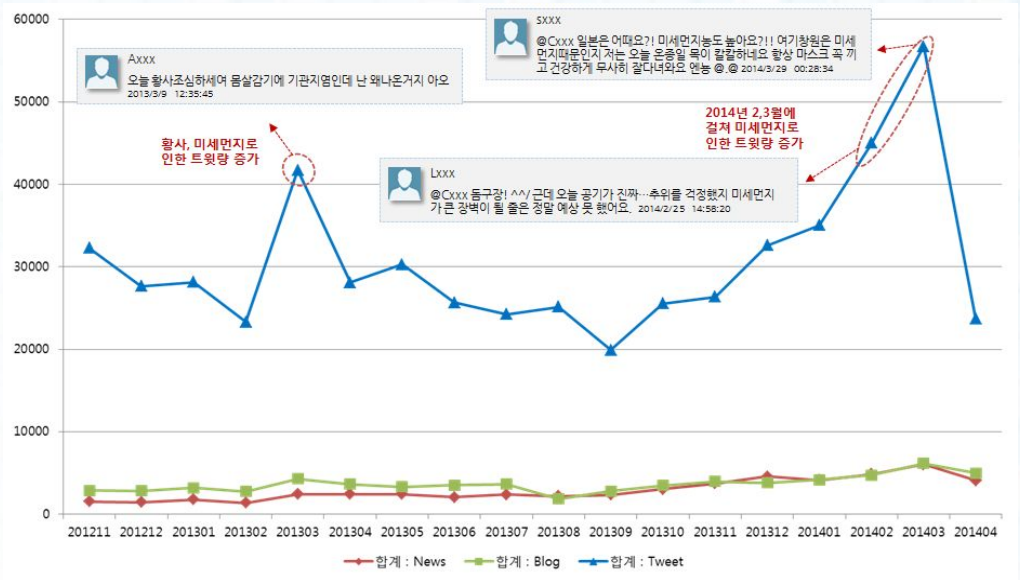
[그림 7] 기후 분야에 대한 연관어 분석 결과

명사						서술어					
순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도
1	RT	232001	26	상황	27864	1	오다	69298	26	쏟아지다	17755
2	눈	65323	27	환경	27512	2	내리다	58292	27	가지다	17591
3	지역	55276	28	정부	27118	3	보다	56156	28	떨어지다	17509
4	피해	51882	29	문제	26354	4	많다	54771	29	넘다	17489
5	비	48137	30	예상	26334	5	크다	54074	30	찾다	17401
6	물	45648	31	안전	26286	6	따르다	46658	31	필요하다	17212
7	사람	42625	32	위	26047	7	보이다	42857	32	쓰다	16929
8	발생	41054	33	필요	25553	8	받다	37373	33	사용하다	16301
9	변화	39891	34	면	24974	9	만들다	36790	34	들어가다	16102
10	한국	39770	35	바람	24728	10	가다	36064	35	가능하다	15954
11	시작	37542	36	최고	24478	11	좋다	35793	36	먹다	15886
12	이상	37262	37	녹색	24270	12	나오다	35296	37	알리다	15371
13	서울특별시	36279	38	사진	24144	13	통하다	34513	38	열리다	15101
14	시간	35355	39	대비	24113	14	인하다	33235	39	이어지다	14895
15	사업	32251	40	가스	24033	15	밝히다	31693	40	이용하다	14674
16	집	32039	41	계획	23376	16	들다	31159	41	내다	14401
17	차	31587	42	관리	23347	17	발생하다	27470	42	타다	13867
18	세계	30483	43	동안	22962	18	말하다	27439	43	열다	13409
19	날씨	29388	44	개발	22513	19	높다	23339	44	강하다	13313
20	길	29370	45	기후변화	22310	20	맞다	21919	45	느끼다	13277
21	관계	28750	46	미국	21924	21	주다	21205	46	불다	13241
22	일본	28390	47	예정	21751	22	나다	20163	47	이루다	13174
23	영향	28243	48	효과	21680	23	알다	19775	48	시작하다	13126
24	지방	28240	49	진행	21454	24	살다	19050	49	막다	12765
25	지구	28181	50	경제	21450	25	다양하다	17988	50	새롭다	12732

