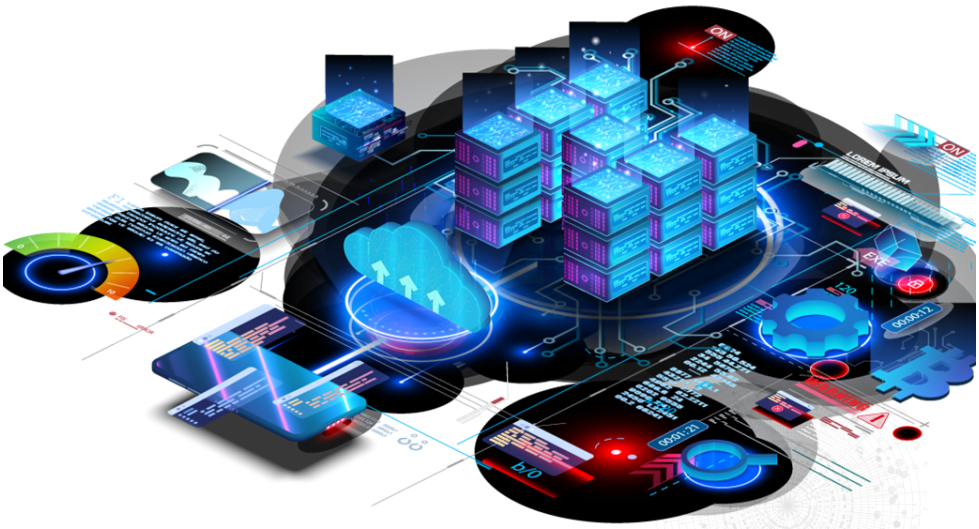
빅데이터 기술 교류 세미나

빅데이터와 여론조사



2022. 5.19 (목) 13:30~17:30

국회의사당 국회도서관 대강당 (지하1층)

- 주최 : 한국경영정보학회
- 주관 : 국민의힘 김영식 의원실
  더불어민주당 윤영찬 의원실
- 후원 : (주)바이브컴퍼니

한국경영정보학회

Program

진행: 박아름 (용인예술과학대 교수)

13:30~14:00	환영사 축 사	양희동 (이화여대 교수 . 2022 한국경영정보학회 회장) 김영식 (국민의힘 국회의원) 윤영찬 (더불어민주당 국회의원)
14:00~14:30	주제발표	**비정형 빅데이터 플랫폼을 활용한 감성 분석 및 실험 사례** 윤준태(바이브컴퍼니 AI 빅데이터 연구소장)
14:30~15:00	발표1	**MZ 세대의 마음을 읽으려면, "커뮤니티"를 공략하라! 6월 지방선거 결과 예측을 위한 분석연구** 양희동(이화여대 교수), 최한별(KAIST 연구교수), 김수림(KAIST 박사과정 연구원)
15:00~15:30	발표2	**Sometrend를 활용한 공공 메타버스 플랫폼에 대한 여론 분석과 정책 제언** 윤혜정(이화여대 교수)
15:30~16:00		휴식 및 개별 토론 시간
16:00~16:30	발표3	**Sometrend "유튜브 분석"을 활용한 20대 대선 여론 분석** 양성병(경희대 교수)
16:30~17:00	발표4	**과거 대선 및 서울시장 선거 회고적 분석: 썸트렌드 vs. 포털 트렌드** 이동원(고려대 교수)
17:00~17:30	발표5	**빅데이터와 AI 기반 여론 분석 연구와 교육 사례: 썸트렌드 vs. 코딩** 이경전(경희대 교수), 박아름(용인예술과학대 교수)

한국경영정보학회
Korea Society of Management Information Systems
www.kmis.or.kr
office@kmis.or.kr

빅데이터 기술 교류 세미나

빅데이터와 여론조사

축사

국민의힘 국회의원 김영식 의원님의 축사는 영상으로 준비됩니다.

빅데이터 기술 교류 세미나
빅데이터와 여론조사
축사

안녕하세요.
국회 과학기술정보방송통신위원회
더불어민주당 국회의원 윤영찬입니다.

한국경영정보학회의 「빅데이터와 여론조사」 빅데이터 기술 교류 세미나 개최를 진심으로 축하합니다.

오늘은 6월 1일 전국동시지방선거 선거운동 개시일입니다. 수많은 후보가 선거 운동을 하고 전략을 짤 때, 여론조사 결과를 기준으로 유권자의 마음을 파악합니다.

하지만 응답률 1%, 2%, 3%의 전화 면접과 ARS 방식의 여론조사 신뢰성에 대한 의문이 제기되고 있습니다. 여론조사가 민심과 여론을 들여다보기 위함임에도 지금 같은 방식은 전체 유권자를 대표하기보다 편향된 조사 결과로 1등 후보에게 민심이 쏠리는 밴드왜건(bandwagon) 효과를 유발 할 수 있기 때문입니다. 결국 잘못된 여론조사로 인해 대세론이 형성되고 그에 따라 표심이 움직인다면 여론조사가 자칫 민주주의를 훼손하는 지경에 이를 수 있습니다.

최근에는 SNS와 다양한 커뮤니티 등을 통해, 많은 국민께서 여론을 형성하고 정치적 표현과 활동을 하는 경우가 많습니다. 하지만 이 역시 모든 민심을 대변하는 것은 아닙니다. 때문에 정확한 데이터 분석과 검증을 통해, 미디어 트렌드에 나타난 데이터와 실제 유권자의 행동 양식의 관계성을 명확히 파악해야할 것입니다.

그런 의미에서 한국경영정보학회가 지방선거를 앞두고 개최하는 이번 세미나는 매우 뜻깊다 할 수 있습니다. 유권자의 다양한 감정과 행동 분석 데이터를 통해 전체 유권자의 민심을 분석하고 예측할 수 있는 시스템이 마련된다면, 기술을 이용해 민주주의 발전을 모색하는 일이라고 할 수 있을 것입니다.

「빅데이터와 여론조사」라는 주제로 진행되는 이번 세미나를 통해, 빅데이터 기술이 인간과 우리 민주주의 발전에 도움이 되는 기술로 활용될 수 있는 귀한 자리이길 소망합니다. 다시 한번 세미나를 준비하신 양희동 한국경영정보학회 회장님을 비롯한 학회 관계자 여러분 그리고 주제발표를 맡아주신 윤준태 바이브컴퍼니 AI빅데이터 연구 소장님께 감사와 격려의 말씀 드립니다.

오늘 세미나를 시작으로 우리나라 빅데이터 기술 분야 발전과 여론조사 분석 방법의 고도화를 통한 민주주의 발전의 지혜가 도출되길 기대합니다.
저도 국회 과학기술정보방송통신위원회 위원으로 빅데이터 기술 교류 세미나와 한국경영정보학회 활동을 응원하겠습니다. 감사합니다.

빅데이터 기술 교류 세미나

빅데이터와 여론조사

주제발표

비정형 빅데이터 플랫폼을 활용한 감성 분석 및 실험 사례

윤준태 (바이브컴퍼니 AI 빅데이터 연구소장)

비정형 빅데이터 플랫폼을 활용한 감성분석 및 실험 사례

바이브컴퍼니 인공지능&빅데이터연구소
윤준태

VAIV

바이브컴퍼니
AI & 빅데이터 플랫폼 소개

데이터에서 정보, 지식, 지혜를 발굴하는 (DIKW) 인공지능 빅데이터 분석 플랫폼

- SOFIA는 데이터의 수집에서 분석, 인사이트 발굴 그리고 의사결정에 이르는 비즈니스 전 과정을 도와주는 빅데이터 인공지능 플랫폼으로 58종의 원시데이터, 71종의 지식베이스, 60종의 세부 모듈로 이루어진 기반 기술과 7,500억 개의 키텀(키워드, 키프레이즈, 키팩트) 및 15조 개의 키텀간 구문 및 의미 연관관계를 포함

SOFIA

인공지능 기술

Explanatory Agent

Core NLP
- 대화 처리
- 자연어생성
- 의미 분석
- 구문 분석
- 어휘 분석

Deep AI
- 연관관계 학습
- Pre-training
- 데이터 분류
- 추론
- 클러스터링
- 시계열 분석
- 감성 분석
- 예측
- 요약
- MRC, QA

지식베이스
- 언어지식
- 실세계지식
- 감성지식

빅데이터 & 지식맵 기술

Visual Repository

트렌드맵
- Big Index
- Knowledge Inference

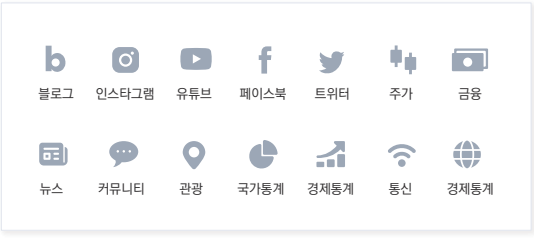
Big Data Processing
- Zookeeper
- Spark
- Kafka
- Impala
- HBase
- Solr
- Sqoop
- MapReduce
- YARN
- HDFS
- Airflow

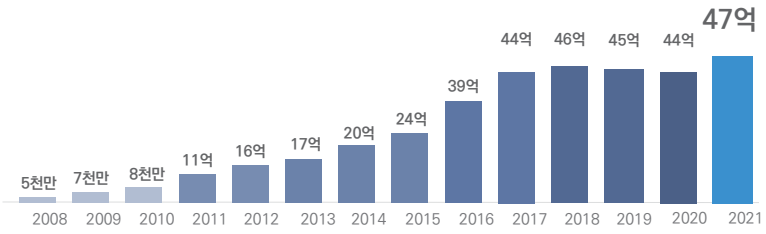
데이터
- 비정형 데이터
- 정형데이터
- Mining Spider

목적성, 최신성, 포괄성, 안정성을 담보한 국내 최다 비정형데이터 수집

- AI 시스템 구축 수요를 충족하기 위하여 목적에 부합하는 데이터를 최대한 포괄적으로 수집 (약 400억 건의 문서)
- 또한 신속한 수집을 바탕으로 최신성을 유지하되 끊김없는 데이터 수집이 가능하도록 안정성 담보

데이터 수집 현황





구분		수집대상	일일 수집	누적 수집
비정형	인스타그램	팔로워 상위 30만 계정	20만 ~ 30만 건	2014년 ~ 현재 (23억 건 이상)
	블로그	4,400만+ 계정	30만 ~ 40만 건	2008년 ~ 현재 (16억 건 이상)
	트위터	6,600만+ 계정	1,000만 ~ 1,500만 건	2010년 ~ 현재 (321억 건 이상)
	뉴스	국내 145+,해외 3만	3만 ~ 5만 건	2008년 ~ 현재 (1억 건 이상)
	유투브	구독자 1,000명 이상 114만+ 계정	2만 ~ 3만 건	2018년 ~ 현재 (2,300만 건 이상)
금융시장	주가/시장지수	3,090개 종목	전체	-
	채권금리/경제	22개/23종	전체	-
	환율 및 유가	36개/WTI(유가)	전체	-
기업재무	기업/섹터 재무재표	59종/48종	전체	-
	산업지수	247종 주요지표	전체	-

사람들의 생각과 행동, 감정까지 이해할 수 있는 NLP

- 단순 키워드뿐 아니라 사람들의 생각, 행동, 감정까지 분석할 수 있도록 함
- 형태소 분석, 구문 분석, 패러프레이징, 의미 분석 등 심층적 자연어처리

Core NLP

자연어처리는 중의성 해소 과정

형태소분석, 띄어쓰기, 개체명, 구문청킹, 구문구조, 의미, 화자 의도 등 전 과정에 걸쳐 발생하는 중의성 해소를 통해 정확한 의미 이해

1. 형태소 분석
   대학생선교회 대학생+선교회 or 대학+생선+교회
   에어팟끼고소리크게크게→ 에어팟(명)+끼(동)+고(어)+소리(명)+크(형)+게(어)+크(형)+게(어)
2. 품사태깅 및 후처리
   산다→ '사다' or '살다' - 전체 의미에 중요
3. 구문 청킹

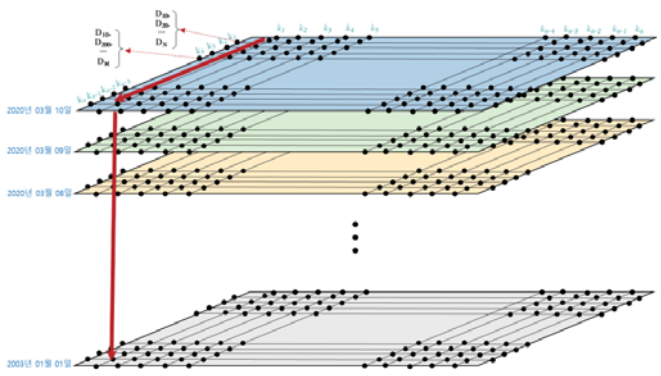
왼쪽 문맥	단어	오른쪽 문맥
기름에 묻힌 솜을 썼다.	[여름	밤]에 물고기들이 몰가로 …
그만 잠이 들었다.	[여름	밤]안긍이나 잤었지만 …
나는 처음에는	[여름	방학 후]에 복학을 하기로 …
사귀게 되었다.	[여름	방학 때]에는 부산에 …
교수들에게 묻고싶다. [지난	여름]	나는 10년의 유학 …
생각한다. [그 해	여름]	마지막 눈더위가 …
않는군요. [이번	여름]	전투는 누가 이겼죠?

4. 구문 분석
   디자인은 맘에 드는데 카메라가 영 좋지 않네요
5. 의미 분석
   이보다 더 만족스러울 수 없다→ 만족스럽다, 긍정
   만족스러울 수 없다→ 불만족스럽다, 부정

소셜 빅데이터로부터 사회 전체의 스토리와 히스토리를 지식화

- 사람들의 생각과 행위는 언어라는 도구에 의해 미묘하게 다른 의미를 가진 표현들로 표출되며 소셜데이터는 이러한 사람들의 생각과 행위가 고스란히 드러나있음
- 트랜드맵은 소셜데이터로부터 사람들의 삶과 역사를 지식으로 저장하고 사회가 어떻게 변화하고 있는지 측정할 수 있도록 하는 거대한 지식맵.

트렌드맵



기술 특징

1. 개개인이 발현한 하나하나의 감성과 행위를 인식하기 위해서는 단순히 키워드(keyword)를 추출하는 것을 넘어 키프레이즈(key phrase), 키팩트(key fact)를 추출
   - 프레이즈나 팩트를 추출하기 위해 명사구, 서술어를 포함한 표현들까지 추출
   - 단순한 형태소분석을 넘어 구문구조의 분석을 하고 많은 표현을 정규화
   - 하나의 키텀(key word, phrase, fact)은 그 키텀이 가지고 있는 표층 형태만으로 파악되는 것이 아니라 다른 많은 연관된 어휘들에 의해 그 의미가 명확해짐
   - 예를들어, 스마트폰이라는 키워드는 스마트폰을 만드는 제조사, 다른 제품, 다양한 감성들과 연관되어 있으며 이들 속에서 스마트폰과 관련된 의미를 찾을 수 있음
   - 키텀이나 키팩트의 시간에 따른 관심도 변화 혹은 연관관계의 변화는 특정 브랜드, 제품에 대한 소비자의 인식의 변화를 추적하는 데 매우 중요한 요소
2. 대상을 효율적으로 관찰하기 위해서는 관련 있는 의미 그룹별로 살펴볼 필요가 있기 때문에 300만 개 이상의 항목으로 구성된 온톨로지를 활용

트렌드맵 (1/2)

연관어 분석

막걸리

No	연관어	빈도
1	술	1934
2	김치	1020
3	파전	1010
4	소주	898
5	맥주	879
6	고기	865
7	밥	855
8	고추	360

맛집

No	연관어	No	연관어
1	나운터횟집	1	앤치즈
2	칠십리	2	레드락
3	쿠우쿠우	3	다운타우너
4	돈부리	4	나리의집
5	마라향	5	자니덤플링
6	돼지집	6	타이펑
7	청년다방	7	라페름
8	반반족발	8	바토스

온톨로지 기반의 연관검색
막걸리의 연관어 중 음식 분류 결과

문맥의 분여
맛집의 연관어 중 음식점 (좌)과 이태원이 포함된 문서 중 맛집의 연관어(우)

제주여행

No	연관어
1	승마체험
2	카트체험
3	감귤체험
4	제주도자기체험
5	해녀체험
6	족욕체험
7	먹이주기체험
8	다도체험
9	가죽공예체험

핵심어 패턴기반 연관관계
제주여행에서 가장 많이 하는 체험

마트 가다

No	연관어
1	장 보러
2	물 사러
3	아이스크림 사러

의미 추론
마트에 가는 이유는 무엇인가에 대한 답

순위	먹고 싶다	먹다
3	삼겹살	술
	–	–
7	디저트	치즈
9	곱창	과일

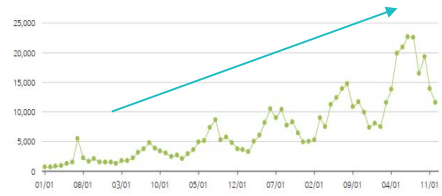
어휘의 미묘한 의미까지 구분

스타벅스

사다	텀블러
	파우치
	럭키백
	원두
	머그컵
좋다	원두
	텀블러
	퀄리티
	향

감성 및 표현 분석
스타벅스에서 사는 것과 좋다고 이야기하는 것

"캠핑" 관련 월별 버즈 추이



감성 문서 검색

'예쁘게 플레이팅 되다' 라는
감성을 가진 문서를 검색한 예시

VAIV

트렌드맵 (2/2)

커피와 케이크의 연관어 변화

2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
치즈케이크	치즈케이크	치즈케이크	치즈케이크	치즈케이크	치즈케이크	치즈케이크	치즈케이크	치즈케이크	치즈케이크
컵케이크	컵케이크	팬케이크	팬케이크	팬케이크	팬케이크	팬케이크	팬케이크	당근케이크	파운드케이크
팬케이크	팬케이크	롤케이크	롤케이크	당근케이크	당근케이크	당근케이크	당근케이크	팬케이크	당근케이크
조각케이크	조각케이크	조각케이크	조각케이크	조각케이크	초코케이크	파운드케이크	파운드케이크	파운드케이크	팬케이크
롤케이크	롤케이크	당근케이크	당근케이크	초코케이크	조각케이크	초코케이크	조각케이크	조각케이크	조각케이크
파운드케이크	당근케이크	컵케이크	초코케이크	롤케이크	파운드케이크	조각케이크	초코케이크	초코케이크	초코케이크
초콜릿케이크	파운드케이크	파운드케이크	컵케이크	컵케이크	딸기케이크	딸기케이크	딸기케이크	딸기케이크	딸기케이크
생크림케이크	핫케이크	생크림케이크	생크림케이크	생크림케이크	롤케이크	생크림케이크	레몬케이크	생크림케이크	바스크치즈케이크
핫케이크	생크림케이크	초콜릿케이크	파운드케이크	파운드케이크	생크림케이크	롤케이크	생크림케이크	레몬케이크	생크림케이크
커피케이크	맛있는케이크	핫케이크	핫케이크	티라미수케이크	컵케이크	컵케이크	롤케이크	롤케이크	롤케이크
당근케이크	초콜릿케이크	맛있는케이크	맛있는케이크	딸기케이크	핫케이크	수플레팬케이크	컵케이크	바스크치즈케이크	홀케이크
크리스마스케이크	커피케이크	블루베리치즈케이크	티라미수케이크	핫케이크	티라미수케이크	레드벨벳케이크	수플레팬케이크	홀케이크	레몬케이크
모카케이크	블루베리치즈케이크	달달한 케이크	달달한 케이크	맛있는케이크	맛있는케이크	핫케이크	맛있는케이크	얼그레이케이크	얼그레이케이크
맛있는케이크	티라미수케이크	커피케이크	초콜릿케이크	달달한 케이크	레드벨벳케이크	떡케이크	핫케이크	수플레팬케이크	수제케이크
생일케이크	달달한 케이크	달콤한 케이크	무지개케이크	초콜릿케이크	초콜릿케이크	티라미수케이크	얼그레이케이크	수제케이크	레터링케이크
고구마케이크	크리스마스케이크	크리스마스케이크	고구마케이크	레드벨벳케이크	수제케이크	맛있는케이크	티라미수케이크	컵케이크	레몬파운드케이크
떡케이크	종류 맛있는케이크	수제케이크	딸기케이크	무지개케이크	달달한 케이크	레몬케이크	레드벨벳케이크	티라미수케이크	딸기생크림케이크
라이스케이크	달콤한 케이크	딸기생크림케이크	블루베리치즈케이크	수제케이크	홀케이크	수제케이크	고구마케이크	핫케이크	티라미수케이크
티라미수케이크	모카케이크	티라미수케이크	크레이프케이크	딸기생크림케이크	딸기생크림케이크	생일케이크	홀케이크	레터링케이크	컵케이크
무스케이크	고구마케이크	딸기케이크	딸기생크림케이크	크레이프케이크	얼그레이케이크	홀케이크	딸기생크림케이크	맛있는케이크	맛있는케이크
스펀지케이크	크림치즈케이크	고구마케이크	달콤한 케이크	다른 케이크	크레이프케이크	얼그레이케이크	생일케이크	딸기생크림케이크	송도케이크
달콤한 케이크	민낯 연주가생일케이크	크림치즈케이크	커피케이크	크리스마스케이크	다른 케이크	초콜릿케이크	단호박케이크	레드벨벳케이크	송도레터링케이크
티라미슈케이크	떡케이크	다른 케이크	레드벨벳케이크	홀케이크	크리스마스케이크	달달한 케이크	레터링케이크	단호박케이크	인천레터링케이크
수플레치즈케이크	스펀지케이크	미니케이크	다른 케이크	고구마케이크	쉬폰케이크	딸기생크림케이크	김해레터링케이크	레몬파운드케이크	인천송도케이크
커피크림롤케이크	쉬폰케이크	스트로베리치즈케이크	크림치즈케이크	커피케이크	무스케이크	다른 케이크	달달한 케이크	고구마케이크	송도신도시케이크
호두롤케이크	딸기생크림	레드벨벳케이크	다양한 케이크	달콤한 케이크	레몬케이크	고구마케이크	수제케이크	컵케이크	빅토리아케이크
호두커피크림롤케이크	무스케이크	떡케이크	무스케이크	블루베리치즈케이크	떡케이크	무스케이크	수플레케이크	인천송도케이크	수플레팬케이크
달달한 케이크	빕풍케이크	무스케이크	수제케이크	떡케이크	떡케이크	쉬폰케이크	크림치즈케이크	송도레터링케이크	단호박케이크
미니케이크	미니케이크	미니케이크	미니케이크	미니케이크	미니케이크	미니케이크	초콜릿케이크	미니케이크	미니케이크
귓진 케이크	떡케이크	떡케이크	떡케이크	떡케이크	떡케이크	떡케이크	마카롱케이크	떡케이크	떡케이크

변화가 없다면 특정 기간의 스냅 사진으로
충분하지만 사회는 늘 변화하고 있음

10년간 커피의 연관어로 당근케이크는 늘고 있고
초콜릿 케이크는 줄어들고 있음

이러한 초 대용량의 지식 탐색이
실시간에 수행 가능함

VAIV

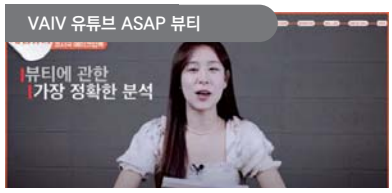
활용 사례

트렌드 분석 및 전달

썸트렌드 – 서비스

트렌드노트 – 책

트렌드 전달 – 미디어



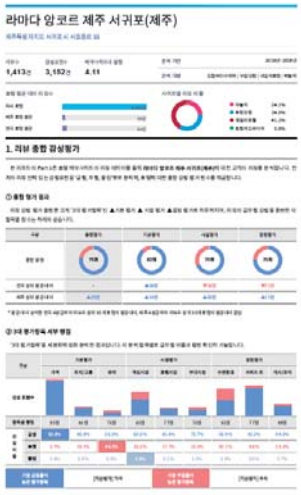
활용 사례 (계속)

AI Report (자동 보고서)

주식 분석 보고서

국내 관광 분석 보고서

호텔 분석 보고서

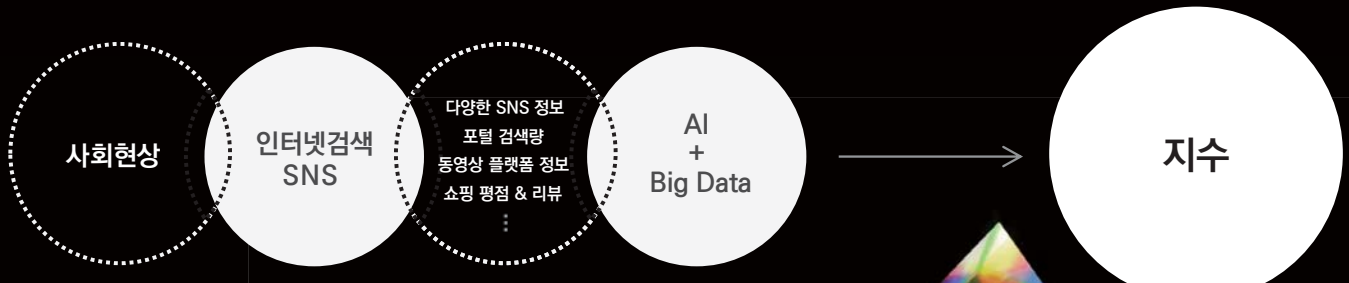


활용 예시 - 지수

비정형데이터를 기반으로 사회현상 해석

현상의 지수화

- 사회라는 현상은 인터넷을 통해 데이터로 표현
- 바이브는 데이터를 AI와 Big Data 기술을 이용해 하나의 지수로 인코딩
  - ✓ 소셜은 사람들의 생각이 자유롭게 표현되는 공간
  - ✓ 지수를 통해 사회현상을 모델링하고 해석할 수 있도록 합니다.



VAIV's Prism

기술적 고려사항

데이터 수집

- 목적에 맞는 데이터를 **(목적성)**
- 현상을 반영할 충분한 정도의 양으로 **(포괄성)**
- 시의 적절하게 반영해야 함 **(실시간성)**

데이터 가공

- 비정형 데이터는 수많은 스팸과 오류 포함
- 정형 데이터 역시 잡음과 이상치가 많음
- 데이터 클린징

모델링

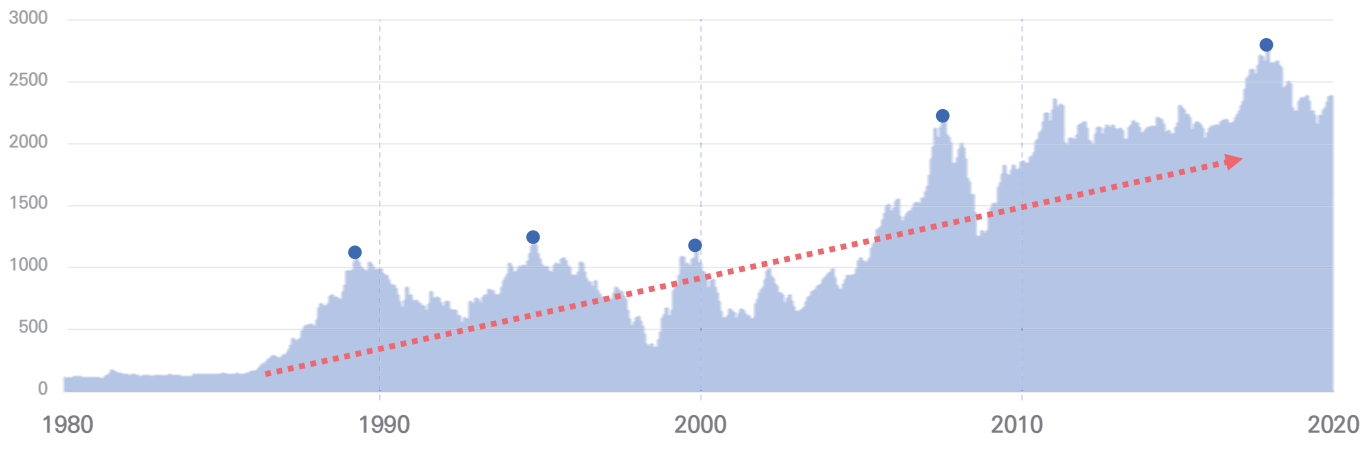
- 분야별 경험 많은 전문가
- 적절한 모델링 기법 선택
- 반복적 실험

분석

- 언어 중의성
  정확한 지수를 위해서는 언어의 중의성 제거 필요
- 현상을 정확히 반영할 수 지식베이스 도출
  금융, 정치 등 도메인 특성 반영
- 초대용량 (일 1,000만 건 이상) 문서 분석
- 분석된 정보의 신속한 제공

EWS (Early Warning System)

- 금융 시장은 유동성, 신용, 수익률 등 위험이 존재하며 이러한 위험을 사전에 감지하고 이에 대응하고자 함
- 금융 조기 경보 시스템 (EWS) - 위험 분석, 모니터링 및 경고, 전파 및 통신, 대응 능력을 지원
- 금융 시장에 참여하고 있는 사람들의 감성이 향후 리스크를 예측하는 데 도움이 될 것이라고 가정
  - 구글 학술검색에서 'sentment analysis in finance'로 검색하면 약 548,000개 결과 나옴 (2022. 5)
  - 뉴스 또는 소셜미디어로부터 감성 분석 시도
  - 하지만 '좋다', '나쁘다' 등 일반적인 감성키워드는 금융시장의 과열이나 패닉 같은 금융 감성을 정확히 반영하기 어려움



금융 감성 지식

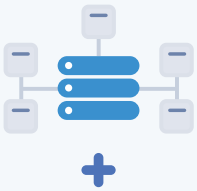
- 금융감성분석이란, 현재 뉴스와 소셜미디어등에서 발견되는 금융에 관련된 감성 정보에 대한 분석을 말함
- 이를 위해서 금융 관련 감성을 14개의 카테고리로 분류하고 해당 분류에 상태, 감성, 긍부정 감성어 매핑하여 지식 사전을 구축

지식사전 구축 프로세스

금융 감성어 분류 체계 및 사전 구축

정치 감성어를 추출하기 위한 분류 체계 및 사전 구축

금융 감성어 분류 체계



키워드별 어휘사전 구축

경기불안	경제요인 – 경기
유가급등	금융시장 – 원자재
긍정적 실적	기업요인 – 사업내용

지식사전 구축 예시

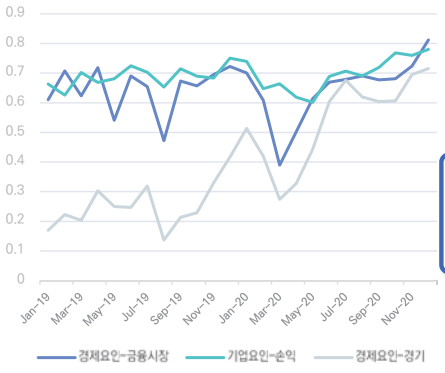
대대분류	대분류	중분류(소분류)
경제요인	경기	경기전반, 고용/임금/실업,국제수지, 물가, 부동산시장, 소비, 투자, 미분류
	금융시장	국내금융기관, 금융시장상태, 대출, 보험, 예금/적금, 외환시장, 원자재 및 상품시장, 주식시장, 증권지수, 채권시장, 파생상품시장, 펀드/신탁, 미분류
	정책	금융정책, 재정정책, 정책기조, 통상정책, 미분류
기업요인	경영내용	계열화, 영업규모, 인사, 인수합병, 미분류
	사업내용	가격, 배당, 비용, 상장, 생산, 수요, 수주, 수출, 실적, 업황, 전략, 출시, 판매, 미분류
	손익	매출, 손실, 이익, 미분류
	재무상태	부채, 자본, 자산, 미분류
	현금흐름	현금흐름
	미분류	미분류
기타	기타	가격/주가, 가치평가/컨센서스, 이벤트명, 미분류
해외요인	경기	경기전반, 고용/임금/실업, 국제수지, 물가, 부동산시장
	금융시장	금융시장상태, 외환시장, 원자재 및 상품시장, 주식시장, 증권지수, 채권시장, 파생상품시장, 펀드/신탁, 미분류
	사업내용	업황
	정책	금융정책, 재정정책, 정책기조, 통상정책, 미분류
	미분류	미분류

빅데이터 분석과 결합한 투자모델(MP) 리스크 관리

- 바이브컴퍼니의 감성분석을 통한 시장 감성와 매크로 분석이 결합된 주식시장 조기경보시스템(EWS) 구축
- 삼성/한화자산운용 → 기관고객 대상 투자자문 • 운용, 리스크 관리 등에 사용

분류 별 감성 지수

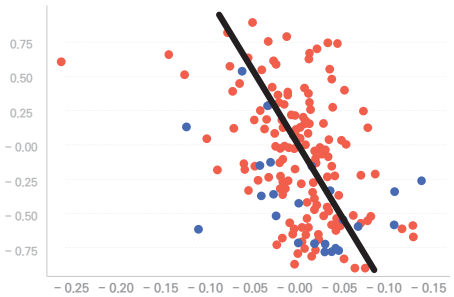
- 2019년 ~ 2020년 대상분류별 감성추이(긍정률) 변화



금융시장 데이터(정형 데이터) 결합
1Month Later 하락가능성 예측

EWS 지수

- EWS지수와 코스피 지수의 수익률 모델링
- 월별 EWS 지수 별 코스피 지수 수익률 측정결과(아래)



X축 : 코스피 지수 월별 수익률
y축 : EWS 지수, (RED : Train , BLUE : Test)

빅데이터 분석과 결합한 투자모델(MP) 리스크 관리

- 바이브컴퍼니의 감성분석을 통한 시장 감성와 매크로 분석이 결합된 주식시장 조기경보시스템(EWS) 구축
- 삼성/한화자산운용 → 기관고객 대상 투자자문 • 운용, 리스크 관리 등에 사용

분류 별 감성 지수

- 2019년 ~ 2020년 대상분류별 감성추이(긍정률) 변화

금융시장 데이터(정형 데이터) 결합
1Month Later 하락가능성 예측

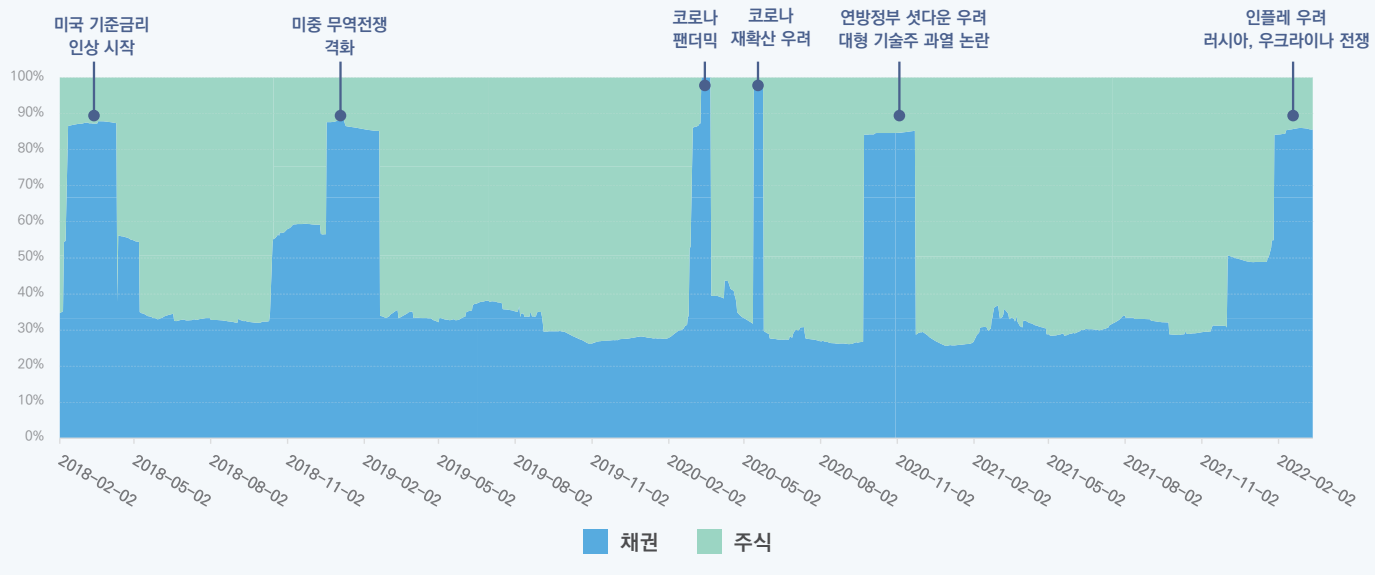
감성 지수 기반 시장 예측

- 감성지수를 활용한 시장 수익률 예측

X축 : 종합 감성 지표 (RED : Train , BLUE : Test)
y축 : 코스피 지수 월별 수익률

안정적인 자산배분 모델(MP)

- 시장하락/변동성 확대 위험을 선제적으로 탐지해 탄력적으로 주식/채권 비중 조정 (시장이 위험할수록 채권 비중 상승)
- 변동성 제어를 위한 다양한 퀀트전략을 혼합(멀티 전략) → 최대손실율을 낮추고 샤프지수를 높이는데 중점



모델 포트폴리오 실제 운용 성과

- 로보 어드바이저 테스트 베드에 실계좌 등록 • 운용 (14회차, 2021.5.3일 등록~)
- 인플레 우려와 우크라이나 사태로 변동성이 확대된 시장에서 안정적인 성과를 시현

구분	누적수익률(%)	연환산(%)	Sharpe	최대손실율(1M,%)
퀀팃 글로벌 자산배분 국내EMP_적극투자	4.33	4.65	0.99	-3.41
SNP500	7.02	7.56	0.58	-9.73
코스피	-13.65	-14.58	-0.99	-12.65

SNP500
퀀팃 SAIV-ROBO 글로벌 자산배분 - 해외
코스피

※ 국내 상장 해외 ETF로 운용되는 로보 MP 비교
※ 샤프지수 : 일간수익률 기준, 무위험 이자율은 편의상 미고려

(2021.5.3~2022.4.8일, 11개월)

정치 감성 지수

- 정치 감성 지수이란, 현 시대 정치인들에게 요구되는 다양한 감성들이 소셜 상에서 얼마나 부합하고 있는가를 측정하는 지수
- 특히 SNS가 발달하면서 SNS 상에서 이야기되는 정치인에 대한 감성이 여론과 상당히 관련이 있을 것으로 예상
- 하지만 대체적으로 SNS에서는 정치성향에 따라 언급량에 차이가 있어 단순 빈도로는 정확하 사회현상을 반영하기 어려움

윤석열

휴가 중 반려견과 찍은 사진을 **인스타그램**을 통해 공개한 윤석열 전 검찰 총장

홍준표

홍준표 의원의 정책 비전을 담은 ‘ jp 희망편지 ’가 연재되는 **페이스북** 계정

이재명

초등학교 동창을 만난 개인적인 이야기를 **페이스북**에 공개한 이재명 경기도지사

박용진

공약 메시지를 30초 내 영상으로 **유튜브** 방송에서 설명하는 박용진 의원

출처) https://www.donga.com/news/article/all/20210813/108547749/1

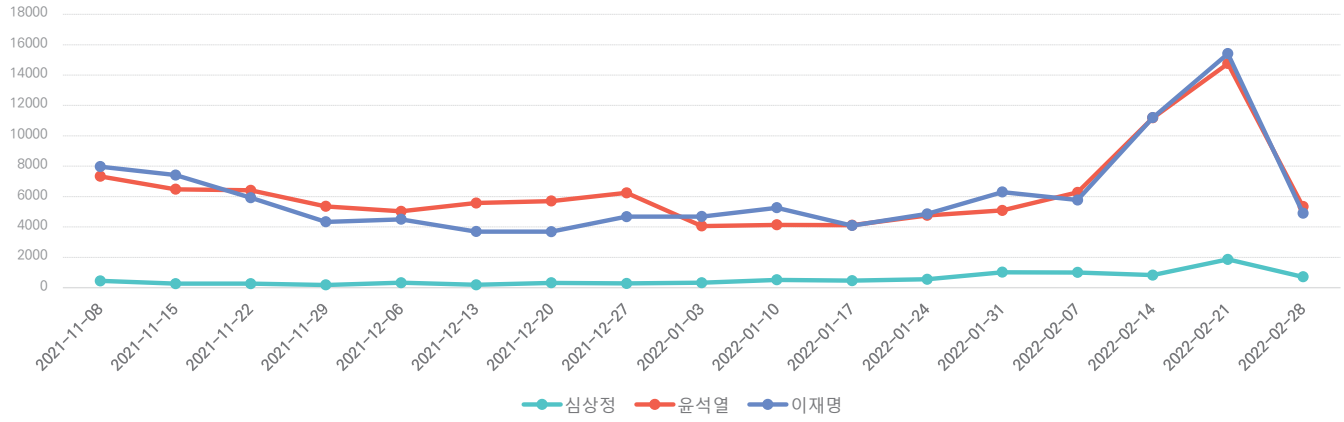
정치 감성 지수

- 소셜 상에서의 정치인의 언급량은 해당 시점의 정치인에 대한 관심 정도를 나타내지만, 그 정치인의 선호 정도를 파악할 수는 없음

후보별 SNS 버즈량 추이 (2021.11.05 ~ 2022.03.06)

분석 후보 : 대선 후보 (윤석열, 이재명, 심상정)

분석 채널 : 뉴스, 블로그, 커뮤니티, 트위터, 유튜브, 인스타그램



VAIV

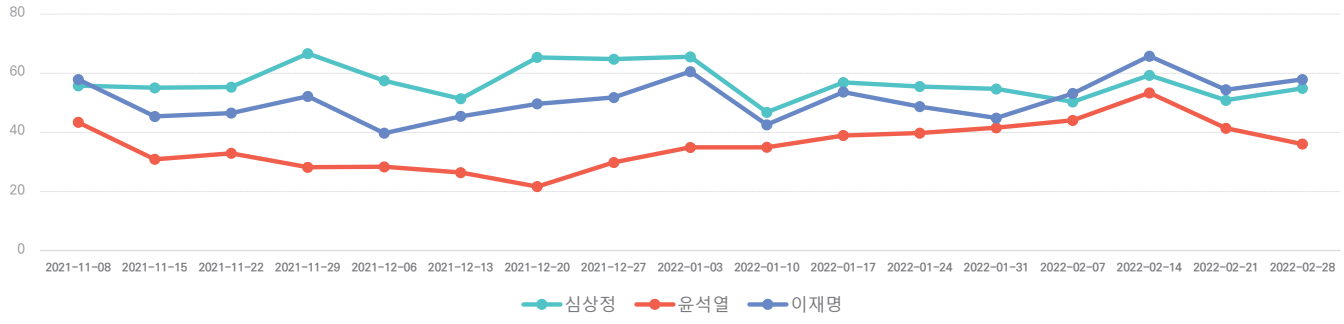
정치 감성 지수

- 소셜 상의 언급량의 절대적인 양을 고려하지 않은 상대적인 긍정, 부정 비율만으로 평가하게 되면, 해당 정치인의 긍정비율이 높다고 하여 다른 정치인 보다 더 긍정적으로 판단 할 수 없음

후보별 긍정 비율 추이 (2021.11.05 ~ 2022.03.06)

분석 후보 : 대선 후보 (윤석열, 이재명, 심상정)

분석 채널 : 뉴스, 블로그, 커뮤니티, 트위터, 유튜브, 인스타그램



앞선 장표에서 낮은 언급량을 가진 심상정 후보는 긍정 비율로 볼때 높은 스코어를 가지고 있다. 이는 긍정비율 ($\frac{\text{긍정}}{\text{긍정+부정}}$)로 계산될 때, 언급량의 절대적인값이 무시된 상대적인 값으로 계산되어지게 되므로 긍정비율 높다고하여 다른 후보보다 더 긍정으로 볼수는 없다. 즉, 긍부정 비율은 온라인 상에서의 긍부정이 여론과 일치하지 않는 경향을 보임

VAIV

정치 감성 분석 기술

- 현 시대적 관점에서 정치인들에게 요구되는 감성만을 고려하기로 하고 이를 정치감성이라 정의
- 정치 감성에 대해 9개 카테고리로 분류하고 각 분류에 해당되는 긍정, 부정 감성어를 매핑하여 지식 사전을 구축

▪ 정치 감성 지식 구축

인물 키워드 사전 구축 및 검증

인물 분석을 위한 필터링 키워드 사전 구축 및 분석 데이터 검증

분석지식 정의	분석 인물 대상 정의	
사전 구축	주제어	분석 대상 키워드셋
	포함어	분석 대상 표현 키워드셋
	제외어	데이터 정확도를 위한 스팸 처리 키워드셋

