



TOBIGS TEAM PROJECT

감성분석을 통한
키워드 기반
대한민국 정치흐름 파악

2017.07.15

구혜인 김서연 연다인 허능호

INDEX



주제선정배경



데이터 수집
및 전처리



데이터 분석



결론 및 제언



주제선정배경

1 | 주제선정배경

한겨레
HANI.CO.KR

서울신문

東亞日報

국민일보



세계일보

J 중앙일보

朝鮮日報

한국일보

경향신문

“

각 신문사들은 대한민국의 사건들을
어떻게 표현하고 있을까?

”

1 | 주제선정배경

한겨레
HANI.CO.KR

서울신문

東亞日報

국민일보



세계일보

J 중앙일보

朝鮮日報

한국일보

경향신문

“그렇다면 그 신문사들의 기사를 읽으면
대한민국의 정치흐름을 볼 수 있지 않을까?”

2

데이터 수집

1) 신문사 수집 기준

전국 신문사 발행 부수 순위와 네이버의 뉴스스탠드 목록을 바탕으로 **총 8개의 신문사**를 선정하였다.
→ 조선일보, 동아일보, 중앙일보, 문화일보, 국민일보, 서울신문, 한겨레, 프레시안

선정된 8개 신문사의 **사설들을 수집**하였다.

∴ 사설이 일반 기사보다 해당 신문사의 주장이나 의견을 더 잘 반영할 것이라 판단

2017년 일간신문 발행 유료부수 인증결과

구분	매체명	발행부수	유료부수
1	조선일보	1,513,073	1,254,297
2	동아일보	946,765	729,414
3	중앙일보	978,798	719,931
4	매일경제	705,526	550,536
5	한국경제	529,226	352,999
6	문화일보	177,887	163,090
7	국민일보	185,787	138,819

예) 사드 배치 시작에 관한 사설의 신문사간 시각 차이


[한겨레 사설] 한·미 정부의 무책임한 '사드 대못 박기'

VS

[중앙일보 사설] 사드 배치 시작...국론분열 없이 마무리해야

출처:

<http://news.heraldcorp.com/view.php?uid=20170621000925>
<http://www.hani.co.kr/art/society/schooling/787278.html>



데이터 수집

감성사전 구축

Süddeutsche Zeitung

2

데이터 수집

2) 크롤링

데이터 수집 기간 : 2016.01.01 ~ 2017.06.26

- 문화일보의 경우, 최근 6개월간의 사설만 제공하기 때문에 2017년 1월 1일부터 2017년 6월 26일까지의 사설만 크롤링
→ 아래 크롤링한 사설의 개수 역시 현저히 적은 것 확인 가능

신문사별 크롤링한 사설 개수



3

감성사전 구축

프로젝트를 진행하면서 서울대학교 언어학과 연구진이 구축한 KOSAC 감성사전을 활용하였다.

KOSAC내의 여러 사전 중에서는 **Polarity**와 **Intensity** 라는 두 개의 사전을 이용하였다.

위 감성사전은 **조선일보의 생활, 사회면과 한국일보, 한겨레**에서 총 332개의 기사, 7744개의 문장을 선정, 3명의 연구진들이 주석하여 구축하였다.

→ 프로젝트의 주제가 **신문**과 깊은 관련이 있다는 점을 바탕으로 해당 감성사전이 가장 적합하여 채택함.

3

감성사전 구축

→ 해당 형태소의 **긍정 / 부정** 정도를 의미하는 가장 중요한 사전!