(3) 대기 분야에 대한 분석 결과

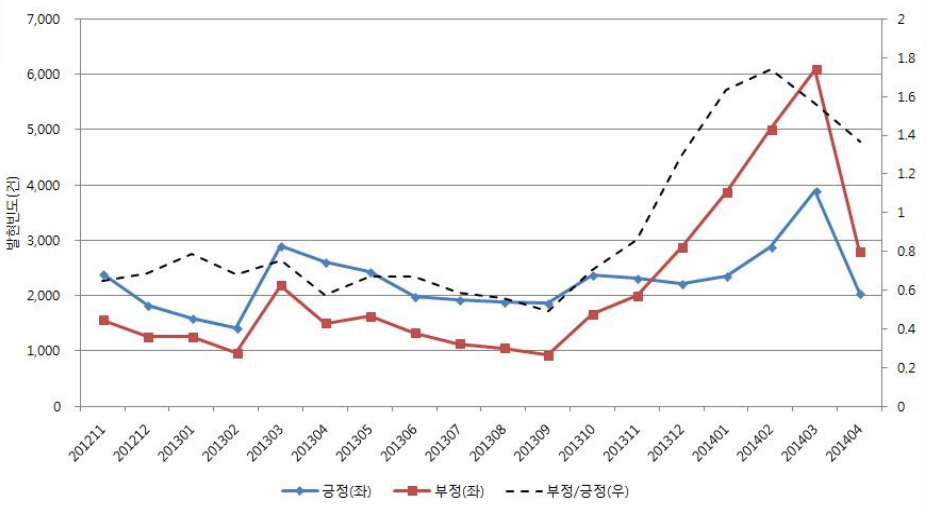
대기 분야에 대한 빈도 분석 결과[그림 8], 문서의 대부분을 차지하는 트위터 채널에서는 봄철인 2013년 3월과 2014년 2~3월에 발현 건수가 큰 폭으로 증가하는 현상을 나타낸다. 해당 시기의 원문을 검토한 결과 실제로 2013년 3월에 황사 및 미세먼지로 인한 트윗량이 크게 증가하였으며, 특히 미세먼지에 대한 문제가 심각하게 대두되었던 2014년 봄철에 관련 트윗이 급증하면서 국민들의 관심이 크게 높아졌던 것을 확인할 수 있었다.

[그림 8] 대기 분야에 대한 빈도 분석 결과



대기 분야에 대한 감성 분석 결과[그림 9] 역시 빈도 분석 결과와 유사하다. 황사 및 미세먼지 문제가 대두되었던 2013년 3월과 2014년 2~3월에 긍정 감성과 부정 감성이 모두 증가하였으며, 특히 2014년에는 부정 감성이 긍정 감성 대비 큰 폭으로 증가하여 미세먼지에 대한 국민들의 우려가 심각한 수준이었음을 알 수 있다.

[그림 9] 대기 분야에 대한 감성 분석 결과



[그림 10] 대기 분야에 대한 연관어 분석 결과

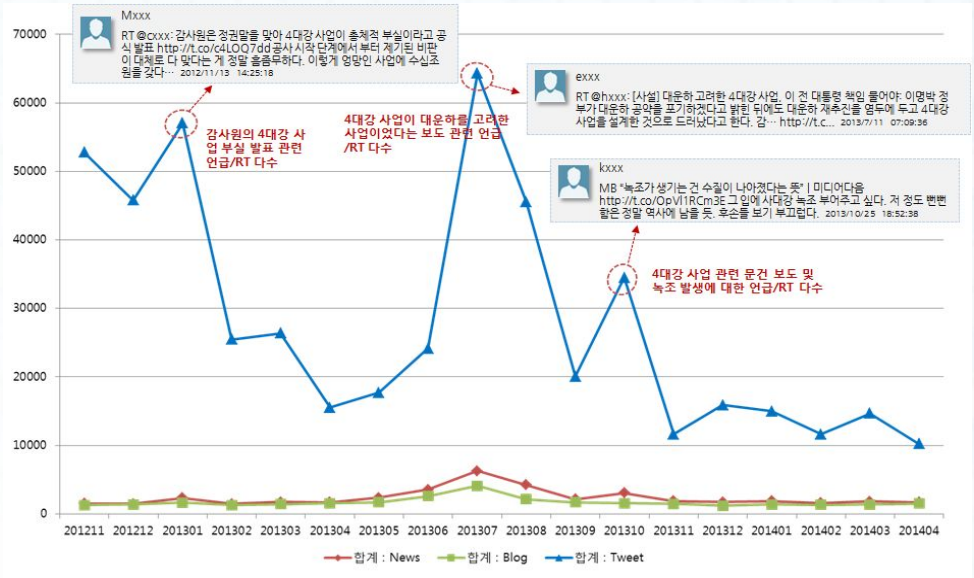
명사						서술어					
순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도
1	RT	204016	26	제품	27446	1	좋다	100722	26	내리다	23380
2	먼지	115979	27	지역	26660	2	보다	71321	27	가능하다	22335
3	미세	107260	28	마음	26542	3	많다	62550	28	필요하다	21997
4	물	59569	29	손	26183	4	보이다	52578	29	넣다	21978
5	사람	56329	30	발생	25660	5	들다	48038	30	맞다	21923
6	집	52814	31	관리	25501	6	만들다	45264	31	가지다	21842
7	눈	49148	32	효과	24882	7	가다	45146	32	차갑다	21301
8	시간	47327	33	상태	24873	8	오다	44842	33	말하다	21288
9	서울특별시	43824	34	한국	23782	9	나오다	44795	34	타다	21243
10	차	43590	35	아이	23716	10	받다	44759	35	떨어지다	21189
11	날씨	39886	36	마스크	23633	11	크다	41879	36	차다	20972
12	비	39759	37	이용	23509	12	따르다	37285	37	다양하다	20650
13	위	37535	38	농도	23154	13	높다	34550	38	열다	19970
14	필요	35388	39	해	23035	14	쓰다	32209	39	인하다	19573
15	시작	35293	40	문제	22640	15	사용하다	32058	40	이용하다	19203
16	중국	33964	41	면	22629	16	먹다	31355	41	들어오다	19086
17	건강	33903	42	발	22575	17	느끼다	30587	42	찾다	18955
18	몸	32483	43	소리	22414	18	맑다	29572	43	생기다	18304
19	바람	32225	44	동안	22303	19	통하다	29455	44	작다	17615
20	이상	30540	45	다양	22206	20	주다	29433	45	따뜻하다	17252
21	길	28404	46	방법	22136	21	알다	28645	46	밝히다	17132
22	사진	28154	47	기분	22109	22	들어가다	27138	47	심하다	17100
23	오염	28135	48	최고	21973	23	살다	27084	48	발생하다	16889
24	환경	27965	49	물질	21417	24	마시다	26948	49	생각하다	16610
25	하늘	27628	50	주의	21348	25	나다	24047	50	좋아하다	16361