정치 감성어 분류 체계 및 사전 구축

정치 관련 감성을 9개의 카테고리로 분류하고 해당 분류에 긍부정 감성어 매핑하여 지식 사전을 구축

대분류	중분류	감성어 개수
인지도	지지율,인지도, 인기, 당선, 유력, 화제 …	472
능력/업적	능력, 업무능력, 정책, 전문가, 비전	430
정책평가	정책	76
신뢰도	신뢰, 믿음, 책임감, 공약	286
공정/청렴	논란, 청렴, 정의감, 공정, 비리, 표절	275
이미지	성격, 겸손, 화, 호감, 비호감, 인성	61
소신/뚝심	소신, 고집, 뚝심	517
업무관련성향평가	성실, 게으름, 꼼꼼함, 성급함	203
포용/공감	포용, 소통, 공감	41

긍정 도미넌스

- 정치인의 소셜 상 절대적인 언급량과 상대적인 긍정, 부정 비율을 고려하여 특정 시점에서 다른 정치인들과의 상대적인 선호 정도를 평가하기 위한 지수를 개발

긍정 도미넌스

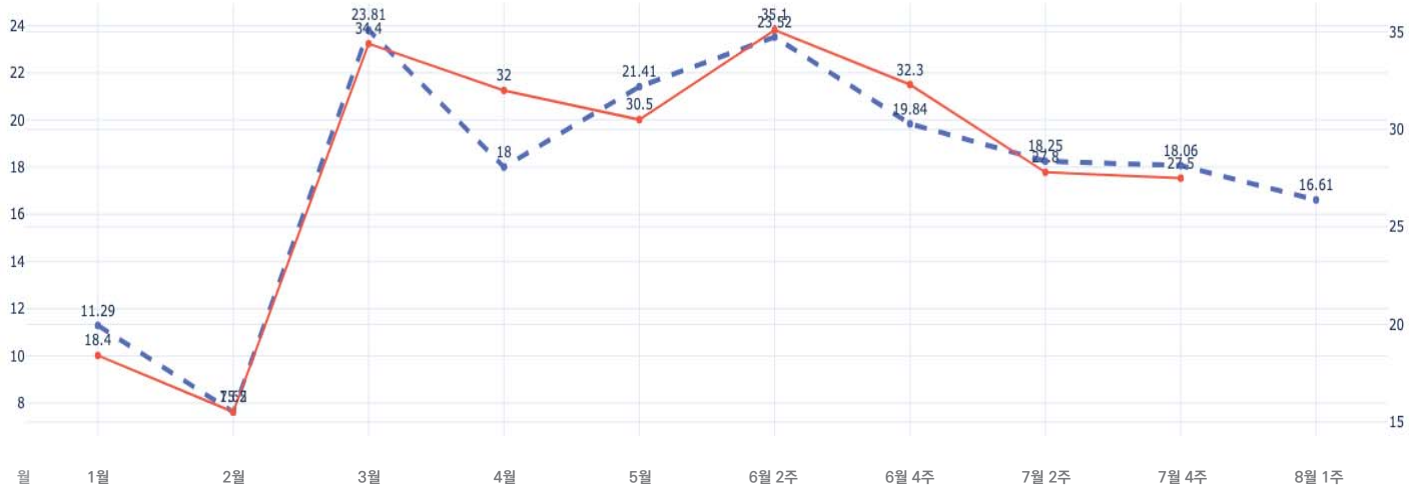
- 전체 정치 감성 긍정 버즈량 중 각 정치인의 긍정 버즈량이 차지하는 비율 의미
- 긍정 감성은 정치 감성 분류를 활용
- 일별로는 변화가 너무 크므로 변동성을 완화하기 위해 주별로 측정

$$\text{긍정도미넌스} = \frac{\text{특정 정치인의 긍정 버즈량}}{\text{해당 주차 모든 정치인의 긍정 버즈량}} * 100$$

- 긍정 도미넌스와 여론조사 대선 후보 선호도의 상관관계 분석 결과
- – 상관계수가 0.9 이상으로 강한 선형관계 및 유의확률 0.05 이하로 통계적으로 유의미함을 보임

후보별 선호도 조사와 긍정 도미넌스 추이

- 분석 후보 : 윤석열
- 분석 기간 : 2021년 1월 ~ 2021년 8월



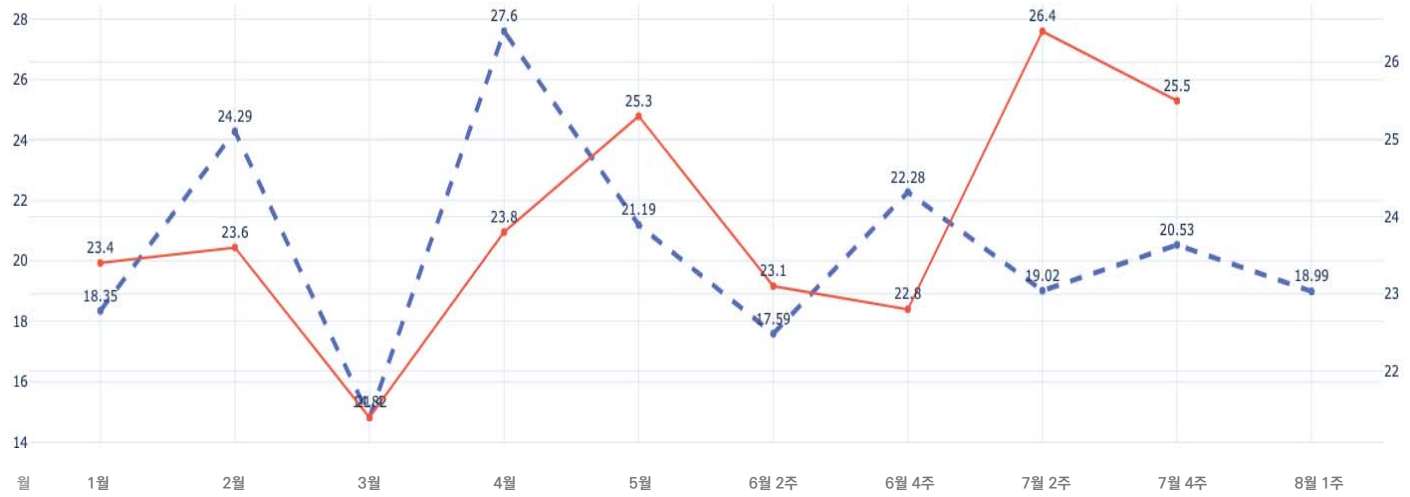
긍정 도미넌스 — 긍정 도미넌스 — 리얼리티 선호도조사 — 선호도 조사

VAIV

04. Score 산출

후보별 선호도 조사와 긍정 도미넌스 추이 (계속)

- 분석 후보 : 이재명
- 분석 기간 : 2021년 1월 ~ 2021년 8월

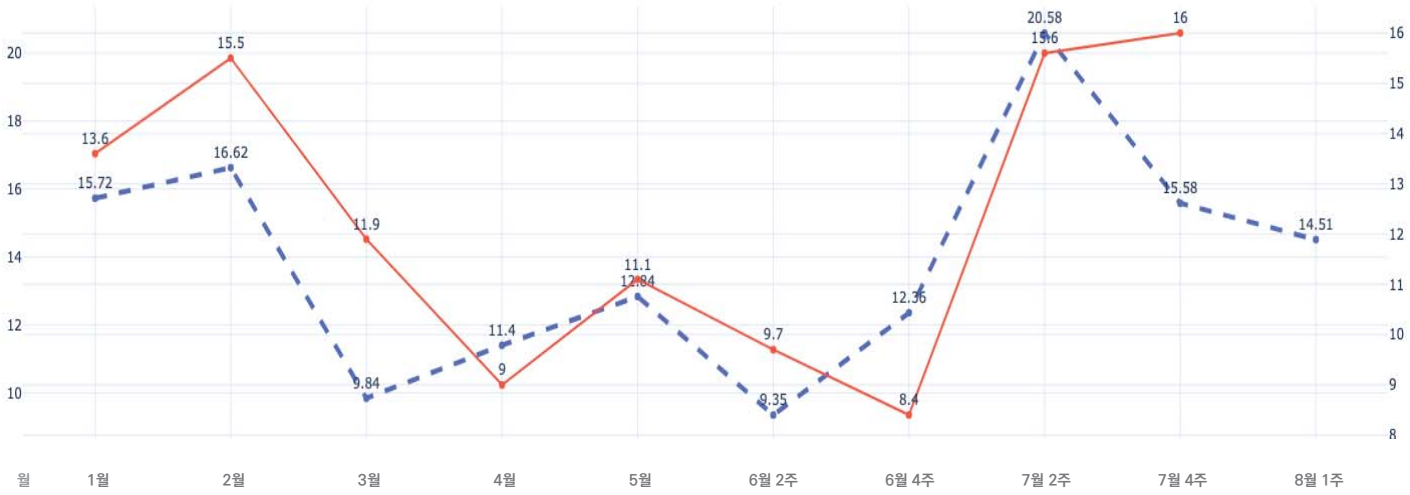


긍정 도미넌스 — 긍정 도미넌스 — 리얼리티 선호도조사 — 선호도 조사

VAIV

후보별 선호도 조사와 긍정 도미넌스 추이 (계속)

- 분석 후보 : 이낙연
- 분석 기간 : 2021년 1월 ~ 2021년 8월



긍정 도미넌스 — 긍정 도미넌스 — 리얼리티 선호도조사 — 선호도 조사

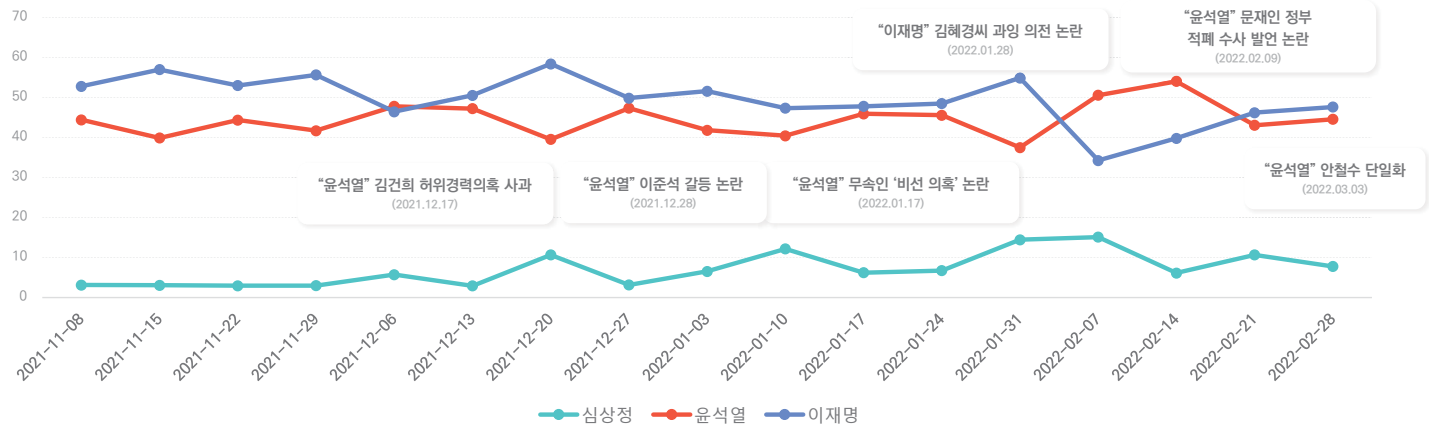
긍정 도미넌스 추이

- 긍정 도미넌스 지수가 소셜 상의 여론을 모두 반영하지는 못하나 특정 시점에서 정치인들의 긍정, 부정 정도에 대한 추이에 대해 어느 정도 유의미함을 보임

"대선" 후보별 긍정 도미넌스 추이 (2021.11.05 ~ 2022.03.06)

분석 후보 : 대선 후보 (윤석열, 이재명, 심상정)

분석 채널 : 뉴스, 블로그, 커뮤니티, 트위터, 유튜브, 인스타그램



기타 소셜 기반 지수

방송 화제성 지수

서울시민 라이프스타일 지수

맛집 랭킹

VISS 지수

화장품 브랜드 디지털 지표

주식 종목 점수화

VAIV

Thank you

VAIV Company inc.
97, Dokseodang-ro, Yongsan-gu, Seoul, 04419, Republic of Korea
Tel 02-565-0531 / Fax 02-565-0532 / www.vaiv.kr

빅데이터 기술 교류 세미나

빅데이터와 여론조사

발표1

MZ 세대의 마음을 읽으려면, “커뮤니티”를 공략하라!
6월 지방선거 결과 예측을 위한 분석연구

양희동(이화여대 교수), 최한별(KAIST 연구교수),
김수림(KAIST 박사과정 연구원)

<6월 지방선거 결과 예측을 위한 분석연구>

MZ 세대의 마음을 읽으려면, “커뮤니티”를 공략하라

이화여대 양희동 교수
카이스트 최한별 연구교수
카이스트 김수림 박사과정

2022.05.19

목차

연구질문 <1>

왜 MZ세대 여론은 커뮤니티에서 보아야 하는가?

MZ세대와 뉴스 댓글

네이버 데이터랩 통계에 따르면 3.9일 대선 전 일주일간 정치 뉴스 댓글에 참여한 연령대는 40대와 50대가 주요하였음

연령별 N포털 정치 뉴스 댓글 수 통계

		2/28 (월)	3/1 (화)	3/2 (수)	3/3 (목)	3/4 (금)	3/5 (토)	3/6 (일)	3/7 (월)	3/8 (화)
10대	남	522	513	507	564	487	538	623	615	611
10대	여	86	82	87	123	101	124	117	161	117
20대	남	4,872	4,566	4,774	6,440	5,242	5,430	5,928	6,594	6,147
20대	여	1,013	1,013	1,446	1,911	2,109	2,191	2,253	4,349	2,949
30대	남	16,920	16,913	16,546	23,856	19,136	18,927	19,929	22,091	20,865
30대	여	4,768	4,731	4,688	7,487	6,051	6,468	7,138	7,512	7,035
40대	남	31,077	32,059	31,271	48,023	37,403	36,543	35,828	40,232	36,352
40대	여	13,161	12,939	12,854	20,324	15,367	15,995	16,952	17,039	15,519
50대	남	32,417	32,640	32,181	45,509	36,128	35,329	33,916	37,290	34,627
50대	여	11,887	11,814	12,358	16,923	12,997	13,562	13,688	14,373	12,996
60대	남	19,293	18,824	19,108	23,014	19,488	18,949	18,885	20,009	18,852
60대	여	5,110	4,875	5,193	6,177	5,034	5,184	5,409	5,452	5,056
총합		700,587	643,457	608,560	767,527	651,654	640,686	651,581	812,705	703,410

10, 20, 30대의 댓글 수 총합에 비하여 40, 50대 댓글 수가 더 많음

MZ세대의 주요 소통 플랫폼

대학내일연구소의 '21년 5월 설문조사에 따르면, MZ세대의 70% 이상이 의견을 공유하기 위해 뉴스 댓글이 아닌 온라인 커뮤니티를 이용하는 것으로 나타남

MZ세대가 주로 이용하는 온라인 커뮤니티 플랫폼 설문 결과

<복수응답, 단위 %>

구분	전체	연령대			
		10대	20대	30대	40대
포털 기반 카페 커뮤니티* (예: 다음, 네이버)	49.8%	**25.8%**	**44.7%**	**58.3%**	74.2%
자체 웹사이트 커뮤니티* (예: 디시인사이드, 뽐뿌 등)	24.3%	**8.1%**	**29.6%**	**22.4%**	22.6%
카카오톡 오픈채팅방	44.9%	62.9%	38.5%	49.8%	29.0%
페이스북 그룹	16.5%	37.1%	16.8%	10.8%	19.4%
에브리타임 (대학교 커뮤니티 및 시간표 서비스)	22.4%	4.8%	44.3%	4.6%	0
블라인드(Blind) (직장인 커뮤니티)	9.2%	4.8%	6.5%	14.3%	0
대학교 학내 커뮤니티 웹사이트	8.9%	3.2%	14.8%	4.6%	0

오픈 커뮤니티 기반 서비스 내에서는 카페 & 자체 웹사이트 커뮤니티에서 MZ세대 이용률이 가장 높음

* 현재 바이브컴퍼니가 썸트랜드(Sometrend) 플랫폼을 통해 데이터 수집 및 통계 서비스 제공 중임

썸트렌드 제공 데이터

썸트렌드에서는 인스타그램, 트위터 등의 소셜미디어 플랫폼 데이터와 국내 6,850개 커뮤니티 게시판의 데이터를 수집 및 분석하여 결과를 제공함

썸트렌드 제공 및 분석 데이터 요약

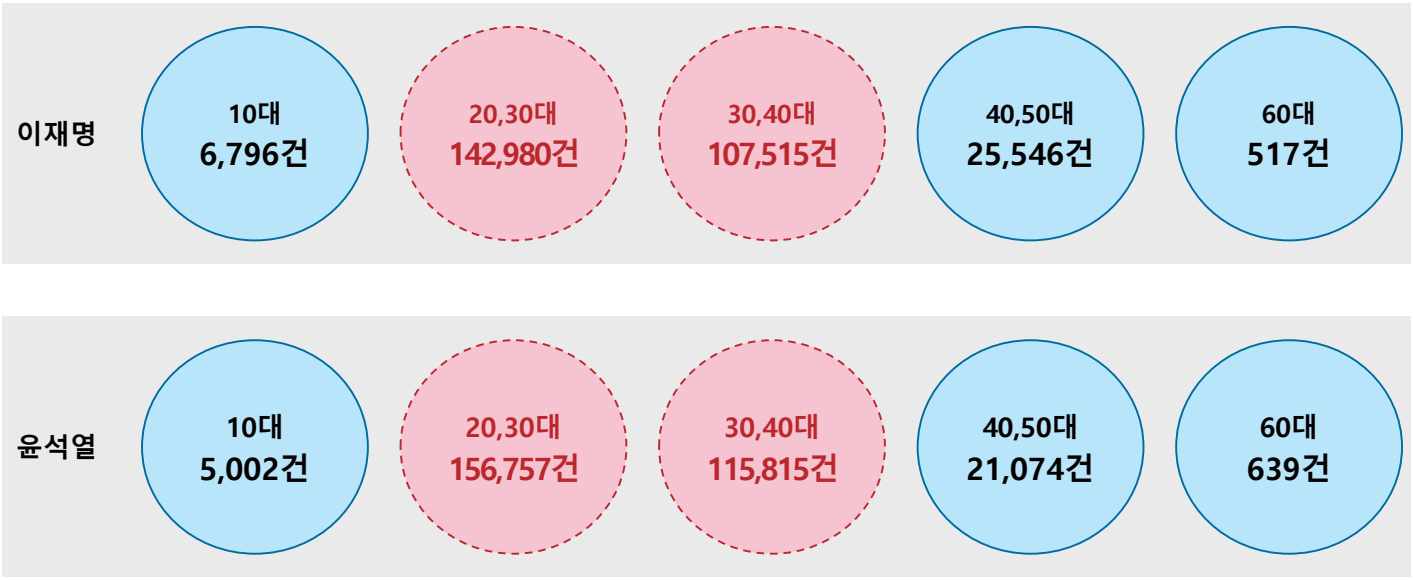
본 연구의 사용 데이터

데이터 수집 플랫폼	제공하는 분석 결과
인스타그램	키워드 언급량
트위터 (RT 제외)	키워드 감성지수 (긍정, 부정, 중립)
블로그	키워드 연관어
커뮤니티(게시판) (D사, N사 포털 카페 및 자체 웹사이트 커뮤니티들의 게시판 6,850개 대상)	키워드 연관 감성지수 단어 (긍정, 부정, 중립)
뉴스	감성 트리맵
유튜브	유튜브 채널 랭킹

커뮤니티 언급량

커뮤니티 내에서 MZ세대의 각 대선후보자별 언급량은 윤석열 후보가 조금 더 많은 것으로 나타남

MZ세대는 커뮤니티에서 각 대선후보에 대해 얼마나 언급했는가?



연구질문 <2>

MZ세대 커뮤니티 사이트 언급 키워드 분석

연구 모델 설명

연구에 사용된 수식과 변수는 하기와 같음

수식

$$\ln(\textit{여론조사결과}_{it}) = \ln(\textit{긍정키워드언급량}_{it-n}) + \ln(\textit{부정키워드언급량}_{it-n}) + \ln(\textit{중립키워드언급량}_{it-n}) + \varepsilon_{it-n}$$

변수

변수명	설명
여론조사결과	각 후보 i에 대한 20대, 30대의 여론조사 결과를 t 일별로 나타냄
긍정키워드 언급량	각 후보 i에 대한 20대, 30대의 긍정키워드 언급량을 t - n일별로 나타냄
부정키워드 언급량	각 후보 i에 대한 20대, 30대의 부정키워드 언급량을 t - n 일별로 나타냄
중립키워드 언급량	각 후보 i에 대한 20대, 30대의 중립키워드 언급량을 t - n 일별로 나타냄

이재명 부정 키워드

이재명 후보의 경우 커뮤니티 내에서 MZ 세대가 언급한 부정적 연관 키워드의 대부분은 후보의 사생활과 관련된 키워드들이었음

10대

랭킹	연관 키워드
1	싫다
2	욕하다
3	논란
4	망하다
5	범죄
6	욕
7	의혹
8	음주운전
9	포퓰리즘
10	부정선거
11	차별
12	비판하다
13	무섭다
14	불쌍하다
15	관심없다

20대

랭킹	연관 키워드
1	의혹
2	논란
3	범죄
4	욕설
5	싫다
6	망하다
7	부정선거
8	비판
9	포퓰리즘
10	무섭다
11	음주운전
12	위기
13	최악
14	빨갱이
15	불법

30대

사생활 관련 키워드

랭킹	연관 키워드
1	의혹
2	논란
3	욕설
4	범죄
5	싫다
6	비판
7	위기
8	부정선거
9	음주운전
10	분노
11	망하다
12	싫어하다
13	최악
14	갈등
15	내로남불

후보자의 사생활과 연관된 부정적 키워드가 많음

* 상위 150위권 연관 키워드 중에서 중복, 인물, 대선 기본정보와 관련된 키워드 제외하고 상위 15개 연관 키워드 추출함

윤석열 부정 키워드

윤석열 후보의 경우 커뮤니티 내에서 MZ 세대가 언급한 부정적 키워드의 대부분은 TV 토론, 기자와의 대화 등에서 노출된 후보의 언행과 관련된 키워드들이었음

10대

랭킹	연관 키워드
1	싫다
2	논란
3	욕하다
4	범죄
5	의혹
6	망하다
7	차별
8	멍청하다
9	비판
10	망언
11	갈등
12	부정선거
13	최악
14	무식하다
15	여혐

20대

랭킹	연관 키워드
1	의혹
2	논란
3	범죄
4	망하다
5	싫다
6	비판
7	욕하다
8	망언
9	부정선거
10	적폐
11	갈등
12	무식하다
13	빨갱이
14	최악
15	불법

30대

역량 관련 키워드

랭킹	연관 키워드
1	의혹
2	논란
3	범죄
4	비판
5	망언
6	욕하다
7	싫다
8	최악
9	갈등
10	부정선거
11	적폐
12	분노
13	허위
14	위기
15	불법

후보자의 평소 언행, 능력과 연관된 부정적 키워드가 많음

* 상위 150위권 연관 키워드 중에서 중복, 인물, 대선 기본정보와 관련된 키워드 제외하고 상위 15개 연관 키워드 추출함

이재명 긍정 키워드

이재명 후보의 경우 커뮤니티 내에서 MZ 세대가 언급한 긍정적 연관 키워드의 주요 내용은 성남시장, 경기도지사와 관련된 후보자의 역량에 대한 긍정 반응이 많았음

10대

랭킹	연관 키워드
1	지지하다
2	진심
3	잘하다
4	믿다
5	낫다
6	좋다
7	일 잘하다
8	간절하다
9	최선
10	좋아하다
11	희망
12	사랑하다
13	칭찬받다
14	멋지다
15	똑똑하다

20대

랭킹	연관 키워드
1	지지하다
2	잘하다
3	믿다
4	낫다
5	진심
6	희망
7	유능한
8	좋다
9	평화
10	좋은
11	일 잘하다
12	대세
13	합리적
14	호감
15	좋아하다

30대

역량 관련 키워드

랭킹	연관 키워드
1	지지하다
2	잘하다
3	믿다
4	진심
5	희망
6	좋은
7	유능한
8	평화
9	낫다
10	적극적
11	최선
12	일 잘하다
13	대세
14	유능하다
15	합리적

후보자의 역량과 연관된 긍정적 키워드가 많음

* 중복, 키워드 제외하고 상위 15개 연관 키워드 추출함

윤석열 긍정 키워드

윤석열 후보의 경우 커뮤니티 내에서 MZ 세대가 언급한 부정적 키워드의 대부분은 TV 토론, 기자와의 대화 등에서 노출된 후보의 언행과 관련된 키워드들이었음

감정 관련 키워드

10대

랭킹	연관 키워드
1	지지하다
2	진심
3	잘하다
4	낫다
5	좋다
6	원하다
7	지지받다
8	좋아하다
9	웃다
10	호감
11	최선
12	희망
13	똑똑한
14	공감
15	진정한

20대

랭킹	연관 키워드
1	지지하다
2	잘하다
3	진심
4	믿다
5	좋다
6	희망
7	원하다
8	낫다
9	평화
10	바라다
11	대세
12	호감
13	공감
14	최선
15	유능한

30대

랭킹	연관 키워드
1	지지하다
2	잘하다
3	진심
4	믿다
5	희망
6	원하다
7	좋은
8	바라다
9	평화
10	최선
11	확실하다
12	기대하다
13	적극적
14	공감
15	해결하다

후보자에 대한 개인적인 관심, 추상적 가치에 대한 긍정 키워드가 많음

* 중복, 키워드 제외하고 상위 15개 연관 키워드 추출함

기간별 감성지수 x 전통적 여론조사 상관관계 분석 <20, 30대>

커뮤니티 내 MZ 세대의 대선후보별 언급 키워드 분석 결과, 부정적인 키워드의 언급량이 높을수록 전통적인 여론조사의 각 대선후보 지지율에 부정적인 영향을 미침을 확인함

MZ세대가 커뮤니티에서 대선후보와 관련된 감성지수 단어 언급 후, 전통적인 여론조사에 영향을 미쳤는가?

MZ세대

<1일 후 여론조사>

변수	결과 (Coefficient)
긍정 키워드 언급량	1.41
부정 키워드	**-8.31***
중립 키워드	2.23

<3일 후 여론조사>

변수	결과 (Coefficient)
긍정 키워드 언급량	0.58
부정 키워드	**-6.02***
중립 키워드	0.81

<5일 후 여론조사>

변수	결과 (Coefficient)
긍정 키워드 언급량	0.56
부정 키워드	**-5.09***
중립 키워드	-0.11

40대 이상

<1일 후 여론조사>

변수	결과 (Coefficient)
긍정 키워드 언급량	-0.47
부정 키워드	0.54
중립 키워드	-0.05

<3일 후 여론조사>

변수	결과 (Coefficient)
긍정 키워드 언급량	-0.46
부정 키워드	-0.01
중립 키워드	0.28

<5일 후 여론조사>

변수	결과 (Coefficient)
긍정 키워드 언급량	-0.34
부정 키워드	-0.14
중립 키워드	0.46

긍정적 키워드 영향 없음

MZ세대 부정적 키워드 영향 있음

MZ세대 부정적 키워드 영향 최대 5~7일까지

MZ세대 부정적 키워드 효과는 시간순에 따라 작아짐

40대 이상 전혀 영향 없음

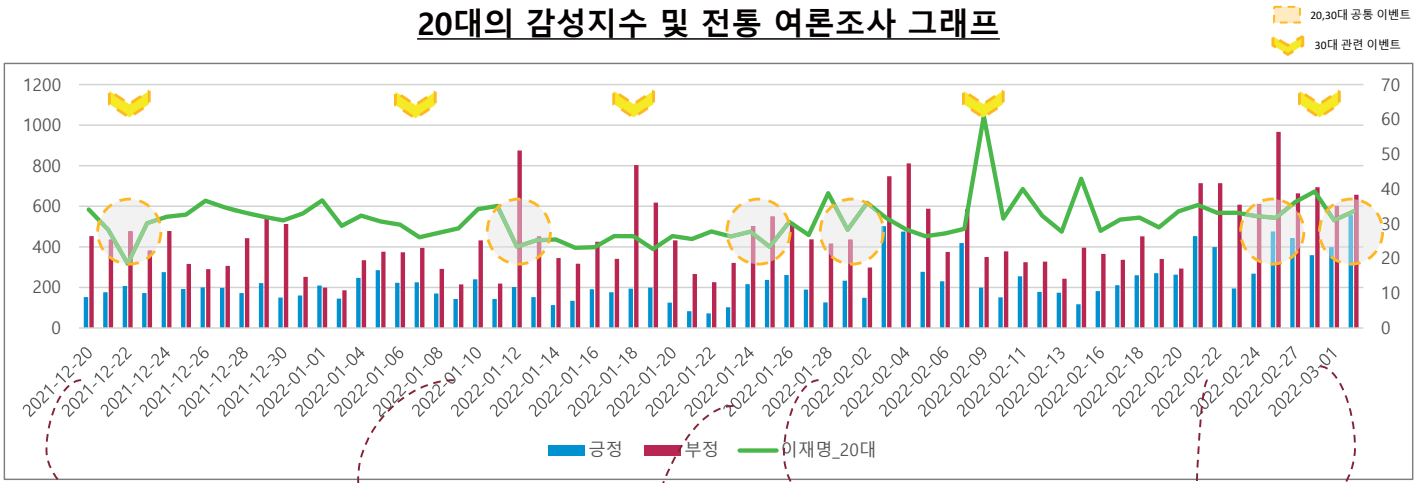
연구질문 <3>

MZ세대가 커뮤니티에서 대선 후보자를 향해 나타낸 여론에 영향을 끼친 Event Analysis

이재명 감성지수 & 전통 여론조사 결과 패턴 <20대>

20대는 커뮤니티 내에서 이재명 후보의 '부정적 이슈'가 발생할 때마다, 여론에 변동이 크게 나타남

20대의 감성지수 및 전통 여론조사 그래프



2021-12-22: 대장동 핵심 실무자 2명 숨짐

2022-01-12: 이재명 변호사비 대납의혹 녹취록 제보자 숨짐

2022-01-26: 상대 후보자 네거티브 중단 선언 & 눈물 유세

2022-02-03: 후보자 성남도시개발공사 사장 사퇴 강요 무혐의

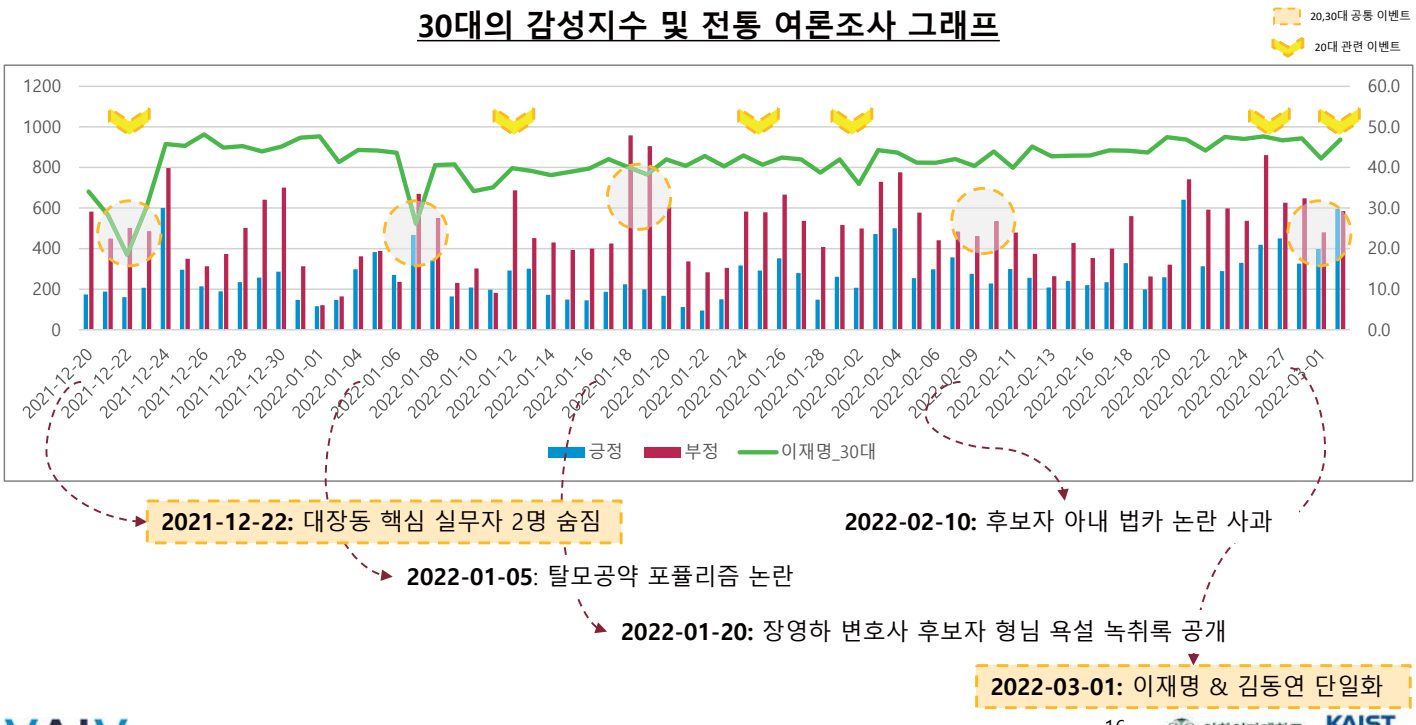
2022-02-25~26: TV 토론회

2022-03-01: 이재명 & 김동연 단일화

이재명 감성지수 & 전통 여론조사 결과 패턴 <30대>

30대는 커뮤니티 내에서 이재명 후보의 '부정적 이슈'에 대해 여론이 크게 변동하지 않으나, 주요 이슈들에 대해 단기적으로 영향을 받음

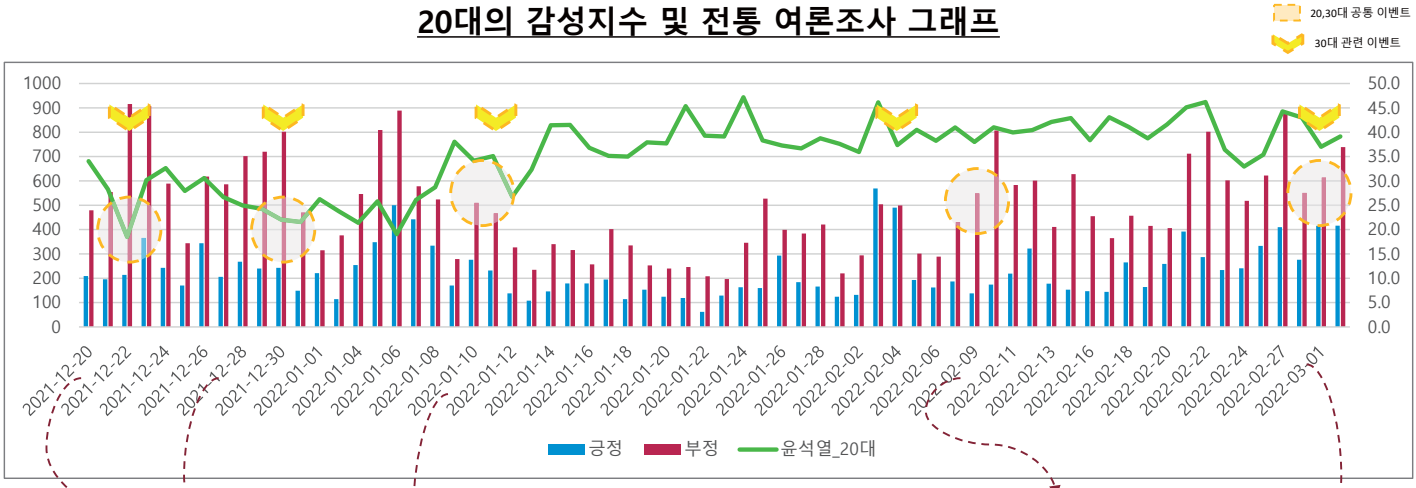
30대의 감성지수 및 전통 여론조사 그래프



윤석열 감성지수 & 전통 여론조사 결과 패턴 <20대>

20대는 커뮤니티 내에서 윤석열 후보의 '부정적 이슈'가 발생할 때마다, 여론에 변동이 크게 나타남

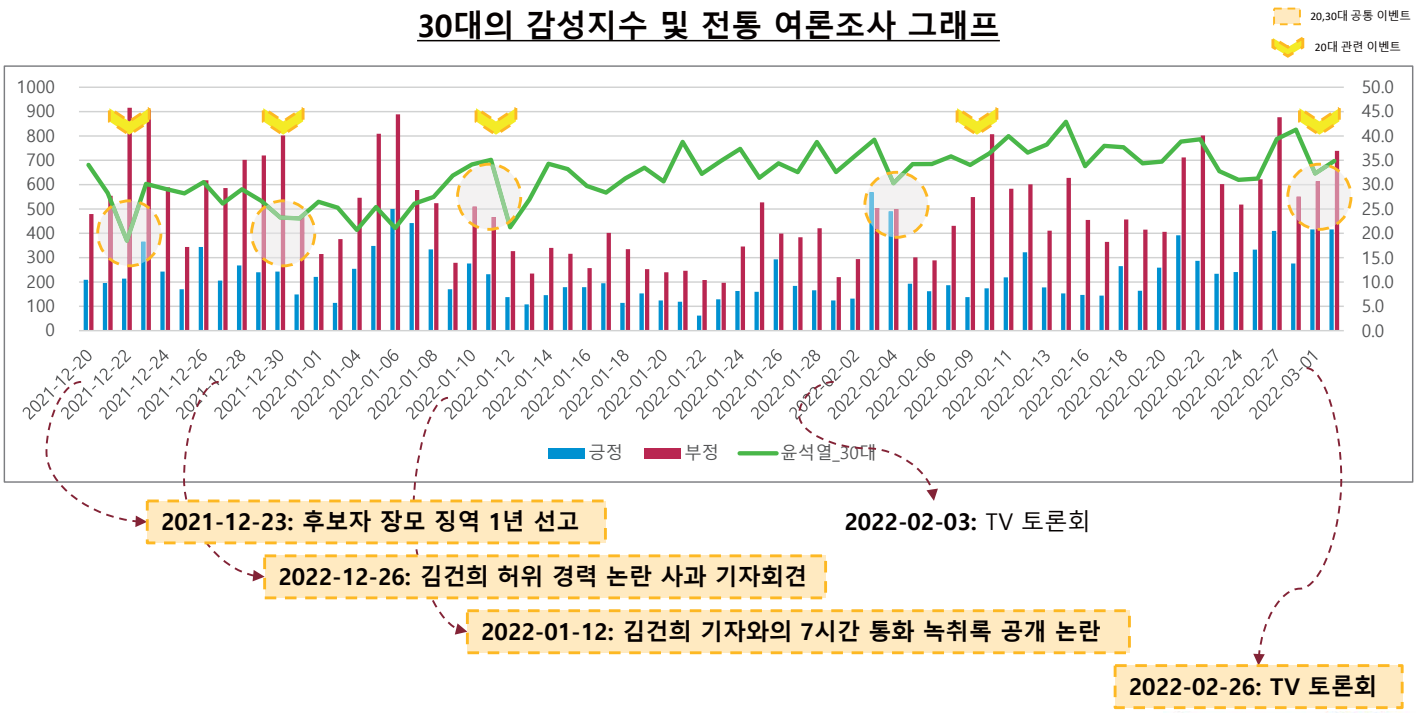
20대의 감성지수 및 전통 여론조사 그래프



윤석열 감성지수 & 전통 여론조사 결과 패턴 <30대>

30대는 커뮤니티 내에서 윤석열 후보의 대통령 후보로써의 역량과 관련된 '부정적 이슈'에 대해 더 집중함

30대의 감성지수 및 전통 여론조사 그래프



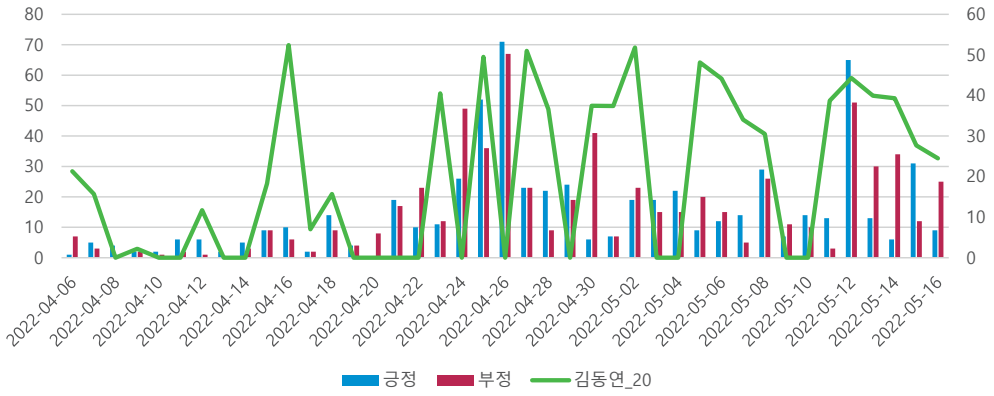
연구질문 <4>

상관관계 분석 결과를 토대로 6월 지방선거 중
주요 격전지의 후보자들 간 현재 여론 상황은 어떠한가?