Polarity 사전

ngram	freq	COMP	NEG	NEUT	None	POS	max.value	max.prop
비값/XR	1	0	1	0	0	0 NEG		1
비값/XR;하/XSA	1	0	1	0	0	0 NEG		1
비값/NNG	1	0	1	0	0	0 NEG		1
비값/NNG;하/XSV	1	0	1	0	0	0 NEG		1
비교/NNG	1	0	0	0	0	1 POS		1
비교/NNG;적/XON	1	0	0	0	0	1 POS		1
비교/NNG;적/XON;명칭/XR	1	0	0	0	0	1 POS		1
비교적/MAG	1	0	0	0	0	1 POS		1
비교적/MAG;종/VA	1	0	0	0	0	1 POS		1
비교적/MAG;종/VA;아/EC	1	0	0	0	0	1 POS		1
비극/NNG	7	0	0.857143	0.142857	0	0 NEG	0.857143	
비극/NNG;참/XON	2	0	1	0	0	0 NEG		1
비극/NNG;의/IKG	1	0	1	0	0	0 NEG		1
비극/NNG;의/IKG;형태/NNG	1	0	1	0	0	0 NEG		1
비극/NNG;적/XON	1	0	0	1	0	0 NEUT		1
비극/NNG;적/XON;이/VCP	1	0	0	1	0	0 NEUT		1
비탄/NNG	1	0	1	0	0	0 NEG		1

✓ 각 형태소가 어떤 극성을 띄고 있는지
COMP, NEG, NEUT, NONE, POS의
속성 중 하나로 표현하였다.

✓ 계산 방식 :
POS / NEG 점수를 각각 **+1, -1**로 잡고
사전 내의 확률과 곱하여 계산

3

감성사전 구축

→ 해당 형태소에 어느 정도의 **주관성**이 개입되는지 설명해주는 사전

Intensity 사전

ngram	freq	High	Low	Medium	None	max.value	max.prop
감내/NNG	1	1	0	0	0	0 High	1
감독/NNG	6	0	0.666667	0.333333	0	0 Low	0.666667
감동/NNG	2	0	0	1	0	0 Medium	1
감사/NNG	1	1	0	0	0	0 High	1
감칠맛/NNG	1	0	0	1	0	0 Medium	1
감탄사/NNG	1	0	0	1	0	0 Medium	1
갑자기/MAG	4	0	0.5	0.25	0.25	0 Low	0.5
갑작스레/MAG	1	0	0	1	0	0 Medium	1
값/NNG	5	0	0.2	0.8	0	0 Medium	0.8
갑싸/VA	1	0	0	1	0	0 Medium	1
강군/NNG	1	0	0	1	0	0 Medium	1
강단/NNG	1	0	1	0	0	0 Low	1
강도/NNG	2	0.5	0	0.5	0	0 High	0.5
강력/XR	2	0.5	0	0.5	0	0 High	0.5
강렬/XR	1	0	0	1	0	0 Medium	1

✓ 각 형태소의 주관성의 정도를 High, Low, Medium, None의 속성으로 표현하였다.

✓ 계산 방식 :

각 속성의 확률에 (NONE)1점, (LOW)4점, (MEDIUM)7점, (HIGH)10점을 곱하여 계산

3

감성사전 구축

더 구체적인 감성사전을 구축하고자 8개 신문사에서 사설 25개씩을 뽑아 총 200개의 사설을 먼저 맞춰본 후,
기존의 감성사전에 없는 **6063개의 단어를 추가**하였다. → 전체 **22423개!**
기존의 사전이 구축된 방법 그대로 새로운 형태소에 대한 각각의 polarity와 intensity 점수를 계산하였다.