대기 분야의 연관어 분석 결과[그림 10]는 빈도 및 감성 분석과 같은 맥락에서 먼지, 미세, 중국, 오염, 마스크, 농도, 물질 등의 주요 연관어가 상위권을 차지하고 있다는 점이 특징적이다. 이러한 현상은 크게 두 가지로 해석될 수 있는데, 먼저 대기 분야에서 황사 및 미세먼지 문제가 차지하는 중요도가 실제로 높을 수 있다. 그리고 대기 분야에서 황사 및 미세먼지 문제가 차지하는 비중이 크지는 않지만 다른 문제에 비해서 국민들의 관심이 많이 표현되는 분야일 수 있다. 소셜 빅데이터 분석을 통해 이러한 두 가지 원인을 명확히 구분하는 것은 불가능하므로 결과를 해석할 때 유의할 필요가 있다.

(4) 수질 분야에 대한 분석 결과

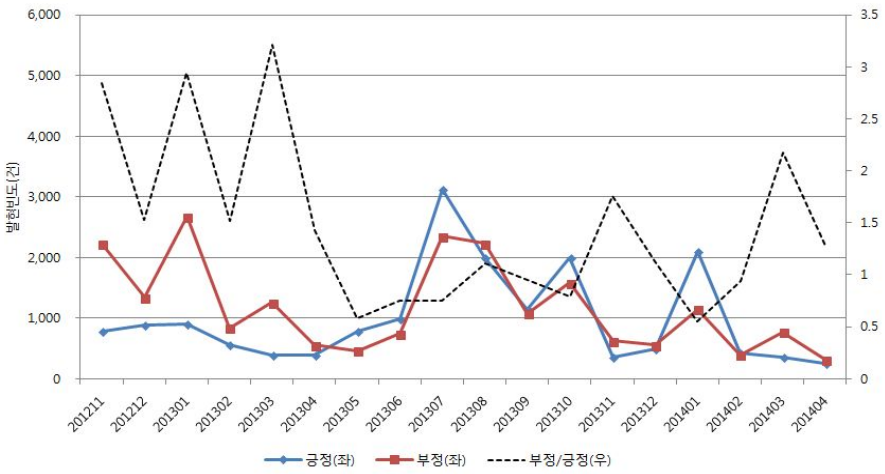
수질 분야의 빈도 분석 결과에서는 트위터 채널에서의 감소세가 두드러진다[그림 11]. 2013년 7~8월에는 트위터, 블로그, 뉴스 전 채널에 걸쳐 언급량이 급증하였는데, 이 시기에 4대강 사업 관련 언론 보도가 증가하면서 이에 대한 국민들의 관심이 집중된 것으로 추측된다. 그 외에도 수질 분야의 트위터 문서량이 급증한 시기에 대해 검토한 결과 주로 4대강 관련 문제가 대두된 시점임을 확인할 수 있었으며, 국민들이 4대강 문제와 정치적 이슈를 연계하여 민감하게 반응하는 것을 확인할 수 있다.