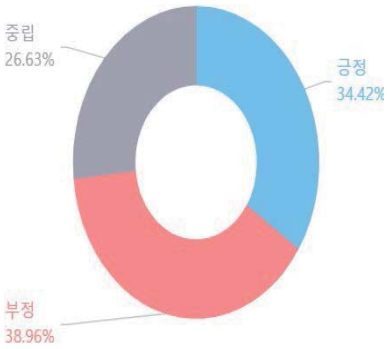
김동연 감성지수 & 전통 여론조사 결과 패턴 <20대>

김동연 경기도지사 후보자의 경우 커뮤니티 내에서 20대로부터 꾸준히 언급되는 양상을 보이고 있으며, 언급 키워드 분석 결과, 부정적인 감성지수가 상대적으로 높은 편임

20대의 감성지수 및 전통 여론조사 그래프



긍,부정 비율



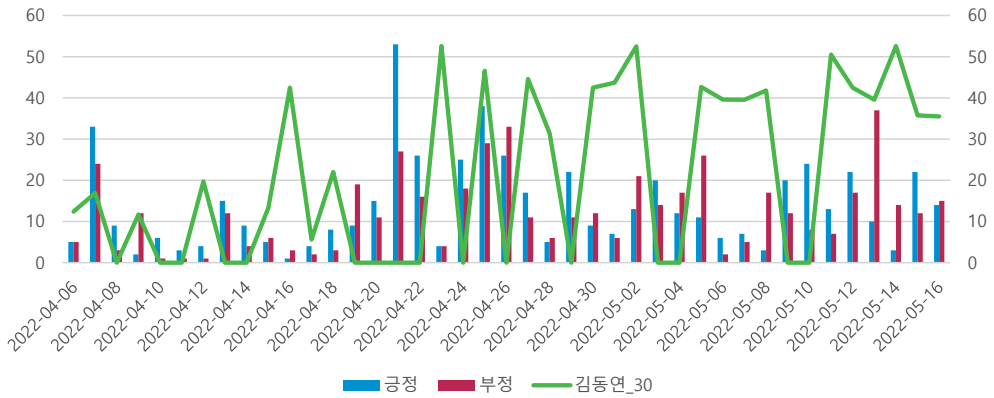
커뮤니티 내에서 후보자에 대한 총 언급량이 꾸준하나, 부정적인 감성지수가 상대적으로 높은 편임

* 김동연 후보 출마 선언 3월 31일 이후, 김은혜 후보 출마 선언 4월 6일 이후인, 4월 6일 부터 5월 16일까지 기간 대상으로 분석함

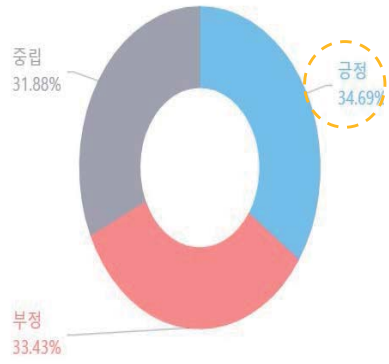
김동연 감성지수 & 전통 여론조사 결과 패턴 <30대>

김동연 경기도지사 후보자의 경우 커뮤니티 내에서 30대로부터 꾸준히 언급되는 양상을 보이고 있으며, 언급 키워드 분석 결과, 긍정적인 감성지수가 상대적으로 높은 편임

30대의 감성지수 및 전통 여론조사 그래프



긍,부정 비율



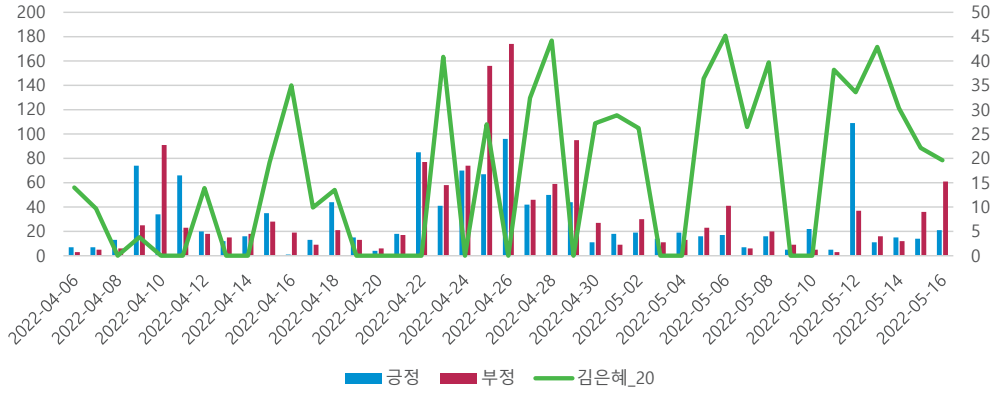
커뮤니티 내에서 후보자에 대한 총 언급량이 꾸준한 편이며, 긍정적인 감성지수가 상대적으로 높은 편임

* 김동연 후보 출마 선언 3월 31일 이후, 김은혜 후보 출마 선언 4월 6일 이후인, 4월 6일 부터 5월 16일까지 기간 대상으로 분석함

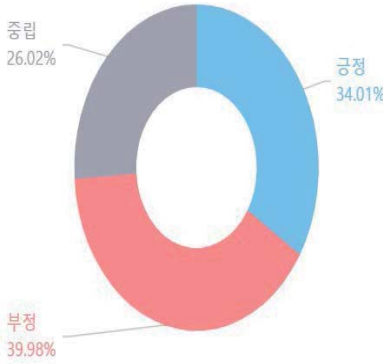
김은혜 감성지수 & 전통 여론조사 결과 패턴 <20대>

김은혜 경기도지사 후보자의 경우 커뮤니티 내 20대로부터의 언급이 적은 것으로 보이며, 언급 키워드 분석 결과, 부정적인 감성지수가 상대적으로 높은 편임

20대의 감성지수 및 전통 여론조사 그래프



긍,부정 비율



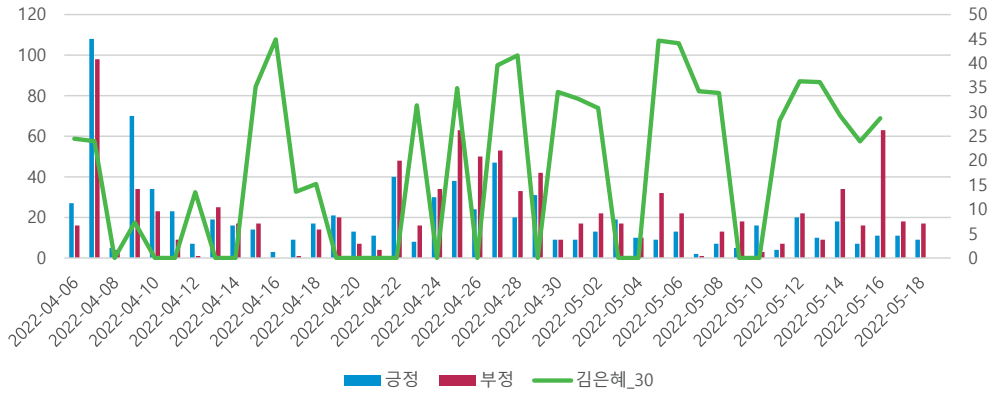
커뮤니티 내에서 후보자에 대한 언급량이 줄어들고 있으나, 부정적인 감성지수가 상대적으로 높은 편임

* 김동연 후보 출마 선언 3월 31일 이후, 김은혜 후보 출마 선언 4월 6일 이후인, 4월 6일 부터 5월 16일까지 기간 대상으로 분석함

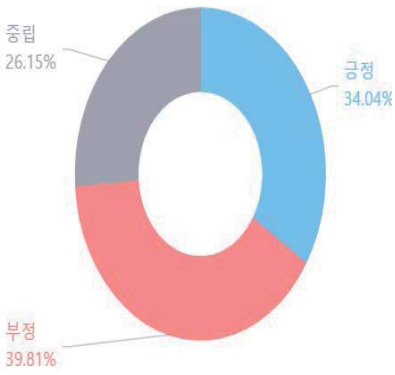
김은혜 감성지수 & 전통 여론조사 결과 패턴 <30대>

김은혜 경기도지사 후보자의 경우 커뮤니티 내 30대로부터의 언급이 적은 것으로 드러나며, 언급 키워드 분석 결과, 부정적인 감성지수가 상대적으로 높은 편임

30대의 감성지수 및 전통 여론조사 그래프



긍,부정 비율



커뮤니티 내에서 후보자에 대한 언급량이 꾸준하나, 부정적인 감성지수가 상대적으로 높은 편임

* 김동연 후보 출마 선언 3월 31일 이후, 김은혜 후보 출마 선언 4월 6일 이후인, 4월 6일 부터 5월 16일까지 기간 대상으로 분석함

MZ세대 커뮤니티 전략

MZ 세대의 여론을 공략하기 위해서는 커뮤니티를 집중해야 함

사생활과 관련된 이슈를 최소화 해야 함
(특히, 20대가 부정적 이슈에 대한 민감도가 높았음)

긍정적인 언급에 현혹되지 말 것!
(부정적 이슈는 단기적이고, 동질성을 야기시킴)

결론

부정적 표현의 중요성

1. 부정적 표현, 가짜뉴스 → 관심경제(Attention economy)

- 놀라움, 흥미진진, 참신함, 새로움 → 단기적 (반면, 중장년들은 관심경제 영역 밖에 존재한다)

2. 부정적 키워드는 Type 1 Error를 줄일 수 있는 긍정적 효과 有

- Type 1 Error = False Positive 떨어져야 하는 후보자가 당선되는 경우

3. Social media, Community site는 "감정"의 Media (특히, "부정적인 감정"을 교류하는 미디어)

4. 동질성 선호(Homophily) 야기:

- Local network effect: 네트워크 가치는 "전체" 사람수가 아니라, "연결된" 사람들의 수가 중요. 개인 계정도 팔로워 수보다 팔로잉 하는 수가 중요 (예, 트럼프: 7,200만 vs. 47명, 테일러 스위프트: 8,500만 vs. 0)
- 부정적 표현이 "사회적 신호"(Social signal) → Trend로 자리잡는 원인
- 정치적 Micro-targeting의 기준 및 도구

5. 선거 결과 예측 & 선거결과 몰아가기 이중 효과

- 사회적 존재(James Surowiecki, "Wisdom of Crowd")의 한계 증폭: Collective Judgment 한계 (대중의 광기)
  - ✓ 서로 영향 줄수록, 집단은 현명해지지 않는다 (집단이 가장 똑똑해지는 방법은 개인들이 각자 최대한 독립적으로 생각하는 것)
- 집단은 너무 똑같아도 문제지만, 양극화도 문제 → "필터 버블"로 편향성 증폭 (판단이 좁아진다)

6. 향후 과제:

- Gender 이슈, 19대 대선 등을 추가 분석하여 어떤 그룹이 부정적 표현에 영향 받는 지를 분석하고자 함

감사합니다.

빅데이터 기술 교류 세미나

빅데이터와 여론조사

발표2

Sometrend를 활용한 공공 메타버스 플랫폼에 대한 여론 분석과 정책 제언

윤혜정 (이화여대 교수)

2022-05-19

Sometrend를 활용한 공공 메타버스 플랫폼에 대한 여론 분석과 정책 제언

이화여자대학교 신산업융합대학 윤혜정 교수
연세대학교 정보대학원 안재영 박사과정

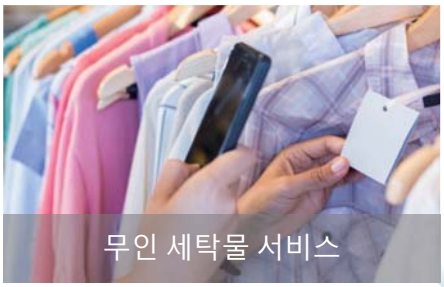
한국경영정보학회 VAIV

Contents

I 연구 배경 및 목적

연구 배경

- 비대면 서비스는 이제 선택이 아닌 필수
  - 코로나19로 인해 비대면 서비스가 일상이 되었고, 가상현실, 증강현실, 디지털 트윈 등 **메타버스에 대한 관심이 촉발됨**
  - 다양한 언택스 서비스가 나타났으며, 개방형 생태계를 갖춘 메타버스를 미래 중심기술로 전망(이상우, 2021)


무인 세탁물 서비스


비대면 식자재 새벽 배송


원격 강의 및 면접

NEXT STEP

Metaverse

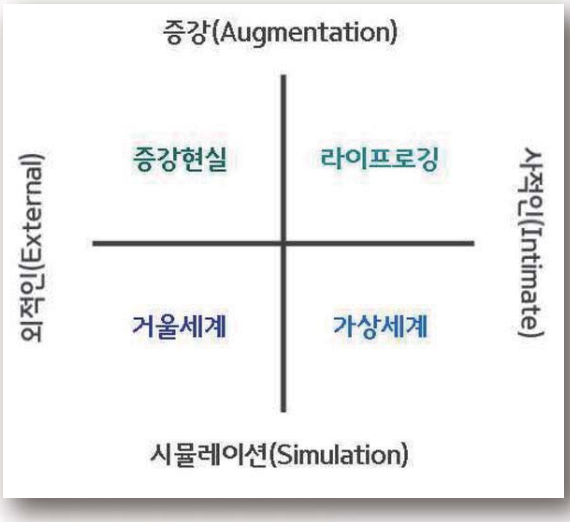
이미지 자료: https://gt-corp.tistory.com/235

참고자료: 이상우. 2021, 인터넷, 모바일 다음은 메타버스...2021 MAOC 개막.
https://www.ajunews.com/view/20211026143034338

연구 배경

❑ 메타버스 개념과 4대 유형

- **메타버스란, 자신을 대리하는 아바타를 통해 활동하는 3D 가상세계를 의미**
- 메타버스는 구현 공간과 정보 형태에 따라 4가지 형태로 구분
  - 현실에 외부 환경정보를 증강하여 제공하는 **증강현실(Augmented Reality)**
  - 현실 생활정보를 기반으로 구현된 **라이프로깅(lifelogging)**
  - 현실의 경제사회적 환경과 유사하게 구축된 **가상 세계(virtual worlds)**
  - 외부 환경정보를 기반으로 현실을 모방하는 **거울 세계(mirror worlds)**
- **각 메타버스 유형은 독립적이기보다 융복합된 형태로 발전할 것으로 전망**



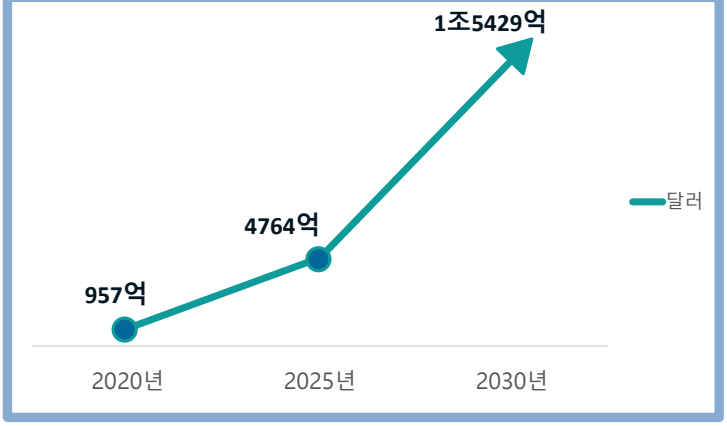
이미지 자료: http://www.m-economynews.com/mobile/article.html?no=30893
참고자료: 한혜원 (2008). 메타버스 내 가상세계의 유형 및 발전방향 연구. 한국디지털콘텐츠학회 논문지, 9(2), 317-323

연구 배경

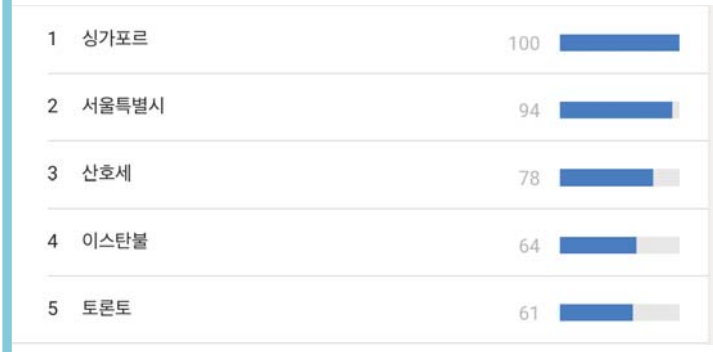
❑ 메타버스의 열풍

- PwC의 보고에 따르면 2030년까지 글로벌 메타버스 시장 규모는 1조 5429억 달러로 전망
- 구글 트렌드에서 'Metaverse' 검색어로 전세계 도시의 관심을 확인한 결과 서울특별시가 2순위로 나타남

글로벌 메타버스 시장 규모 전망

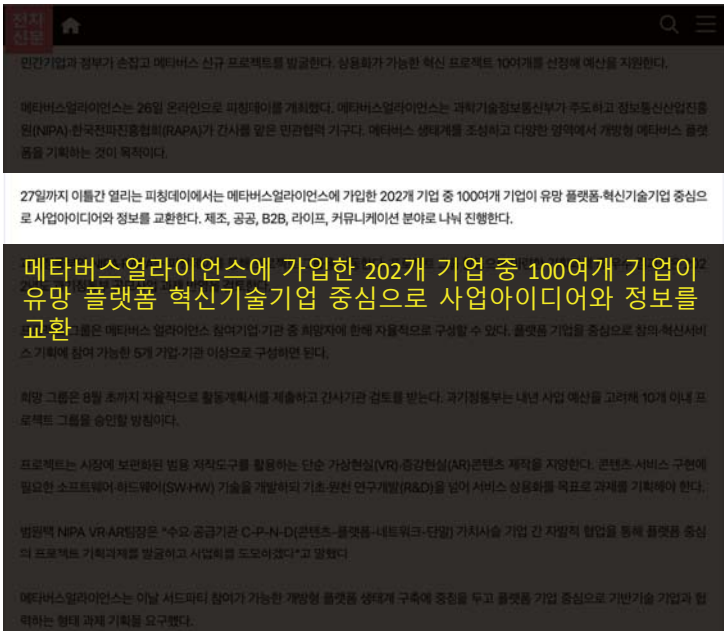


최근 5년 간 전세계 도시별 메타버스 관심 순위



연구 배경

- 메타버스 확산을 위한 정부 지원
  - 비대면 인프라 고도화와 초연결 신사업 육성을 목표로 **'개방형 메타버스 플랫폼'** 구축 진행
  - 2025년까지 2조 6천억원 예산 투입
  - 민관 얼라이언스를 구축하고, 참여 기업과 영역을 확장해 나가고 있음
  - 하지만, 성공적인 **'개방형 메타버스 플랫폼'**을 진행하기 위해서는 거버넌스가 필요하고, 이를 위해서는 명확한 계층별 요소별 정의와 각각의 책임주체와 전체 플랫폼의 컨트롤타워가 필요함
  - **이는 융복합서비스의 장점이자 약점!**



자료출처: https://m.etnews.com/20210726000149

연구 배경

- 윤석열 정부 110대 국정과제를 통한 메타버스 플랫폼의 중요성

54 전 국민 생애단계별 직업능력개발과 일터학습 지원 (고용부)

□ 과제목표

○ 모든 국민에게 평생에 걸쳐 필요한 직업능력 개발 기회 확대

○ 온·오프라인, 일과 학습이 융합된 통합적 직업능력개발 체계 구축

□ 주요내용

○ (프로그램 재구조화) 재학-구직-재직-이·전직 등 생애단계별로 필요한 직업훈련이 충분히 공급될 수 있도록 훈련프로그램 개편

- 청년층을 대상으로 디지털 신기술 등 기업수요를 반영한 훈련을 확대하여, 미래유망 분야로 조기에 노동시장 참여 지원

- 산업전환 업종 종사자에게는 경력 재설계 컨설팅과 훈련을 선제적으로 제공하고, 경단여성·중장년 친화적인 맞춤형 훈련 확충

- 개인별 훈련계좌인 「국민내일배움카드」를 통해 전 생애에 걸쳐 필요한 포괄적 직무기초능력 훈련과 경력설계서비스까지 확대 제공

- 훈련·자격·경력 등 개인이 습득한 다양한 직무능력 정보를 통합해서 관리·활용하는 「직무능력은행제」 구축

○ (온·오프라인 훈련 생태계 구축) 메타버스, VR, AR 등 신기술을 접목한 원격훈련 플랫폼 구축 검토 및 스마트직업훈련플랫폼(STEP)과 연계

- 민간의 혁신 훈련기관을 통해 강사·교재가 없이 실제 기업 프로젝트를 기반으로 훈련하는 혁신적 훈련모델 확산

○ (일터학습 인프라) 기업-학교를 오가며 훈련을 받는 일학습병행 업종 다양화, 재직 중 능력개발 기회를 확대하여 전문인력으로 성장 지원

- 체계적으로 인재를 육성하는 모범 기업에 대한 HRD 우수기업 인증 인센티브 강화(훈련비 지원 확대 등)

55 중소기업·자영업자 맞춤형 직업훈련 지원 강화 (고용부)

□ 과제목표

○ 중소기업 재직자 등의 직업훈련 참여 확대

○ 직업능력개발에 참여하는 기업·직업훈련기관의 자율성·혁신성 제고

□ 주요내용

○ (중소기업 맞춤형 훈련 지원) 현장별 업무 프로세스 및 문제해결에 적합한 훈련과정을 맞춤형으로 지원하는 현장맞춤형 체계적 훈련(S-OJT) 확대

- 기업별로 정해진 한도 내에서 자유롭게 훈련을 실시할 수 있도록 지원하는 기업직업훈련카드(기업직업훈련바우처) 도입

- 영세사업장 공동훈련 지원, 훈련기관의 다양한 컨텐츠를 일정기간 구독형으로 제공받는 패키지형 원격구독훈련 등 다양한 훈련방식 도입

○ (능력개발전담주치의 도입) 기업별 여건을 진단하고, 맞춤형 훈련프로그램을 설계해주는 '능력개발전담주치의'를 도입, 현장중심 지원 확대

○ (플랫폼종사자·자영업자 지원 강화) 플랫폼종사자에게 직종·수준별 특화훈련을 제공하고, 자영업자 국민내일배움카드 발급 시 소득기준 완화

○ (미래지향적 방식으로 전환) 사전통제 중심의 운영방식에서 벗어나 훈련기관의 자율·혁신성을 제고하여 훈련프로그램의 질적 수준 향상

- 훈련기관과 과정에 대한 사전평가 간소화 및 프로젝트 학습(PBL) 등 새로운 교육법 및 메타버스 등 신기술을 접목한 훈련 확산 지원

연구 배경

❑ 윤석열 정부 110대 국정과제를 통한 메타버스 플랫폼의 중요성

59 국민과 동행하는 디지털·미디어 세상 (방통위)

□ 과제목표

○ 소외 계층을 포함한 전국민 대상 미디어 접근성 및 활용도를 제고하고, 디지털 플랫폼사업자·이용사업자·이용자 간 상생 생태계 구축

□ 주요내용

○ (미디어 교육) (유아)찾아가는 놀이형 교육, (청소년)초중고 교육과정 연계 교육, (중·장년층·노인·장애인)특화교육 등 국민 생애주기별 격차없는 맞춤형 교육 제공

○ (미디어 접근권) 시청자미디어센터 전국화 및 찾아가는 서비스(미디어 나눔버스)로 지역민의 미디어 체험기회 확대

- 장애인방송 의무편성(한국수어방송 5→7%) 확대,장애인방송 품질평가제 도입, 재난방송의 수어제공 의무를 확대를 통해 재난방송 접근성 강화

○ (미디어 플랫폼의 신뢰성·투명성) 추천 알고리즘으로 인한 확증편향적 미디어 소비 등의 해결을 위해 기사·동영상 배열에 대한 책임성·신뢰성 제고

- 포털의 뉴스서비스 제공 방식·절차의 투명성을 제고하고, 미디어 플랫폼의 이용자 권익 보호를 위해 불만처리 체계 및 리터러시 교육 강화

○ (디지털 신산업 이용자 보호) 디지털 플랫폼·메타버스·모빌리티 등 디지털 신산업 분야에서의 이용자 보호 기반 마련

- 이용자 보호 업무 평가 등 자율규제 체계 구축을 지원하고, 이용자 불편 해소 및 권익 보호를 위한 필요최소한의 제도적 장치 등을 마련

- 메타버스 산업 진흥 시 디지털 공동체 윤리원칙 등 협력적 자율규제 체계를 마련하고, 모빌리티 산업 진흥 및 이용자 보호를 위한 위치정보법 개편

○ (디지털 폭력 피해구제 강화) 피해신고 절차 간소화 및 분쟁조정 도입 등 디지털 폭력 원스톱 지원체계 구축 및 제도적 기반 마련

77 민·관 협력을 통한 디지털 경제 패권국가 실현 (과기정통부)

□ 과제목표

○ 전 세계적인 디지털 전환과 기술패권 경쟁 속에서 민·관의 역량을 결집하여 국가·사회 디지털 혁신의 근간인 AI·데이터·클라우드 등 핵심기반을 강화하고, 메타버스·디지털플랫폼 등 신산업을 육성하여 디지털 경제 패권국가로 도약

□ 주요내용

○ (초일류 인공지능 국가) 최고 수준의 인공지능 기술 확보를 위해 대규모의 도전적 AI R&D를 추진하고, AI의 핵심 두뇌인 AI반도체 육성 추진('22~)

- 대학·중소기업 등의 AI 활용을 지원하는 세계적 컴퓨팅 인프라를 구축(광주 AI특화 데이터센터 및 차세대 슈퍼컴 도입, '23~)하고, 재난안전·교육·복지 등 全 분야에 AI 전면 적용('22~)을 통해 AI 융합 확산

○ (공공·민간데이터 대통합) 국가 데이터정책 컨트롤타워를 확립('22)하고, 민간이 필요로 하는 데이터의 개방 확대, 이용자가 편리하게 검색·활용 가능한 산업기반('23~'24) 조성 등을 통해 데이터 혁신강국 도약

○ (클라우드·SW 육성) AI·데이터의 핵심인프라인 클라우드·SW 경쟁력 강화를 위해 공공분야에서 민간 클라우드 및 상용SW를 우선 이용하도록 하고, 서비스형 SW(SaaS) 중심 생태계 조성 및 SW 원천기술 확보('22~) 등 추진

○ (한계돌파 新기술 확보) 국가 전략자산으로서 기술 축적을 위해 민·관 공동으로 핵심전략분야에 선택·집중한 대규모 R&D 추진으로 기술혁명 선도('22~)

○ (메타버스 경제 활성화) 메타버스 특별법 제정, 일상·경제활동을 지원하는 메타버스 서비스 발굴 등 생태계를 활성화하고, 블록체인을 통한 신뢰기반을 조성('22~)

○ (혁신·공정의 디지털플랫폼) 플랫폼의 건전한 혁신·성장 촉진 및 사회적 가치 창출 극대화를 위해 발전전략 수립 및 민간 주도의 자율규제체계 확립('22)

※ 범부처 민간과 함께하는「디지털 국가전략」수립 및 민관 합동 디지털혁신위원회 신설 검토

연구 목적

연구목적

Sometrend를 활용하여, 공공 메타버스 플랫폼에 대한 여론을 파악하고, 관련 정책 수립에 있어서 우선적으로 고려해야 하는 요소 파악

첫째, 공공 메타버스 플랫폼에 대한 여론은 어떠한가?

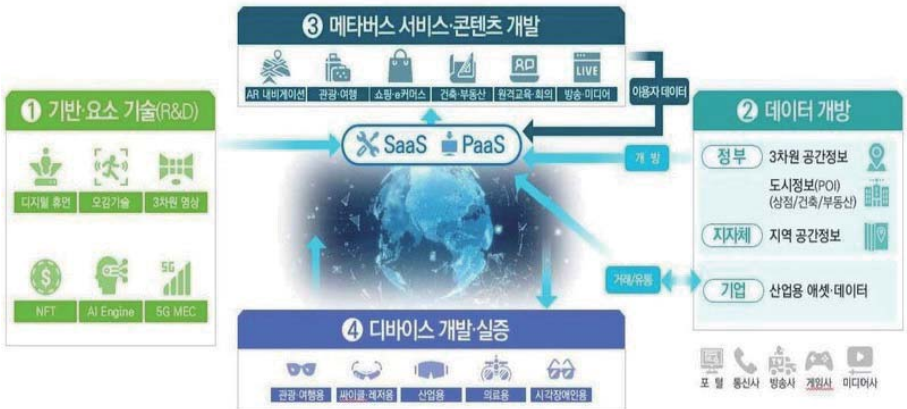
둘째, 공공 메타버스 플랫폼과 관련한 정책 수립시 어떠한 요소들을 고려해야 하며, 요소들 간의 상대적인 우선순위는 어떠한가?

Ⅱ 개념적 배경

개념적 배경

❑ 개방형 메타버스 플랫폼 소개

- 개방형 메타버스 플랫폼은 정부(3차원 공간정보, 도시정보), 지자체(지역 공간정보), 기업(산업용 애셋 및 데이터)의 데이터를 SaaS, PaaS로 개방하여 메타버스 서비스 및 콘텐츠를 개발할 수 있는 생태계를 조성

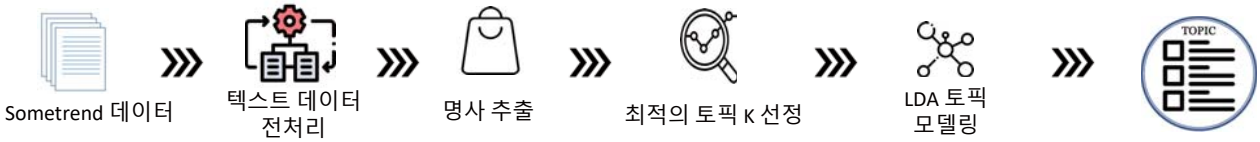


이미지 출처: https://www.irsglobal.com/bbs/rwdboard/15093
참고자료: 한국판 뉴딜 2.0 보고서

Ⅲ 연구방법 및 모형

연구 절차

Sometrend를 통한 공공 메타버스 여론에 대한 감성분석, 네트워크 분석 확인



AHP 프레임워크 제안 → 문헌 연구를 통한 각 계층 요소 도출 → 토픽 모델링에서 얻은 주제어와 문헌 연구에서 도출된 요인으로 AHP 프레임워크 설정 → FGI를 통한 타당성 확보

AHP 요소의 중요도 순위 결과에 대한 토의

Sometrend 여론 분석 결과

❑ **Sometrend 네트워크분석** : 여론의 단어들이 어떤 중심단어들로 연결되어 있는지 파악

- 기간: 2021. 07. 01 – 현재
- 검색 키워드: 메타버스, 공공, 정부

[메타버스] 연관어 순위

순위	연관어	건수
1	메타	3,662
2	기업	3,549
3	기술	3,476
4	플랫폼	3,447
5	디지털	2,968
6	사업	2,942
7	서비스	2,791
8	산업	2,721
9	개발	2,448
10	시장	2,402
11	계획	2,025
12	공간	1,825
13	교육	1,799
14	콘텐츠	1,754
15	데이터	1,738

Sometrend 여론 분석 결과

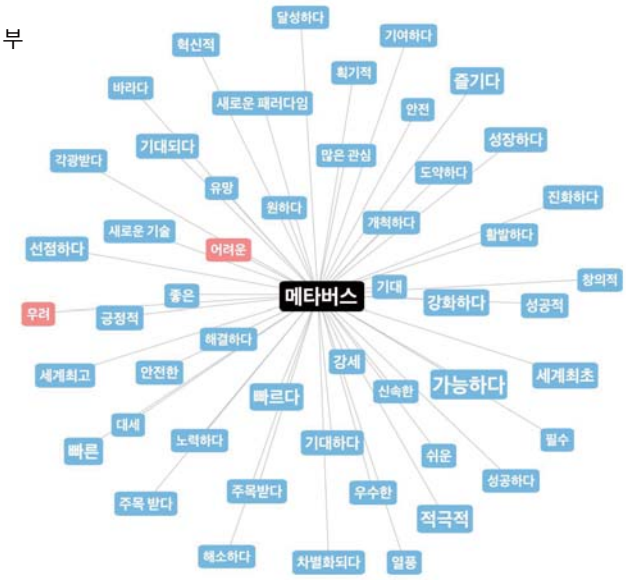
❑ **Sometrend 감성분석 (트리맵)** : 여론의 긍정/부정 정도와 긍정, 중립, 부정 단어 파악

- 기간: 2021. 07. 01 – 현재
- 검색 키워드: 메타버스, 공공, 정부
- 긍정/부정에 대한 감성 분석

Sometrend 여론 분석 결과

❑ **Sometrend 감성분석 (종합)** : 여론의 긍정/부정 정도와 긍정, 중립, 부정 단어 파악

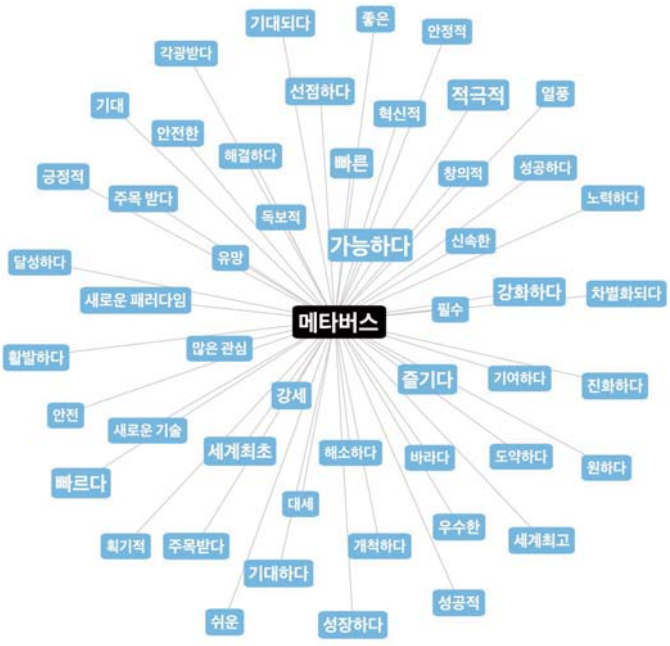
- 기간: 2021. 07. 01 – 현재
- 검색 키워드: 메타버스, 공공, 정부
- 긍정/부정에 대한 감성 분석



Sometrend 여론 분석 결과

❑ **Sometrend 감성분석 (긍정)** : 여론의 긍정/부정 정도와 긍정, 중립, 부정 단어 파악

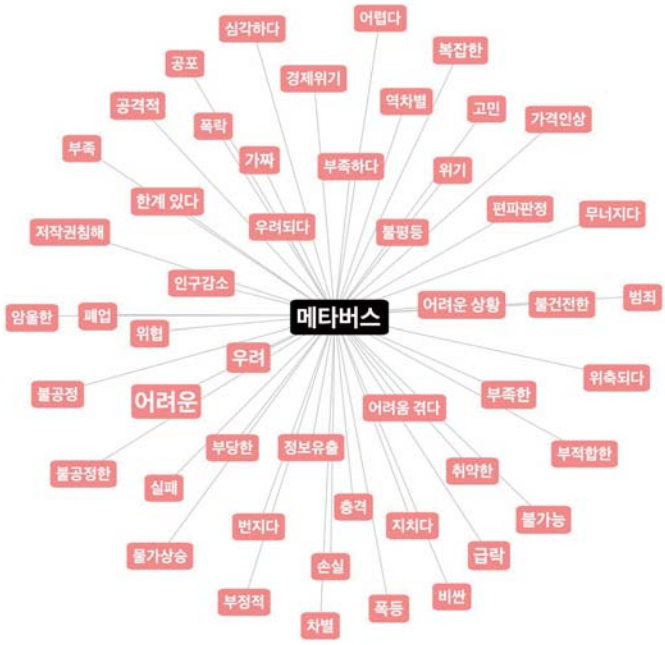
- 기간: 2021. 07. 01 – 현재
- 검색 키워드: 메타버스, 공공, 정부
- **Top 10 긍정: 가능하다, 적극적, 빠른, 빠르다, 즐기다, 세계최초, 강화하다, 성장하다, 선점하다, 강세**



Sometrend 여론 분석 결과

❑ **Sometrend 감성분석 (부정)** : 여론의 긍정/부정 정도와 긍정, 중립, 부정 단어 파악

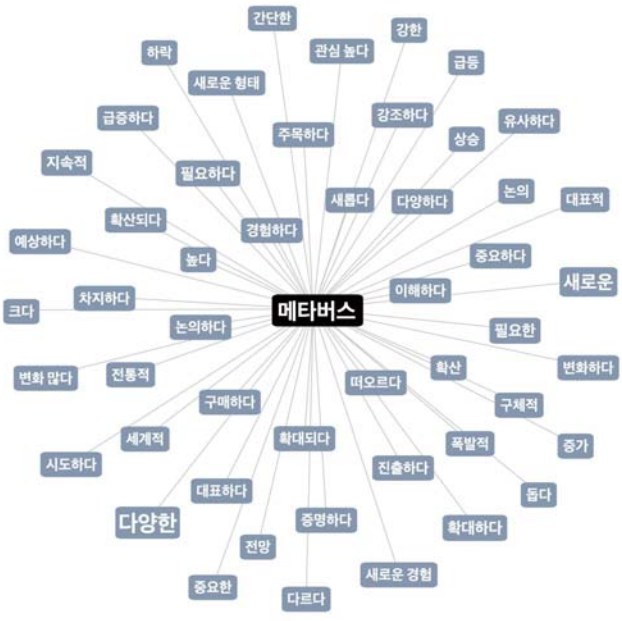
- 기간: 2021. 07. 01 – 현재
- 검색 키워드: 메타버스, 공공, 정부
- **Top 10 부정: 어려운, 우려, 급락, 어려운 상황, 복잡한, 폭등, 한계있다, 불가능, 폭락, 공격적**



Sometrend 여론 분석 결과

❑ **Sometrend 감성분석(중립)** : 여론의 긍정/부정 정도와 긍정, 중립, 부정 단어 파악

- 기간: 2021. 07. 01 – 현재
- 검색 키워드: 메타버스, 공공, 정부
- **Top 10 중립: 다양한, 새로운, 확대하다, 필요하다, 지속적, 필요한, 전망, 새롭다, 세계적, 중요한**



LDA 기반 토픽 모델링

- Sometrend의 텍스트 데이터 분석
  - 기간: 2021. 07. 01 – 2022. 04. 30
  - 검색 키워드: 메타버스, 공공, 정부
  - 제공 받은 데이터: 공공 메타버스 관련 블로그, 뉴스, 트위터
  - **분석 데이터 수: 6,955개 뉴스(중복 데이터 제거)**

```
[ ]  1 import pandas as pd
     2 rdata = pd.read_csv('Merged file(Non-duplication).csv')

[ ]  1 data_ct = rdata.loc[:, ("CT")]
     2 data_ct

0          동영상 뉴스                    [뉴스투데이] ◀ 앵커 ▶ '뉴스 열어보기...
1          /사진제공=과기정통부 정부가 '초연결 신산업 육성'을 목표로 오는 2025년까지 2...
2          [파이낸셜뉴스]정부가 '초연결 신산업'을 육성하기 위해 2025년까지 메타버스와 블...
3          [이데일리 이지현 기자] 메타버스(확장가상세계) 대장주 맥스트가 다시 상승세다. 정...
4          정부가 '초연결 신산업'을 육성하기 위해 2025년까지 메타버스와 블록체인 등 핵심...
                                    ...
6950       [경향신문]  28일 세계 이동통신 전시회 MWC 개막 삼성전자는 '갤럭시 북' 공...
6951       KT가 28일부터 3월 3일(현지시간)까지 스페인 바르셀로나에서 열리는 MWC 20...
6952       메타버스 기반 '클라우드 원격개발 지원플랫폼'이 연내 가동된다. 한국SW산업협회와 ...
6953       MWC2022 개막 SKT, 메타버스 '이프랜드' 공개 KT, AI·첨단 로봇 기술...
6954       '모바일 월드 콩그레스 2022'…28일, 3년 만에 정상 개최 [경향신문]  SK...
Name: CT, Length: 6955, dtype: object
```


LDA 기반 토픽 모델링

- Sometrend의 텍스트 데이터 분석
  - 정규 표현식의 데이터 전처리

```
[ ]  1 cleanText = list()
     2 for sent in dataList:
     3
     4     review = re.sub(r'\d+','', str(sent))# remove number
     5     review = re.sub(r'[^가-힣\s]','', review) #remove english spell
     6     #review = re.sub(r'\s+', ' ', review) #remove extra space
     7     review = re.sub(r'<[^>]+>','',review) #remove Html tags
     8     #review = re.sub('  ', ' ', review)
     9     review = re.sub(r'\s+', ' ', review) #remove spaces
    10     # review = re.sub(r"^\s+", '', review) #remove space from start
    11     # review = re.sub(r'\s+$', '', review) #remove space from the end
    12
    13     cleanText.append(review.split('\t'))
    14 cleanText

[['동영상 뉴스 뉴스투데이 앵커 뉴스 열어보기 시작합니다 앵커 먼저 서울경제부터 볼까요 앵커 정부가 지급한 긴급재난지원금으로 상품을 구매한 뒤
 ['사진제공과기정통부 정부가 초연결 신산업 육성을 목표로 오는 년까지 조억원의 예산을 투자해 메타버스클라우드블록체인 등 유망 분야 육성에 나선
 ['파이낸셜뉴스정부가 초연결 신산업을 육성하기 위해 년까지 메타버스와 블록체인 등 핵심 유망 분야에 약 조억원의 예산을 투입한다는 소식에 성호
 ['이데일리 이지현 기자 메타버스확장가상세계 대장주 맥스트가 다시 상승세다 정부가 메타버스 산업에 대한 육성 방안을 논의했다는 소식이 호재로
 ['정부가 초연결 신산업을 육성하기 위해 년까지 메타버스와 블록체인 등 핵심 유망 분야에 약 조 천억 원의 예산을 투입하기로 했습니다 과학기술정
 ['정부의 초연결 신산업 육성 계획 과기정통부 제공 뉴스 서울뉴스 이기범 기자 정부가 메타버스로 디지털 뉴딜 추진을 위한 회의를 열고 메타버스
 ['임혜숙 과학기술정보통신부 장관이 일 오후 서울 여의도 국회에서 열린 과학기술정보방송통신위원회 전체회의에 출석하고 있다 김명국 선임기자 정부
 ['디지털데일리 채수웅기자 확장가상세계인 메타버스 쓰임새가 확장되고 있다 코로나 장기화로 비대면 사회로의 전환이 빨라지고 있는 가운데 정부의
 ['권칠승 중소벤처기업부 장관과 수상기업 간 유닛미팅 진행 전자신문이 기술자립도를 높인 소재부품장비소부장 분야 히든 챔피언을 찾았다 전자신문에
 ['새로운 기회의 땅으로 여겨지는 메타버스에 제조 유통 등 전통기업들의 관심이 뜨겁다 세대밀레니얼세대년대 출생로 대표되는 새로운 소비자들이 메
 ['임혜숙 과학기술정보통신부 장관이 일 오전 서울 중구 서울중앙우체국에서 메타버스 플랫폼을 이용해 제차 디지털 뉴딜반 회의 를 주재하고 있다
 ['한국방송통신전파진흥원은 일부터 정보통신 무선설비 등 분야 개 국가자격증에 대해 모바일 자격증 서비스를 제공한다고 제공 한국방송통신전파진흥
 ['테크비즈코리아 정부출연연구기관출연연과 과학기술특성화대학 등의 연구개발 성과를 한자리에 모아 소개하는 테크비즈코리아 행사가 일 이틀간 온라
 ['연구기획팀 구성인공지능융합산업 등 특화된 광주형 전략 마련 파이낸셜뉴스 광주황태종 기자광주광역시가 메타버스 융합신산업 육성에 본격 나선다
 ['이데일리 권효중 기자 국내 증시 사상 최고 청약 경쟁률을 기록했던 메타버스 플랫폼 기업 맥스트가 코스닥 상장 첫 날 따상시초가를 공모가 배
 ['월 상장 자이언트스텝은 시총 조 넘봐 장기적으로 고부가가치 기술인지 따져봐야 지난 월 권칠승 중소벤처기업부 장관이 솔루션업체 맥스트를 방문
 ['웰레콤이 개발한 메타버스 플랫폼 에서 구현된 운동회 메타버스에 대한 관심이 높아지고 있다 정부는 메타버스 산업을 집중 육성하기 위해 디지털
 ['맥스트상진 맥스트 홈페이지 캡처 메타버스 플랫폼을 개발하는 기업인 맥스트가 코스닥 상장 첫날인 일 따상을 달성했다 이날 맥스트의 주가는 시
```

LDA 기반 토픽 모델링

- Sometrend의 텍스트 데이터 분석
  - KoNLpy의 Mecab으로 명사만 추출하는 형태소 분석

```
[ ]  1 noun_T = preP.pos_tag(cleanText)

     1 noun_T

     '금융',
     '센터',
     '금융',
     '센터',
     '예치',
     '은행',
     '자산',
     '예치',
     '관리',
     '메타',
     '버스',
     '공간',
     '금융',
```


LDA 기반 토픽 모델링

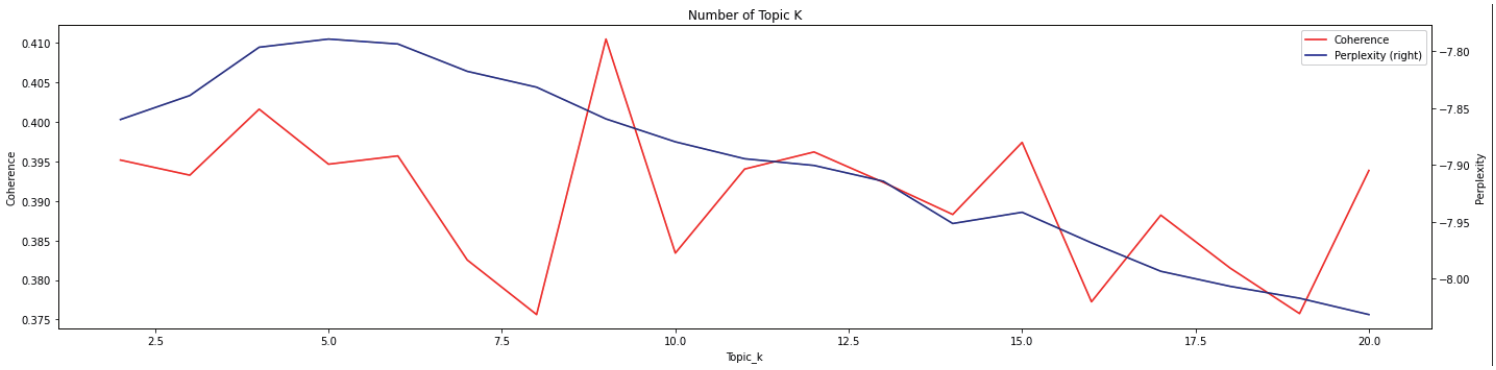
- Sometrend의 텍스트 데이터 분석
  - 불용어: 국가명, 지역명, 정치자 이름 등을 제거
  - LDA 기반 토픽 모델링을 반복 수행하여 불용어 취사선택

```
1 stopWord = ['정부', '후보', '공약', '글로벌', '한국', '서울', '메타', '버스', '세계', '독도', '시장', '대표', '지역',
2              '은행', '고객', '추진', '부산', '시티', '대전', '코로나', '국내', '해외', '행사', '시간', '대선', '기관',
3              '위원회', '방송', '의원', '교수', '카카오', '네이버', '매출', '올해', '분기', '행사', '랜드', '건설', '도
4              '현실', '필요', '국가', '대통령', '이재명', '우리', '윤석열', '서울시', '민원', '플랫', '기반', '수원',
5              '전망', '주가', '탄소', '민주당', '학교', '학생', '이번', '공공', '진행', '과제', '게임']
6
7
8 finalT = preP.stop_word(noun_T, stopWord)
9 finalT
```