실제 추가한 형태소 사전 및 예시

오염원/NNG	NEG	NEG	NEG	LOW	MEDIUM	HIGH
강공책/NNG	POS	NEG	NEG	MEDIUM	LOW	MEDIUM
영원/NNG	POS	POS	POS	MEDIUM	MEDIUM	LOW
페이샤오통/UN	NEG	NEG	NEG	NONE	MEDIUM	HIGH
열심/NNG	POS	POS	POS	MEDIUM	MEDIUM	HIGH
지분/NNG	POS	POS	NEG	LOW	LOW	LOW
정비/NNG	POS	POS	POS	LOW	MEDIUM	LOW
지역구/NNG	NEUT	NEUT	NEUT	LOW	NONE	NONE
친분/NNG	POS	POS	NEG	LOW	LOW	MEDIUM
옥포신형/NNG	POS	POS	POS	MEDIUM	NONE	LOW
철새/NNG	NEG	NEG	NEG	LOW	NONE	HIGH
홍채/NNG	NEUT	NEUT	POS	NONE	NONE	LOW
약진/NNG	NEG	POS	POS	LOW	MEDIUM	HIGH
드리우/VV	NEG	NEG	NEG	MEDIUM	MEDIUM	MEDIUM

샘플 기사 200개를 적용시켰을 때의 정확도

원래의 사전으로 구한 정확도 : $80/200 = 40\%$
 새로 추가했을 때의 정확도 : $140/200 = 70\%$

✓ 강공책

polarity : $(+1-1-1)/3 = -0.667$ intensity : $4*(1/3) + 7*(2/3) = 6$

✓ 열심

polarity : $(+1+1+1)/3 = 1$ intensity : $7*(2/3) + 10*(1/3) = 8$



감성사전 적용

4

감성사전 적용

감성사전이 **n-gram**으로 이루어져 있기 때문에, 각 사설의 단어를 3-gram/2-gram/1-gram으로 나누어 사전에서 검색한 후, 각 점수를 합산하는 방식을 채택하였다.

* 명확한 이해를 위하여 다음 예시를 통하여 알아보도록 하자.

이때, Int*Pol은 Intensity*Polarity 점수를 사용하였음을 의미하며, 이 점수가 적용된 형태소는 따로 주황색으로 표시하였다.

예시) 나는 밥을 맛있게 먹는다. → 나/NP+는/JX+밥/NNG+을/JKO+맛있/VA+게/ECD+먹/VV+는/EPT+다/EFN

전체 Polarity 적용

나/NP+는/JX+밥/NNG+
을/JKO+맛있/VA+게
/ECD+먹/VV+는/EPT+
다/EFN

141/200 = **70%** ✓

전체 Int*Pol 적용

나/NP+는/JX+밥/NNG+
을/JKO+맛있/VA+게
/ECD+먹/VV+는/EPT+
다/EFN

129/200 = **64.5%**

서술어 Int*Pol +
나머지 Polarity

✓ 서술어 {동사, 형용사, 종결어미, 연결어미}

나/NP+는/JX+밥/NNG+
을/JKO+맛있/VA+게
/ECD+먹/VV+는/EPT+
다/EFN

127/200 = **63.5%**

명사 Int*Pol +
나머지 Polarity

✓ 명사 {보통명사, 고유명사}

나/NP+는/JX+밥/NNG+
을/JKO+맛있/VA+게
/ECD+먹/VV+는/EPT+
다/EFN

126/200 = **63%**

4

감성사전 적용

“역대 정부의 비선 실세는 제왕적 대통령의 어두운 그림자다.”



{역대/NNG + 정부/NNG + 의/JKG + 비선/NNG + 실세/NNG + 는/JK + 제왕/NNG + 적/XSN + 대통령/NNG + 의/JKG + 어둠/VA + ㄴ/ETD + 그림자/NNG + 이/VCP + 다/EFN}



{polarity 점수}

{0 + 1 + 0.581920904 + -1 + -0.666666667 + 0 + -1 + -0.555555556 + 0 + 0.581920904 + -1 + -0.409090909 + -0.666666667 + 0.385620915 + -0.435483871}

-2.677428034

“이혜훈 대표는 보수 철학과 소신이 뚜렷하고 열정 넘치는 합리적 정치인으로 평가된다.”