[그림 11] 수질 분야에 대한 빈도 분석 결과



수질 분야는 부정적 감성과 긍정적 감성의 비중이 시기별로 크게 달라지는 결과를 보인다[그림 12]. 2013년 4월까지는 대체적으로 부정 감성이 높게 나타났으나 이후 2013년 5~7월동안 긍정 감성의 비중이 더 높아졌다. 2013년 7월과 10월의 문서량 급증은 긍정 감성과 부정 감성에 모두 영향을 주었지만 긍정 감성의 증가율이 약간 더 우세하였다.

[그림 12] 수질 분야에 대한 감성 분석 결과



마찬가지로 연관어 분석에서도 4대강사업, 강, 공사, 운하 등 4대강 사업 관련 연관어가 최상위에 나타나고 있다[그림 13]. 또한 이슈의 특성상 이명박, 정부, 국민,

박근혜, 대통령, 새누리 등 정치적 키워드도 많이 나타나는 특성이 보인다. 결과적으로 수질 분야에서는 4대강과 관련된 수질 문제가 중요한 비중을 차지함과 동시에, 4대강 키워드에 대한 국민의 관심도가 상당히 높은 수준임을 확인할 수 있다.

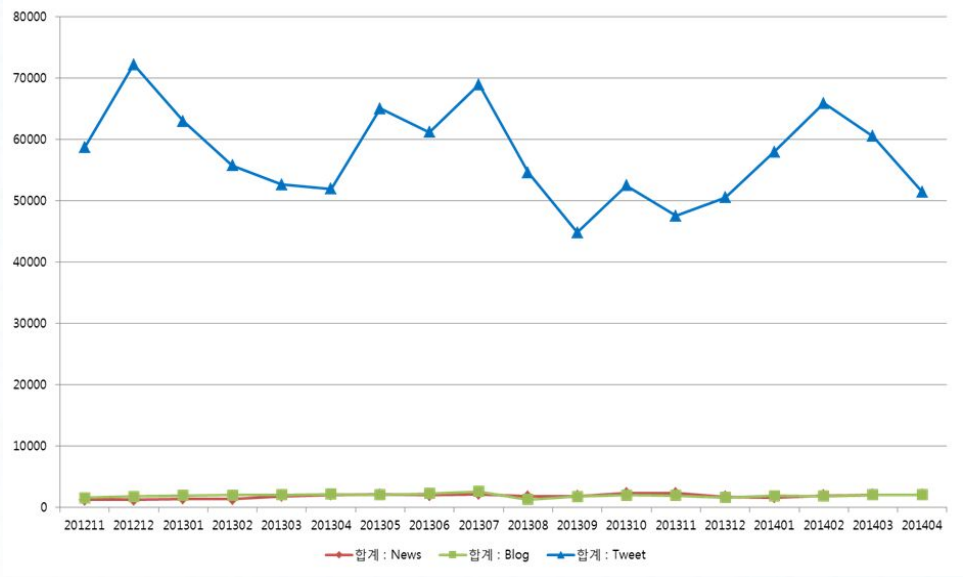
[그림 13] 수질 분야에 대한 연관어 분석 결과

명사						서술어					
순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도
1	대강	376244	26	예산	25091	1	보다	49081	26	가능하다	14807
2	RT	374791	27	시작	24746	2	많다	37090	27	생기다	14119
3	사업	152743	28	시간	23965	3	받다	35479	28	살다	14050
4	이명박	122046	29	국토	23853	4	만들다	34724	29	발생하다	13916
5	4대강사업	115042	30	한국	23572	5	나오다	33882	30	가지다	13914
6	강	60932	31	윤하	23347	6	가다	31131	31	나다	13702
7	정부	57092	32	조사	22779	7	크다	29595	32	찾다	13528
8	국민	55243	33	정권	22718	8	좋다	29291	33	내다	13389
9	환경	51040	34	서울특별시	22537	9	보이다	26511	34	올리다	13245
10	놀이	45122	35	발생	22052	10	따르다	26481	35	이용하다	13021
11	박근혜	43794	36	수도	21963	11	들다	26450	36	막다	12614
12	사람	41719	37	경제	21711	12	통하다	25363	37	높다	12310
13	공사	41470	38	위	21436	13	밝히다	25356	38	열다	11883
14	문제	36680	39	사기	21367	14	오다	23773	39	반대하다	11761
15	감사	35686	40	사진	21194	15	먹다	21897	40	모르다	11666
16	대통령	32891	41	국가	21144	16	알다	20873	41	추진하다	11395
17	지역	28406	42	의원	21035	17	쓰다	20516	42	다양하다	11219
18	결과	28066	43	이상	21011	18	들어가다	19956	43	타다	10821
19	반대	26908	44	낙동강	20847	19	말하다	19094	44	내리다	10489
20	건설	26867	45	낙동	20829	20	인하다	17045	45	안되다	10488
21	새누리	26575	46	확인	20586	21	사용하다	16477	46	알리다	10450
22	필요	26340	47	시설	20349	22	주다	16466	47	생각하다	10414
23	감사원	26316	48	나라	20293	23	살리다	16379	48	찍다	10197
24	추진	26221	49	법	20270	24	필요하다	16233	49	보내다	10128
25	수돗	25395	50	집	20013	25	맞다	15098	50	깨끗하다	10005