LDA 기반 토픽 모델링

❑ Sometrend의 텍스트 데이터 분석

- 최적의 토픽 K: 9 (응집도 기준 0.410)
- Silhoutte Score: k가 9일 때 0.641(타당성 검증)
- 혼잡도(perplexity)는 절대 최소치를 가지지 못하기 때문에 응집도를 기준으로 최적의 토픽 K를 선정하고 이에 대한 타당성을 확보하기 위해 실루엣 계수를 확인(Krasnov & Sen, 2019)
- 실루엣 계수는 -1부터 1까지이며, 1에 가까울 수록 좋은 군집을 의미(Subramani et al., 2018)

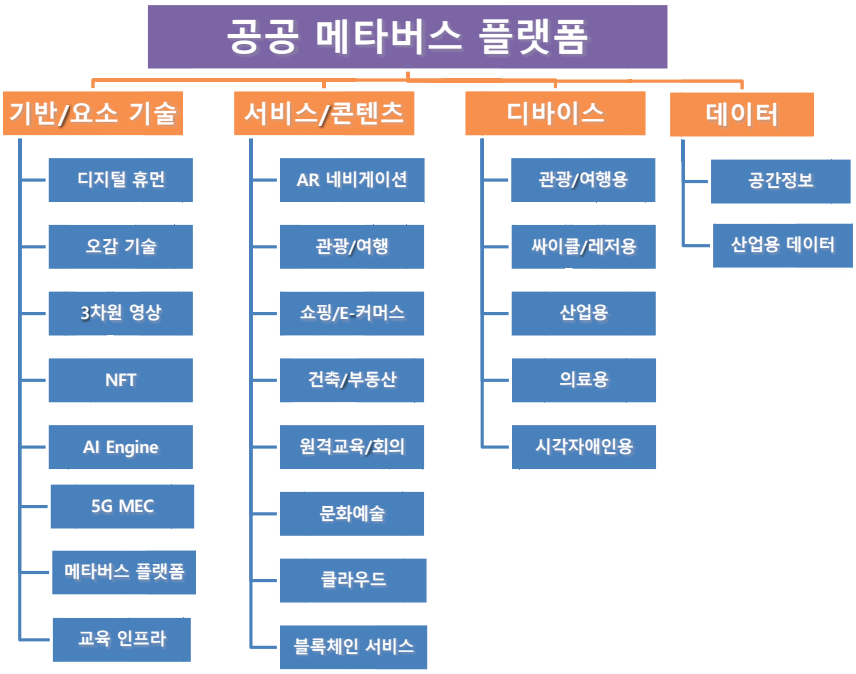


LDA 기반 토픽 모델링

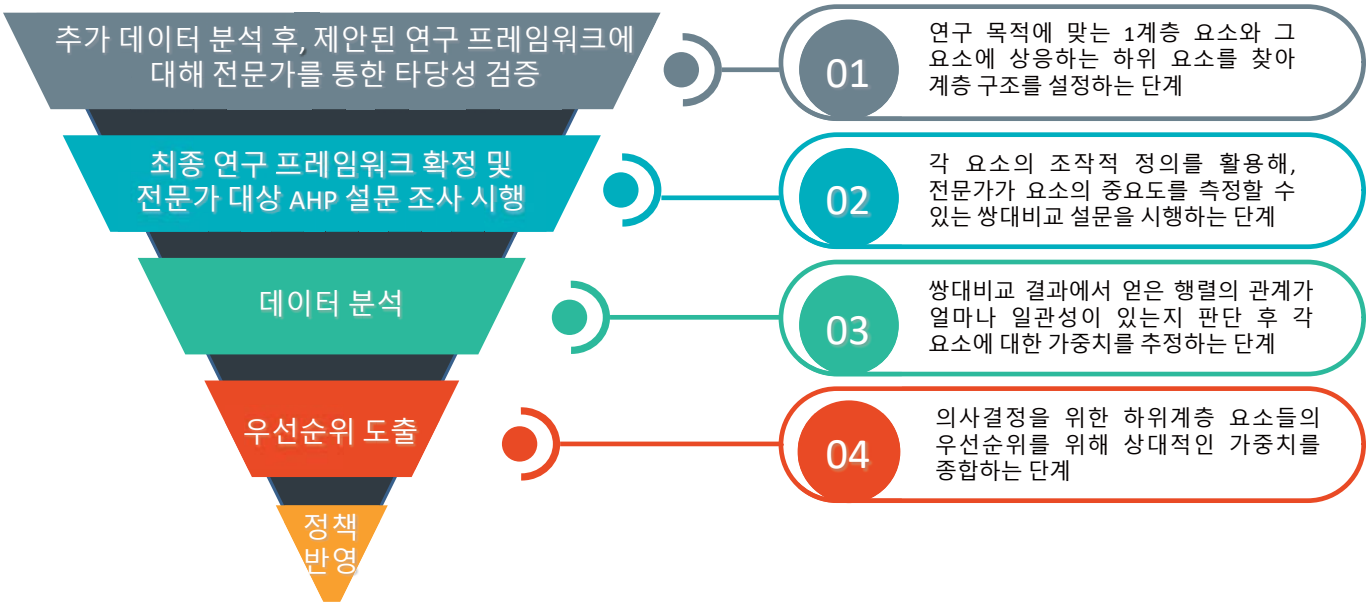
❑ 토픽 도출

토픽	키워드
문화예술 콘텐츠	문화, 콘텐츠, 사업, 관광, 예술, 산업, 구축, 계획, 센터, 한류
NFT	가상, 기업, 자산, 투자, 거래, 규제, 블록체인, 가능, 관련, 서비스
통신 / 네트워크	기술, 산업, 기업, 디지털, 정보, 통신, 개발, 연구, 서비스, 과학
메타버스 플랫폼	공간, 가상, 교육, 활용, 온라인, 제공, 아바타, 참여, 서비스, 소통
디지털 트랜스포메이션	디지털, 산업, 교육, 전환, 기술, 기업, 혁신, 경제, 전략, 뉴딜
블록체인 서비스	사업, 기업, 개발, 서비스, 투자, 기술, 성장, 계획, 구축, 블록체인
클라우드 서비스	데이터, 서비스, 디지털, 정보, 클라우드, 스마트, 기술, 구축, 시스템, 제공
교육 / 취업	정책, 경제, 청년, 산업, 사회, 혁신, 문제, 미래, 일자리, 교육
의료 / 바이오	의료, 기술, 개발, 병원, 기업, 바이오, 백신, 헬스, 케어, 건강

연구 프레임워크 제안(안)



연구 계획



Ⅳ 기대하는 시사점

기대하는 시사점

학술적 시사점

빅데이터를 활용한 연구와 기존의 사회과학적인 연구를 융합하여, 각 연구방법의 장점을 살리고 단점을 보완할 수 있음

Sometrend의 분석 및 시각화 기술과 데이터를 학술적 연구에 활용

여론을 정책에 반영할 수 있는 통합적인 연구 프레임워크로 활용 가능

실무적 시사점

디지털 플랫폼 정부의 중요한 수단이 될 공공 메타버스 플랫폼에 대한 여론을 파악하고, 향후 정책 수립 시 우선 순위를 파악할 수 있음

새로운 일자리 창출과 창업을 기획하는 이해관계자들에게 정책의 우선순위를 알리고 함께 만들어 나갈 수 있음

감사합니다

빅데이터 기술 교류 세미나

빅데이터와 여론조사

발표3

Sometrend “유튜브 분석”을 활용한 20대 대선 여론 분석

양성병 (경희대 교수)

한국경영정보학회
The Korea Society of Management Information Systems

VAIV 경희대학교 KYUNG HEE UNIVERSITY

빅데이터 기술 교류 세미나:

빅데이터와 여론조사

Sometrend "유튜브 분석"을 활용한 20대 대선 여론분석

2022. 05. 19

경희대학교 빅데이터응용학과 양성병 교수
경희대학교 빅데이터응용학과 강은경 박사과정
경희대학교 빅데이터응용학과 양선욱 석사과정
경희대학교 빅데이터응용학과 권지윤 석사과정

VAIV 경희대학교

목차

VAIV 경희대학교

기존 여론조사의 한계(1/2)

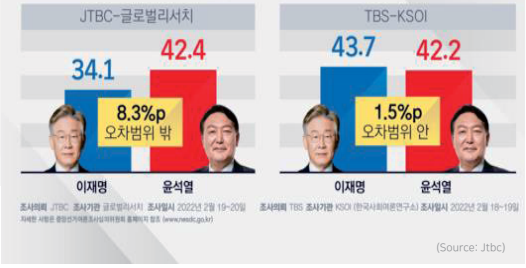
여론조사는...
선거운동의 강력한 수단이자, 언론의 가장 중요한 기사거리로 자리잡음(권혁남, 2001)

- 각 정당은 공천의 주요 기준으로, 후보들은 선거운동의 전략 수단으로 활용
- 유권자들은 당선 가능성이 있는 후보와 가능성이 없는 후보에 대한 정보를 제공받아 후보 선택에 도움을 받음

But, 지속적으로 제기되는 **여론조사의 문제점**(김양원 & 송경재, 2021)

- 여론조사 언론 보도가 국민들에게 혼란 야기
- 여론조사 기관의 전문성 및 신뢰성 문제
- 후보들의 공약과 정책에 대한 검증보다, 당선 가능성이나 지지도에 관한 여론조사만 반복적으로 실시(경마식 보도)

2022년 2월 21일 발표 여론조사 대선후보 지지도 (%)



VAIV 경희대학교

기존 여론조사의 한계(2/2)

대선 전 100일간 소셜 분석 긍·부정 언급 비율

'여론조사의 한계'를 보완하고자...
여론조사 메타분석, 검색어 트렌드 분석, 소셜 분석 등 시도(임예인, 2022)

- MBC, SBS 등 여론조사 메타분석 사이트는 동적선형모형을 이용, 관측방정식과 상태방정식으로 분해한 후 Bias가 최소가 되는 여론 추세의 움직임을 추출
- 구글, 네이버, 카카오 등 검색엔진의 검색어 트렌드 분석을 이용해 여론의 흐름을 알아보려는 움직임
- 트위터, 블로그, 커뮤니티, 인스타그램 등 소셜 분석을 이용해 여론의 흐름을 알아보려는 움직임

여론조사 메타분석: "콩심은 데 콩난다!"

- 다양한 여론조사 결과를 종합하는 것으로 여론조사 결과가 잘못될 경우, 메타분석 결과의 신뢰성도 떨어짐

검색어 트렌드 분석: "악플이 무플보다 좋다?"

- 단순 키워드 검색량(키워드 서치 & 클릭)으로만 여론을 예측하므로, 긍·부정 검색 목적을 걸러내지 못하는 한계
- 과도한 마케팅, 정치적 '실검전쟁' 등의 폐해로 2021년 2월 네이버와 다음 카카오가 실검 서비스 폐지(이지민, 2021)

소셜 분석: "소통·불통·일방통행 SNS!"

- 인플루언서들의 일방성과 폐쇄성으로 개인적인 의견, 특히 부정적 의견이 많음("악플의 배설구")
- 개인의 정치적인 의견보다 특정 정당이나 후보를 지지하는 용도로 사용하는 경우가 많고, 동일한 데이터가 다수 생성(예: 리트윗)되는 문제로 자료수집 및 처리에 한계(하상현 & 노태협, 2020)

Source: WISEAPP (2019)

WHY 유튜브?

Source: dokdok (2021)

01 2005년 탄생한 세계 최대 플랫폼 (이강유 & 성동규, 2018)

- 국내 온라인 동영상 이용자의 80%가 이용
- 검색 시장이 유튜브로 이동하면서 여론 선도의 패권이 바뀜(김종훈, 2019)

02 Mehrabian(1981)의 "821 법칙"

- **821 법칙: 눈으로 본 것 80%**, 읽은 것 20%, 들은 것 10% 기억(Mehrabian, 1981)
- 시각정보로 가득 찬 동영상에 달린 댓글의 내용적 특성과 댓글망의 패턴은 선거캠페인의 효과 측정에 큰 가치(김찬우 외, 2017)

03 콘텐츠 소비가 전파수신 중심에서 미디어 플랫폼 중심으로 넘어 감(박상현 외, 2020)

- **50대 이상**: 재테크, 부동산, 정치적 성향의 콘텐츠가 대폭 늘어남
- **MZ 세대:** "스낵컬쳐" 현상으로 텍스트보단 이미지, 동영상 선호

04 유튜브 정치·시사 채널: 높은 이용률에 주목하여 이용자 관점에서 현상 분석 필요(박상현 외, 2020)

- '신의 한수': 신혜식 독립신문 대표가 운영, 구독자 145만 명(2022년 5월 현재)
- '딴지방송국': 김어준 딴지일보 편집장이 운영, 구독자 101만 명(2022년 5월 현재)
- '이재명': 계양구을 국회의원 보궐선거 후보 운영, 63.5만 명(2022년 5월 현재) / 대한민국 정치인중 누적조회수 1억뷰 돌파
- '윤석열': 윤석열 대통령 공식 채널, 45.9만 명(2022년 5월 현재)

유튜브를 활용한 여론분석

학계에서도 유튜브를 활용해 여론을 분석하려는 시도가 일부 있어 왔으나,
빅데이터 활용을 통해 전체 여론을 예측해 보려는 시도 보다는 일부 대표 동영상 및 관련 댓글 활용에만 그침

Sometrend Biz !

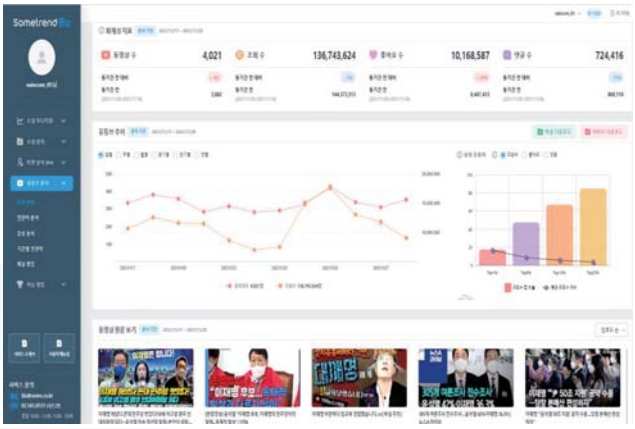
저자(연도)	연구방법	연구내용
Krishna et al. (2013)	감성분석(Naïve Bayes 분류 기법 및 Weka 활용)	웹 '버즈', 주식시장 동향, 박스오피스 결과 및 정치적인 선거와 관련된 400만 개 이상의 각 유튜브 동영상에서, 동영상당 1,000개의 댓글을 수집하고, 'Federer', 'Nadal', 'Obama'와 같은 특정 키워드에 대한 데이터와 timestamp, 작성자 이름 수집 후, Naïve Bayes 분류 기법을 사용하여 감성분석 진행 ⇒ **키워드 관련 감성의 추세와 실제 이벤트 간의 상관관계 식별**
Shevtsov et al. (2020)	감성분석	2020년 미국 대통령 선거 기간(2020년 7월 ~ 9월까지 6가지 다른 시점) 중 가장 인기 있는 헤시태그를 이용 750만개의 트위터 데이터와 유튜브 동영상 및 메타데이터(좋아요, 댓글, 작성자 등)를 추출하여 감성분석 실시 ⇒ **트럼프에 대한 긍정적인 감성(트위터: 45.7%, 유튜브: 14.55%)이 바이든에 대한 긍정적인 감성(트위터: 33.8%, 유튜브: 8.7%) 보다 높게 나타남**
김찬우 외 (2017)	댓글망 분석, 댓글 단어 빈도 분석, 단어쌍분석	2017년 대통령 후보자 이름과 후보수락 연설문을 검색어로 설정하여 선별한 유튜브 동영상 중 각 후보별로 조회수나 댓글수가 가장 많은 동영상을 각 2편 씩 선정(단, 홍준표 후보는 수집 과정에서 두번째 영상이 삭제되어 한 편만 분석)하여 댓글을 수집하고, 댓글망 분석, 댓글 단어 빈도 분석, 단어 쌍 분석을 실시 ⇒ **대선 메시지가 댓글상에 나타나는 것이 중요**, 유튜브 연구 및 선거캠페인 분석에 새로운 분석지표와 연구방향 제시
송화영 외 (2020)	오피니언 마이닝(감성분석), 단어빈도 분석, 의미연결망 분석	2020년 총선 공식 선거 운동 기간(4/2~4/14) 중에 4개 정당(더불어민주당, 미래통합당, 더불어시민당, 미래한국당)의 업로드 된 유튜브 선거 캠페인에 대한 대중의 반응을 영상목록, 조회 수, 댓글 수 등으로 수집하여 오피니언 마이닝(감성분석), 단어빈도 분석, 의미연결망 분석 등을 실시 ⇒ **유권자들의 반응과 앞으로의 정치적 선택, 행동 변화 등을 파악**

VAIV 경희대학교

연구 목적

VAIV Sometrend Biz 의 "유튜브 분석" 기능을 활용한 20대 대선 여론 분석 방법 제시

- 기존의 여론조사(여론조사 메타분석) 결과 보완
- 기존의 검색어 트렌드 분석 및 소셜 분석 결과 보완



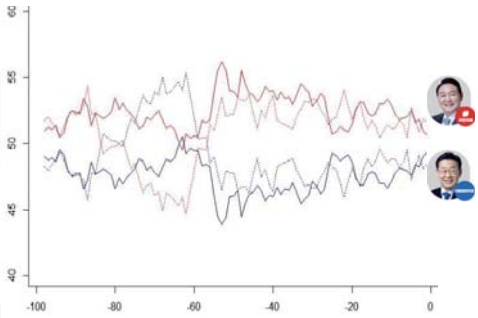
Study 1 "유튜브 분석"을 활용한 **20대 대선 결과 회고적 예측**
⇒ 결과를 기존의 여론조사, 검색어 트렌드 분석, 소셜 분석 결과와 비교

Study 2 "유튜브 분석" 데이터를 활용한 여론조사 "깜깜이 기간" 및 **대선 결과 예측모델** 구축
⇒ 예측모델 성능을 기존의 검색어 트렌드 데이터, 소셜 분석 데이터를 활용한 결과와 비교

VAIV 경희대학교

Study 1 개요

Study 1 "유튜브 분석"을 활용한 **20대 대선 결과 회고적 예측**
⇒ 결과를 기존의 여론조사, 검색어 트렌드 분석, 소셜 분석 결과와 비교



유튜브 분석 데이터

- 키워드 관련 **긍정/부정/중립** 언급량 데이터
- 키워드 관련 **동영상/조회/좋아요/댓글 수** 데이터

소셜 분석 데이터

- 트위터, 블로그, 커뮤니티, 인스타그램에서 키워드 관련 **긍정/부정/중립** 언급량 데이터

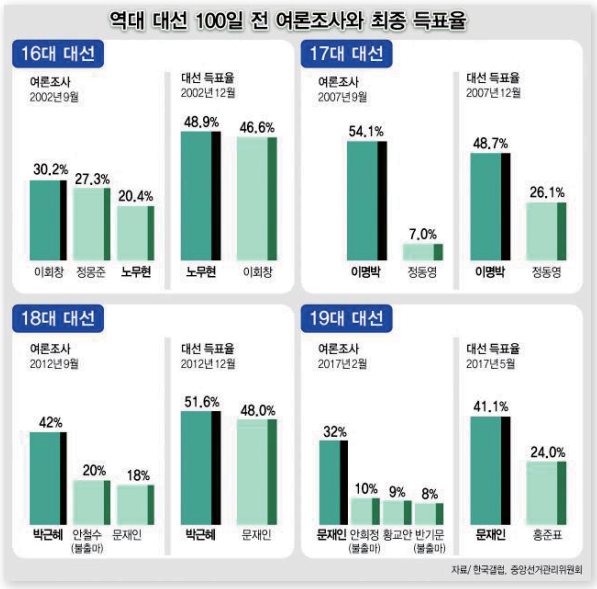
검색 트렌드 분석 데이터

- 구글 트렌드_웹 검색량
- 네이버 데이터랩_검색량
- 카카오 데이터 트렌드_검색량

여론조사 메타분석 데이터

- MBC '[여론M] 여론조사를 조사하다'의 여론조사 통합데이터 (http://poll-mbc.co.kr/)

데이터 수집 기간:
선거 100일 전



분석기간

2021.11.29 ~ 2022.03.08**(100일)**

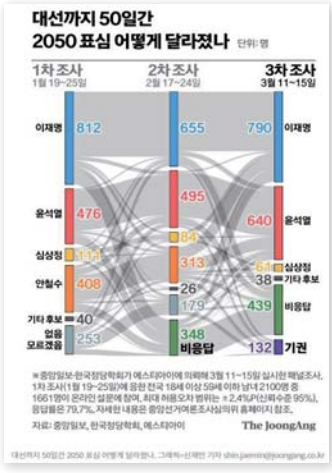
기간 설정이유

- 모든 후보자 선출 이슈 이후로 선정
- 충분한 데이터 확보 가능한 기간
- 모든 후보에게 일정기간 주요 이슈가 없는 기간을 시작점으로 선정
- 선거 100일 전 웃은 후보가 결국 이겼다(전창훈, 2021)

선거 100일 전 웃은 후보가 결국 이겼다

입력:2021-11-28 17:57:55 | 수정:2021-11-28 19:55:23 부산일보

데이터 수집 대상:
이재명 & 윤석열 양강구도(안철수 제외)



Source: 나무위키(2022)

The JoongAng 정치

▶ 단일화 발표 이후, 안철수 지지층 38.3%는 이재명, 37.7%는 윤석열 찍었다(고정애, 2022)

- "안철수 후보에 대한 60대 이상의 지지가 높지 않다고 추정한다면 막판 단일화 발표가 윤 후보의 당선에 미친 영향은 크지 않은 것으로 판단"(아주대 강신구 교수)

VAIV 경희대학교

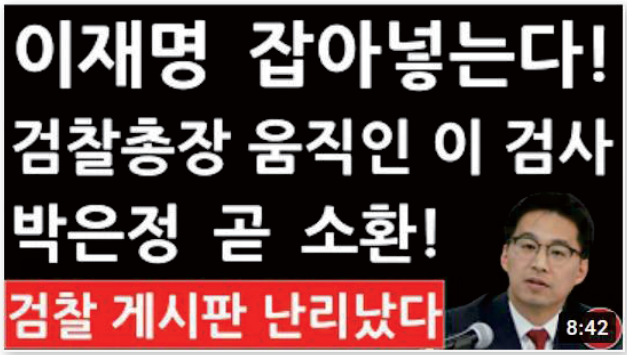
파생변수 생성:
긍정/부정 데이터 분류(1/2)

'이재명' 키워드 긍정 동영상

(OhmynewTV, 2022.02.20)

(https://www.youtube.com/watch?v=QLPLoq-Ipa8)

'이재명' 키워드 부정 동영상



(진성호방송, 2022.01.27)

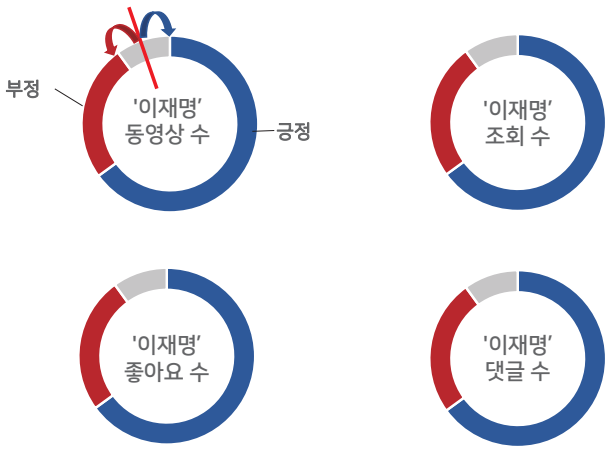
(https://www.youtube.com/watch?v=wU5M6_wz7Mw)

- 유튜브 동영상, 조회, 좋아요, 댓글 수가 꼭 긍정적으로 작용하지 않음
- 긍정적인 동영상의 조회, 좋아요, 댓글은 후보자에 대한 지지를, 부정적인 동영상의 조회, 좋아요, 댓글은 후보자에 대한 반대를 의미

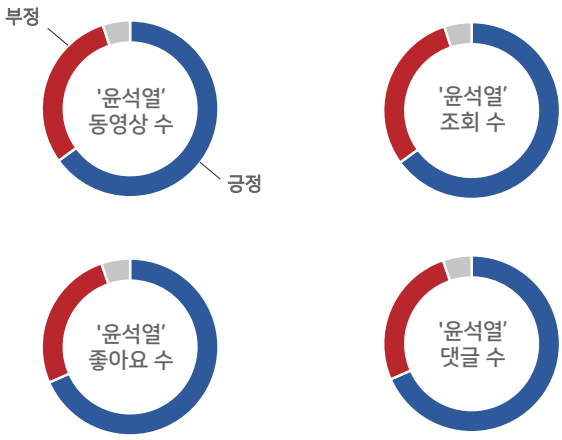
VAIV 경희대학교

파생변수 생성:
긍정/부정 데이터 분류(2/2)

'이재명' 키워드 긍·부정 비율



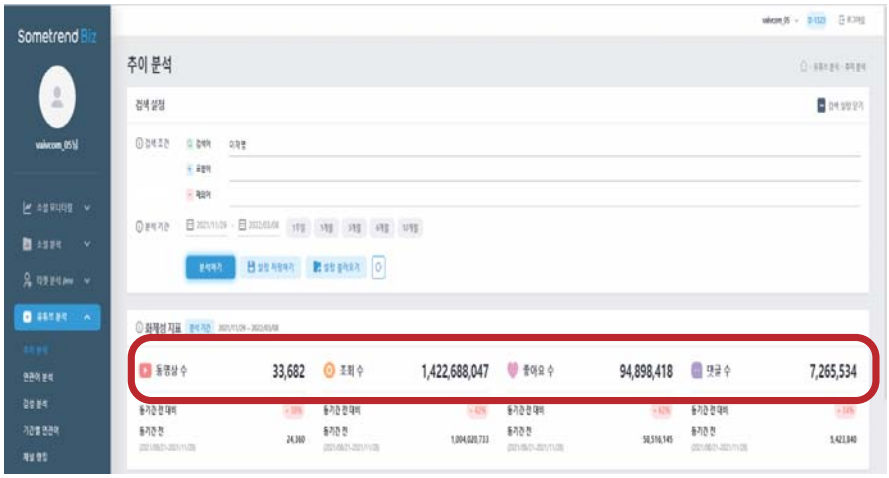
'윤석열' 키워드 긍·부정 비율

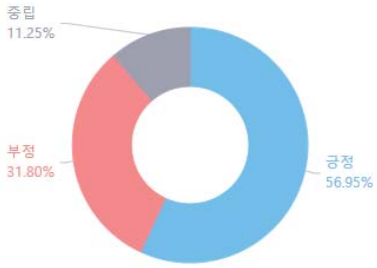


- 각 후보자별 키워드 긍·부정 비율에 맞춰 유튜브 동영상/조회/좋아요/댓글의 긍·부정 개수 파악

VAIV 경희대학교

파생변수 생성:
긍정/부정 데이터 분류(이재명)

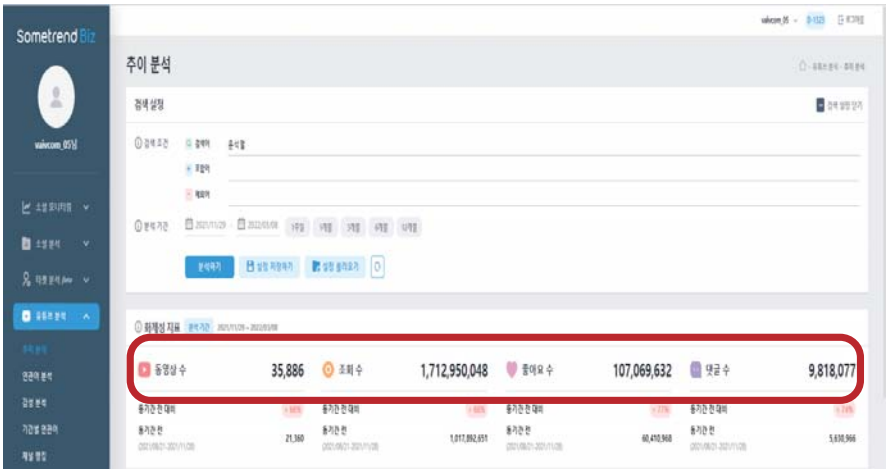


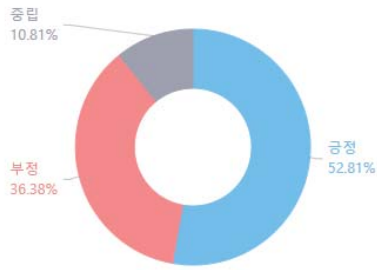


	동영상	조회	좋아요	댓글
긍정 데이터 수	21,077	890,247,045	59,382,485	4,546,408
부정 데이터 수	12,605	532,441,002	35,515,733	2,719,126

VAIV 경희대학교

파생변수 생성:
긍정/부정 데이터 분류(윤석열)

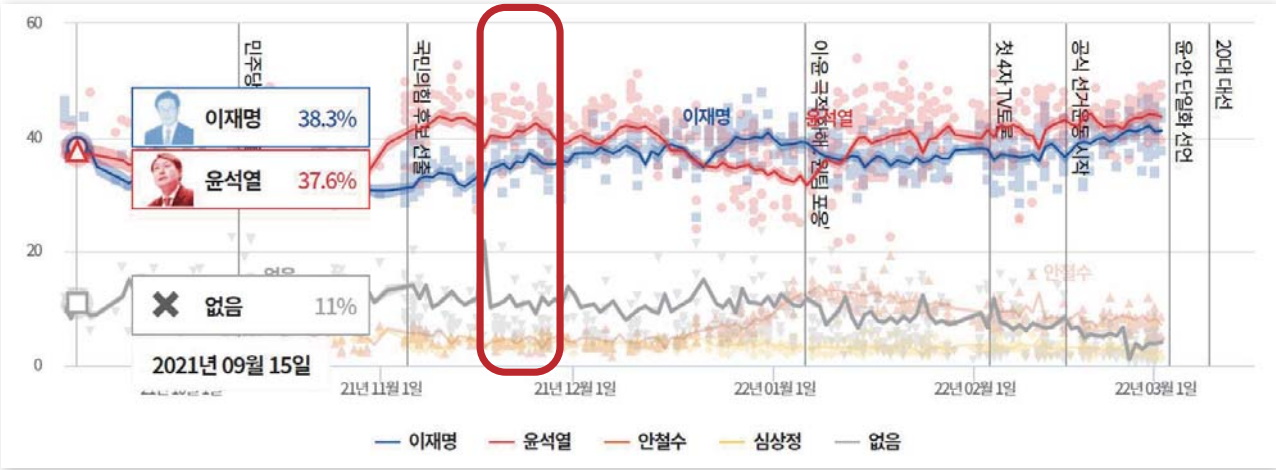




	동영상	조회	좋아요	댓글
긍정 데이터 수	20,891	997,193,930	62,330,587	5,715,594
부정 데이터 수	14,995	715,756,118	44,739,045	4,102,483

VAIV 경희대학교

파생변수 생성:
가중치 도출 및 파생변수 생성(1/2)



MBC '[여론M] 여론조사를 조사하다' (http://poll-mbc.co.kr/)

▶ **데이터 수집 직전 12일간의 데이터(21.11.17 ~ 21.11.28)**

- 두 후보 간 지지율의 격차가 평행하게 함께 움직이는 구간 선정
- 해당 기간 각 후보 별 데이터의 **조회 수** 대비 **동영상 수, 좋아요 수, 댓글 수** 비율을 가중치로 하여 선거일 전 100일 간의 일별 파생변수 값(긍정/부정 동영상/조회/좋아요/댓글 수) 계산

VAIV 경희대학교

파생변수 생성:
가중치 도출 및 파생변수 생성(2/2)

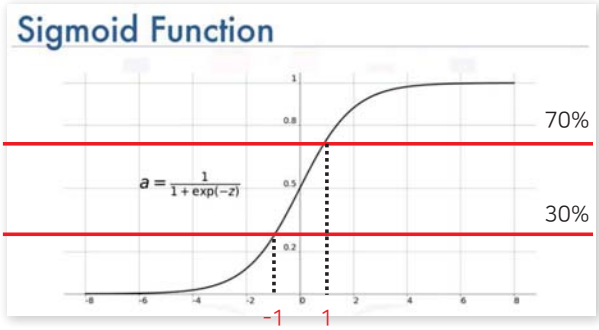
	동영상 수	조회 수	좋아요 수	댓글 수
이재명 가중치	34,007.37	1	13.45	188.76

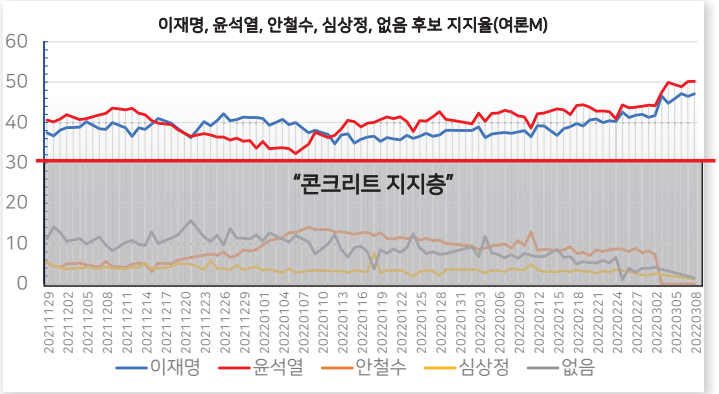
	동영상 수	조회 수	좋아요 수	댓글 수
윤석열 가중치	42,607.86	1	13.34	182.01

▶ More Passive Use **조회 < 좋아요 < 댓글 < 동영상** More Active Use(Yu, 2016)

▶ 가장 수동적인 반응인 **조회 수**를 가중치의 기준으로 삼아 후보 별 **동영상/좋아요/댓글 수**의 가중치 도출 ⇒ 선거일 전 100일 간의 일별 파생변수 값(긍정/부정 동영상/조회/좋아요/댓글 수) 계산

일별 지지율 수식 도출





▶ 지지율 수식

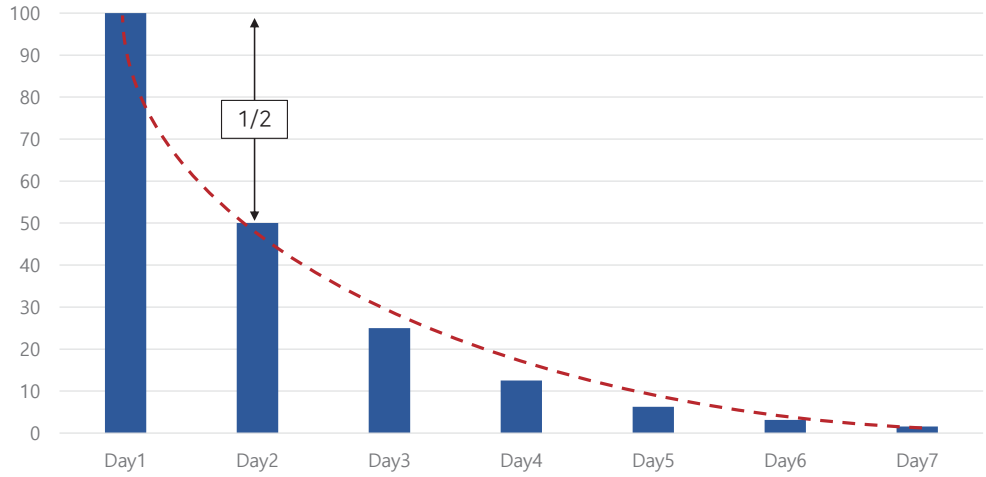
1차 수식: (긍정 데이터량* – 부정 데이터량)/전체 데이터량 x 100
⇒ 긍정 혹은 부정 데이터량이 극단일 경우, 지지율이 최소 -100%, 최대 100%까지 나오는 한계 발생

2차 수식: Sigmoid(1차 수식)
⇒ 긍정 혹은 부정 데이터량이 극단이더라도, 지지율이 최소 30%, 최대 70%로 제한(**콘크리트 지지층 30%** 고려)

* 데이터량 = $w_{동영상}$ x 동영상 수 + $w_{조회}$ x 조회 수 + $w_{좋아요}$ x 좋아요 수 + $w_{댓글}$ x 댓글 수 (w = 가중치)

누적 일별 지지율 수식 도출

Forgetting Curve

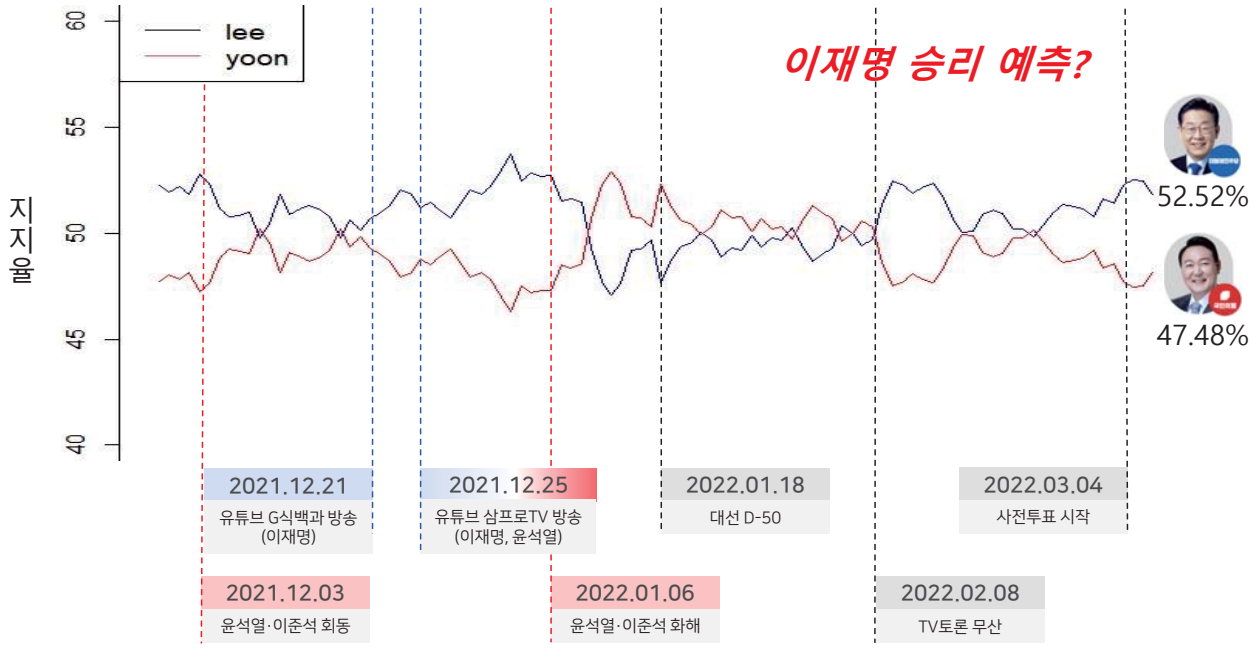


▶ 반감기 7일 누적 함수

- 이슈에 대한 기억은 하루가 지날 때마다 절반(50%)만 남게 되어, 7일이 지나면 1% 미만으로 남음
- 주요 이슈는 지속적으로 언급되어 사람들의 기억 계속 유지

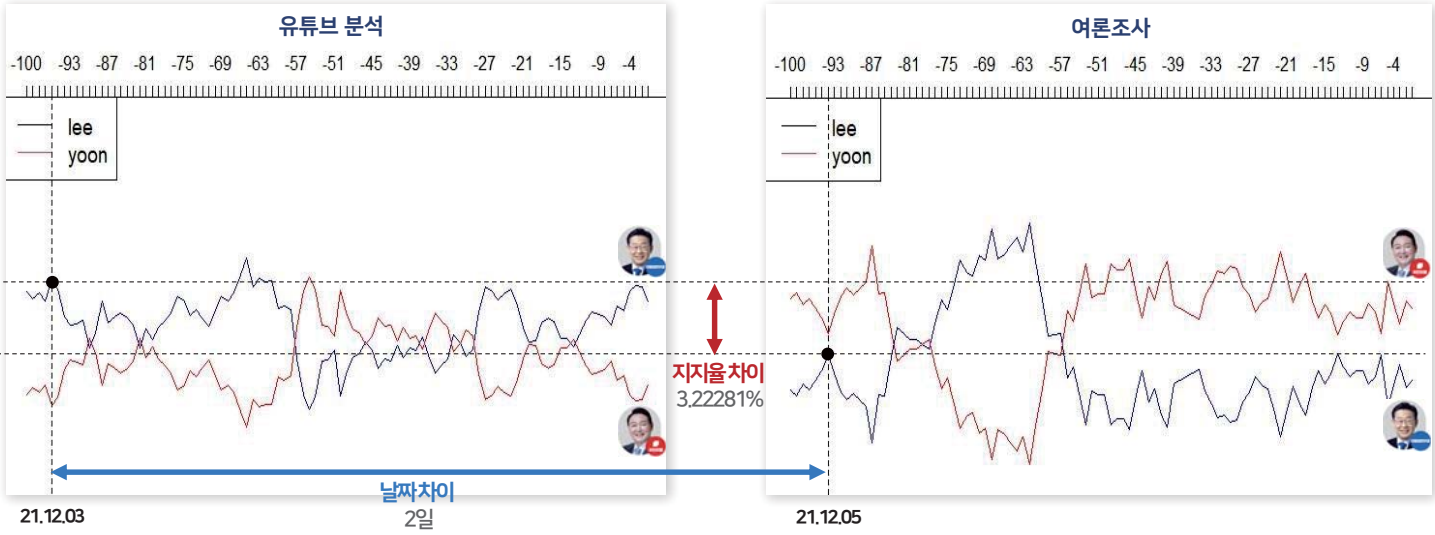
VAIV 경희대학교

유튜브 분석 결과:
누적 일별 지지율(보정 전)



VAIV 경희대학교

혹시, 결과 보정이 필요할까?