{이혜훈/NNG + 대표/NNG + 는/JX + 보수NNG + 철학/NNG + 과/JC + 소신/NNG + 이/JKS + 뚜렷/XR + 하/XSA + 고/ECE + 열정/NNG + 넘치/VV + 는/ETD + 합리적/NNG + 정치인/NNG + 으로/ETD + 합리적/NNG + 정치인/NNG + 으로/JKM + 평가/NNG + 되/VV + ㄴ다/EFN}



5.03155871

궁고구마

중앙일보

2020년 창간 100주년
역청여자대학교

joongang.co.kr

朝鮮日報

chosun.com

통상임금 아니다

분석 결과

한겨레

hani.co.kr

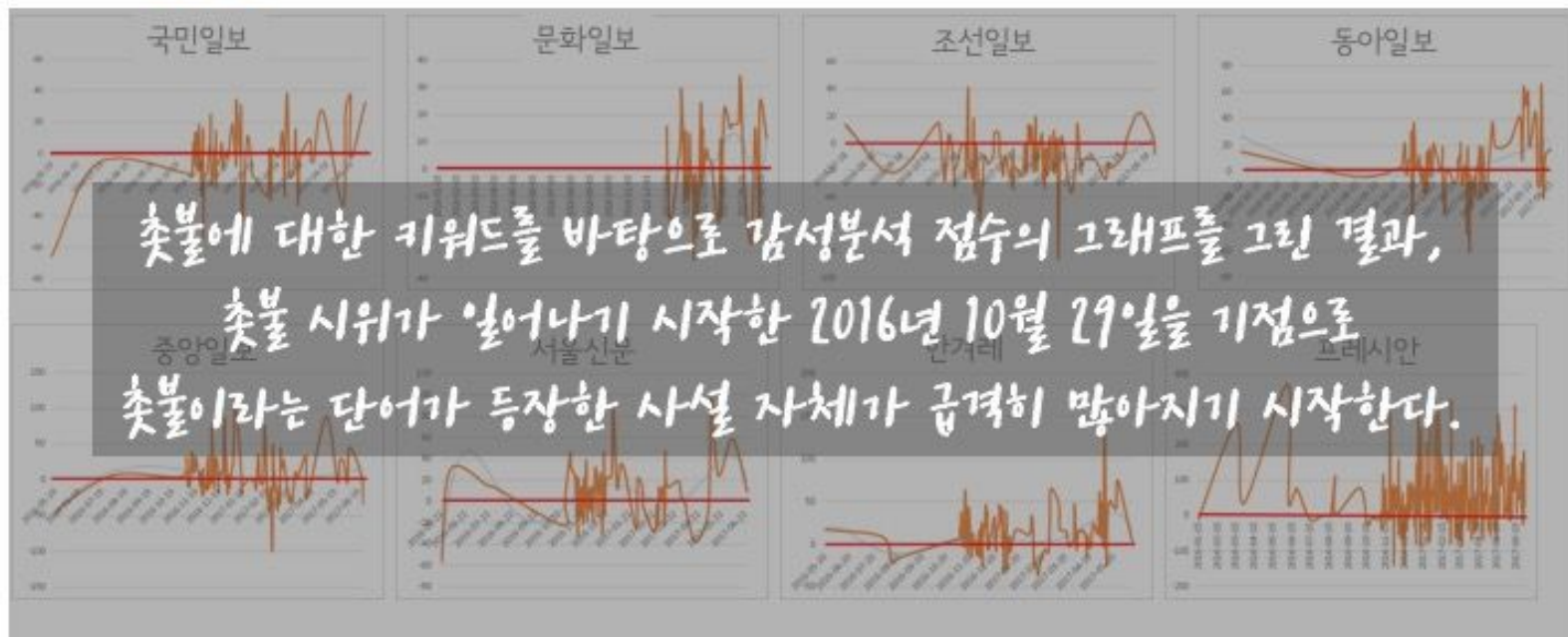
궁고구마

비루한 흡연의 풍경들

유인원...비인간적채널 아



② 촛불 키워드에 대한 그래프



대한민국의 1년 6개월 동안의 정치 키워드를 파악하기 위해 크롤링한 사설 본문의 제목을 바탕으로 **빈도 분석을 진행**하였다.