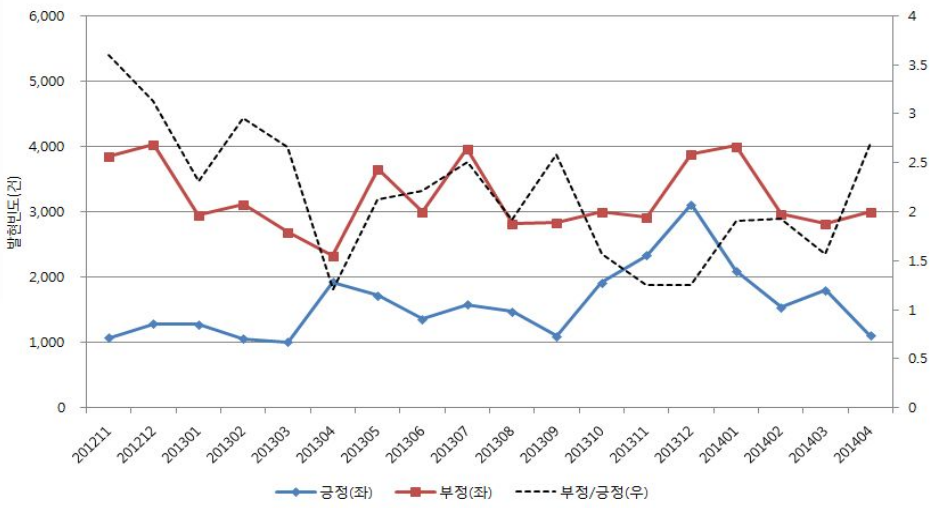
(5) 쓰레기(폐기물) 분야에 대한 분석 결과

쓰레기 분야의 키워드에 대한 발현 빈도는 전반적으로 일정한 수준으로 나타내며, 다른 환경 분야에 비해 절대적인 빈도수가 매우 높다[그림 14]. 쓰레기와 관련된 키워드는 중의적인 사용을 배제하기 위해 최대한의 정제 과정을 거쳤지만, 전수조사 자체가 어렵기 때문에 부적합 문서들이 포함되어 있을 가능성이 있다. 따라서 이처럼 소셜 빅데이터에 대한 빈도 분석을 수행했음에도 불구하고 직접적인 결과 해석에는 한계가 존재하는 경우가 있음을 염두에 두어야 한다.

[그림 14] 쓰레기(폐기물) 분야에 대한 빈도 분석 결과



[그림 15] 쓰레기(폐기물) 분야에 대한 감성 분석 결과



쓰레기 문제의 특성상 감성 분석 결과[그림 15]에서는 부정 감성이 긍정 감성에 비해 큰 비중을 차지하지만, 각각의 증감 추이는 일정하지 않다. 빈도 분석 결과와 마찬가지로 다수의 부적합 문서가 포함되어 있음을 고려할 때 일시적인 감성 변화에 의미를 부여하기는 어려울 것으로 판단된다. 다만 2012년 11월부터 2013년 8월까지의 부정 감성 빈도 추이가 총 빈도 추이와 유사하게 나타나 해당 시기에 발현된 문서들이 부정적인 감성의 발현에 영향을 미쳤음을 짐작할 수 있다.

쓰레기 분야의 연관어 분석 결과[그림 16]에서 쓰레기통, 활용, 음식물, 수거, 처리, 분리, 봉투 등 환경 분야에서 의미 있는 용어들이 상위권에 나타났으며 버리다,

줍다, 치우다, 모으다, 재활용하다 등의 서술어도 많이 나타나는 것을 확인할 수 있다.