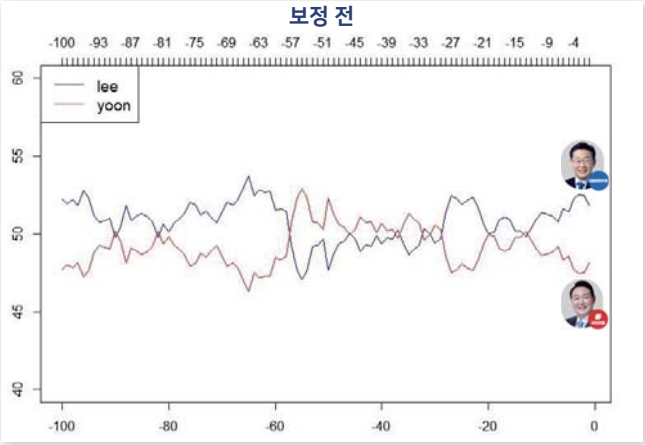


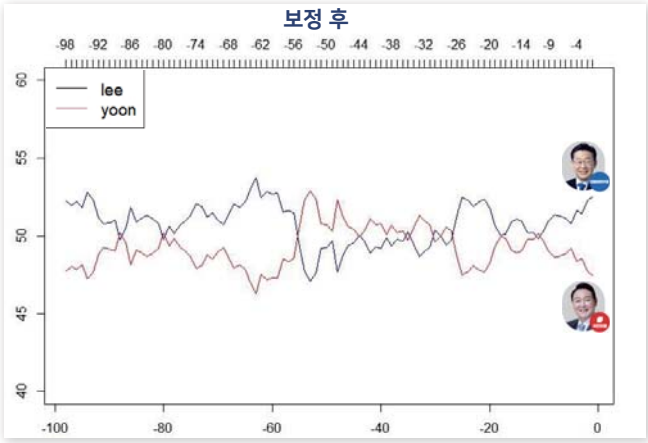
▶ 유튜브 분석 12월 3일 기준

- 12월 3일, 윤석열·이준석 회동 이후 두 후보자 지지율 급변
- 12월 3일 이후, 유튜브 분석과 여론조사의 지지율 일정기간 유사한 추세 유지

VAIV 경희대학교

결과 보정:
날짜 +2일



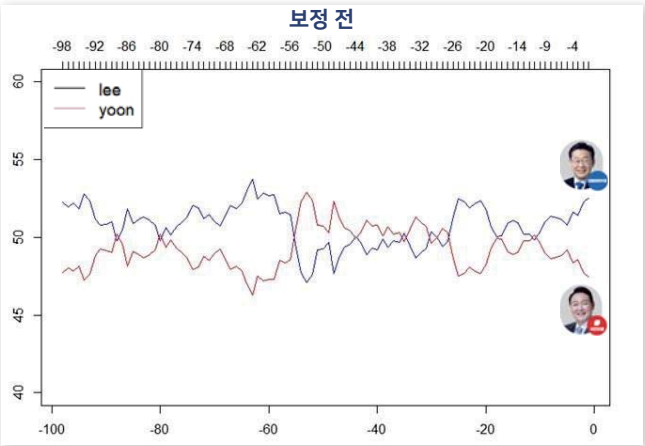


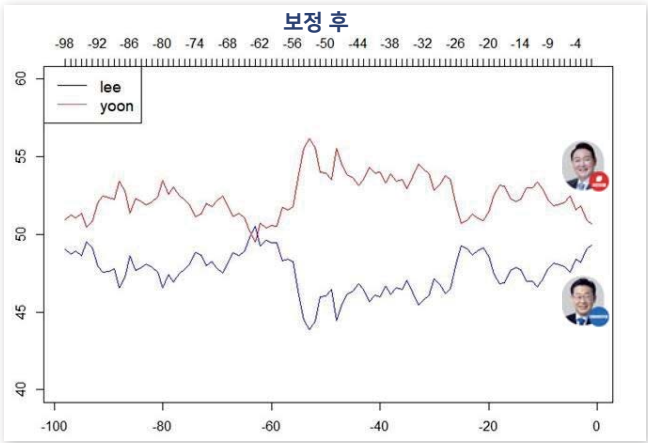
상관계수(r)	여론조사 결과
유튜브 분석 결과	0.4642
유튜브 분석+1일 결과	0.5110
유튜브 분석+2일 결과	0.5200
유튜브 분석+3일 결과	0.4933

유튜브 분석 결과가 여론조사 결과보다 "2일 선행"
- 여론조사는 1~3일 간격 조사 후, 다음날 발표
- 유튜브는 즉각(real-time)적인 여론 반영
- 즉, 유튜브 여론이 여론조사 결과에 영향을 미침

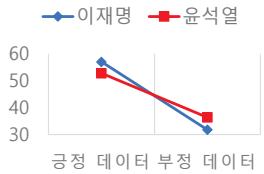
VAIV 경희대학교

결과 보정:
지지율 ± 3.22281





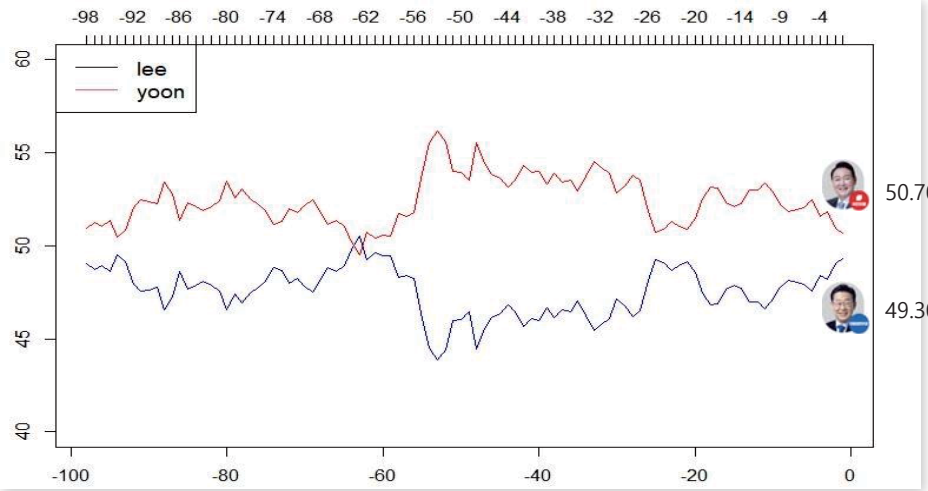
	이재명	윤석열
긍정 데이터 비율	57.0%	52.8%
부정 데이터 비율	31.8%	36.4%



유튜브는 여론조사에 비해 "진보쪽으로 기울어진 운동장"
- 진보 성향 콘텐츠를 생산·소비하는 비율이 상대적으로 높은 것으로 나타남
- 여론조사 결과에 비해 이재명 후보에게 3.22281% 유리한 결과 도출
- 진보쪽으로 기울어진 정도가 SNS에 비해 심하진 않음

VAIV 경희대학교

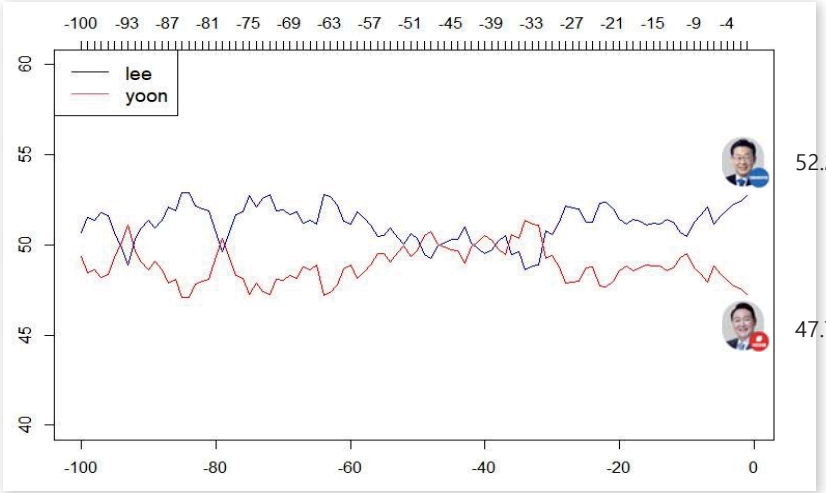
유튜브 분석 결과:
누적 일별 지지율(보정 후)



최종일 결과	이재명	윤석열
유튜브 분석(3월 6일)	48.18%	51.81%
유튜브 분석(3월 8일)	49.30%	50.70%
실제 대선(3월 9일) (양강 구도로 비중 조정)	49.62%	50.37%

VAIV 경희대학교

소셜 분석 결과:
누적 일별 지지율(보정 전)



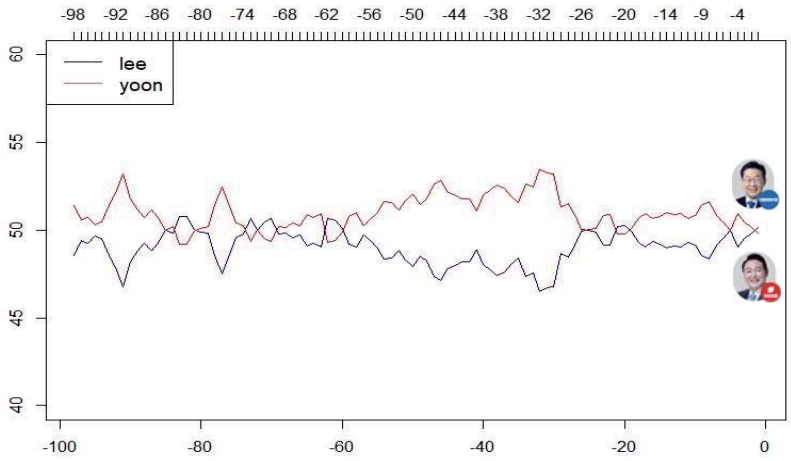
최종일 결과	이재명	윤석열
소셜 분석	52.27%	47.73%
실제 대선	49.62%	50.37%

1) 이재명, 윤석열 두 후보에 대한 긍정/부정/중립 키워드 언급량 데이터 수집
(트위터, 블로그, 커뮤니티, 인스타그램)
2) 지지율 수식에 맞춰 후보별 지지율 계산
3) 두 후보의 지지율 비교분석

VAIV 경희대학교

소셜 분석 결과:
누적 일별 지지율(보정 후)

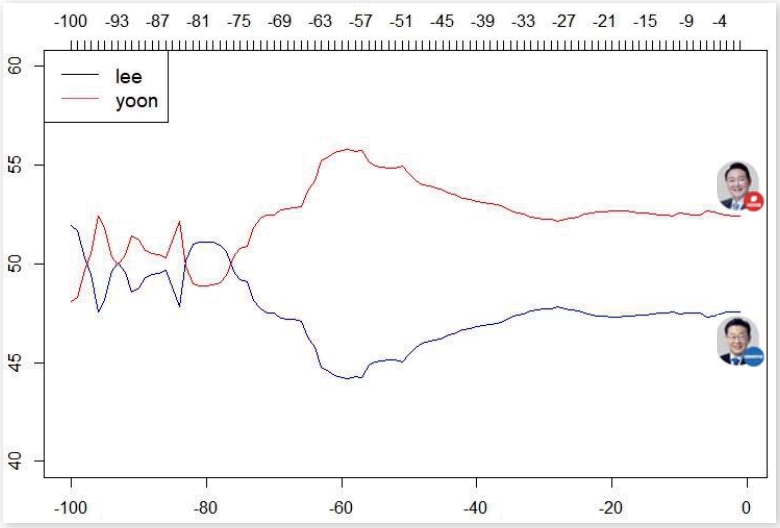
날짜 +2일
지지율 ± 2.104457



최종일 결과	이재명	윤석열
소셜 분석(3월 6일)	49.54%	50.46%
소셜 분석(3월 8일)	50.16%	49.83%
실제 대선(3월 9일) (양강 구도로 비중 조정)	49.62%	50.37%

VAIV 경희대학교

검색 트렌드 분석 결과*



※ 선거전 100일 기간동안 검색엔진 시장점유율 반영
- 네이버(54.14%),
- 구글(35.13%),
- 다음(5.93%)

* 반감기 7일 누적 함수 대신 전체 누적 함수 적용

최종일 결과	이재명	윤석열
검색 트렌드 분석	47.58%	52.42%
실제 대선	49.62%	50.37%

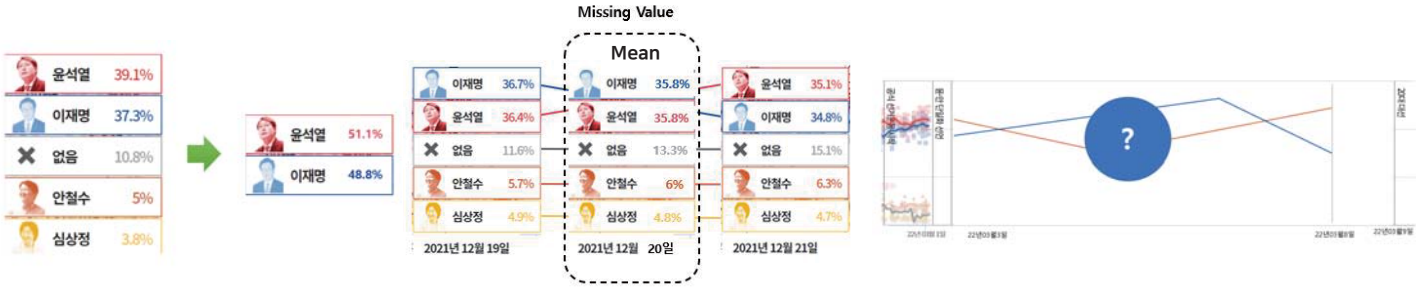
1) 이재명, 윤석열 두 후보에 대한 구글, 네이버, 다음(카카오) 검색량 데이터 수집
2) 데이터 수집기간 동안의 검색엔진 시장점유율 반영하여 가중합산
3) 두 후보의 지지율 비교분석

여론조사 분석 절차

후보 간소화(양강) 및 비율 조정

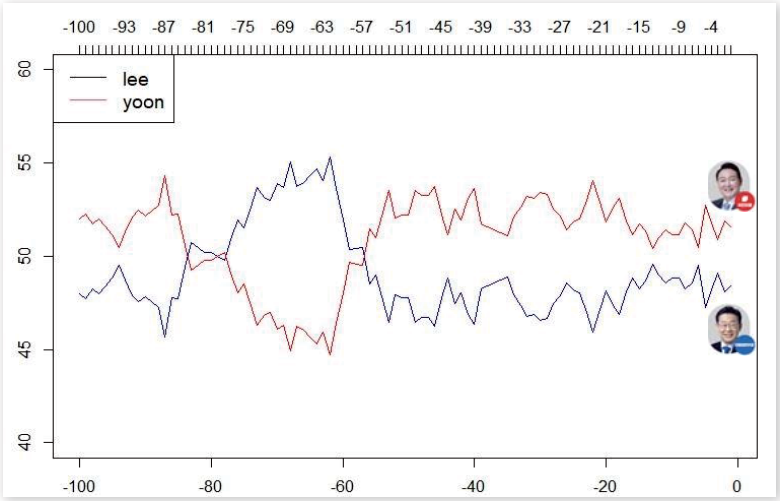
결측값 보완

지지율 그래프



1) 여론조사 통합 결과 데이터 수집(여론M)
2) 후보 간소화 및 비율 조정
3) N일의 결측값은 N-1일과 N+1의 평균으로 보완
4) 여론조사 공표 금지기간 데이터 보완(리얼미터)
5) 두 후보의 지지율 그래프 생성

여론조사 분석 결과

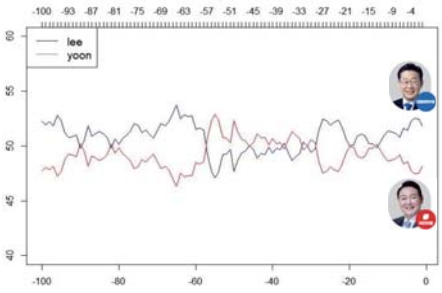


51.59%

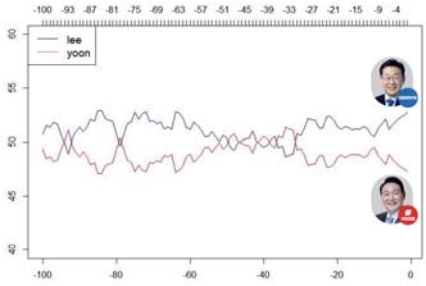
48.41%

최종일 결과	이재명	윤석열
여론조사	48.41%	51.59%
실제 대선	49.62%	50.37%

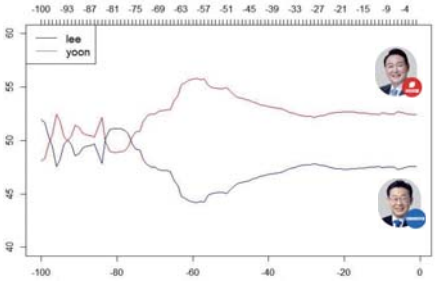
보정 전 결과 비교



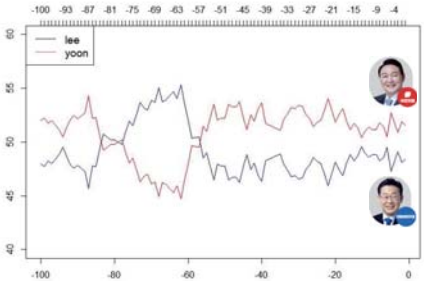
유튜브 분석



소셜 분석



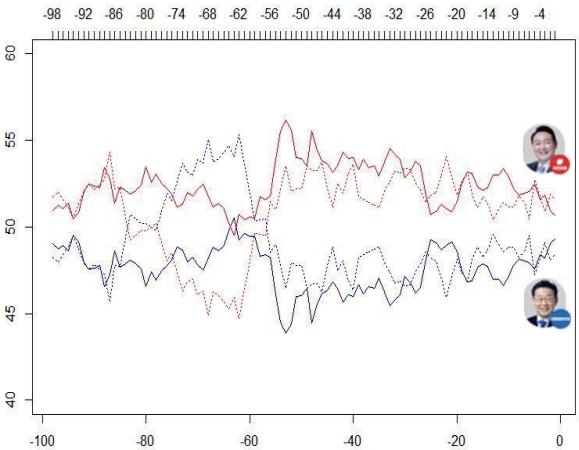
검색 트렌드 분석



여론조사

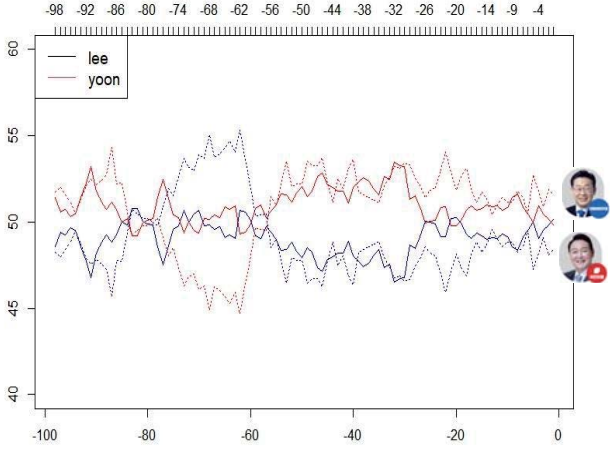
최종일 결과	이재명	윤석열
유튜브 분석(3/8)	52.52%	47.48%
소셜 분석(3/8)	52.27%	47.73%
검색 트렌드 분석 (3/8)	47.58%	52.42%
여론조사(3/8)	48.41%	51.59%
실제 대선(3/9)	49.62%	50.37%

보정 후 결과 비교



유튜브 분석

※ 유튜브 분석(실선) / 소셜 분석(실선) / 여론조사(점선)



소셜 분석

상관계수(r)	유튜브 분석	소셜 분석
여론조사	0.5106	0.5124

	유튜브 분석	소셜 분석
RMSE	2.5524	2.0113

최종일 결과	이재명	윤석열
유튜브 분석(3/8)	49.30%	50.70%
소셜 분석(3/8)	50.16%	49.83%
여론조사(3/8)	48.41%	51.59%
실제 대선(3/9)	49.62%	50.37%

Study 1 결과 요약

VAIV Sometrend Biz 활용

결과

- 긍·부정 동영상/조회/좋아요/댓글 수 변수만을 활용한 **유튜브 분석 결과는 기존 다른 방법의 결과를 충분히 보완**할 수 있음
- 유튜브 분석 결과만으로도 실제 선거 결과를 성공적으로 예측할 수 있음(p. 30)
- 유튜브 분석 결과와 소셜 분석 결과는 여론조사 메타분석 결과와 비슷한 패턴(높은 상관관계)과 일치도(낮은 RMSE)를 보임(p. 30)
- 유튜브 분석 결과는 실제 여론의 흐름을 실시간(real-time)으로 알려주는 반면, 여론조사 결과는 약 2일 전의 여론을 공표(p. 21)
- 유튜브 채널과 SNS 채널은 실제 여론(혹은 여론조사 결과)에 비해 상대적으로 진보쪽으로 편향(p. 29)
- 반면 포털 서비스 채널(검색 트렌드 분석)은 실제 여론(혹은 여론조사 결과)에 비해 상대적으로 보수쪽으로 편향(p. 29)

시사점

- 유튜브 등 동영상 플랫폼의 비중이 점차 커지는 시점에서, 간단한 유튜브 분석 결과만으로도 실제 여론(혹은 여론조사 결과)에 근접한 결과를 쉽게 얻을 수 있을 것으로 기대
- 유튜브 분석에 동영상 콘텐츠에 대한 댓글 분석을 추가할 수 있으면, 좀더 정확하고 유용한 결과가 도출될 것으로 기대
- 유튜브 분석 결과에 소셜 분석 결과, 검색 트렌드 분석 결과 등을 종합적으로 분석한다면, 실제 여론(혹은 여론조사 결과)에 가장 근접한 결과를 도출할 수 있을 것으로 기대

Study 2 개요

Study 2

"유튜브 분석" 데이터를 활용한 여론조사 "깜깜이 기간" 및 **대선 결과 예측모델** 구축
⇒ 예측모델 성능을 기존의 검색어 트렌드 데이터, 소셜 분석 데이터를 활용한 결과와 비교

유튜브 데이터

머신러닝 모델

- 후보자 관련 데이터와 여론조사 결과 데이터
- K-NN, 의사결정나무, ANN 등 다양한 모델 학습

"깜깜이 기간" 예측

- 22.03.03 ~ 22.03.08 기간의 여론 예측 및 성능평가

학습 데이터

- 유튜브 분석 데이터
- 소셜 분석 데이터
- 검색 트렌드 분석 데이터(유튜브 검색 제외)
- 검색 트렌드 분석 데이터(유튜브 검색 포함)

대선결과 예측

- 22.03.09 대선결과 예측 및 성능평가

머신러닝 모델 학습

모델 학습 기간	예측 기간
Train Data Set 80% Test Data Set 20%	깜깜이 기간 + 대선 결과
독립변수: 21.11.29 ~ 독립변수: 22.02.28 종속변수: 21.12.01 ~ 종속변수: 22.03.02	22.03.03 ~ 22.03.09

학습에 사용된 모델 목록

모델명	설명
표준 선형 회귀모형	성능 대조군
릿지 선형 회귀모형	고차원 데이터 세트에서 사용, 라쏘와 다르게 가중치를 유지함(성능대조군)
K-NN	다중종속변수를 학습할 수 있는 모델
의사결정나무	다중종속변수를 학습할 수 있는 모델
SVM	수치 예측이 가능한 SVR(Support Vector Regression)이 존재
ANN	다층 퍼셉트론을 이용한 비선형 문제 해결 가능
앙상블 모형	성능이 좋은 모델 세 가지를 선정하여 보팅(voting) 방식으로 학습

변수 설명:
유튜브 데이터

Date	후보자	독립변수							종속변수
2021.11.29 ~ 2022.02.28	이재명	긍정 언급량	부정 언급량	중립 언급량	동영상 수	동영상 조회 수	좋아요 수	댓글 수	2일 뒤 여론조사 지지율
	윤석열	긍정 언급량	부정 언급량	중립 언급량	동영상 수	동영상 조회 수	좋아요 수	댓글 수	2일 뒤 여론조사 지지율

변수명	설명
긍정 언급량	후보자가 포함된 해당 날짜 유튜브 동영상의 제목, 설명에서 분석된 긍정적인 언급의 수
부정 언급량	후보자가 포함된 해당 날짜 유튜브 동영상의 제목, 설명에서 분석된 부정적인 언급의 수
중립 언급량	후보자가 포함된 해당 날짜 유튜브 동영상의 제목, 설명에서 분석된 중립적인 언급의 수
동영상 수	후보자가 포함된 해당 날짜 유튜브 동영상의 수
동영상 조회 수	후보자가 포함된 해당 날짜 유튜브 동영상의 조회 수
좋아요 수	후보자가 포함된 해당 날짜 유튜브 동영상의 좋아요 수
댓글 수	후보자가 포함된 해당 날짜 유튜브 동영상의 댓글 수

VAIV 경희대학교

성능 평가 및 결과:
유튜브 데이터

	모델 RMSE	후보	03/03(목)	03/04(금)	03/05(토)	03/06(일)	03/07(월)	03/08(화)	03/09(대선)	예측 RMSE
일반 선형회귀	1.9573	이재명	45.73*	49.68	46.93	46.94	47.47	46.99	48.52	2.0747
		윤석열	54.27	50.32	53.07	53.06	52.53	53.01	51.48	
릿지 회귀모형	1.9390	이재명	45.89	49.96	47.28	46.93	47.74	47.41	48.32	2.0303
		윤석열	54.11	50.04	52.72	53.07	52.26	52.59	51.68	
K-NN	1.6056	이재명	47.86	48.57	48.55	48.28	48.57	48.40	48.55	0.9754
		윤석열	52.14	51.43	51.45	51.72	51.43	51.60	51.45	
의사결정 나무	2.3056	이재명	48.21	48.66	48.58	48.58	48.21	48.58	47.82	1.0269
		윤석열	51.79	51.34	51.42	51.42	51.79	51.42	52.18	
SVR	1.9679	이재명	46.87	49.42	46.66	46.54	47.71	47.43	48.25	1.8352
	1.9291	윤석열	53.13	50.58	53.34	53.46	52.29	52.57	51.75	1.8397
ANN	1.9223	이재명	45.95	50.34	47.55	46.76	48.03	47.92	48.20	2.0844
		윤석열	54.05	49.66	52.45	53.24	51.97	52.08	51.80	
앙상블	1.7834	이재명	47.28	49.03	47.75	48.04	48.10	47.72	47.96	1.3459
	1.7329	윤석열	52.72	50.97	52.25	51.96	51.90	52.28	52.04	1.3501
여론 조사		이재명	49.52	47.27	48.20	49.11	48.08	48.40	49.62**	
		윤석열	50.47	52.72	51.80	50.88	51.91	51.59	50.37**	

* 결과는 이재명 후보 지지율+윤석열 후보 지지율=100%로 환산된 값
** 03/09(대선)의 투표 결과를 100%합산으로 환산한 값

VAIV 경희대학교

변수 설명:
소셜 데이터

Date	후보자	독립변수								종속변수
2021.11.29 ~ 2022.02.08	이재명	트위터				블로그				2일 뒤 여론조사 지지율
		게시글 수	긍정 언급량	부정 언급량	중립 언급량	게시글 수	긍정 언급량	부정 언급량	중립 언급량	
		커뮤니티				인스타그램				
		게시글 수	긍정 언급량	부정 언급량	중립 언급량	게시글 수	긍정 언급량	부정 언급량	중립 언급량	
	윤석열	트위터				블로그				2일 뒤 여론조사 지지율
		게시글 수	긍정 언급량	부정 언급량	중립 언급량	게시글 수	긍정 언급량	부정 언급량	중립 언급량	
		커뮤니티				인스타그램				
		게시글 수	긍정 언급량	부정 언급량	중립 언급량	게시글 수	긍정 언급량	부정 언급량	중립 언급량	

변수명	설명
긍정 언급량	후보자가 포함된 해당 날짜 해당 소셜 채널에서 분석된 긍정적인 언급의 수
부정 언급량	후보자가 포함된 해당 날짜 해당 소셜 채널에서 분석된 부정적인 언급의 수
중립 언급량	후보자가 포함된 해당 날짜 해당 소셜 채널에서 분석된 중립적인 언급의 수
게시글 수	후보자가 포함된 해당 날짜 해당 소셜 채널의 총 게시글의 수

VAIV 경희대학교

성능 평가 및 결과:
소셜 데이터

	모델 RMSE	후보	03/03(목)	03/04(금)	03/05(토)	03/06(일)	03/07(월)	03/08(화)	03/09(대선)	예측 RMSE
일반 선형회귀	2.6943	이재명	46.96	49.58	62.70	59.94	52.20	49.99	51.68	7.2013
		윤석열	53.04	50.41	37.30	40.06	47.80	50.01	48.32	
릿지 회귀모형	1.8474	이재명	48.77	48.72	55.44	52.43	49.79	48.59	49.75	3.1401
		윤석열	51.23	51.28	44.56	47.57	50.21	51.41	50.25	
K-NN	1.8142	이재명	48.68	48.52	48.66	48.66	48.66	48.66	48.66	0.7579
		윤석열	51.32	51.47	51.33	51.33	51.33	51.33	51.33	
의사결정 나무	2.2115	이재명	48.55	48.55	48.55	48.55	48.55	48.55	48.55	0.7938
		윤석열	51.45	51.45	51.45	51.45	51.45	51.45	51.45	
SVR	1.9686	이재명	47.18*	50.15	52.50	55.97	50.67	50.41	47.70	3.6597
	1.9726	윤석열	52.82	49.85	47.50	44.03	49.33	49.59	52.30	3.6558
ANN	1.8659	이재명	49.17	49.42	58.84	53.84	50.13	47.77	50.30	4.5560
		윤석열	50.83	50.58	41.16	46.16	49.87	52.23	49.70	
앙상블	1.9318	이재명	47.92	49.26	53.83	52.91	49.81	49.33	50.22	2.8504
	1.8927	윤석열	52.08	50.74	46.17	47.09	50.19	50.67	49.78	2.8467
여론 조사		이재명	49.52	47.27	48.20	49.11	48.08	48.40	49.62**	
		윤석열	50.47	52.72	51.80	50.88	51.91	51.59	50.37**	

* 결과는 이재명 후보 지지율+윤석열 후보 지지율=100%로 환산된 값
** 03/09(대선)의 투표 결과를 100% 합산으로 환산한 값

VAIV 경희대학교

변수 설명:
검색 트렌드 데이터

Date	후보자	독립변수				종속변수
2021.11.29 ~ 2022.02.28	이재명	구글 트렌드 웹 검색	구글 트렌드 유튜브 검색	네이버 데이터랩	카카오 데이터트렌드	2일 뒤 여론조사 지지율
	윤석열	구글 트렌드 웹 검색	구글 트렌드 유튜브 검색	네이버 데이터랩	카카오 데이터트렌드	2일 뒤 여론조사 지지율

변수명	설명
구글 트렌드 웹 검색	구글 트렌드에서 조사된 해당 날짜에 후보자가 웹에서 검색된 횟수
구글 트렌드 유튜브 검색	구글 트렌드에서 조사된 해당 날짜에 후보자가 유튜브에서 검색된 횟수
네이버 데이터랩	네이버 데이터랩에서 조사된 해당 날짜에 후보자가 네이버에서 검색된 횟수
카카오 데이터트렌드	카카오 데이터트렌드에서 조사된 해당 날짜에 후보자가 다음에서 검색된 횟수

VAIV 경희대학교

성능 평가 및 결과:
검색 트렌드 데이터(유튜브 검색 제외)

	모델 RMSE	후보	03/03(목)	03/04(금)	03/05(토)	03/06(일)	03/07(월)	03/08(화)	03/09(대선)	예측 RMSE
일반 선형회귀	2.3581	이재명	49.45	51.00	49.67	47.62	46.27	47.32	48.15	1.8869
		윤석열	50.55	49.00	50.33	52.38	53.73	52.68	51.85	
릿지 회귀모형	2.2916	이재명	49.19	50.44	50.12	47.97	46.62	47.36	48.44	1.6805
		윤석열	50.81	49.56	49.88	52.03	53.38	52.64	51.56	
K-NN	2.3757	이재명	48.62	48.44	49.12	48.73	48.73	48.44	48.44	0.8439
		윤석열	51.38	51.56	50.88	51.27	51.27	51.56	51.56	
의사결정 나무	2.6285	이재명	48.79	49.57	46.78	46.78	48.23	48.79	48.23	1.4820
		윤석열	51.21	50.43	53.22	53.22	51.77	51.21	51.77	
SVR	2.2783	이재명	48.65*	49.73	50.28	47.87	46.70	47.05	48.74	1.5633
	2.2495	윤석열	51.35	50.27	49.72	52.13	53.30	52.95	51.26	1.5693
ANN	2.2547	이재명	48.92	50.02	50.23	48.01	46.61	47.17	48.54	1.6089
		윤석열	51.08	49.98	49.77	51.99	53.39	52.83	51.46	
앙상블	2.3135	이재명	48.96	49.67	48.54	47.72	47.94	48.19	48.47	1.1654
	2.2700	윤석열	51.04	50.33	51.46	52.28	52.06	51.81	51.53	1.1667
여론 조사		이재명	49.52	47.27	48.20	49.11	48.08	48.40	49.62**	
		윤석열	50.47	52.72	51.80	50.88	51.91	51.59	50.37**	

* 결과는 이재명 후보 지지율+윤석열 후보 지지율=100%로 환산된 값
** 03/09(대선)의 투표 결과를 100%합산으로 환산한 값

Ⅲ. Study 2

VAIV 경희대학교

성능 평가 및 결과:
검색 트렌드 데이터(유튜브 검색 포함)

	모델 RMSE	후보	03/03(목)	03/04(금)	03/05(토)	03/06(일)	03/07(월)	03/08(화)	03/09(대선)	예측 RMSE
일반 선형회귀	2..1067	이재명	49.62	50.52	48.95	47.49	47.36	48.92	50.05	1.4498
		윤석열	50.38	49.48	51.05	52.51	52.64	51.08	49.95	
릿지 회귀모형	2.0667	이재명	49.34	50.00	49.47	47.88	47.66	48.85	50.19	1.2705
		윤석열	50.65	50.00	50.53	52.12	52.34	51.15	49.8	
K-NN	1.7581	이재명	48.52	48.82	47.89	47.89	47.89	48.86	48.86	0.9121
		윤석열	51.48	51.18	52.11	52.11	52.11	51.14	51.14	
의사결정 나무	1.5510	이재명	48.53	48.53	49.57	49.57	49.57	48.53	48.53	1.0736
		윤석열	51.47	51.47	50.43	50.43	50.43	51.47	51.47	
SVR	2.2133	이재명	48.68*	49.56	49.78	47.8	47.25	47.78	49.15	1.2793
	2.2024	윤석열	51.32	50.44	50.22	52.2	52.75	52.22	50.85	1.2813
ANN	2.0421	이재명	49.16	49.69	49.4	47.79	47.66	48.73	50.26	1.1863
		윤석열	50.84	50.31	50.6	52.21	52.34	51.27	49.74	
앙상블	1.5996	이재명	48.79	49.63	48.82	48.34	48.30	49.13	49.52	1.0468
	1.5485	윤석열	51.21	50.37	51.18	51.66	51.70	50.87	50.48	1.0445
여론 조사		이재명	49.52	47.27	48.20	49.11	48.08	48.40	49.62**	
		윤석열	50.47	52.72	51.80	50.88	51.91	51.59	50.37**	

* 결과는 이재명 후보 지지율+윤석열 후보 지지율=100%로 환산된 값
** 03/09(대선)의 투표 결과를 100%합산으로 환산한 값

Study 2 결과 요약

VAIV Sometrend Biz 활용

결과

- **유튜브 분석 데이터**들은 다른 분석 데이터 보다 뛰어난 머신러닝 학습 모델 성능과 예측 결과를 보여줌(소셜 데이터에 비해 조회 수, 좋아요 수, 댓글 수 변수를 포함)(p. 35)
- 검색 트렌드 분석 데이터를 학습한 모델의 경우 유튜브 검색 데이터를 포함하면 머신러닝 모델의 성능이 향상됨(pp. 39-40)
- 유튜브 분석 데이터가 다른 분석 데이터 보다 "깜깜이 기간" 대선 여론의 흐름 또한 잘 반영하고 있음(p. 35)

시사점

- 후보자별 유튜브 관련 7개 변수(긍정/중립/부정 언급량, 동영상/조회/좋아요/댓글 수)만으로도 성능이 좋은 여론 예측모델을 구축할 수 있을 것으로 기대
- 유튜브 동영상에 대한 댓글 분석을 추가할 수 있으면, 좀더 정확하고 유용한 예측모델 구축이 가능할 것으로 기대
- 기존의 여론 분석 데이터(소셜 데이터, 검색 트렌드 분석 데이터 등)에 유튜브 분석 데이터를 포함하면 좀더 정확한 예측모델 구축이 가능할 것으로 기대

결론 및 제언

결론

- **유튜브 분석**은 기존 여론조사(여론조사 메타분석 포함), 소셜 분석, 검색어 트렌드 분석 결과를 충분히 보완할 수 있는 유용한 방법임
- 유튜브 분석 결과만으로도 실제 여론을 실시간으로 예측할 수 있음(여론조사 결과에 영향). 단, 편향성이 존재할 수 있음
- (댓글 분석 없이) 크롤링 가능한 유튜브 데이터만으로도 성능 좋은 선거 예측 모형 개발이 가능함
- 유튜브 분석 결과와 함께 다른 분석 결과를 종합한다면 좀 더 정확한 결과 도출이 가능함

분석의 한계

- 유튜브 구독자수 1,000명 이상 채널 대상 동영상 설명 부분의 텍스트 데이터만을 크롤링 ⇒ 댓글 데이터 제외
- 실제 여론(진실값)을 알 수 없기 때문에, 여론조사 메타분석 결과를 reference 값으로 이용 ⇒ 여론조사 메타분석 결과의 신뢰성 이슈
- 유튜브 분석 과정에 많은 가정이 포함됨(예: 분석기간 100일, 양강구도, 가중치 부여, 지지율 수식 도출 등) ⇒ 결과의 robustness 확보를 위해 가정의 변화에 대한 추가적인 sensitivity analysis 필요
- 본 분석 결과는 대선과 같은 전국단위의 관심도가 큰 선거에서 양강구도가 형성될 때 가장 여론을 잘 분석해 낼 수 있음

제언

- **유튜브 분석에 관심을 기울여야!** ⇒ 활용범위가 넓어질 것으로 예상, 향후 댓글 데이터 추가 분석에서 나아가 궁극적으로 Voice/Image/Video Mining 기술이 활발히 적용될 것으로 예상
- **급할 땐 유튜브 분석을 적절히 활용하자!** ⇒ 실시간 여론의 흐름을 확인할 수 있으며, 여론조사 비용을 Save하면서도 좀 더 대표성 있는 결과를 도출할 수도 있음
- **MZ세대가 열광하는 차세대 플랫폼에도 관심을!** ⇒ 대중 매체에서, 검색 포털로, SNS로, 유튜브로, 그 다음은? Metaverse Platform?

참고문헌

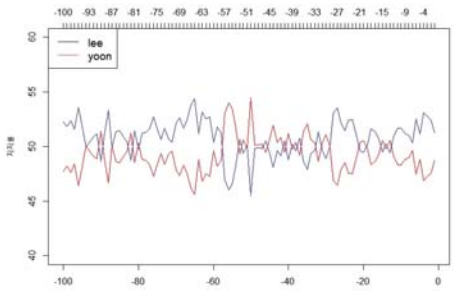
1. 권혁남. (2001). 16 대 총선 여론조사의 문제점 및 개선방안: 출구조사와 무응답률 문제를 중심으로. *언론과학연구*, 1(1), 46-74.
2. 김양원, 송경재. (2021). *[열린라디오 YTN] 여론조사 과잉시대…대선후보 여론조사 믿어도 될까*. Available online: https://www.ytn.co.kr/_ln/0101_202111150859285538 (accessed on 14 MAY 2022)
3. 안지현. (2022). *'오락가락' 여론조사 결과 왜?... '정치 저관여층 ' 에 달렸다*. Available online: https://news.jtbc.joins.com/article/article.aspx?news_id=NB12048215 (accessed on 14 MAY 2022)
4. 임예인. (2022). *2022대선, 여론조사의 패배를 묻다: 괴물 인터뷰 2/2*. Available online: https://ppss.kr/archives/254019 (accessed on 14 MAY 2022)
5. 이지민. (2021). *[뉴튜분석] 네이버 실시간 검색어 폐지의 3가지 이유*. Available online: https://www.news2day.co.kr/article/20210208500251. (accessed on 18 MAY 2022).
6. 하상현, & 노태협. (2020). SNS 기반 여론 감성 분석. *The Journal of the Convergence on Culture Technology (JCCT)*, *6*(1), 111-120.
7. 이강유, & 성동규. (2018). 유튜브 이용자의 몰입경험과 만족에 영향을 미치는 요인 연구. *한국콘텐츠학회논문지*, *18*(12), 660-675.
8. 김종훈. (2019). *연령불문 인기 1위 '유튜브' 50대 증가하는 까닭은?*. Available online: http://www.dizzotv.com/site/data/html_dir/2019/05/16/2019051680230.html (accessed on 14 MAY 2022
9. Mehrabian, A. (1981). Silent messages: implicit communication of emotions and attitudes, Wadsworth Pub.
10. 김찬우, 박효찬, & 박한우. (2017). 2017 년 대통령 후보수락 연설 유튜브 동영상의 댓글망과 의미망 분석. *Journal of The Korean Data Analysis Society (JKDAS)*, 19, 1379-1390.
11. 박상현, 김성훈, & 정승화. (2020). 유튜브 정치· 시사 채널 이용이 정치사회화에 미치는 영향. *한국콘텐츠학회논문지*, *20*(9), 224-237.
12. Krishna, A., Zambreno, J., & Krishnan, S. (2013, December). Polarity trend analysis of public sentiment on YouTube. In *Proceedings of the 19th international conference on management of data* (pp. 125-128).
13. Shevtsov, A., Oikonomidou, M., Antonakaki, D., Pratikakis, P., & Ioannidis, S. (2020). Analysis of Twitter and YouTube during US elections 2020. *arXiv preprint arXiv:2010.08183*.
14. 송화영, 박세정, & 박한우. (2020). 2020 년 국회의원 선거 기간의 유튜브 빅데이터 분석. *Journal of The Korean Data Analysis Society*, *22*(5), 2063-2074.
15. 나무위키. (2022). *제 20대 대통령 선거/출구조사*. Available online: https://namu.wiki/w/%EC%A0%9C20%EB%8C%80%20%EB%8C%80%ED%86%B5%EB%A0%B9%20%EC%84%A0%EA%B1%B0/%EC%B6%9C%EA%B5%AC%EC%A1%B0%EC%82%AC (accessed on 18 MAY 2022).
16. 전창훈. (2021). *선거 100일 전 웃은 후보가 결국 웃었다*. Available online: http://www.busan.com/view/busan/view.php?code=2021112817571734900. (accessed on 18 MAY 2022).
17. 김가현. (2022). *'깜깜이 기간' 쿠으로 기운 표심… 지지 격차는 0.9% ~ 5.2%P 요동*. Available online: https://www.seoul.co.kr/news/newsView.php?id=20220310006005 (accessed on 14 MAY 2022)
18. 김종훈. (2019). *연령불문 인기 1위 '유튜브' 50대 증가하는 까닭은?*. Available online: http://www.dizzotv.com/site/data/html_dir/2019/05/16/2019051680230.html (accessed on 14 MAY 2022
19. Yu, R. P. (2016). The relationship between passive and active non-political social media use and political expression on Facebook and Twitter. *Computers in Human Behavior*, *58*, 413-420.
20. 장승진, 한정훈.(2021). 유튜브는 사용자들을 정치적으로 양극화시키는가 주요 정치 및 시사 관련 유튜브 채널 구독자에 대한 설문조사 분석. *현대정치연구*, 14(2), 5-35.

감사합니다!