사설의 제목은 가장 중요하고 요점이 되는 부분을 포함한다는 가정 하에, 제목에서 많이 언급된 키워드는 당시 이슈라 판단하였기 때문이다.

그 결과는 다음과 같다.

세월호, **대통령+박근혜**, 촛불, 사드

→ 관련 사설의 개수가 너무 많아 의미 있는 분석이 어렵다고 판단

위 3개의 키워드를 기반으로 신문사 간 비교 진행하였다.

5

분석결과

2) 세월호 키워드

① 세월호 키워드에 대한 그래프 ***문화일보의 경우, 세월호에 대한 사설의 개수가 10개뿐이라 제외

중앙일보

서울신문

조선일보

국민일보

세월호에 대한 키워드를 바탕으로 그린 감성분석 점수의 그래프 중에서,
가장 만족스러운 결과인 동아일보의 그래프에 대하여 알아보도록 하자!

동아일보

하계레

프레시안

① 세월호 키워드에 대한 그래프



2016년 10월쯤 박근혜 사건을 계기로 세월호에 대한 사실들이 **부정적**으로 변하기 시작하였다. 이는 대부분 세월호 사건 당일 박근혜의 행적을 비판했기 때문이다.

[동아일보 사설] '비선 실세'의 단골 의사에게 대통령 건강 맡기다니 ...세월호 참사가 일어난 2014년 4월 16일 7시간 동안... 노화방치전문 김모 원장이 ...대통령 자문의'를 맡아 수시로 대통령에게 주사제를 처방했다는 것은 사실로 확인됐다...

2017년 4월쯤 세월호 인양과 함께 본격적으로 **긍정적**인 사실들이 많이 작성되었다. 이는 세월호 인양 및 정권 교체를 통한 새로운 변화를 희망하는 이야기가 많았기 때문이다.

[동아일보 사설] 과거보단 미래 향한 통합, 복지 대한민국으로 ...은 대통령은 ... "세월호 ... 다시 좀 제대로 조사되고 진실이 규명되게끔 하는 것이 필요..." 라 말했다. ...

5

분석결과

4) 사드 키워드

③ 사드 키워드에 대한 그래프



③ 사드 키워드에 대한 그래프



실제 국민일보에서 '사드'에 대한 **사설의 개수**가 **기간별**로 급격한 차이를 보였다. 급격히 개수가 많아지는 지점에 대한 이유를 다음과 같이 찾아 볼 수 있었다.

2016년 2월 초 : 사드 협상 시작

2016년 7월 경 : 대한민국 경북 성주에 사드 배치 확정 → **부정적**

2017년 3월 경 : 정권 교체에 따른 사드 재협상 논의 → **긍정적**으로 변화

- A. **프레스시안**의 경우, 타 신문사보다 사설의 양과 길이가 압도적으로 많음을 데이터 수집 단계에서부터 확인하였다. 실제 분석한 결과, 전반적으로 세 키워드에 대한 **긍정적**인 성향을 갖고 있었다.
- B. **문화일보**의 경우, 최근 반년간의 데이터밖에 제공하지 않아 분석의 한계가 있었지만, **'세월호' 키워드** 검색 결과 반년동안 총 **10개**의 사설(전체의 **2.7%**)밖에 없는 것은 눈여겨볼만하다.
- C. **중앙일보**의 경우, JTBC의 영향을 많이 받은 것으로 보인다. 박근혜 전 대통령과 태블릿 PC 사건을 기점으로, **'촛불' 키워드**가 전반적으로 **긍정적**으로 사용되었음을 확인하였다.

usaltos. que se ve en las
buena cantidad del tiempo
izan estos especialistas
la vigilancia, se
ias médicas. I
s llega una
ellos queda

THANK YOU