[그림 16] 쓰레기(폐기물) 분야에 대한 연관어 분석 결과

명사						서술어					
순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도
1	RT	422263	26	위	25597	1	버리다	95146	26	맞다	21494
2	통	116815	27	국민	24792	2	보다	90197	27	줍다	21489
3	쓰레기통	96853	28	이상	24525	3	만들다	61900	28	가지다	20841
4	사람	90486	29	민주	24188	4	쓰다	53709	29	나다	18930
5	활용	69997	30	환경	23935	5	좋다	52524	30	치우다	18122
6	집	56389	31	필요	23732	6	나오다	51538	31	통하다	18071
7	음식	55246	32	시작	23488	7	많다	50538	32	내다	18063
8	인간	48047	33	아이	23193	8	가다	48247	33	좋아하다	17878
9	개	45683	34	학교	22612	9	먹다	44510	34	그렇다	17748
10	폐기	41489	35	마음	22524	10	들다	44335	35	사용하다	17654
11	음식물	38810	36	세상	22338	11	받다	43135	36	사다	17547
12	새끼	38638	37	나라	22292	12	알다	40481	37	듣다	16012
13	물	37307	38	친구	22052	13	보이다	37344	38	보내다	15879
14	수거	36888	39	소리	21854	14	오다	31927	39	잡다	15428
15	사진	32146	40	이유	21159	15	살다	29974	40	밝히다	15288
16	처리	31434	41	법	21104	16	넣다	29923	41	올리다	15129
17	눈	31403	42	차	20291	17	크다	27614	42	치다	14589
18	분리	30999	43	기자	19579	18	말하다	27557	43	만나다	14564
19	시간	30920	44	어머니	19163	19	모르다	25952	44	걸다	14421
20	길	29824	45	밥	18936	20	주다	25649	45	느끼다	14356
21	문제	29237	46	면	18591	21	생각하다	23598	46	끝나다	13883
22	봉투	28730	47	지역	18574	22	찾다	22500	47	찍다	13578
23	한국	26932	48	입	18573	23	들어가다	22037	48	말다	13535
24	손	26308	49	민주통합당	18392	24	따르다	22013	49	모으다	13469
25	청소	26133	50	그림	18261	25	안되다	21675	50	재활용하다	13434

(6) 세부 환경 분야별 특성 비교

앞서 살펴본 바와 같이 세부 환경 분야별 빈도 분석, 감성 분석, 연관어 분석 결과를 통합적으로 검토하여 분야별 특성을 비교할 수 있다. 예를 들어 분석 기간 동안 국민들의 관심이 높았던 분야, 관심이 증가(또는 감소)하고 있는 분야, 긍정 대비 부정 감성이 높은 분야, 긍정 대비 부정 감성이 증가(또는 감소)하고 있는 분야 등을 파악할 수 있을 것이다.

[표 2]는 총 빈도수와 감성 변화 추이에 따라 세부 환경 분야를 구분한 자료이다. 총 빈도수의 측면에서 쓰레기, 대기, 기후, 수질 분야에 대한 국민들의 관심도가 높으며, 그 중에서도 발현빈도가 상승하는 추세에 있는 대기 분야의 중요도가 큰 것으로 파악된다. 특히 국민의 관심도가 높은 '상' 그룹에서는 부정 감성의 비중이 상대적으로 증가한 대기와 기후 분야에 대한 원인 분석과 대책 마련이 필요할 것이다.

추가적으로 토양 분야의 경우 현재 발현빈도는 10만 건 이하로 낮은 수준이지만 국민들의 관심이 점차 증가하는 것으로 나타나 향후 국민들의 정책수요가 높아질 가능성이 있다. 반대로 수질이나 기후, 생물다양성 분야는 정책적으로 중요한

분야임에도 불구하고 점차 국민들의 관심도가 낮아지는 경향을 보이고 있다. 기후변화 대응, 녹조 등 수질 관리, 생물다양성 보존 등 지속적으로 환경정책을 추진하고 지속가능성을 높이기 위해서는 국민들의 지지와 관심을 확보하기 위하여 추가적인 노력이 필요할 것으로 보인다.

[표 2] 세부 환경 분야별 총 빈도수와 감성 변화 추이

구분		총 빈도		
		상(>40만 건)	중(10~40만 건)	하(<10만 건)
부정감성 변화 추이	증가	대기	-	악취
	유지	-	생물다양성, 인적재해	토양
	감소	기후, 수질, 쓰레기	소음, 유해물질(보건)	해양
긍정감성 변화 추이	증가	대기, 쓰레기	소음, 유해물질(보건)	악취
	유지	수질	생물다양성, 인적재해	토양
	감소	기후	-	해양
부정/긍정 변화 추이	증가	대기, 기후	-	-
	유지	-	생물다양성	토양, 해양
	감소	수질, 쓰레기	소음, 인적재해, 유해물질(보건)	악취

3. 소셜 빅데이터 분석 결과의 응용

(1) 환경에 대한 국민의식조사

현재 환경과 관련된 일반 국민의 인식을 파악하기 위해서는 일반적으로 설문조사 기법을 활용하고 있다. 그러나 앞서 살펴본 바와 같이 소셜 빅데이터 분석을 통해 환경 분야별로 국민의 관심과 정서를 파악할 수 있다면, 이는 전통적인 설문조사 기법과 함께 유용한 수단으로 활용될 수 있을 것이다. 예를 들어 환경 분야에서 소셜 빅데이터를 통해 향후 분석할 수 있을 것이라 기대되는 주제로는 성공적인 환경정책, 불필요한 환경정책, 특정 환경정책에 대한 인지도 등을 들 수 있다.