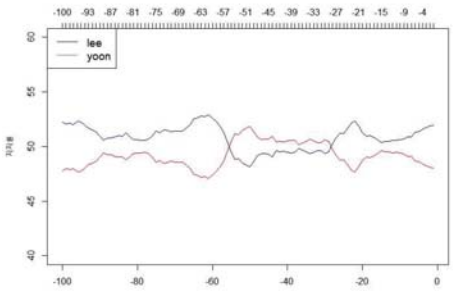
Q & A

부록

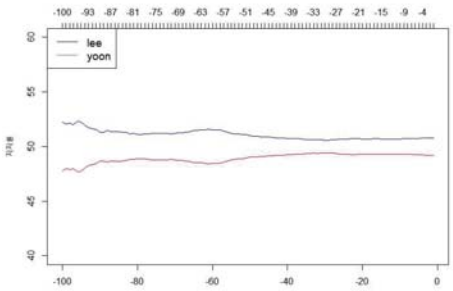
누적 그래프



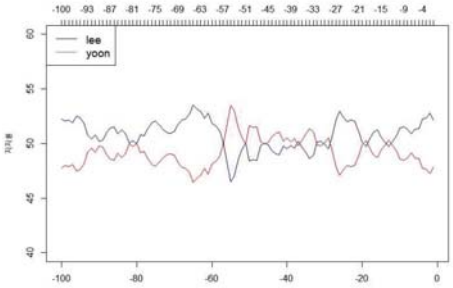
당일



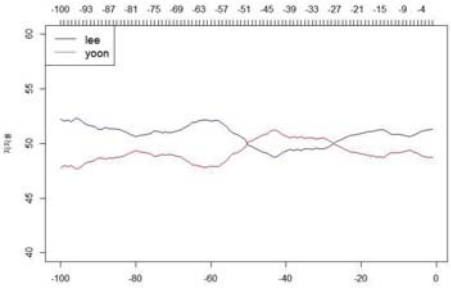
7일 누적



전체 누적



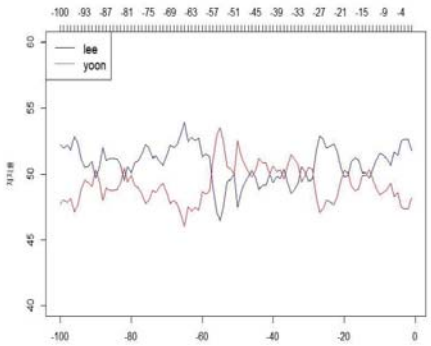
3일 누적



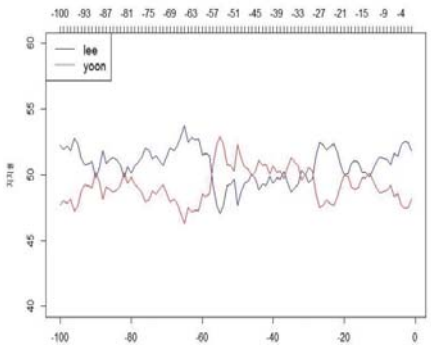
15일 누적

부록

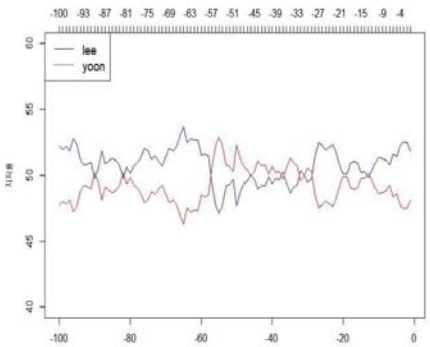
반감기 그래프



반감기 3일 누적



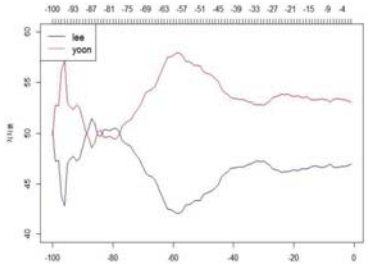
반감기 7일 누적



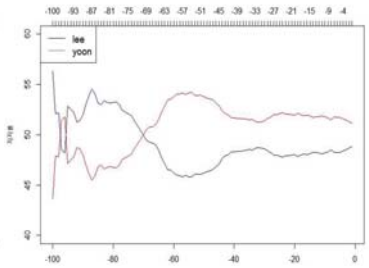
반감기 15일 누적

부록

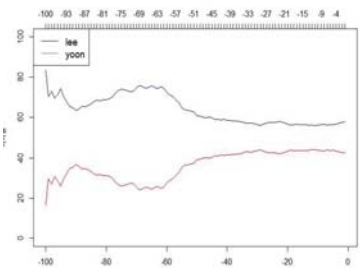
Sigmoid 도출 과정(전체 누적)



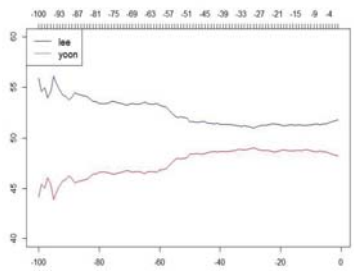
전체 언급량



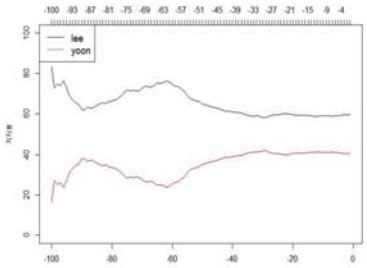
긍정 언급량



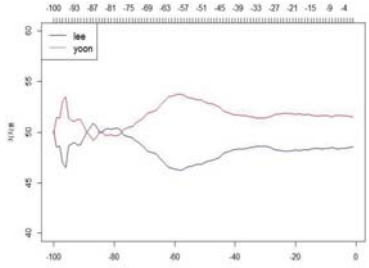
긍정 언급량 – 부정 언급량



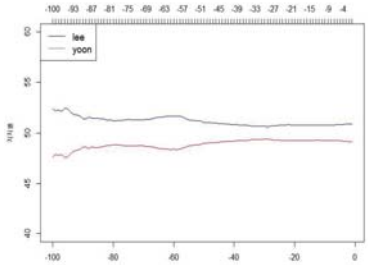
긍정 언급량 + 상대부정 언급량



(긍정 언급량 – 부정 언급량) / 전체 언급량



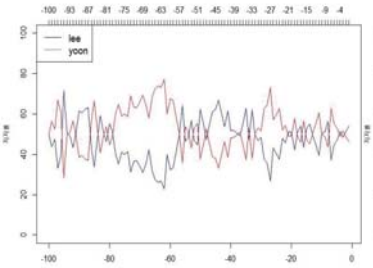
Tan((긍정 언급량 – 부정 언급량) / 전체 언급량))



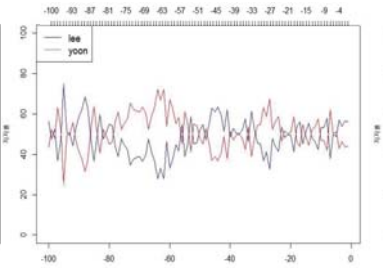
Sigmoid((긍정 언급량 – 부정 언급량) / 전체 언급량))

부록

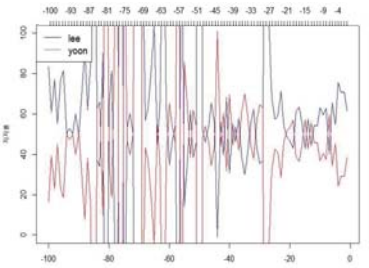
Sigmoid 도출 과정(당일)



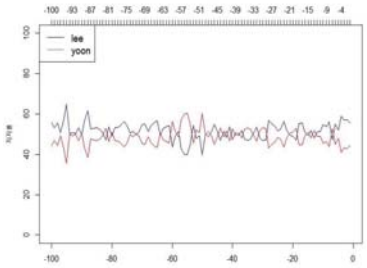
전체 언급량



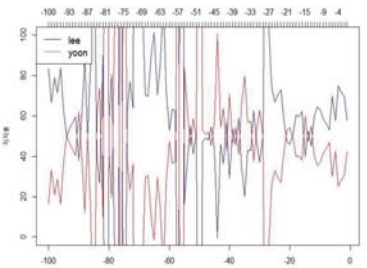
긍정 언급량



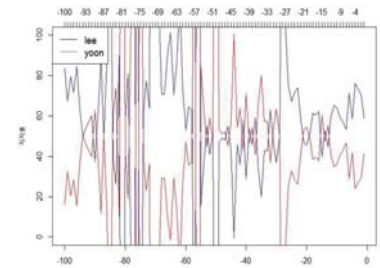
긍정 언급량 – 부정 언급량



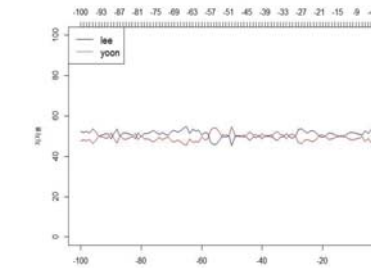
긍정 언급량 + 상대부정 언급량



(긍정 언급량 – 부정 언급량) / 전체 언급량



Tan((긍정 언급량 – 부정 언급량) / 전체 언급량))



Sigmoid((긍정 언급량 – 부정 언급량) / 전체 언급량))

빅데이터 기술 교류 세미나

빅데이터와 여론조사

발표 4

과거 대선 및 서울시장 선거 회고적 분석: 썸트렌드 vs. 포털 트렌드

이동원 (고려대 교수)

고려대학교 경영대학
Korea University
Business School

과거 대선 및 서울시장 선거 회고적 분석

- 썸 트렌드 vs. 검색엔진 트렌드 -

2022-05-19

이동원 교수
고려대학교 경영대학

Social Mobile Analytics Cloud
SoMAC Lab

SoMAC

목차

SoMAC

0.
Intro

빅데이터로 선거 예측이 가능하다?

SoMAC

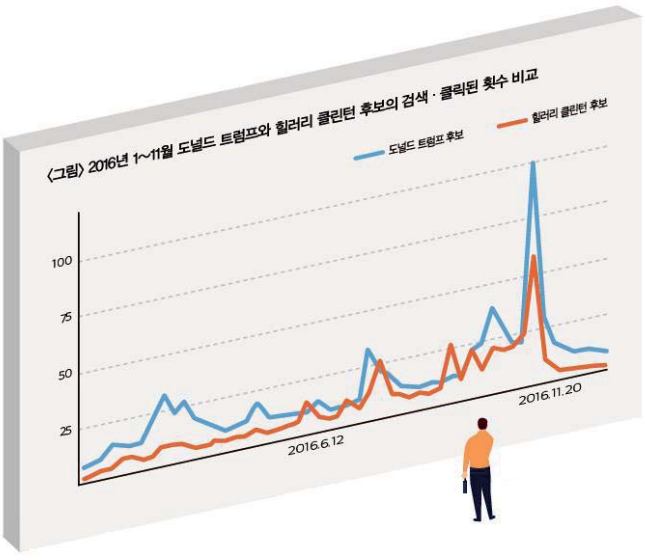
빅데이터로 선거 결과를
예측할 수 있을까?

빅데이터로 선거 예측이 가능하다?

▷ **여론조사**는 **과학적인 통계학**을 바탕으로 하지만, 족집게처럼 선거 **결과를 정확하게 맞출 수는 없음** (예: 2016년 미국 대선, 2016년 한국 총선 등)

▷ 전통적인 전화조사 방법 대신 **빅데이터**를 이용하면 더 **정확한 예측**이 **가능**하지 않을까?

빅데이터로 선거 예측이 가능하다?



주: 구글에서 검색·클릭된 횟수를 각각 일별로 합산한 자료. 조회기간 내 최대 검색량을 100으로 표현해 상대적인 변화를 나타냄.
자료: 구글 트렌드(Google Trends)

참고문헌: 빅데이터로 선거 결과를 예측할 수 있을까? (2018)

▷ 2016년 미국 대선에서 **빅데이터**(구글 트렌드 검색량)를 **활용**하여 **도널드 트럼프**의 **당선**을 **예측**한 사례가 주목

▷ 하지만, **힐러리 클린턴**이 **다수표**를 얻고도 **선거인단 수**에서 **뒤진 탓**에 대통령이 되지 못함

▷ 트럼프 **당선 예측은 맞았지만 다수표 획득**에 대한 **예측**은 **실패**!

1.

미국 및 캐나다 대선 사례

Google Trends as a Predictor?

Google Trends as a Predictor of Presidential Elections: The United States Versus Canada

American Behavioral Scientist
2021, Vol. 65(4) 666–680

DOI: 10.1177/0002764220975067
journals.sagepub.com/home/abs

Camilo Prado
and Carmen

빅데이터(구글 트렌드 검색량)를 활용하여
지난 미국 선거(4번), 캐나다 선거(5번) 결과를
정확히 예측함!

Abstract

The media and election campaign managers conduct several polls in the days leading up to the presidential elections. These preelection polls have a different pred ive ing he he he ate his article, we have taken into account the past four elections in the United States and the past five in Canada, since Google first published its search statistics in 2004. The results show that this method has predicted the real winner in all the elections held since 2004 and highlights that it is necessary to monitor the next elections for the presidency of the United States in November 2020 and to have more accurate information on the future results.

Google Trends as a Predictor (캐나다 대선)

"Google Trends as a Predictor of Presidential Elections: The United States Versus Canada" (2021)

- ▷ **검색엔진 검색량**과 미국 및 캐나다 **대통령 후보자** 사이의 **관계**를 확인한 논문에 따르면,
- ▷ **검색량**이 **더 높은 후보자**가 **당선** (2004-2019, 9개 선거)

지난 5번의 캐나다 대선 결과 정확히 예측!

Election date	10/21/2019	10/19/2015	05/02/2011	10/14/2008	06/28/2004
Blue candidate	Andrew Scheer	Stephen Harper	Stephen Harper	Stephen Harper	Stephen Harper
Red candidate	Justin Trudeau	Justin Trudeau	Jack Layton	Stéphane Dion	Paul Martin
1 Month: Blue average	10	21	28	60	44
1 Month: Red average	23	43	23	34	38
2 Months: Blue average	4	15	21	42	32
2 Months: Red average	10	27	14	24	33
3 Months: Blue average	3	12	16	31	29
3 Months: Red average	8	21	10	18	35
Prediction	Red	Red	Blue	Blue	Red
Winner	Red	Red	Blue	Blue	Red
Blue votes (I)	121	99	166	143	99
Red votes (II)	157	184	103	77	135

Google Trends as a Predictor (미국 대선)

Table 1. United States Results: Period 2004-2016.

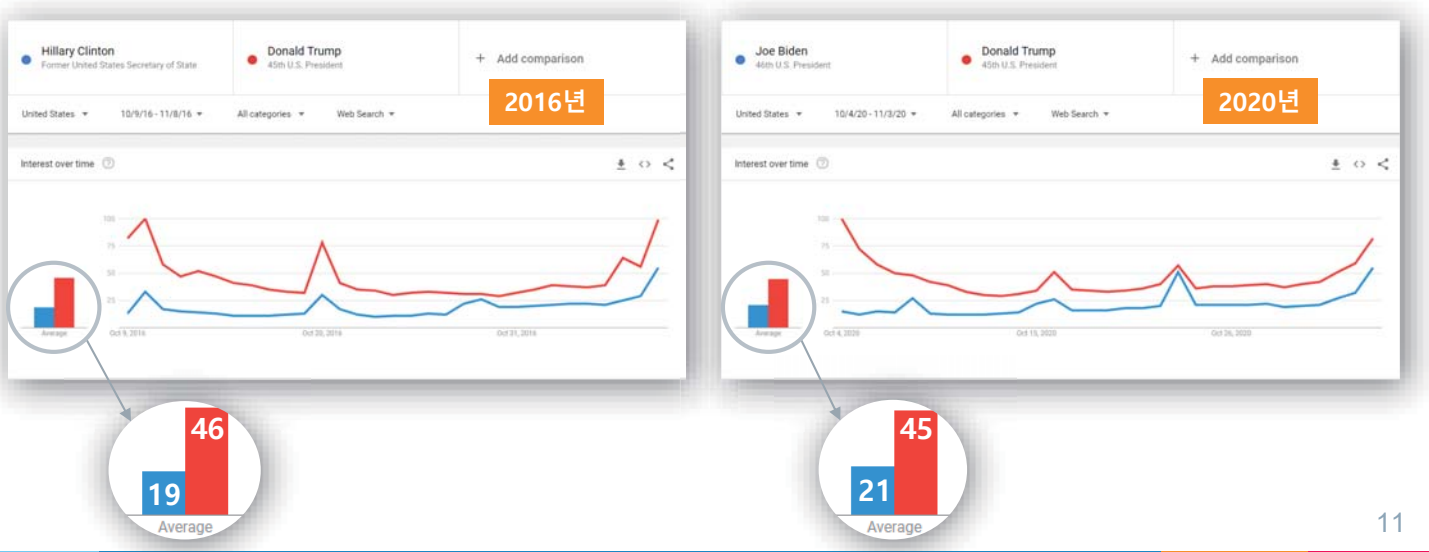
Year	2016	2012	2008	2004	2020
Election date	11/08/2016	11/06/2012	11/04/2008	11/02/2004	11/03/2020
Blue candidate	Hillary Clinton	Barack Obama	Barack Obama	John Kerry	Joe Biden
Red candidate	Donald Trump	Mitt Romney	John McCain	George W. Bush	Donald Trump
2 Months: Red average	34	8	16	43	9
3 Mo					23
3 Mo					12
					29
					8
					21
Prediction	Red	Blue	Blue	Red	Red
Winner	Red	Blue	Blue	Red	Blue
Blue votes (I)	227	332	365	251	306
Red votes (II)	304	206	173	286	232

지난 4번의 미국 대선 결과 정확히 예측

하지만 2020년 미국 대선 결과 예측 틀림!

2016년 vs. 2020년 미국 대선 (검색량 기준)

▷ **2016년, 2020년** 미국 대선은 모두 **트럼프**가 (구글 트렌드) **검색량**은 **우위**였으나, **실제** (전국민) **득표율**은 **힐러리**, **바이든**이 **우위**였음



2016년 vs. 2020년 미국 대선 (출구조사 및 득표율)

2016년 미국 대선

	출구조사	득표율	검색량	선거인단
도널드 트럼프	47.5%	46.09%	**46%**	**304 (당선)**
힐러리 클린턴	**47.7%**	**48.18%**	19%	227

2020년 미국 대선

	출구조사	득표율	검색량	선거인단
조 바이든	**51%**	**51.31%**	21%	**306 (당선)**
도널드 트럼프	47%	46.86%	**45%**	232

▷ **출구조사**, **득표율** 모두 **힐러리**가 앞섰지만 **검색량**이 더 많았던 **트럼프**가 더 많은 선거인단을 가져가며 당선

▷ **출구조사**, **득표율** 모두 앞섰던 **바이든**이 **검색량**이 더 많았던 **트럼프**를 제치고 당선

검색량 ≠당선

2016년 vs. 2020년 미국 대선 (State별 검색량 비교)

미국 50개 주 가운데 2개 주를 제외한 48개 주가 **승자독식** 방식을 채택하고 있으며, 네브래스카주와 메인주는 득표율에 따라 **선거인단**을 나눔

2016년 대선에서는 트럼프가 50개 주 모두에서 검색량에서 앞섰고, **2020년 대선에서 역시 5개 주를 제외한 45개 주에서 트럼프가 앞섬**

2016

		힐러리	트럼프	검색 승자
2	델라웨어주	26	74	트럼프
3	펜실베이니아주	29	71	트럼프
4	뉴저지주	29	71	트럼프
5	조지아주	31	69	트럼프
6	코네티컷주	28	72	트럼프
7	매사추세츠주	28	72	트럼프
8	메릴랜드주	27	73	트럼프
9	사우스캐롤라이나주	32	68	트럼프
10	뉴햄프셔주	27	73	트럼프
11	버지니아주	28	72	트럼프
12	뉴욕주	29	71	트럼프
13	노스캐롤라이나주	32	68	트럼프
14	로드아일랜드주	23	77	트럼프
15	버몬트주	26	74	트럼프
16	켄터키주	32	68	트럼프
17	테네시주	34	66	트럼프
18	오하이오주	31	69	트럼프
19	루이지애나주	32	68	트럼프
20	인디애나주	30	70	트럼프
21	미시시피주	32	68	트럼프
22	일리노이주	29	71	트럼프
23	앨라배마주	33	67	트럼프
24	메인주	24	76	트럼프
25	미주리주	32	68	트럼프
26	아칸소주	34	66	트럼프
27	미시간주	30	70	트럼프
28	플로리다주	30	70	트럼프
29	텍사스주	32	68	트럼프
30	아이오와주	31	69	트럼프
31	위스콘신주	29	71	트럼프
32	캘리포니아주	27	73	트럼프
33	미네소타주	29	71	트럼프

2020

		트럼프	바이든	검색 승자
2	델라웨어주	79	85	바이든
3	펜실베이니아주	88	89	바이든
4	인디애나주	63	64	바이든
5	미주리주	65	66	바이든
6	텍사스주	58	58	None
7	뉴저지주	79	72	트럼프
8	조지아주	61	56	트럼프
9	코네티컷주	81	72	트럼프
10	매사추세츠주	88	75	트럼프
11	메릴랜드주	72	69	트럼프
12	사우스캐롤라이나주	58	57	트럼프
13	뉴햄프셔주	96	77	트럼프
14	버지니아주	74	70	트럼프
15	뉴욕주	71	63	트럼프
16	노스캐롤라이나주	71	68	트럼프
17	로드아일랜드주	81	63	트럼프
18	버몬트주	97	59	트럼프
19	켄터키주	58	56	트럼프
20	테네시주	65	60	트럼프
21	오하이오주	76	73	트럼프
22	루이지애나주	51	49	트럼프
23	미시시피주	45	43	트럼프
24	일리노이주	78	63	트럼프
25	앨라배마주	52	51	트럼프
26	메인주	100	77	트럼프
27	아칸소주	61	55	트럼프
28	미시간주	84	77	트럼프
29	플로리다주	76	70	트럼프
30	아이오와주	73	65	트럼프
31	위스콘신주	85	84	트럼프
32	캘리포니아주	73	61	트럼프
33	미네소타주	93	77	트럼프

2.

한국 대선 및 서울시장 선거 사례

검색엔진 트렌드 vs. 썸트렌드

한국 대선 및 서울시장 선거 요약 (2012 ~ 2021)

	2012년 18대 대선	2017년 19대 대선	2014년 서울 시장	2018년 서울시장	2021년 서울시장
일정	2012.11.25~26 후보자 등록 신청 2012.12.13~14 사전투표 **2012.12.19 본투표**	2017.04.15~16 후보자 등록 신청 2017. 05.04~05 사전투표 **2017.05.09 본투표**	2014.05.15~16 후보자 등록 신청 2014.05.30~31 사전투표 **2014.06.04 본투표**	2018.05.24~25 후보자 등록 신청 2018.06.08~09 사전투표 **2018.06.13 본투표**	2021.03.18~19 후보자 등록 신청 2021.04.02~03 사전투표 **2021.04.07 본투표**
주요 후보자	박근혜 vs. 문재인	문재인 vs. 홍준표	박원순 vs. 정몽준	박원순 vs. 김문수	오세훈 vs. 박영선
출구조사	**박근혜 50.1%** 문재인 48.9%	**문재인 41.4%** 홍준표 23.3%	**박원순 54.5%** 정몽준 44.7%	**박원순 55.9%** 김문수 21.2%	**오세훈 59.0%** 박영선 37.7%
득표율	**박근혜 51.55%** 문재인 48.02%	**문재인 41.08%** 홍준표 24.03%	**박원순 56.12%** 정몽준 43.02%	**박원순 52.79%** 김문수 23.34%	**오세훈 57.50%** 박영선 39.18%

검색엔진 트렌드 vs. 썸트렌드

서비스	데이터 분석 시작	기반	주소
Google Trends	2004년	검색엔진	https://trends.google.com/
NAVER DataLab.	2016년	검색엔진	https://datalab.naver.com/
kakaodatatrend	2018년	검색엔진	https://datatrend.kakao.com/
Sometrend Biz	2014년 (트위터 2011년)	감성분석	https://biz.some.co.kr/

검색엔진 점유율 (미국 vs. 한국)

미국

한국

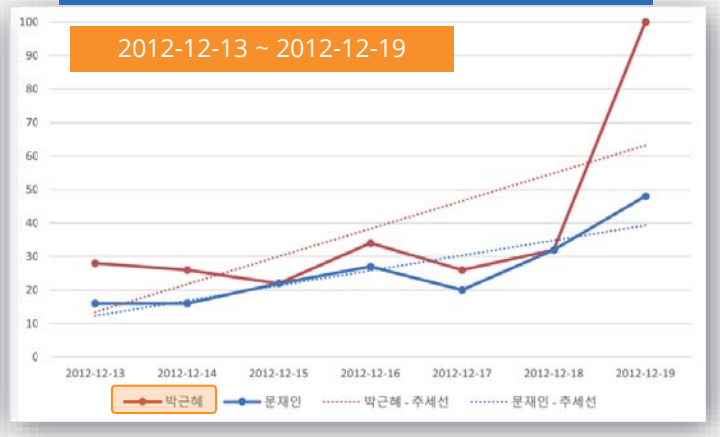
▷ 미국 및 캐나다 사례와 달리 **국내 사례**는 '**썸트렌드**'와 **네이버, 구글, 다음** 3사의 검색엔진의 시장점유율을 반영한 '**검색엔진 트렌드**'를 활용

▷ 검색엔진 트랜드 = $\sum$ 검색량$_t^{SE}$ × 점유율^{SE} (t: 시점, SE: 검색엔진, 점유율: 기간 평균)

2012년 18대 대선 검색량 (박근혜 vs. 문재인)

검색엔진 트렌드(투표일 직전 1개월)

검색엔진 트렌드 (투표일 직전 1주일)

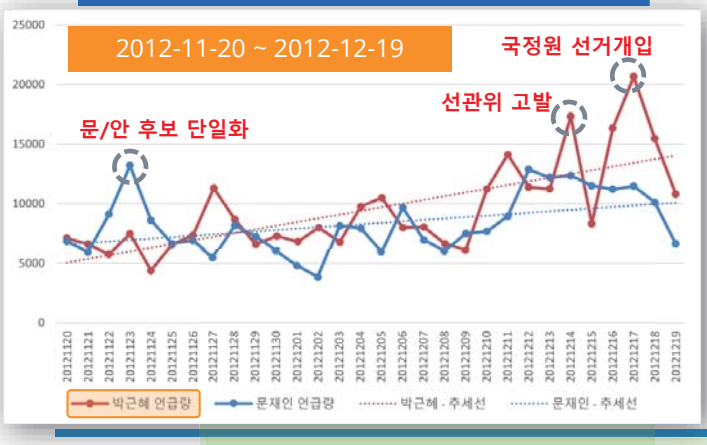


	출구조사	개표결과	검색엔진	4주 전	3주 전	2주 전	1주 전
박근혜	50.1%	51.55%	60.1%	56.7%	66.2%	58.8%	59.7%
문재인	48.9%	48.02%	39.9%	43.3%	33.8%	59.7%	40.3%

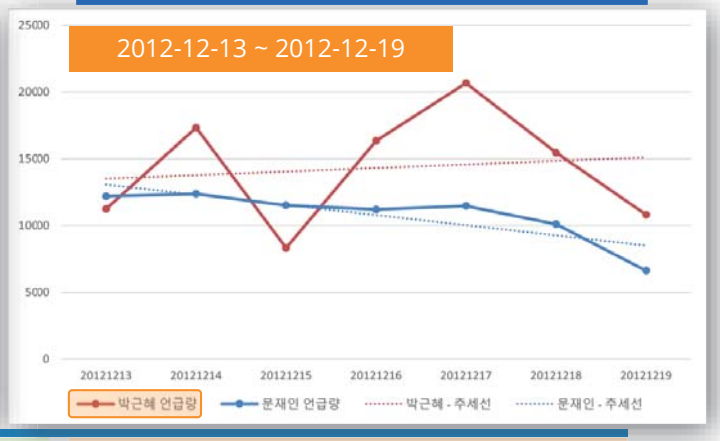
※ 구글 트렌드 데이터 활용

2012년 18대 대선 언급량 (박근혜 vs. 문재인)

썸트렌드 감성 (투표일 직전 1개월)



썸트렌드 감성 (투표일 직전 1주일)

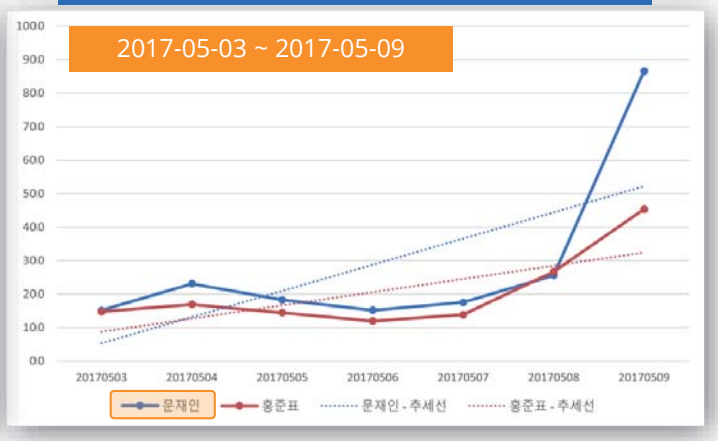


	출구조사	개표결과	검색엔진	썸트렌드	4주 전	3주 전	2주 전	1주 전
박근혜	50.1%	51.55%	60.1%	53.4%	47.9% (65,254)	55.9% (55,712)	52.3% (65,494)	57.0% (100,202)
문재인	48.9%	48.02%	39.9%	44.6%	52.1% (70,956)	44.1% (44,020)	47.7% (59,643)	43.0% (75,539)

2017년 19대 대선 검색량 (문재인 vs. 홍준표)

검색엔진 트렌드(투표일 직전 1개월)

검색엔진 트렌드 (투표일 직전 1주일)

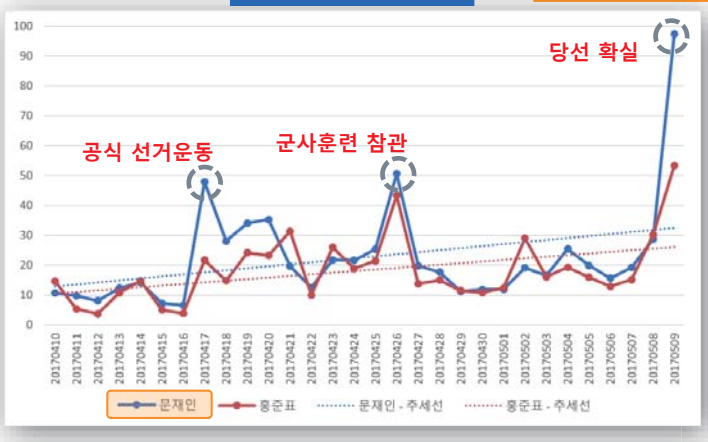


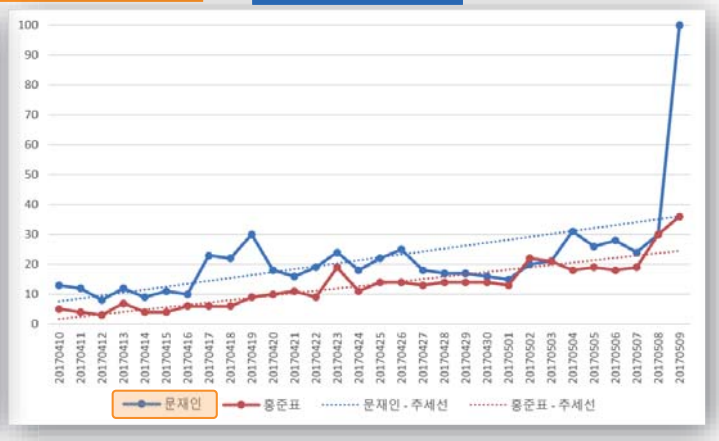
	출구조사	개표결과	검색엔진	4주 전	3주 전	2주 전	1주 전
문재인	41.4%	41.08%	56.1%	61.6%	53.4%	51.6%	58.3%
홍준표	23.3%	24.03%	43.9%	38.4%	46.6%	48.4%	41.7%

※ 시장평균 점유율
네이버(77.60%),
구글(11.02%) 반영

2017년 19대 대선 검색량 비교 (문재인 vs. 홍준표)

네이버 트렌드 | 2017-04-10 ~ 2017-05-09 | 구글 트렌드



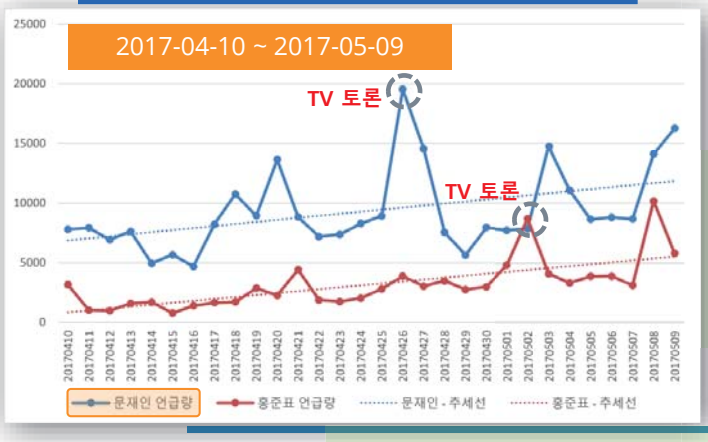


	1개월	4주 전	3주 전	2주 전	1주 전
문재인	55.4%	60.5%	52.3%	51.1%	58.3%
홍준표	44.6%	39.5%	47.7%	48.9%	41.7%

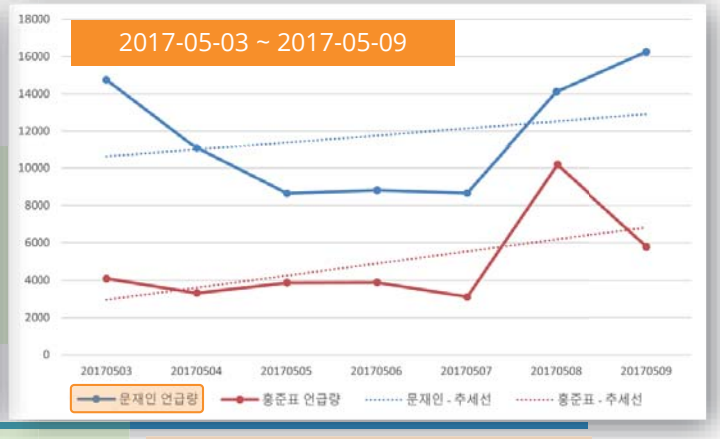
	1개월	4주 전	3주 전	2주 전	1주 전
문재인	62.5%	72.7%	63.9%	55.2%	61.8%
홍준표	37.5%	27.3%	36.1%	44.8%	38.2%

2017년 19대 대선 언급량 (문재인 vs. 홍준표)

썸트렌드 감성 (투표일 직전 1개월) | 2017-04-10 ~ 2017-05-09



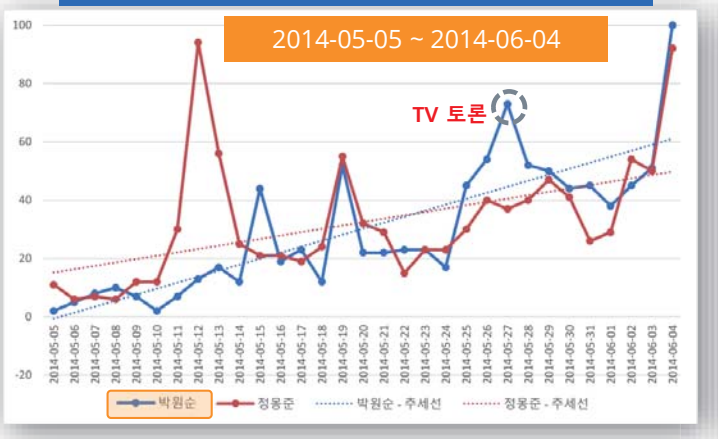
썸트렌드 감성 (투표일 직전 1주일) | 2017-05-03 ~ 2017-05-09



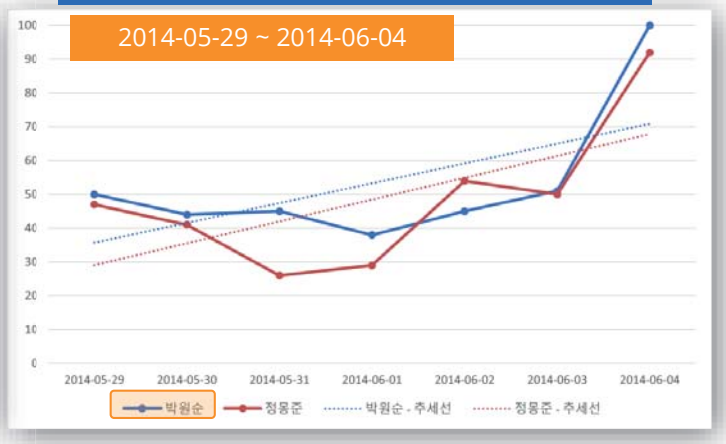
	출구조사	개표결과	검색엔진	썸트렌드	4주 전	3주 전	2주 전	1주 전
문재인	41.4%	41.08%	56.1%	74.6%	82.2% (64,532)	77.8% (63,161)	70.5% (70,802)	70.6% (82,326)
홍준표	23.3%	24.03%	43.9%	25.4%	17.8% (13,983)	22.2% (17,985)	29.5% (29,616)	29.4% (34,251)

2014년 서울시장 선거 검색량 (박원순 vs. 정몽준)

검색엔진 트렌드 (투표일 직전 1개월)



검색엔진 트렌드 (투표일 직전 1주일)

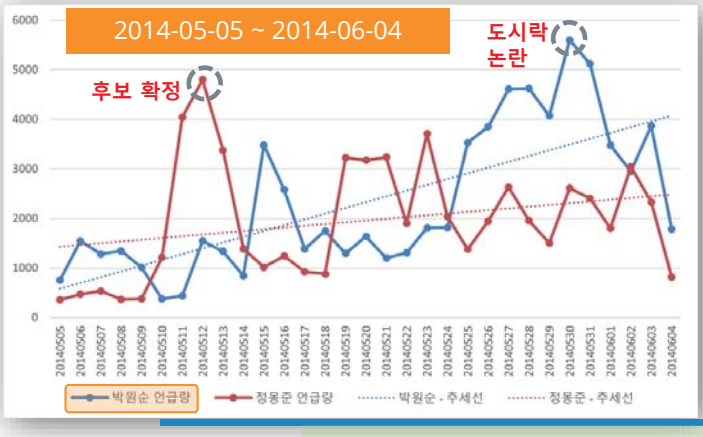


	출구조사	개표결과	검색엔진	4주 전	3주 전	2주 전	1주 전
박원순	54.5%	56.12%	48.2%	24.3%	49.1%	58.0%	52.4%
정몽준	44.7%	43.02%	51.8%	75.7%	50.9%	42.0%	47.6%

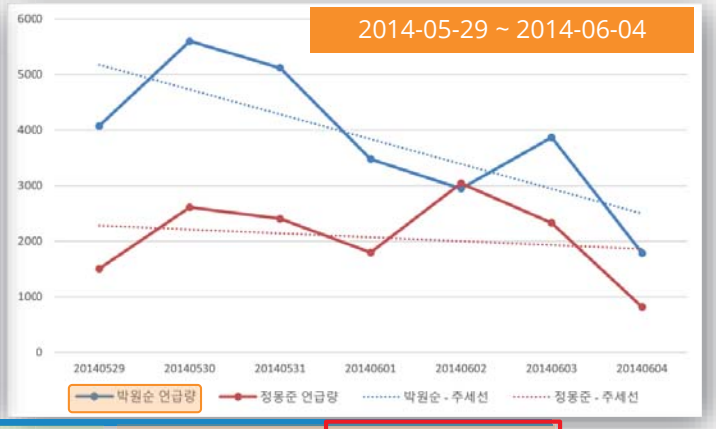
※ 구글 트렌드 데이터 활용

2014년 서울시장 선거 언급량 (박원순 vs. 정몽준)

썸트렌드 감성 (투표일 직전 1개월)



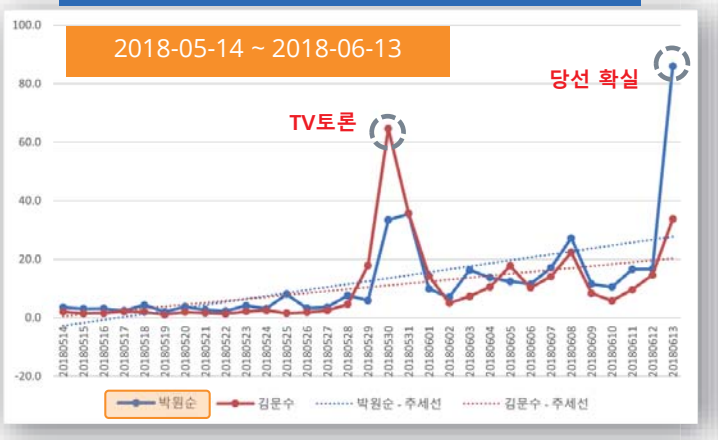
썸트렌드 감성 (투표일 직전 1주일)



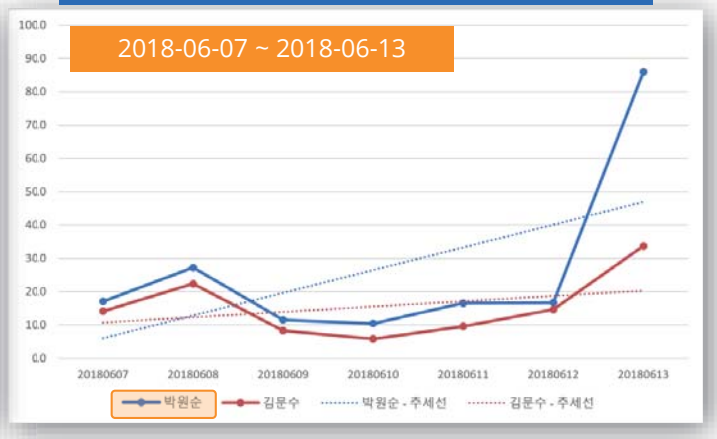
	출구조사	개표결과	검색엔진	썸트렌드	4주 전	3주 전	2주 전	1주 전
박원순	54.5%	56.12%	48.2%	54.3%	38.2% (10,463)	49.3% (13,316)	58.1% (21,537)	64.9% (26,869)
정몽준	44.7%	43.02%	51.8%	45.7%	61.8% (16,919)	50.7% (13,699)	41.9% (15,545)	35.1% (14,500)

2018년 서울시장 선거 검색량 (박원순 vs. 김문수)

검색엔진 트렌드(투표일 직전 1개월)



검색엔진 트렌드 (투표일 직전 1주일)



	출구조사	개표결과	검색엔진	4주 전	3주 전	2주 전	1주 전
박원순	55.9%	52.79%	54.6%	64.1%	40.4%	51.3%	63.1%
김문수	21.2%	23.34%	45.4%	35.9%	59.6%	48.7%	36.9%

※ 시장평균 점유율
네이버(70.72%),
구글(16.66%),
다음(7.94%) 반영

2018년 서울시장 선거 검색량 (박원순 vs. 김문수)

네이버 트렌드

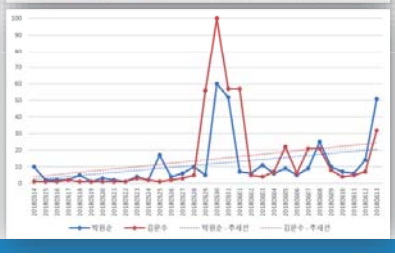
2018-05-14 ~ 2018-06-13

	1개월	4주 전	3주 전	2주 전	1주 전
박원순	52.9%	58.1%	37.4%	51.1%	62.2%
김문수	47.1%	41.9%	62.6%	48.9%	37.8%

구글 트렌드

	1개월	4주 전	3주 전	2주 전	1주 전
박원순	65.2%	69.1%	63.1%	59.4%	68.1%
김문수	34.8%	30.9%	36.9%	40.6%	31.9%

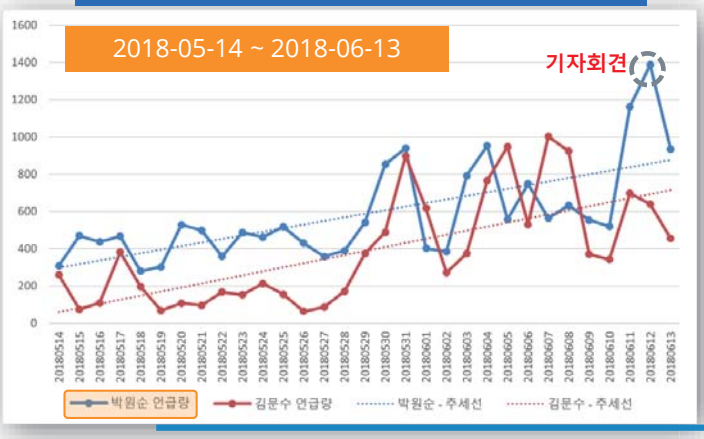
다음 트렌드



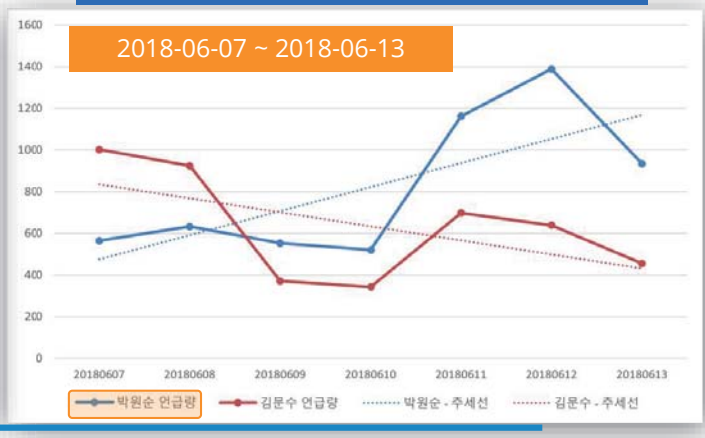
	1개월	4주 전	3주 전	2주 전	1주 전
박원순	44.7%	71.1%	38.1%	37.8%	55.5%
김문수	55.3%	28.9%	61.9%	62.2%	44.5%

2018년 서울시장 선거 언급량 (박원순 vs. 김문수)

썸트렌드 감성 (투표일 직전 1개월)