또한 소셜 빅데이터를 통해 확인할 수 있는 중요한 주제 중의 하나는 환경정책의 공급자와 수요자 간 인식 차이를 분석하는 것이다. 정부는 공급자 측면에서 공익과 관련된 환경정책을 수립·이행하고 있지만, 실제 국민이 요구하는 정책수요와는 차이가 존재할 수 있다. 이를 확인하기 위한 한 가지 방법은 정부나 환경부의 계획에 반영된 환경 관련 주요 키워드를 도출하고 이에 대한 국민의 인식을 소셜 빅데이터로 분석하는 것이다. 예를 들어 정부에서 굉장히 중요한 비중을 두고 추진하는 환경정책이지만 이에 대한 국민의 인식이 저조하거나 부정적인 정서를

나타낸다면 해당 정책의 실효성에 대해서는 다시 검토할 필요가 있다.

그 외에도 세부 환경 분야에 대한 분석은 현재 가장 높은 정책수요를 보이는 분야를 파악하여 정책분야별 우선순위를 설정하는 데 도움이 될 것으로 보인다. 이슈 및 사건 전후의 인식 변화 분석이나 환경문제의 사전 탐지 가능성 검토를 위해서는 민감도 분석이나 선후관계 분석 등이 활용될 수 있다. 특히 세부 환경 분야별로 특정 이슈 및 사건 전후의 발현 빈도가 어떻게 변화했는지 분석하여 이슈 및 사건의 영향력을 파악하는 방법은 향후 필요한 부분에 활용될 수 있을 것이다.

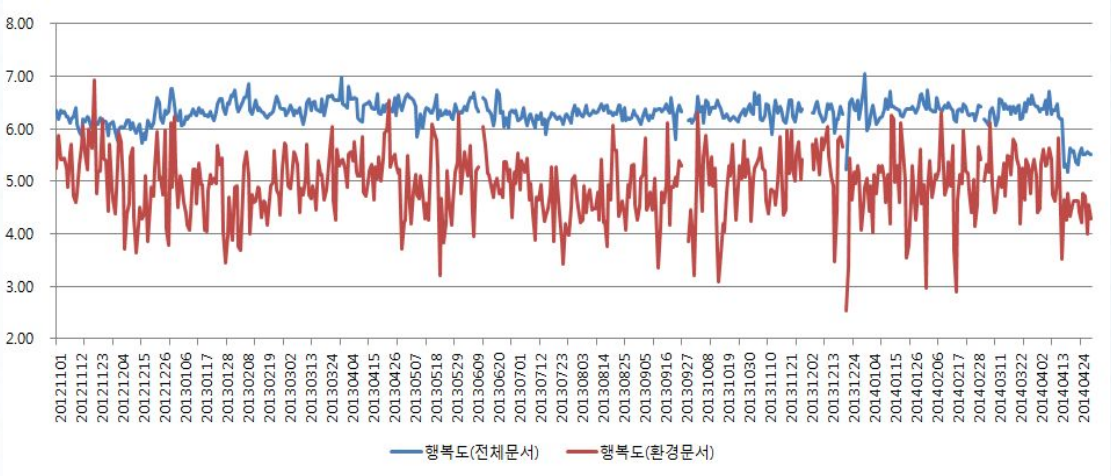
(2) 행복도 측정

소셜 미디어 사용자들은 하루에도 수십 번씩 자신의 생각이나 감정이 담긴 글을 온라인에 게시하고 있다. 비록 인터넷 사용자에 한정되기는 하지만 실시간으로 변화하는 개개인의 감성을 집단적으로 분석할 수 있는 기반이 마련된 것이다. 일반 국민들의 감정은 단순히 사회현상을 분석하는 학자들뿐만 아니라 정책을 계획하고 수립하는 정책입안자들에게 있어서도 중요한 지표가 된다. 특히, 국민들이 느끼는 행복의 정도를 정량화하고 원인을 분석할 수 있다면 국민들이 원하는 정책을 마련하는데 유용하게 활용될 수 있을 것이다.

이러한 점에 착안하여 미국 Vermont 대학의 연구팀은 트위터 데이터를 통해 사용자들의 행복도를 측정할 수 있는 "헤도노미터(Hedonometer)"라는 분석툴을 개발하였다(Dodds et al., 2011). 앞서 추출한 소셜 빅데이터에 헤도노미터 기법을 적용하여 우리나라 국민들의 행복도를 측정한 결과는 [그림 17]과 같다. 행복도 측정을 위해 필요한 데이터는 소셜 빅데이터 중 트위터 채널을 기준으로 수집하였으며, 전체 문서와 환경 관련 문서에 대하여 각각 동일한 방법론을 적용하여 행복도를 산정하였다.

전체 문서를 대상으로 산정한 행복도는 평균 6.30점으로 미국 Vermont대학의 헤도노미터 분석 결과(5.8~6.4점)와 유사한 범위 내에서 신뢰할 수 있는 것으로 판단된다. 특히 세월호 사건이 발생한 2014년 4월 16일 이후 행복도가 급감하여 어느 정도 국민 정서를 대변한다고 볼 수 있다. 반면 환경 문서 기준의 행복도는 평균 4.94점으로, 전체 문서 기준의 행복도에 비해 크게 낮은 것으로 나타났다. 이 결과에 따르면 우리나라 국민들은 분석 기간 동안 전반적으로 행복한 수준을 나타냈지만, 환경분야와 관련해서는 행복하지 않은 감성을 드러낸 것으로 보인다.

[그림 17] 일별 행복도 분석 결과



(3) 외부 통계자료와의 연계

환경 분야에서의 소셜 빅데이터 분석이 보다 입체적이고 다층적인 함의로 이어지기 위해서는 정형화된 외부 데이터와의 연계 분석이 필수적이다. 특히 소셜 빅데이터는 어떤 주제 하에서 어떠한 외부 데이터와 연계되느냐에 따라 매우 다양한 결과를 도출할 수 있을 것이다. 예를 들어 앞서 측정된 행복도와 상관관계가 있을 것으로 추측되는 환경 관련 외부 변수와의 연관성 분석을 통해 의미있는 정보를 도출할 수 있다. 이미숙 외(2014)의 분석 결과에 따르면 서울의 평균 가시거리가 $1km$ 길어질수록 환경 분야에서의 행복도는 0.01점의 크기로 개선되며(p-value=0.043), 습도가 10% 높아질수록 환경 분야에서의 행복도는 0.05점 낮아지는 것으로 나타났다(p-value=0.012). 즉, 가시거리 및 습도와 같은 기후·대기 환경이 사람들의 감정에 영향을 미친다는 것이 수치적으로 확인되었다고 볼 수 있다. 이러한 결과는 단편적인 사례에 불과하지만, 환경 분야의 소셜 빅데이터를 이용해 산정한 행복도 점수가 일상생활에서 직접적으로 느낄 수 있는 대기 관련 변수에 의해 영향을 받을 수 있다는 점을 시사한다. 이처럼 소셜 빅데이터 분석은 외부 데이터와 연계될 때 더욱 유용한 정보를 제공할 수 있을 것이다.

[정리하기]

1. 소셜 빅데이터 분석 절차

- 소셜 빅데이터 분석은 일반적으로 분석 목적 및 관점의 정의, 키워드 선정, 키워드 검증, 데이터 수집 및 분석, 결과 도출의 과정을 통해 수행된다.
- 소셜 빅데이터 분석 결과의 신뢰도를 높이기 위해서는 적절한 검색 키워드를 설정하고 연구 주제에서 벗어나는 문서들을 배제할 수 있도록 데이터를 충분히 정제하는 과정이 필수적이다.

2. 환경 분야 소셜 빅데이터 분석

- 2012년 11월~2014년 4월까지 환경 분야 소셜 빅데이터 분석 사례에 따르면 소셜 미디어 채널 중 트위터를 통한 의견 표출이 지배적이었으며, 환경분야 전반에 대해서는 대체로 국민들의 감성이 부정적인 것으로 나타났다.
- 세부 환경 분야별 분석에 따르면 쓰레기, 대기, 기후, 수질 분야에 대한 국민의 관심이 상대적으로 높았으며, 감성 분석과 연관어 분석을 통해서도 분야별 특징을 파악할 수 있다.
- 특히 기후 분야에서는 계절별 이상 기후 현상에 대해, 대기 분야에서는 봄철 황사 및 미세먼지에 대해 국민의 관심이 집중됨과 동시에 부정적 감성이 발현된다는 특징이 있다.

3. 소셜 빅데이터 분석 결과의 응용

- 환경 분야 소셜 빅데이터 분석은 전통적인 설문조사 방식과 함께 환경에 대한 국민의 인식을 파악하기 위해 유용하게 활용될 가능성이 있다.
- 소셜 미디어를 통해 실시간으로 표출되는 개인의 감성은 국민들이 느끼는 행복의 정도를 측정하기 위해 활용될 수 있으며, 외부 데이터와의 연계를 통해 더욱 유용한 정보를 제공할 수 있다.

[참고문헌]

- 이미숙 외(2014), 빅데이터를 활용한 환경분야 정책수요 분석, 한국환경정책평가연구원.
- Dodds, P.S. et al.(2011), "Temporal patterns of happiness and information in a global social network: hedonometrics and twitter", PLoS ONE, Vol.6, e26752.