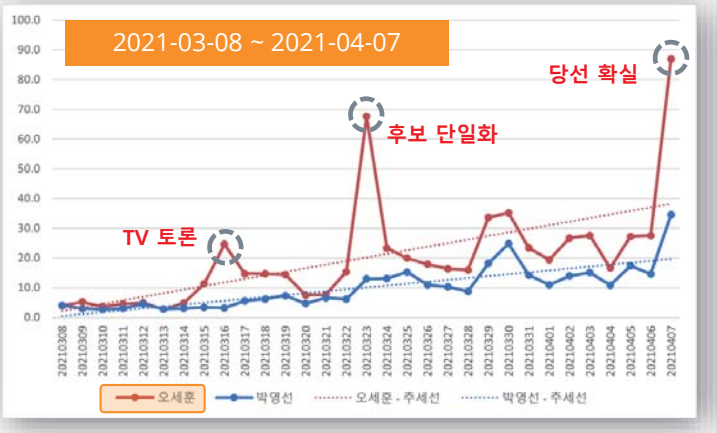
썸트렌드 감성 (투표일 직전 1주일)



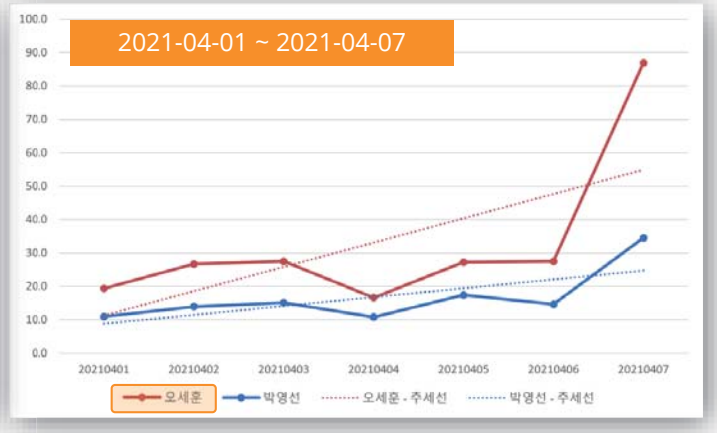
	출구조사	개표결과	검색엔진	썸트렌드	4주 전	3주 전	2주 전	1주 전
박원순	55.9%	52.79%	54.6%	60.3%	71.9% (4,138)	69.6% (3,552)	52.0% (4,779)	56.5% (5,753)
김문수	21.2%	23.34%	45.4%	39.7%	28.1% (1,614)	30.4% (1,554)	48.0% (4,406)	43.5% (4,432)

2021년 서울시장 보궐선거 검색량 (오세훈 vs. 박영선)

검색엔진 트렌드 (투표일 직전 1개월)



검색엔진 트렌드 (투표일 직전 1주일)



	출구조사	개표결과	검색엔진	4주 전	3주 전	2주 전	1주 전
오세훈	59.0%	57.50%	66.7%	69.6%	72.6%	61.3%	66.4%
박영선	37.7%	39.18%	33.3%	30.4%	27.4%	38.7%	33.6%

※ 시장평균 점유율
네이버(54.77%),
구글(38.27%),
다음(5.09%) 반영

2021년 서울시장 보궐선거 검색량 (오세훈 vs. 박영선)

2021-03-08 ~ 2021-04-07

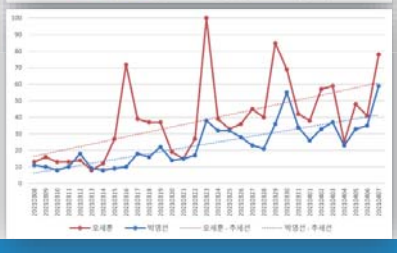
네이버 트렌드

	1개월	4주 전	3주 전	2주 전	1주 전
오세훈	69.8%	70.7%	76.7%	65.0%	68.5%
박영선	30.2%	29.3%	23.3%	35.0%	31.5%

구글 트렌드

	1개월	4주 전	3주 전	2주 전	1주 전
오세훈	65.0%	69.3%	70.1%	58.4%	66.0%
박영선	35.0%	30.7%	29.9%	41.6%	34.0%

다음 트렌드

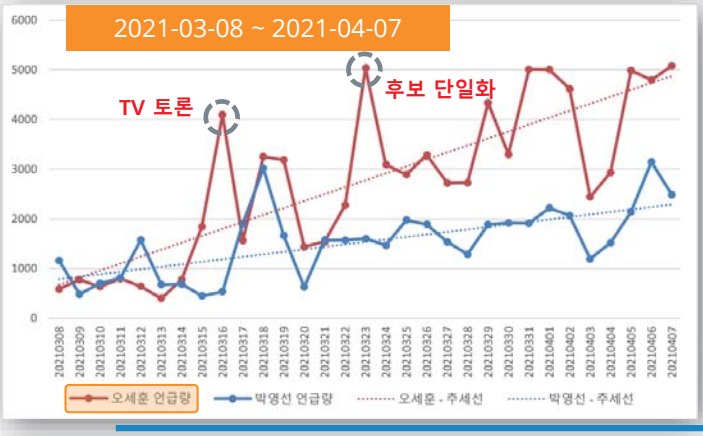


	1개월	4주 전	3주 전	2주 전	1주 전
오세훈	61.8%	67.2%	64.0%	60.4%	58.4%
박영선	38.2%	32.8%	36.0%	39.6%	41.6%

2021년 서울시장 보궐선거 언급량 (오세훈 vs. 박영선)

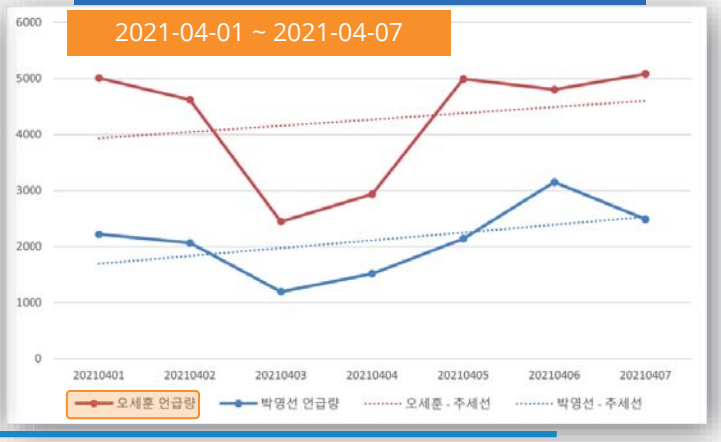
썸트렌드 감성 (투표일 직전 1개월)

2021-03-08 ~ 2021-04-07



썸트렌드 감성 (투표일 직전 1주일)

2021-04-01 ~ 2021-04-07



	출구조사	개표결과	검색엔진	썸트렌드	4주 전	3주 전	2주 전	1주 전
오세훈	59.0%	57.50%	66.7%	64.4%	57.4% (10,128)	63.2% (19,829)	66.2% (24,284)	66.9% (29,870)
박영선	37.7%	39.18%	33.3%	35.6%	42.6% (6,650)	36.8% (11,535)	33.8% (12,419)	33.1% (14,778)

한국 대선 및 서울시장 선거 결과 (2012 ~ 2021)

	2012년 18대 대선	2017년 19대 대선	2014년 서울 시장	2018년 서울시장	2021년 서울시장
주요 후보자	박근혜 vs. 문재인	문재인 vs. 홍준표	박원순 vs. 정몽준	박원순 vs. 김문수	오세훈 vs. 박영선
출구조사	**박근혜 50.1%** 문재인 48.9%	**문재인 41.4%** 홍준표 23.3%	**박원순 54.5%** 정몽준 44.7%	**박원순 55.9%** 김문수 21.2%	**오세훈 59.0%** 박영선 37.7%
검색엔진 트렌드	**박근혜 60.1%** 문재인 39.9%	**문재인 56.1%** 홍준표 43.9%	박원순 48.2% **정몽준 51.8%**	**박원순 54.6%** 김문수 45.4%	**오세훈 66.7%** 박영선 33.3%
썸트렌드	**박근혜 53.4%** 문재인 44.6%	**문재인 74.6%** 홍준표 25.4%	**박원순 54.3%** 정몽준 45.7%	**박원순 60.3%** 김문수 39.7%	**오세훈 64.4%** 박영선 35.6%
득표율	**박근혜 51.6%** 문재인 48.0%	**문재인 41.1%** 홍준표 24.0%	**박원순 56.1%** 정몽준 43.0%	**박원순 52.8%** 김문수 23.3%	**오세훈 57.5%** 박영선 39.2%

☞ 2012년, 2014년 검색엔진 트렌드는 구글 트렌드 데이터만 이용

3.

과거 선거와의 차이점

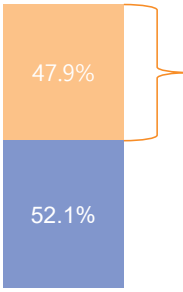
2022년 20대 대선과 썸 트렌드 감성 분석

논문 Replication (국내 사례 적용, 구글 트렌드 활용)

Year	18대	19대	20대
Election Date	2012-12-19	2017-05-09	2022-03-09
Blue Candidate	문재인	문재인	이재명
Red Candidate	박근혜	홍준표	윤석열
1 Month: Blue Average	14	22	20
1 Month: Red Average			
2 Months: Blue Average			
2 Months: Red Average			
3 Months: Blue Average			
3 Months: Red Average	10	6	10
Prediction	**Red**	**Blue**	**Blue**
Winner	**Red**	**Blue**	**Red**

빅데이터(구글 트렌드 검색량)를 활용하여
지난 2번의 대선(18대/19대) 결과를 정확히 예측함!
하지만 2022년 20대 대선 결과 예측은 틀림!

2022년 20대 대선 출구조사 및 개표결과



사전 투표율 36.93%

최종 투표율: 77.1%

	윤석열	이재명
지상파	48.4%	47.8%
JTBC	47.7%	48.4%
개표결과	48.56%	47.83%

표본크기와
보정 작업의 차이

KBS MBC SBS

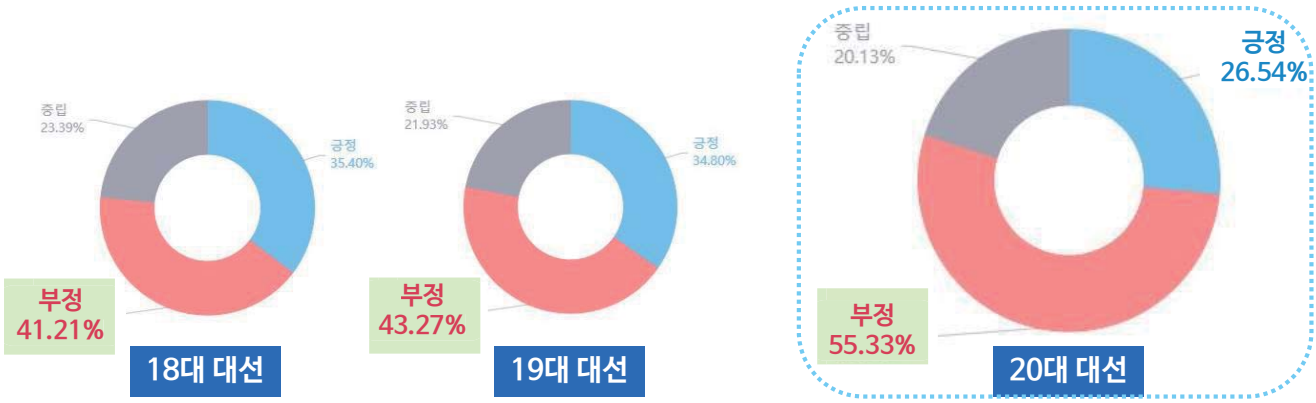
- 의뢰기관: 3곳 (한*리서치, 코**리서치, 입**코리아)
- 출구조사: **330 투표소, 73,297명 대상**
- 사전투표 보정: **1만명 대상 전화조사**
- 95% 신뢰수준, **±0.8%P**

JTBC

- 의뢰기관: 1곳 (글**리서치)
- 출구조사: **140 투표소, 36,000여명 대상**
- 사전투표 보정: **3천명 대상 전화 및 온라인 패널 조사**
- 95% 신뢰수준, **±1.2%P**

'대통령 선거'에 대한 감성 변화

※ 각 선거별 **'대통령 선거 or 대선 or 후보자 이름' 키워드**에 대한 **감성** 분석



연합뉴스 [대선 D-1] 전례없는 **네거티브**...누가 돼도 대야관계·통합 '과제'

동아일보 '비전보다 **네거티브**' '부인은 나홀로 투표'...치열했던 **비호감** 경쟁

⋮

검색 vs 감성 분석

검색

- '**검색**'의 **의도 파악 불가**
- 부정적 관심도 함께 포함되어 **결과 해석**에 **왜곡** 발생 가능

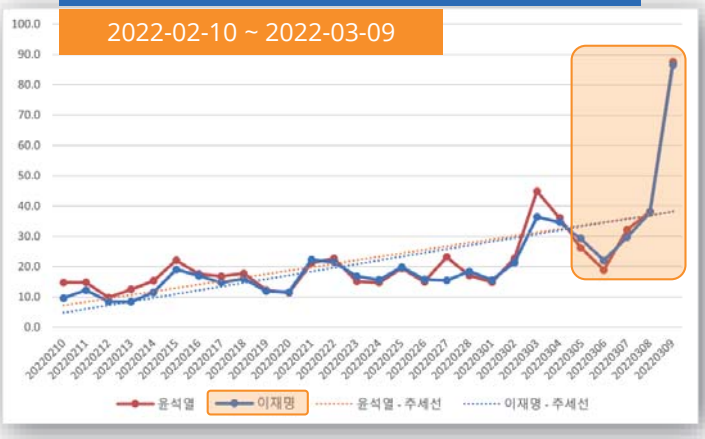
감성 분석

- 작성자의 '태도, 의견, 성향'과 같은 **주관적인 데이터 분석** 가능
- 보통 '**긍정, 부정, 중립**'으로 구분

2022년 20대 대선 검색량 (윤석열 vs. 이재명)

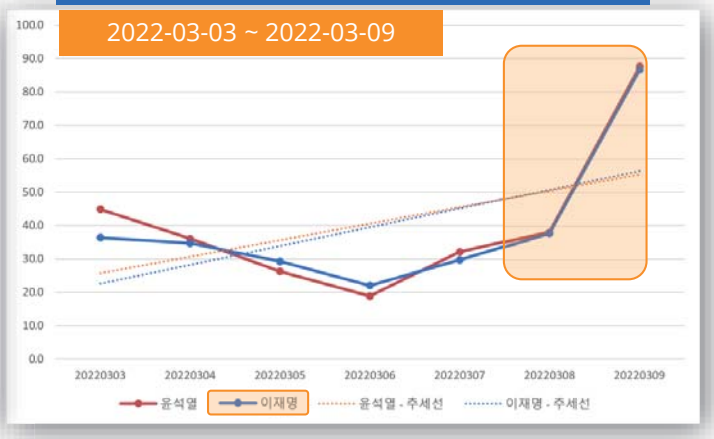
검색엔진 트렌드(투표일 직전 1개월)

2022-02-10 ~ 2022-03-09



검색엔진 트렌드 (투표일 직전 1주일)

2022-03-03 ~ 2022-03-09



	출구조사	개표결과	검색엔진	4주 전	3주 전	2주 전	1주 전
윤석열	48.4%	48.56%	51.4%	55.3%	50.5%	51.0%	50.7%
이재명	47.8%	47.83%	48.6%	44.7%	49.5%	49.0%	49.3%

※ 시장평균 점유율
네이버(62.30%),
구글(26.81%),
다음(4.89%) 반영

2022년 20대 대선 검색량 (윤석열 vs. 이재명)

네이버 트렌드

2022-02-10 ~ 2022-03-09

	1개월	4주 전	3주 전	2주 전	1주 전
윤석열	53.3%	57.4%	52.5%	53.4%	52.0%
이재명	46.7%	42.6%	47.5%	46.6%	48.0%

구글 트렌드

	1개월	4주 전	3주 전	2주 전	1주 전
윤석열	45.1%	47.7%	42.7%	42.2%	46.5%
이재명	54.9%	52.3%	57.3%	57.8%	53.5%

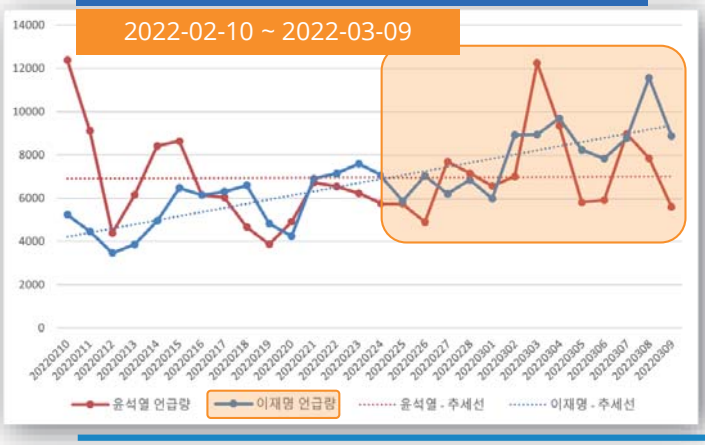
다음 트렌드

	1개월	4주 전	3주 전	2주 전	1주 전
윤석열	50.3%	55.7%	50.0%	55.4%	47.4%
이재명	49.7%	44.3%	50.0%	44.6%	52.6%

2022년 20대 대선 언급량 (윤석열 vs. 이재명)

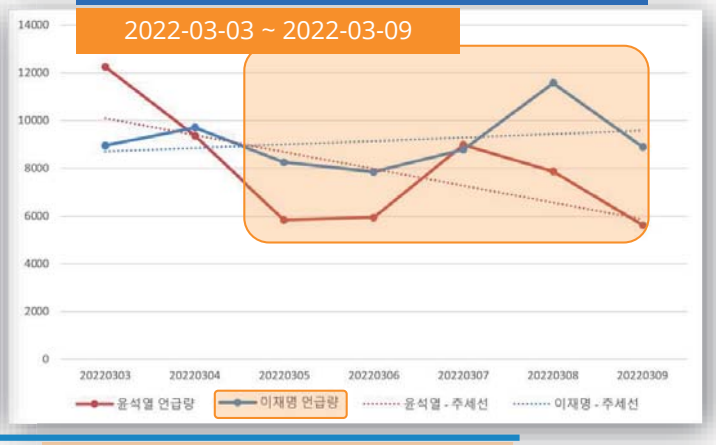
썸트렌드 감성 (투표일 직전 1개월)

2022-02-10 ~ 2022-03-09



썸트렌드 감성 (투표일 직전 1주일)

2022-03-03 ~ 2022-03-09

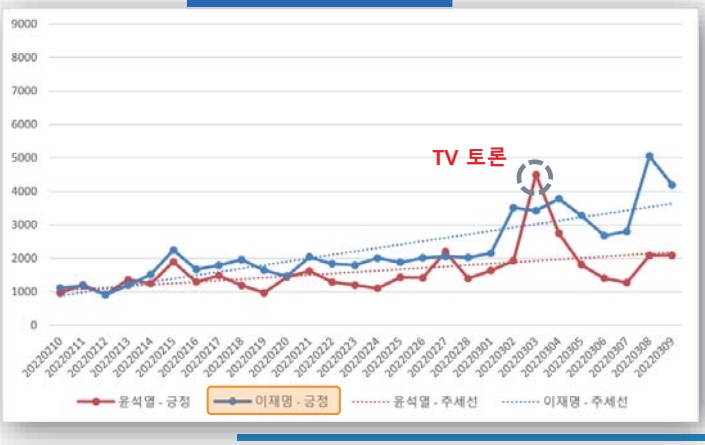


	출구조사	개표결과	검색엔진	썸트렌드	4주 전	3주 전	2주 전	1주 전
윤석열	48.4%	48.56%	51.4%	50.6%	61.5% (55,258)	47.2% (38,980)	48.3% (44,960)	46.6% (55,875)
이재명	47.8%	47.83%	48.6%	49.4%	38.5% (34,617)	52.8% (43,620)	51.7% (48,079)	53.4% (64,002)

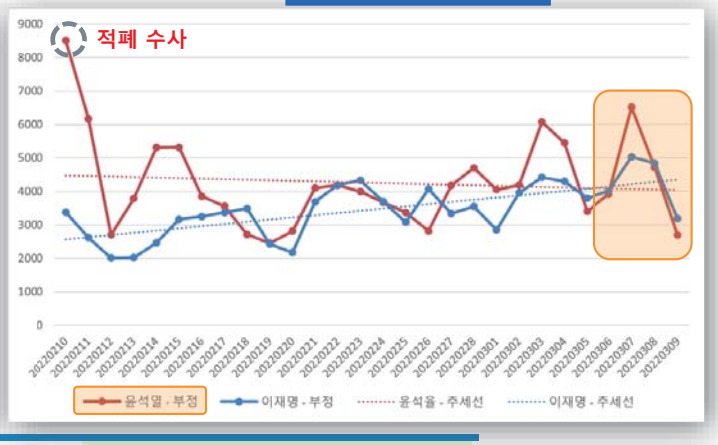
2022년 20대 대선 감성분석 (윤석열 vs. 이재명)

썸 트렌드: 긍정

2022-02-10 ~ 2022-03-09



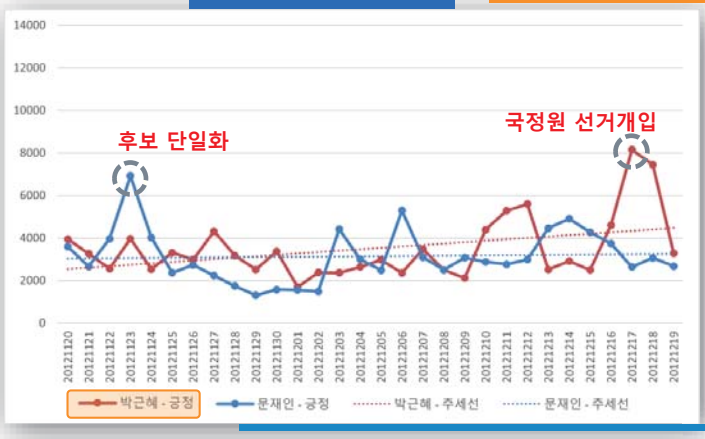
썸 트렌드: 부정

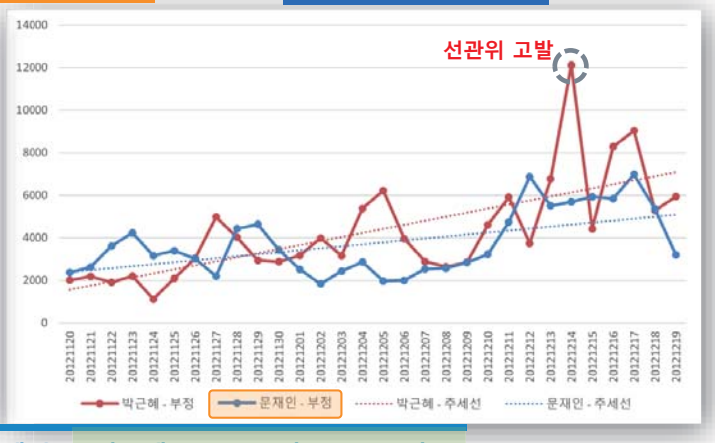


	출구조사	개표결과	검색엔진	썸트렌드	긍정	부정
윤석열	48.4%	48.56%	51.4%	50.6% (195,073)	41.6% (45,115)	55.2% (119,264)
이재명	47.8%	47.83%	48.6%	49.4% (190,318)	58.4% (63,244)	44.8% (96,757)

2012년 18대 대선 감성 분석 (박근혜 vs. 문재인)

썸 트렌드: 긍정 | 2012-11-20 ~ 2012-12-19 | **썸 트렌드: 부정**

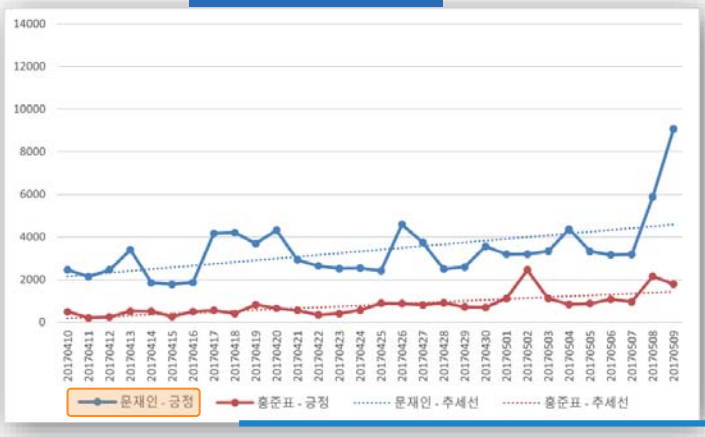


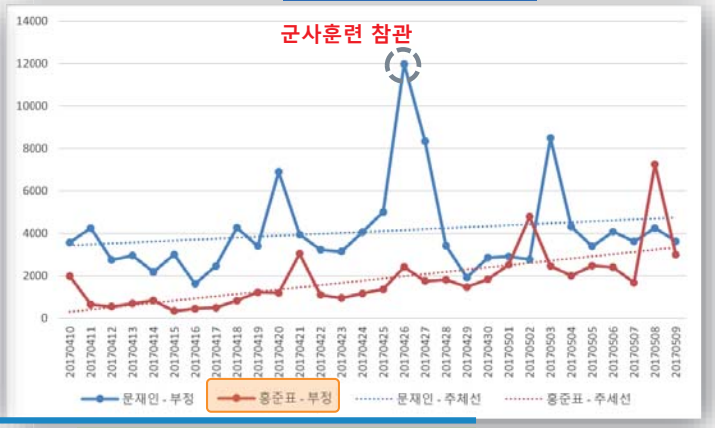


	출구조사	개표결과	검색엔진	썸트렌드	긍정	부정
박근혜	50.1%	51.55%	60.1%	53.4% (286,662)	52.7% (105,174)	53.7% (129,699)
문재인	48.9%	48.02%	39.9%	46.6% (250,158)	47.3% (94,477)	46.3% (111,979)

2017년 19대 대선 감성 분석 (문재인 vs. 홍준표)

썸트렌드: 긍정 | 2017-04-10 ~ 2017-05-09 | **썸트렌드: 부정**

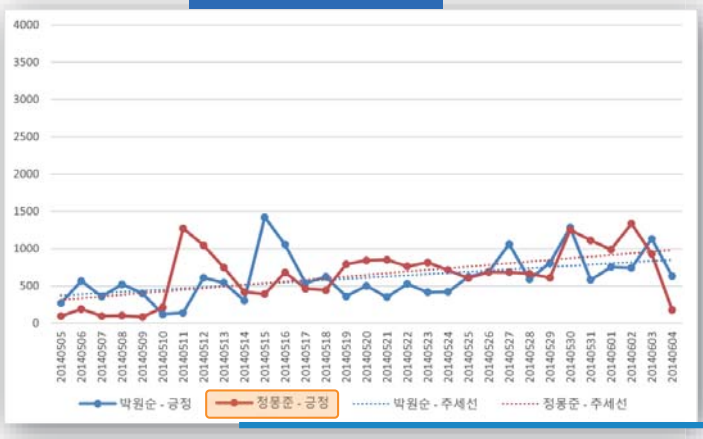


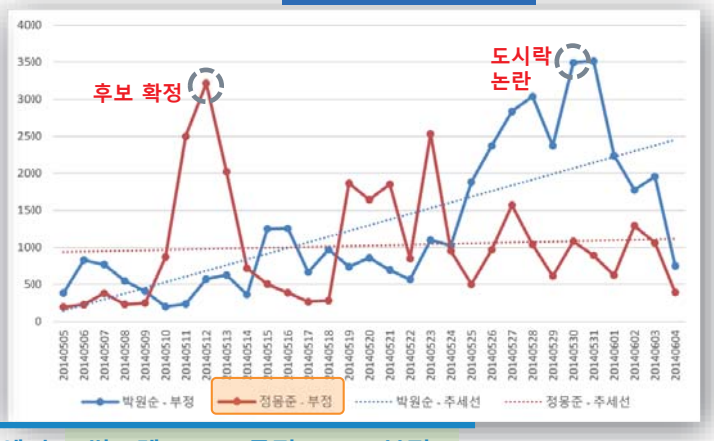


	출구조사	개표결과	검색엔진	썸트렌드	긍정	부정
문재인	41.4%	41.08%	56.1%	74.6% (280,821)	80.6% (101,024)	69.1% (122,815)
홍준표	23.3%	24.03%	43.9%	25.4% (95,835)	19.4% (24,333)	30.9% (54,837)

2014년 서울시장 선거 감성분석 (박원순 vs. 정몽준)

썸트렌드: 긍정 | 2014-05-05 ~ 2014-06-04 | **썸트렌드: 부정**

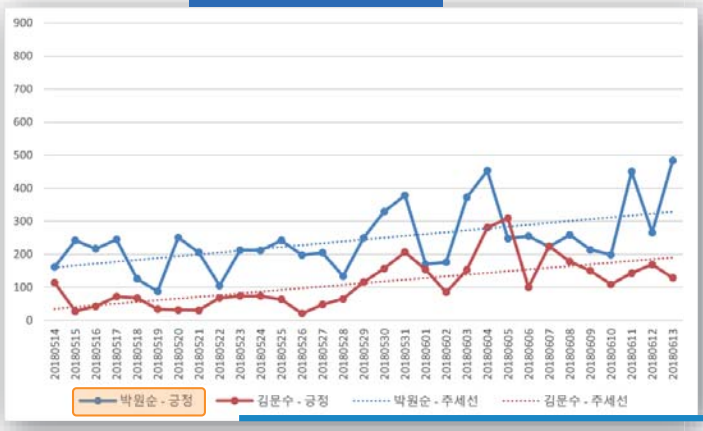


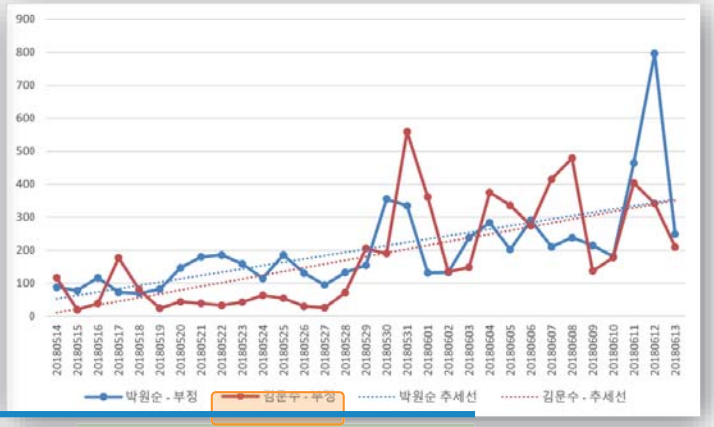


	출구조사	개표결과	검색엔진	썸트렌드	긍정	부정
박원순	54.5%	56.12%	48.2%	54.3% (72,185)	48.6% (18,935)	54.4% (40,303)
정몽준	44.7%	43.02%	51.8%	45.7% (60,663)	51.4% (20,049)	45.6% (31,801)

2018년 서울시장 선거 감성분석 (박원순 vs. 김문수)

썸트렌드: 긍정 | 2018-05-14 ~ 2018-06-13 | **썸트렌드: 부정**

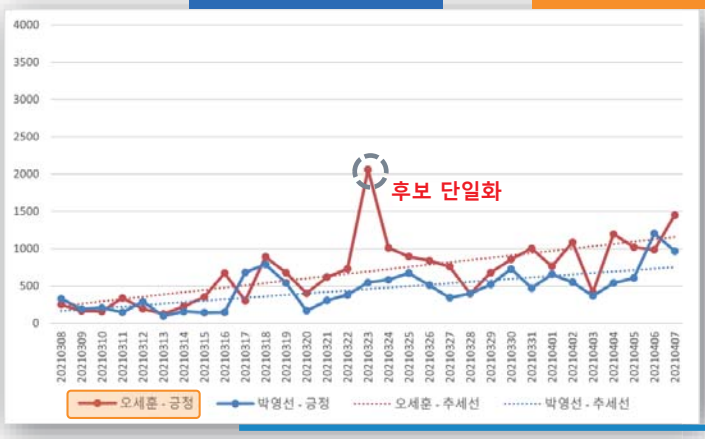


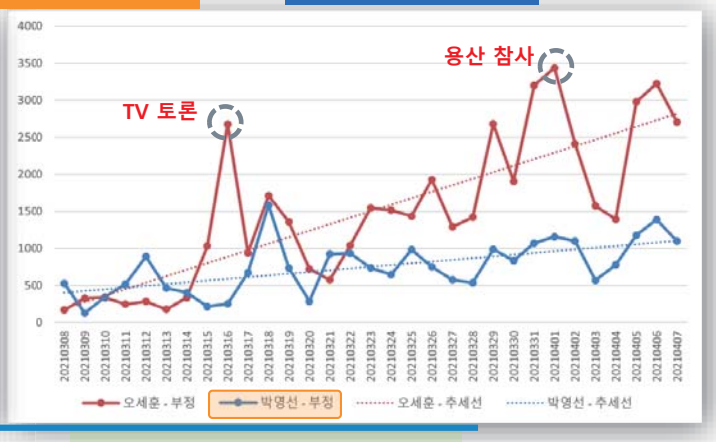


	출구조사	개표결과	검색엔진	썸트렌드	긍정	부정
박원순	55.9%	52.79%	54.6%	60.3% (18,222)	68.4% (7,572)	52.9% (6,317)
김문수	21.2%	23.34%	45.4%	39.7% (12,006)	31.6% (3,501)	47.1% (5,615)

2021년 서울시장 보궐선거 감성분석 (오세훈 vs. 박영선)

썸트렌드: 긍정 | 2021-03-08 ~ 2021-04-07 | **썸트렌드: 부정**

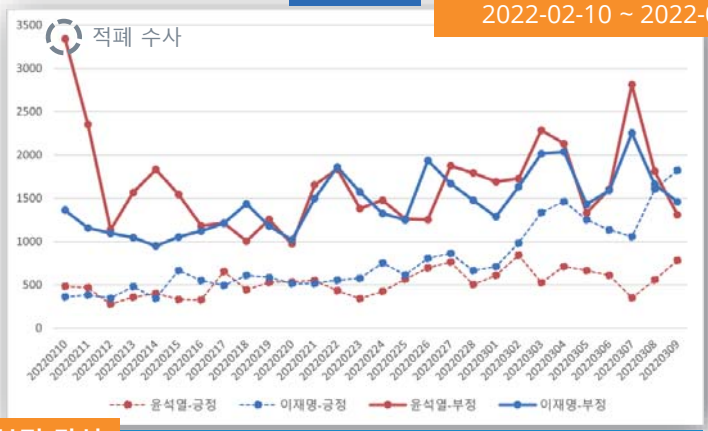


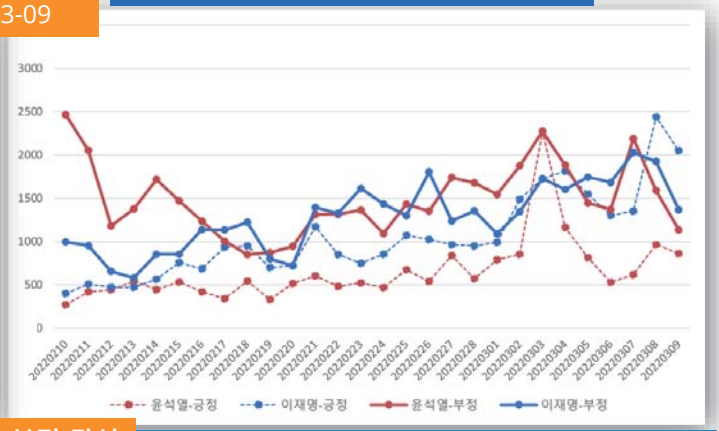


	출구조사	개표결과	검색엔진	썸트렌드	긍정	부정
오세훈	59.0%	57.50%	66.7%	64.3% (86,111)	60.2% (21,510)	66.7% (46,600)
박영선	37.7%	39.18%	33.3%	35.7% (47,731)	39.8% (14,238)	33.3% (23,290)

2022년 20대 대선 감성 분석 (윤석열 vs. 이재명)

트위터 | 2022-02-10 ~ 2022-03-09 | **블로그+커뮤니티+인스타그램**





부정 감성	1개월	4주 전	3주 전	2주 전	1주 전
윤석열	53.5% (46,655)	62.5% (12,962)	48.8% (9,323)	51.2% (11,082)	51.6% (13,288)
이재명	46.5% (40,608)	37.5% (7,790)	51.2% (9,783)	48.8% (10,582)	48.4% (12,453)

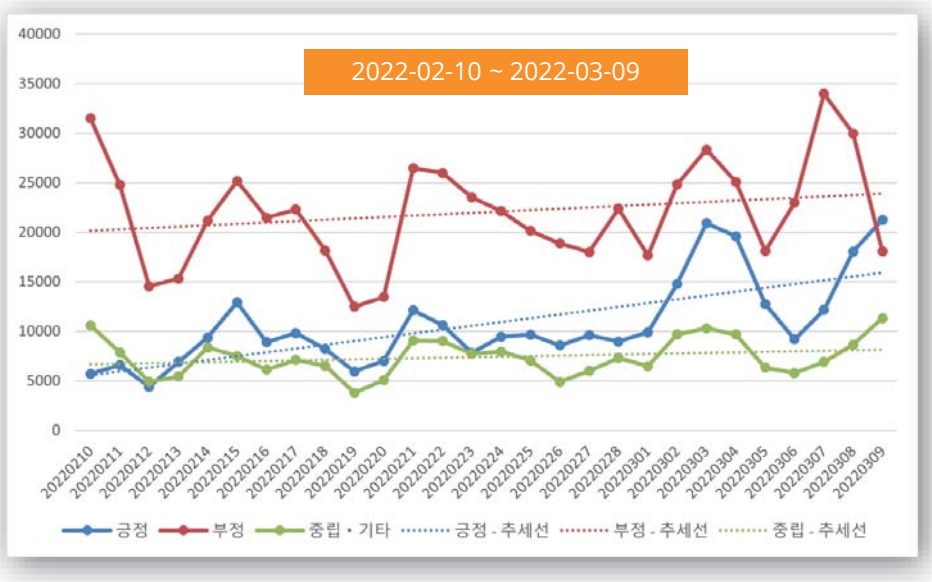
부정 감성	1개월	4주 전	3주 전	2주 전	1주 전
윤석열	53.8% (41,781)	65.6% (11,502)	48.3% (7,665)	52.8% (10,721)	49.6% (11,893)
이재명	46.2% (35,904)	34.4% (6,037)	51.7% (8,214)	47.2% (9,572)	50.4% (12,081)

2022년 20대 대선 공표금지기간 트렌드

		전체	트위터	블로그	커뮤니티	인스타	뉴스	블+커+인
D-0	긍정	이재명 (+13.6%)	이재명 (+14.2%)	**윤석열 (+4.6%)**	이재명 (+18.4%)	**윤석열 (+40.0%)**	이재명 (+3.9%)	이재명 (+9.1%)
	부정	**윤석열 (-2.8%)**	**윤석열 (-1.7%)**	**윤석열 (-19.4%)**	**윤석열 (-6.7%)**	이재명 (-19.1%)	이재명 (-5.2%)	**윤석열 (-6.6%)**
D-1	긍정	이재명 (+14.2%)	이재명 (+15.9%)	**윤석열 (+14.3%)**	이재명 (+16.3%)	**윤석열 (+25.0%)**	**윤석열 (+7.9%)**	이재명 (+2.5%)
	부정	**윤석열 (-1.0%)**	이재명 (-1.6%)	**윤석열 (-1.2%)**	**윤석열 (-6.1%)**	이재명 (-10.5%)	**윤석열 (-1.4%)**	**윤석열 (-2.7%)**
D-2	긍정	이재명 (+8.8%)	이재명 (+10.2%)	**윤석열 (+5.9%)**	이재명 (+10.8%)	**윤석열 (+23.7%)**	**윤석열 (+2.4%)**	이재명 (+3.9%)
	부정	이재명 (-9.2%)	이재명 (-9.5%)	**윤석열 (-11.5%)**	**윤석열 (-0.5%)**	이재명 (-5.9%)	**윤석열 (-22.9%)**	**윤석열 (-2.9%)**

※ 기준: 긍정이 더 높거나, 부정이 더 낮은 후보자

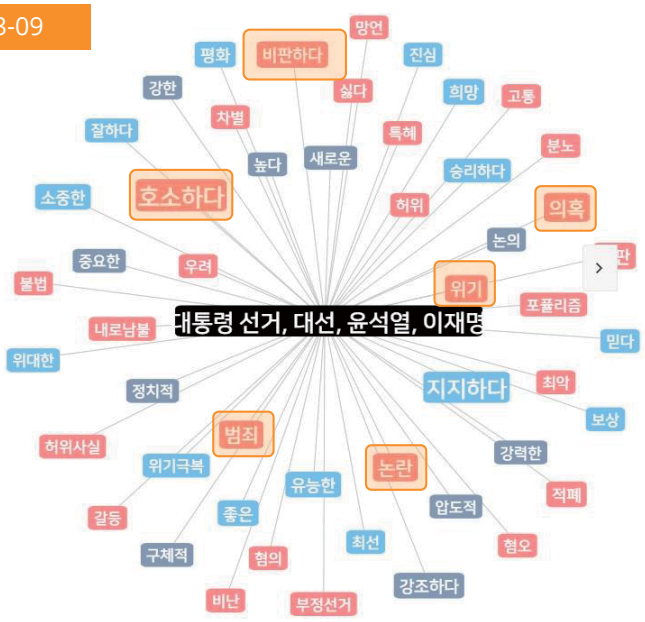
2022년 20대 대선 감성 변화 (썸트렌드)



▷ 선거기간 중 지속적으로 **'부정' 감성**이 **우세**함을 확인

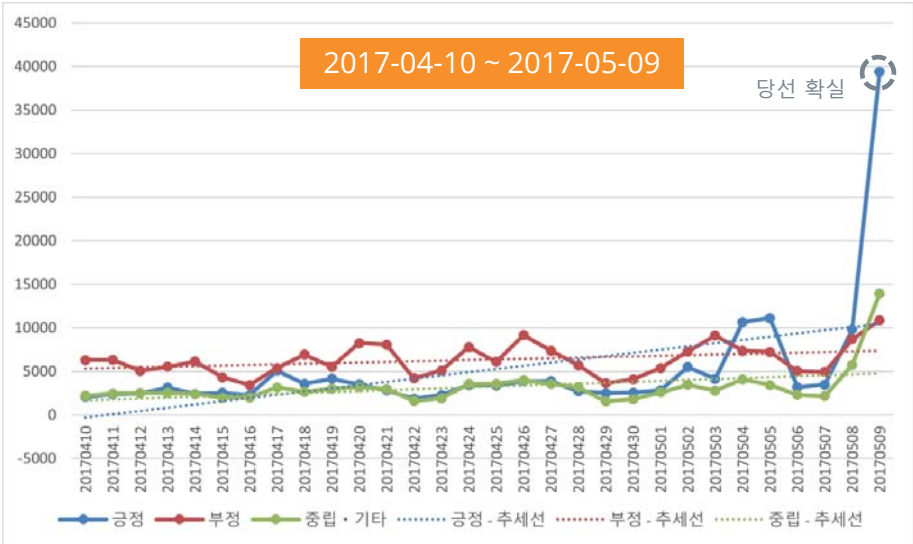
2022년 20대 대선 감성 연관어

2022-02-10 ~ 2022-03-09



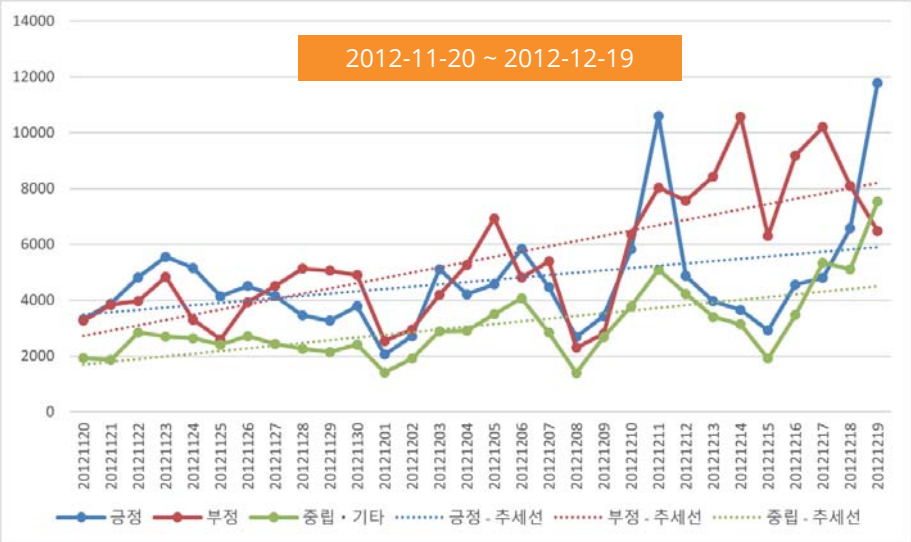
1	호소하다	48,054
2	지지하다	45,400
3	의혹	34,868
4	논란	26,757
5	범죄	25,288
6	비판	20,988
7	위기	20,423
8	비판하다	15,148
9	강조하다	14,354
10	유능하	14,224
11	평화	13,535
12	좋은	11,984
13	희망	11,663
14	소중하	10,663
15	적폐	10,250
16	부정선거	9,964
17	새로운	9,147
18	진심	8,684
19	불법	8,496
20	분노	8,404
21	정치적	8,374
22	차별	7,713
23	허위사실	7,413
24	압도적	6,999

2017년 19대 대선 감성 변화 (썸트렌드)



▷ 선거기간 중 **'부정' 감성**이 우세하였으나 후반에 **'긍정' 감성**이 **우세**

2012년 18대 대선 감성 변화 (썸트렌드)



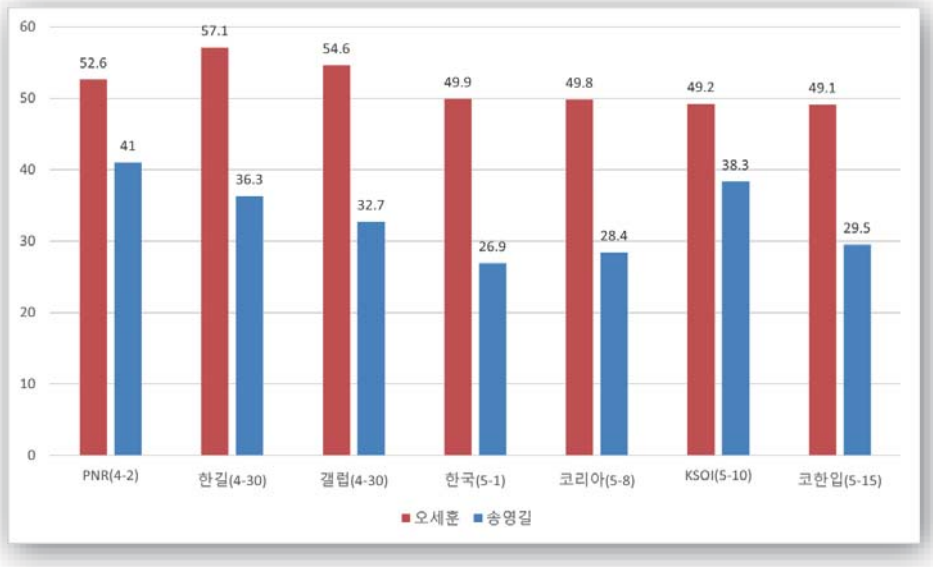
▷ 선거기간 중 **'긍정' 감성**과 **'부정' 감성**이 **교차**됨을 확인

4.

2022년 서울시장 선거 추이 및 예측

썸 트렌드 vs. 포털 트렌드

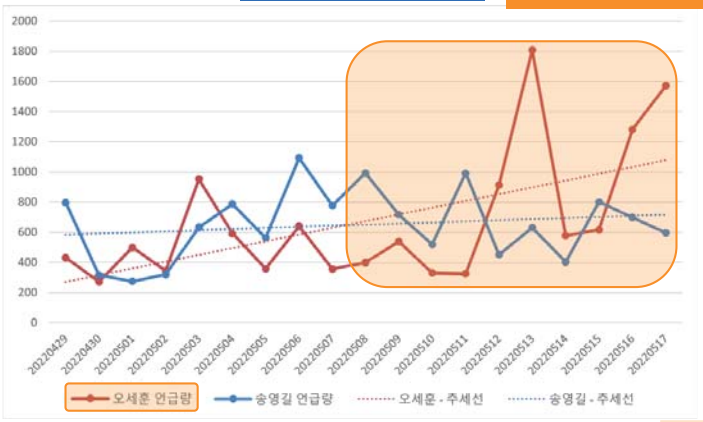
2022년 서울시장 선거 여론조사 (오세훈 vs. 송영길)

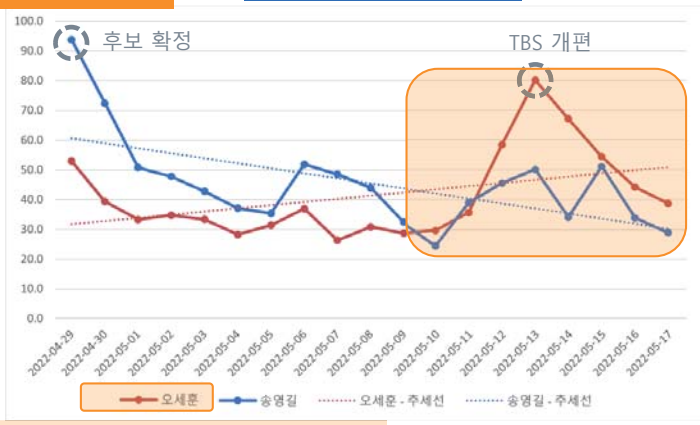


▷ 4월 29일 송영길 후보 확정 이후로 보면 **20% 전후** 차이로 **오세훈** 후보 **우세**

2022년 서울시장 선거 언급량 (오세훈 vs. 송영길)

썸트렌드 감성 2022-04-29 ~ 2022-05-17 **검색엔진 트렌드**



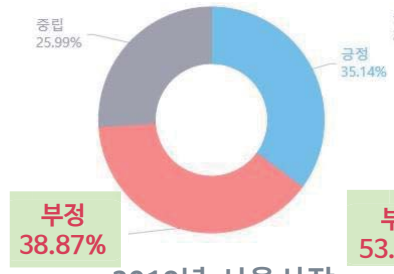


	썸트렌드	검색엔진	썸트렌드(5.8이후)	검색엔진(5.10이후)
오세훈	**50.9%** (12,822)	47.6%	**55.1%** (8,374)	**57.1%**
송영길	49.1% (12,376)	**52.4%**	44.9% (6,819)	42.9%

※ 시장평균 점유율
네이버(63.45%),
구글(25.56%),
다음(5.76%) 반영

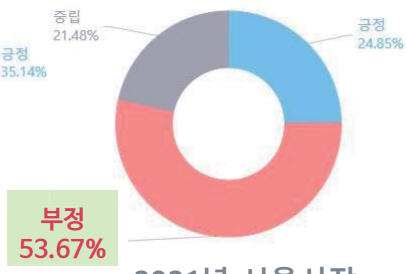
‘서울시장 선거’에 대한 감성 변화

※ 각 선거별 **‘서울시장 선거 or 후보자 이름’ 키워드**에 대한 **감성** 분석



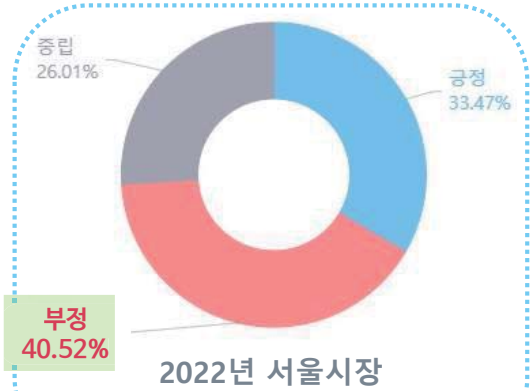
2018년 서울시장

2018-05-14 ~ 2018-06-13



2021년 서울시장

2021-03-08 ~2021-04-07



2022년 서울시장

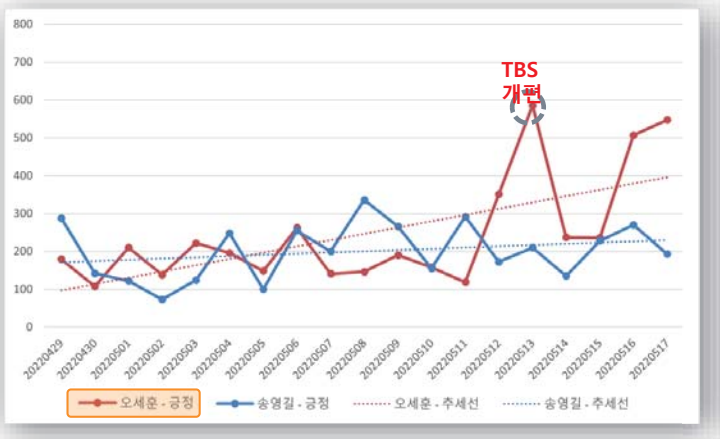
2022-04-29 ~ 2022-05-17

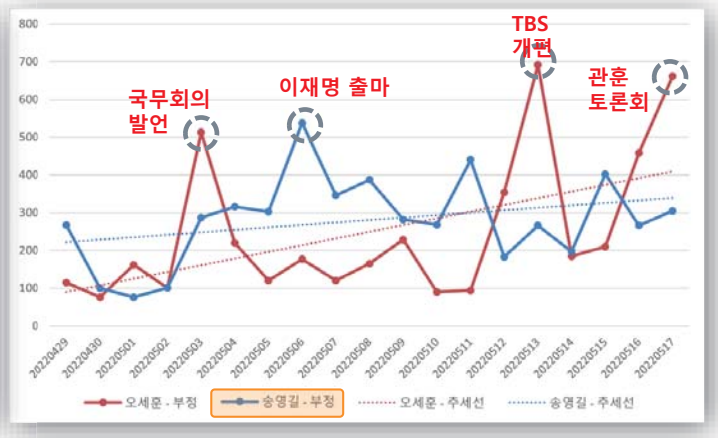
2022년 서울시장 선거 감성분석 (오세훈 vs. 송영길)

썸트렌드: 긍정

2022-04-29 ~ 2022-05-17

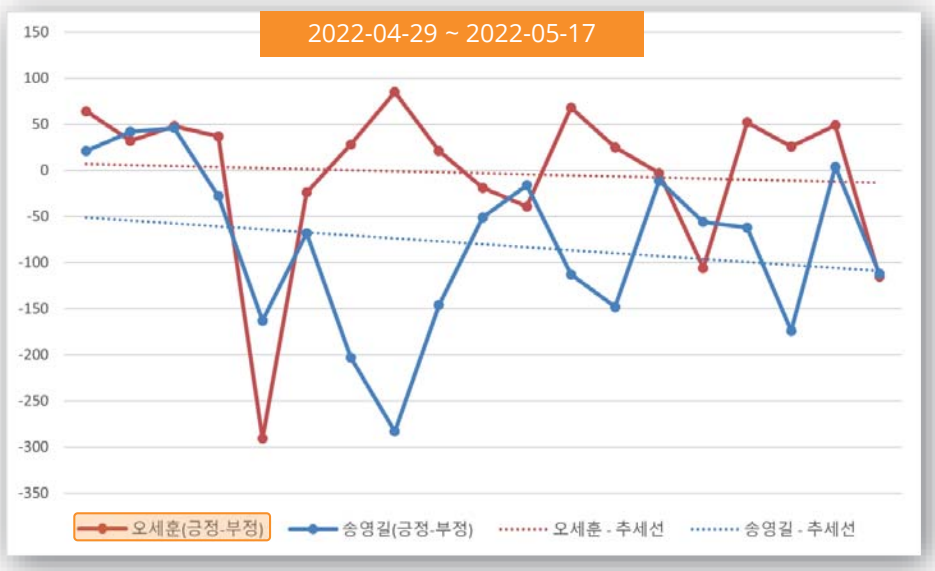
썸트렌드: 부정





▷ **긍정** 감성은 **오세훈** 후보 **우위(+)** **부정** 감성은 **송영길** 후보가 **우위(-)**

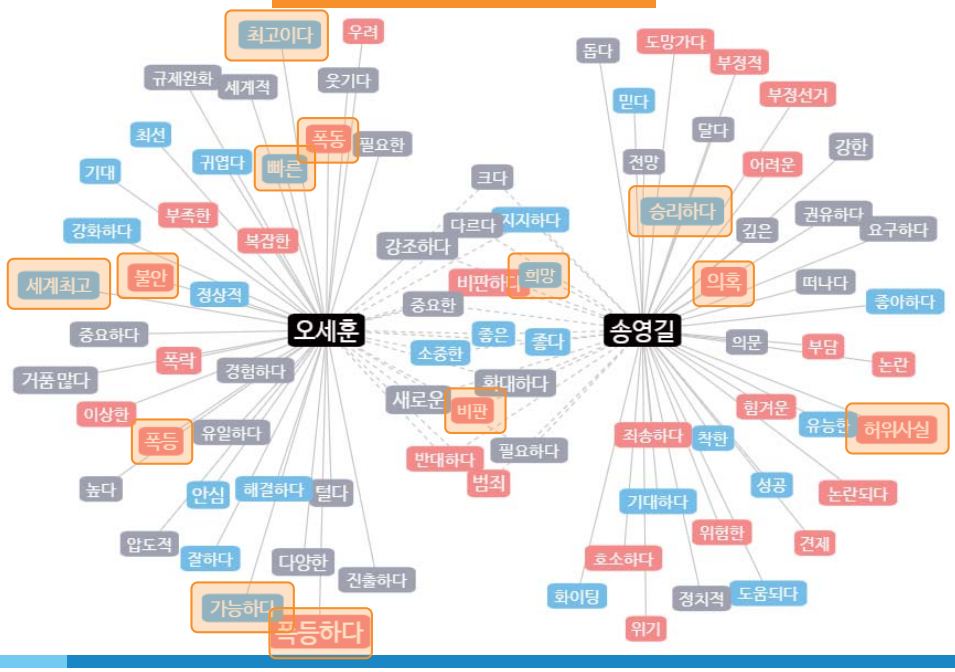
2022년 서울시장 선거 감성분석



▷ **긍·부정통합**(긍정-부정)은 **오세훈** 후보가 **우위**

2022년 서울시장 선거 감성 연관어

2022-04-29 ~ 2022-05-17



[오세훈] 연관어 순위

순위	연관어	건수
1	폭등하다	438
2	새로운	187
3	빠른	174
4	불안	161
5	폭동	159
6	최고이다	158
7	세계최고	153
8	폭등	147
9	가능하다	146
10	털다	146
11	비판하다	139
12	확대하다	139
13	좋다	131
14	폭락	125
15	범죄	121
16	안심	119
17	강조하다	114
18	거품 많다	114
19	다양한	113
20	크다	101
21	정상적	91
22	잘하다	89

[송영길] 연관어 순위

순위	연관어	건수
1	비판	233
2	의혹	217
3	강한	164
4	희망	163
5	허위사실	155
6	새로운	151
7	필요하다	145
8	승리하다	137
9	크다	124
10	강조하다	109
11	좋은	108
12	어려운	104
13	범죄	97
14	확대하다	94
15	다르다	91
16	비판하다	90
17	권유하다	90
18	좋다	83
19	반대하다	81
20	견제	80
21	위기	80
22	중요한	80

빅데이터로 선거 결과를 예측할 수 있을까?

빅데이터로 선거 예측이 가능하다?

- 여론조사도 중요하지만, 여론조사로 확인할 수 없는 부분은 '**빅 데이터**'를 **활용**하여 **보완**이 가능
- 하지만, 단순히 검색량 데이터 만으로는 작성자의 **숨어있는 의도**를 **파악**하기 **힘듦** (예: 2020년 미국 대선, 2022년 20대 대선 등)

빅데이터로 선거 예측이 가능하다?

- 작성자의 감정을 파악할 수 있는 '**감성분석**' **데이터**를 **활용**한다면 선거 예측에 도움
- 더 나아가, '**빅 데이터**'를 **정기적**으로 **분석**해서 여론을 살피고 그에 맞는 **선거 전략**과 **전술**을 **기획**하는 것이 효과적

Thanks!

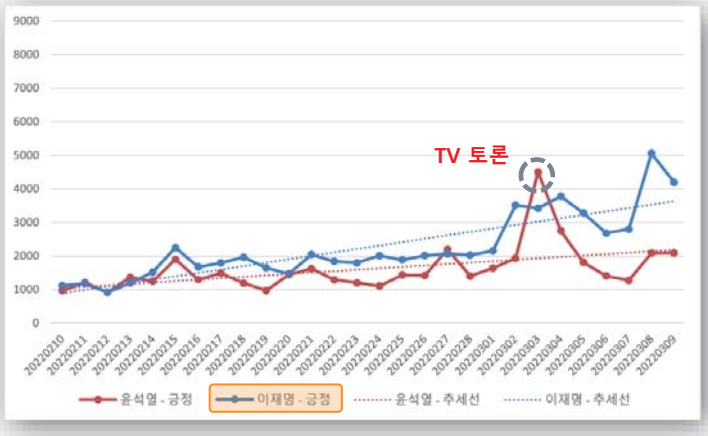
Any questions?

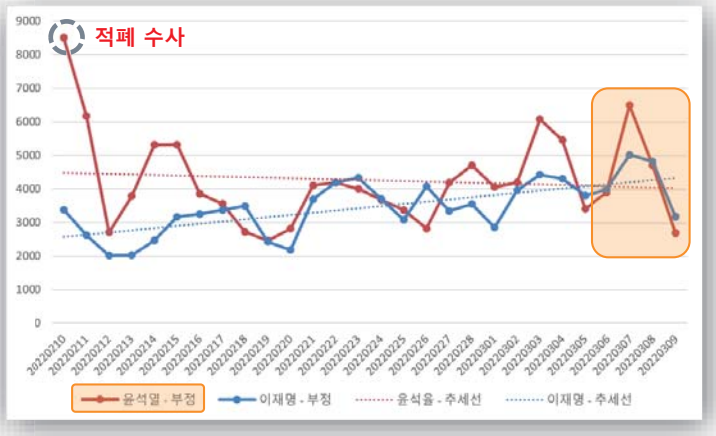
참고문헌

- Sometrend (https://biz.some.co.kr/)
- 인터넷트렌드 (http://www.internettrend.co.kr/trendForward.tsp)
- 네이버 트렌드 (https://datalab.naver.com/keyword/trendSearch.naver)
- 구글 트렌드 (https://trends.google.co.kr/trends/?geo=KR)
- 카카오데이터트렌드 (https://datatrend.kakao.com/)
- 빅데이터로 선거 결과를 예측할 수 있을까? (https://eiec.kdi.re.kr/publish/naraView.do?cidx=11538)
- Google Trends as a Predictor of Presidential Elections: The United States Versus Canada (Camilo Prado-Román, Raúl Gómez-Martínez, and Carmen Orden-Cruz)
- Market share of search engines in United States 2008-2022 (https://www.statista.com/statistics/267161/market-share-of-search-engines-in-the-united-states/)
- "구글 추격 따돌리자"...네이버 검색 지금보다 똑똑해진다 (https://m.mk.co.kr/news/it/view/2021/05/463524/)
- 이미지 (https://www.pngegg.com)

2022년 20대 대선 감성 분석 (윤석열 vs. 이재명)

긍정 | 2022-02-10 ~ 2022-03-09 | 부정





- **네거티브 선거**를 고려하면, 감성 언급량이나 '긍정' 감성 보다는 **더 낮은 '부정' 감성(-)**이 **중요**

2022년 대선 여론조사기관별 비교

리얼미터

조사기간조사의뢰자	이재명	윤석열
12.20~21 YTN	41.3	45.6
12.6~7 YTN	42	46.9
12.4~5 폴리뉴스	44.1	46.4
11.26~27 YTN	39.7	48.6
11.22~23 YTN	39.3	48.9
11.8~9 YTN	37	50

엠브레인퍼블릭

조사기간조사의뢰자	이재명	윤석열
1.24~25 문화일보	37.1	43.2
12.30~31 중앙일보	47.9	35.1
11.26~27 중앙일보	40.7	43.8
11.7~8 news1	38.6	39.4

코리아리서치

조사기간조사의뢰자	이재명	윤석열
1.7~8 MBC	44.5	39.2
12.29~31 MBC	44.4	35.2
12.11~12 MBC	40	44
11.27~28 MBC	38.3	43.9
11.6~7 MBC	38.3	44.3

빅데이터 기술 교류 세미나

빅데이터와 여론조사

발표5

빅데이터와 AI 기반 여론 분석 연구와 교육 사례: 썸트렌드 vs. 코딩

이경전(경희대 교수), 박아름(용인예술과학대 교수)

2022 한국경영정보학회 기술교류 세미나 No.1: 빅데이터와 여론조사

빅데이터와 AI 기반 여론·정책 분석 연구와 교육 사례: 썸트렌드 vs 코딩

2022.5.19

이경전*, 박아름**

*경희대학교 경영대학 & 빅데이터응용학과 교수
**용인예술과학대학교 빅데이터경영과 교수

한국경영정보학회
The Korea Society of Management Information Systems

VAIV

용인예술과학대학교
YONG-IN ARTS & SCIENCE UNIVERSITY

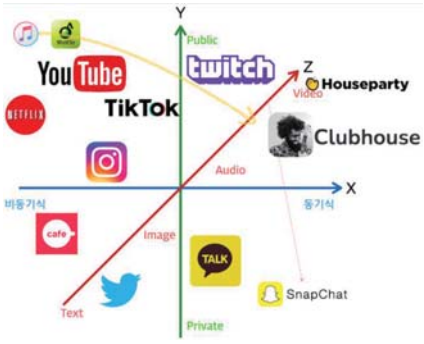
새로운 미디어의 등장과 정책: 빅데이터·AI 활용 분야 및 활용성 확대

- 2002년 12월 대선
  - 오마이뉴스의 등장(vs. 종이 신문)
- 2004년 12월 다음 아고라 서비스 시작
- 2007년 12월 대선
- 2011년 10월 서울시장 보궐선거
  - 2011년 4월 팟캐스트 '나는 꼼수다 ' 첫방송
- 2011년 12월 종편방송 개국 (vs. 공중파TV)
- 2012년 12월 대선
  - 국정원 댓글 부대 사건, '나는 꼼수다' 종영 (12.18)
- 2016년 가을 종편 방송 '최순실 사건' 적극 보도
- 2016년 9월
  - TBS 김어준의 뉴스 '공장 ' 첫 방송
- 2017년 5월 대선
  - 드루킹 사건
  - 여론조사 발표의 영향력 및 오남용(예: 역선택)
- 2017년 8월 청와대 국민청원 신설
- 2018년 12월 아고라 폐지
- 2019년 10월 네이버 악성 댓글 차단 클린봇 서비스 시작
- 2020년 4월 총선
  - 빅데이터 선거: 민주연구원
  - 유튜브 선거
- 2020년 5월: 415총선 부정선거 논란 유튜브 미디어에 의해 촉발
- 2020년 6월: 빅데이터·AI 전문가 여의도 연구원장 영입 추진
- 2020년 7월 & 8월: 포털3사 연예 & 스포츠 댓글 폐지
- 2020년 12월, 2021년1월: 미국 대선 부정선거 논란 및 의사당 점거 폭력 사태
- 2021년 1월, 2월: 클럽하우스 열풍
- 2021년 4월: 서울, 부산 시장 보궐선거
- 2022년 3월 대선
  - AI 선거: 크라켄, AI 윤석열, "매타"버스
  - 봄뿌, 에펨코리아 등 자체 커뮤니티의 영향력
- 2022년 6월 지선: ???
- 2024년 총선: ????
- 2027년 대선: ????

미디어 현황과 빅데이터·AI 활용 정책 연구의 필요성

- 여론이 소셜미디어와 댓글로 표현되어 정책수요/결정자에 영향 미치며, 데이터로 축적, 분석 가능해져, AI/빅데이터 활용 분석 능력이 정책 능력에 영향을 미치는 상황.
- 온라인 국민청원, 인플루언서의 컨텐트 기사화 등 다양한 의제 설정 채널 등장
- 세대간 미디어 사용의 분절로 인식의 격차가 커지고 있고, 소셜 미디어 알고리듬이 사용자의 정보 편식, 쏠림 현상을 초래하고 있으며, 카톡방, 텔레그램방 등을 활용한 폐쇄적 정보 공유로 탈진실의 증폭이 이루어지면서, 정치 성향의 양극화, 정책의 비합리성, 포퓰리즘의 위험이 커져 합리적인 정책합의가 더욱 어려워질 위험이 커지고 있음
- 신속 정확한 여론 분석 능력을 갖추고, 새로운 의제 설정 기회를 활용하며, 세대간, 집단간 분절, 쏠림, 양극화 현상에 대한 대안을 제시하는 정책 연구 방법론 필요

넷플릭스 최대 경쟁자는 여기다, Z세대가 열광하는 '메타버스', 중앙일보 2021.2.18, mnews.joins.com/amparticle/23994539



https://brunch.co.kr/@ioojoo/99

빅데이터·AI기반 정책 연구

- 의제 설정, 정책 결정, 집행, 평가 등 전과정에 빅데이터·AI방법론을 활용
  - 국민의 의견을 신속, 정확히 반영하여 정책을 수립하는 과학적 방법론으로서의 AI
  - 정책 현안과 수요 발굴 및 정책 반응 탐지를 감이나, 여론조사, 전문가에 의존했다면, 비대면 미디어와 AI·빅데이터 등의 도구 활용하여 정책 현안 발굴하고 및 정책 수요자와의 커뮤니케이션을 통해서 수요자의 반응을 고려하여 정책을 튜닝, 미세조정 하여, 정책 수요자에게 적합하고 실질적 도움되는 효과를 기대할 수 있는 정책을 만드는 방법 필요
  - 왜 정책 수요자의 선호도가 바뀌었고, 선호를 바꾸기 위해서 어떤 정책이 필요한가?
  - 정책 성공을 위해 파라미터를 어떻게 조정하여, 맞춤형 정책들을 만들어낼 것인가?
- 참고: 정치나 경영에 AI를 접목한다면?, 전자신문, 2021.2.9.etnews.com/20210209000191
  - "이해관계 대립에 따른 사회적 비용이나 갈등이 증가되는데 여전히 위원회, 청와대 국민청원 등의 불완전한 방법에 의존한다"
  - 사회문제에 대해 미디어를 통해 표출되는 집단 간 의견 수집해 합의 가능 지점 찾는 AI
  - 여론조사, 온라인 청원 등 갈등과 이견 해소 위한 기존 노력 넘어 다양한 의견 간 차별성과 유사성 분석하고, 상호 동의 가능 부분 포착해 집단 간 토론과 의견 공유 지원

빅데이터 · AI 활용 정책 수립 시스템 연구 주제들

- 키워드 분석을 통한 정책 주목도 분석
- 특정 정책 관련 뉴스/댓글 감성 분석을 통한 지지도 분석
- 정책수요자 샘플링과 AI활용을 통한 효율적인 여론 조사
- 댓글러/정책수요자 선호/성향 변화 추이/이유 분석: 사건/사고, 정책, 정치인
- 정책 반응 분석 (뉴스 댓글 반응) 및 피드백을 통한 정책 변화 및 파라미터 조정 연구
- 정책수요자 클러스터링을 통한 맞춤형 정책 설계 및 제시
- 정책 관련 키워드에 대한 빅데이터 분석을 통한 정책 반응 탐지 및 정책 보완 아이디어 도출 방법론
- 특정 정책에 대해서는 도메인 지식에 기반한 모델 및 시뮬레이션, 사례 연구에 의한 정책 대응 방법론
- 정책 관련 미디어 전략의 효과성 평가
- 정책수요자 성향 고려한 감성 분석 고도화: 크롤링, 키워드/유사어 추출, 감성분석, 요약 모델 개발
- 크라우드 소싱(Crowdsourcing)형태의 협업 및 동시 다발적인 데이터 수집과 분석 및 정책 피드백 시스템 구축

빅데이터· AI 기반 여론 조사 연구 사례 및 활용

(사례)

- 특정 키워드 빈도 분석을 통한 주목도 분석
- 특정 정책 관련 뉴스와 댓글의 감성 분석을 통한 지지도 분석 (AI에 의한 감성 분석 필요)
- 특정 정책에 관련한 긍정적 뉴스, 부정적 뉴스에 대한 통계(AI에 의한 감성 분석 필요)
  - 특정 정책에 대한 긍정적 뉴스, 부정적 뉴스의 수
  - 특정 정책에 대한 긍정적 뉴스, 부정적 뉴스에 대한 Like/Dislike 수
  - 특정 정책에 대한 긍정적 댓글 수와 이 댓글에 대한 Like 수, dislike 수
  - 특정 정책에 대한 부정적 댓글 수와 이 댓글에 대한 Like 수, dislike 수

(활용)

- 내부 정보로만 사용
- AI분석 결과 공표에 의한 Signaling Effect에 유의해야 함
  - 왜곡의 유혹, AI 모델 성능의 문제, ‘AI’에 대한 대중의 과신
- 정책 커뮤니티에 공유
- 민간 서비스 및 불법적 사용자, 사용자 그룹의 감시
  - 실행 사례: 크라켄

연구 사례 1: K-정책플랫폼(K-Pol.org) 연구 사례
2021.1 ~ 2021.4

01. 데이터 수집 및 전처리

- 인터넷 포털(네이버, 다음)과 SNS(유튜브, 트위터)에서 데이터 수집
- 오타 검사, 토큰화 등 데이터 분석을 위한 텍스트 전처리

02. 키워드 탐색 및 빈도 분석

- 수집된 데이터에서 관련된 중요 키워드 추출
- 탐색된 키워드의 빈도를 분석하여 여론 반응 탐색

03. 감성 분석

- 수집된 데이터에 대한 감성 분석수행
- 선거 후보자에 대한 이미지 분석

연구 사례 1: K-정책플랫폼(K-Pol.org) 연구 사례

댓글 키워드 빈도 분석

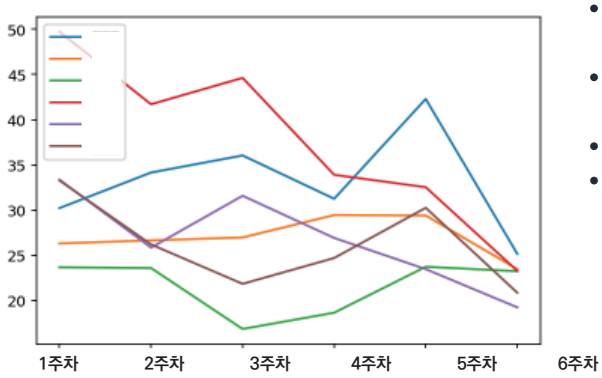
형태소에 따라 나눠진 데이터를 기반으로 빈도 분석 진행

- WordCloud 사용하여 수집된 뉴스 댓글에 대한 워드 랭킹 출력
- 특정 기간 가장 많이 언급된 단어
- 특정 기간동안 순위 변동 체크

	word	count
0	안철수	15052
1	국민	11410
2	사람	7306
3	민주당당	7126
4	생각	6619
5	문재인	6613
6	서울시장	6216
7	대통령	6203
8	정치	6171
9	나라	5159
10	김종인	4795
11	선거	4286
12	후보	4185
13	시장	3901
14	나경원	3876
15	경영	3677
16	국민들	3634
17	정권	3437
18	서울	3108
19	소리	2982
20	국민의힘	2917
21	하나	2854
22	야당	2854
23	단일화	2690
24	인간	2530
25	지지	2488
26	박영선	2442
27	사람들	2352
28	대한민국	2343
29	보수	2296
30	당신	2280
31	정신	2238
32	오세훈	2201

댓글 감성 분석

- 긍정 댓글 vs. 부정 댓글 비율 분석
- KoBERT에 Fine Tuning 과정 거쳐 모델 재학습
- 분석단위 진화: 단어 -> 댓글 -> 댓글러
- 단어 단위 감성 분석은 어느 정도 정확하나 댓글 단위의 감성 분석에 대해서는 더 향상될 여지를 확인하였으며, 댓글러 단위가 오히려 의미가 있음도 확인.



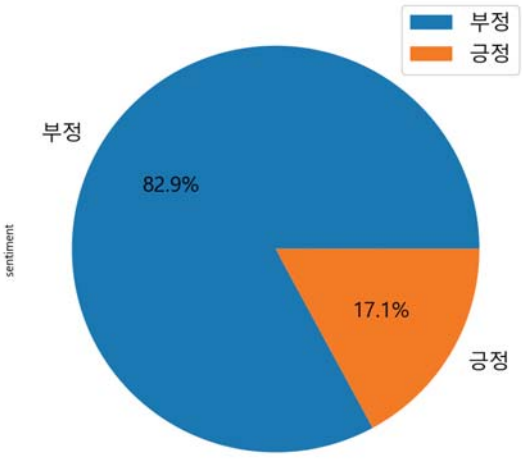
- 감성 분석 결과 전체적으로 각 후보자에 대한 이미지 감소
- OOO' 후보자는 긍정 댓글의 비율이 가장 높았지만 점차 감소하여 5주차에 2위로 하락
- 'OOO' 후보자는 긍정 댓글 비율이 3위 였지만 점차 증가하여 5주차 1위 달성
- OOO-OOO 단일화가 필요했던 이유
- 2021.3.23: 야권 단일 후보 선출

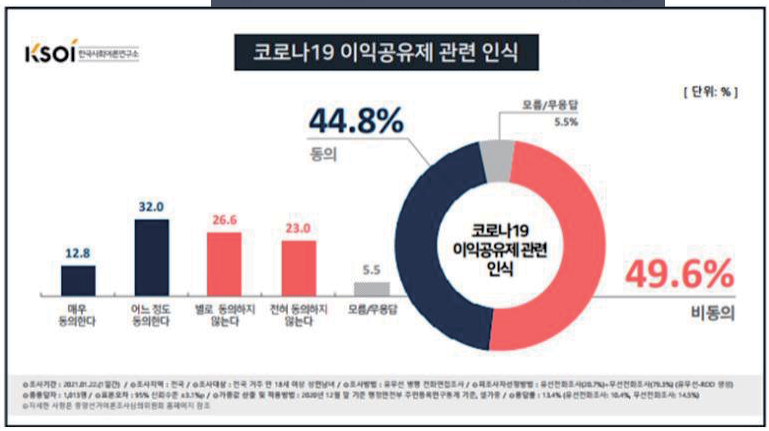
2021.1.17

연구 사례 1: K-정책플랫폼(K-Pol.org) 연구 사례

정책 여론 분석 사례: 이익 공유제

- 뉴스 댓글 분석 결과 이익 공유제 관련 인식과 한국사회여론연구소에서 조사한 이익 공유제 관련 인식에서 큰 차이 확인
- 주주대상 전경련 조사와는 상대적으로 차이가 적음



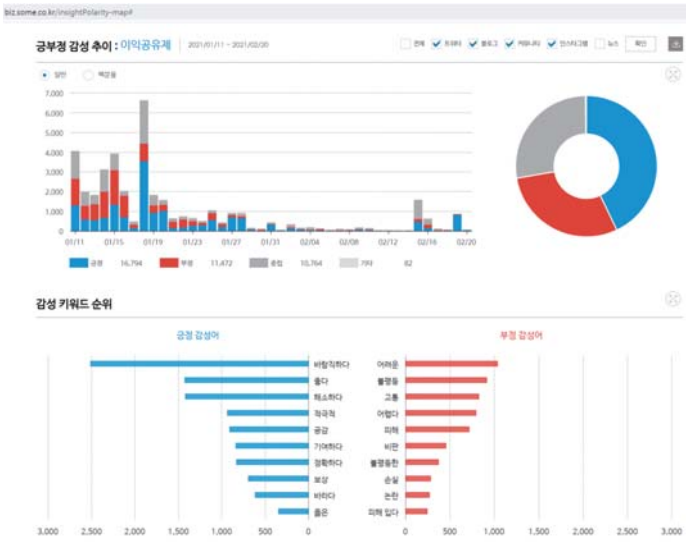


이익 공유제 반응(댓글) 분석 사례

- SomeTrend 활용 감성 분석 수행
  - 뉴스와 SNS가 결과가 다름을 확인

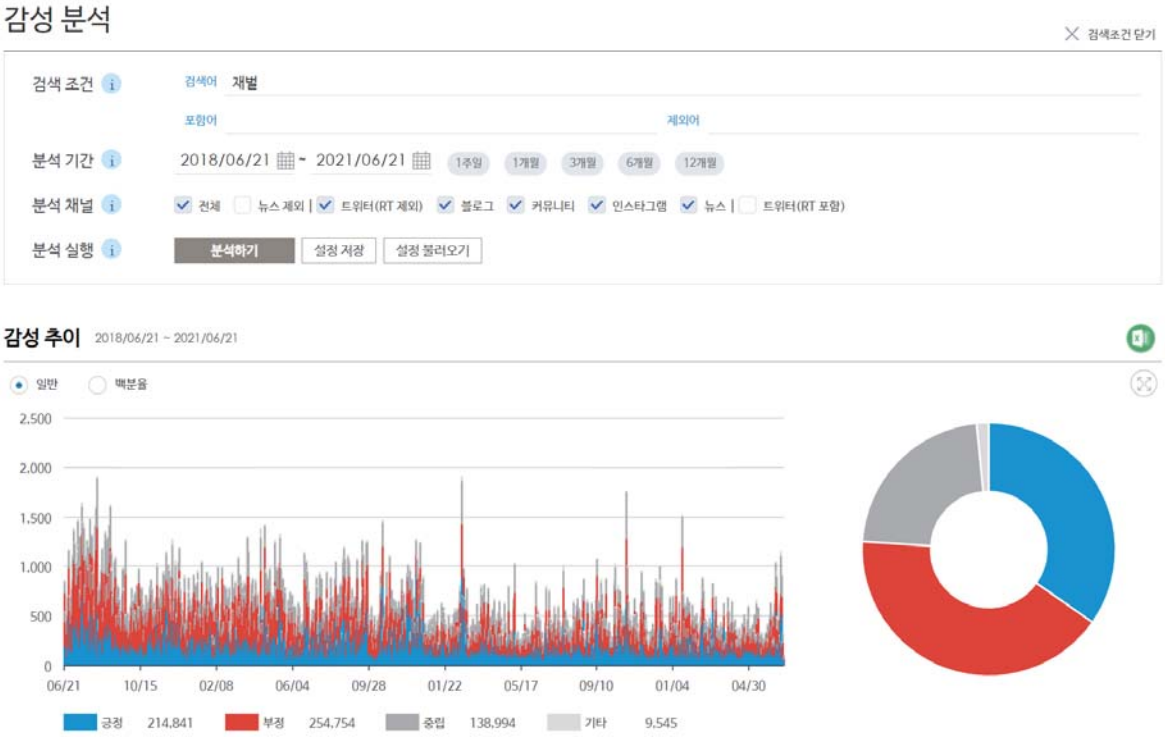
뉴스만 분석

SNS만 분석



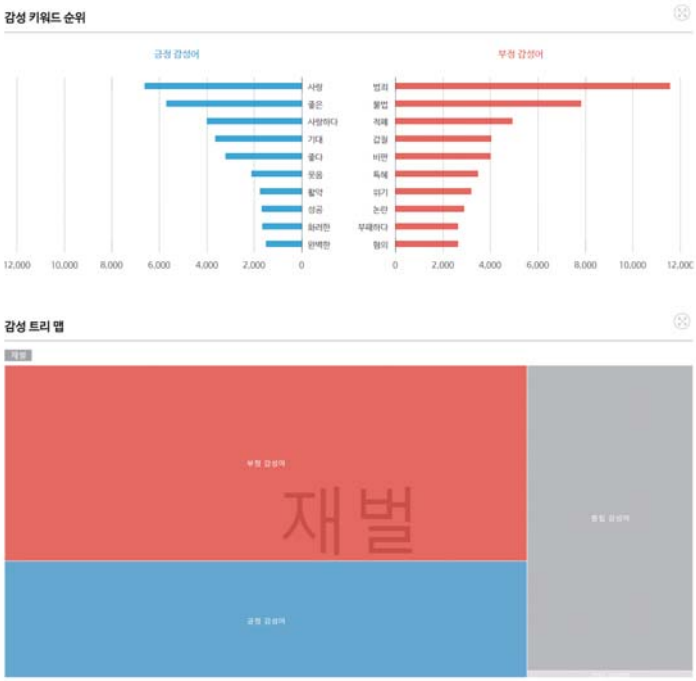
연구 사례 1: K-정책플랫폼(K-Pol.org) 연구 사례

Sometrend 활용 반기업/반재벌 정서 빅데이터 분석 사례



Sometrend 활용 반기업/반재벌 정서 빅데이터 분석 사례

사랑
좋은
기대
좋다
웃음
활약
성공
화려
완벽



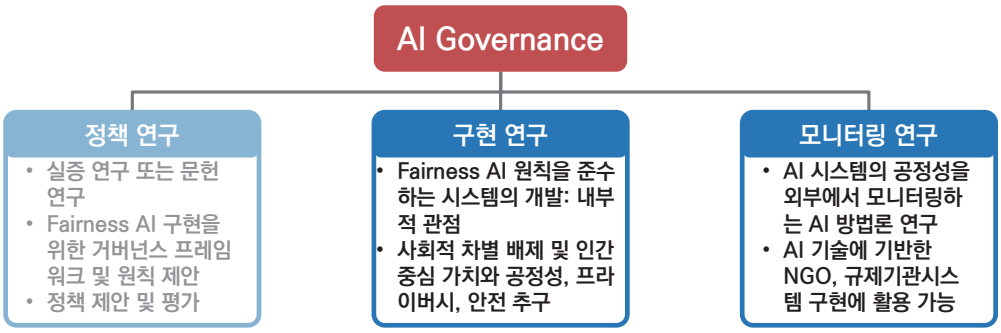
범죄
불법
갑질
특혜
부패
혐의

K-POL(K-정책플랫폼) 빅데이터· AI 기반 여론 조사 연구 시사점

- 댓글 감성 분석 고도화 필요 – 이 부분이 정확하기 전에는 이를 자동적인 여론 조사 결과로 공표하는 것은 무리
- 뉴스, 뉴스 댓글, SNS, 기존 여론 조사의 결과가 상충되는 경우 이를 앙상블하고 대시보드화 하는 방법론 필요
- 빅데이터· AI 기반 여론 조사는 자동화하여 발표하는 Signaling Effect를 주기 보다는 내부적 정책 개선 및 전략 수립에 활용하는 것이 바람직
  - 댓글 수준의 감성 분석이 95%~99%의 정확성을 가지게 될 경우는 자동화 발표도 가능할 것임
  - 댓글러의 성향 변화는 댓글 수준의 감성 분석보다 먼저 더 높은 정확성을 가지게 될 것으로 예상되므로, 댓글러 패널 구성에 의한 여론 조사 방법론을 연구 개발하는 것은 의미가 있을 것임
- 다만 기존 여론 조사가 발표되었을 경우 어느 정도 공신력을 갖춘 기구를 통해 빅데이터· AI 기반 여론 조사 결과가 경쟁 및 보완하는 체제로 갈 필요가 있음
- 빅데이터· AI 기반 조사는 기술, 경험, 데이터가 축적되어서 일종의 데이터 효과를 일으키는 분야
  - 데이터 효과 (Data Effect): 제품과 서비스, 비즈니스 모델에 데이터가 쌓임에 따라 경쟁력과 가치가 커지는 효과(coined by 이경전)
  - K-POL(K-정책플랫폼)과 같은 민간 ThinkTank 등 민간과 여의도연구원, 민주 연구원 등 정당 연구소에서 빅데이터· AI 기반 정책 연구 및 여론 조사 시스템을 제대로 갖추어 '축적의 시간'을 확보할 필요가 있음.
  - 정책가, 정책, 여론 조사 전문가, AI 빅데이터 전문가의 협업 시너지 확인

연구 사례 2: 인공지능 거버넌스 연구(2020)
네이버 클린봇 평가 및 대안 모델 개발

유지웅, 황보유정, 손동성, 이경전, 악성 댓글 분류 시스템 모니터링 연구: 네이버 클린봇 분석(Malicious Comment Classification System Monitoring Research: Naver Cleanbot Analysis), 2020 한국지능정보시스템학회 춘계학술대회, 2020.
유지웅, 인공지능 거버넌스 시스템: 모니터링 및 구현 연구: 네이버 클린봇 분석, 경희대학교 빅데이터응용학과 석사학위논문, 2021.
Yoo, Park, & Lee, Governance of News Comment Filtering AI: Naver Cleanbot Case, Working Paper, 2022.



인공지능을 활용하여 악성 댓글 차단 사례
- 네이버 클린봇: 사용자가 뉴스 기사에 댓글을 작성하면 악성 유무를 판단하여 자동으로 노출 제한
- 뉴욕타임즈 모더레이터
  - 직소의 Perspective API를 활용하여 Moderator 시스템 개발
  - Moderator는 사용자에서 '유해', '스팸', '음란' 세가지 항목에 대해 점수를 할당
  - 뉴욕 타임즈는 Moderator의 점수를 기반으로 최종적으로는 사람이 댓글 노출 여부 판단

네이버 뉴스 댓글 수집 & 레이블링

악성 댓글 분류 모델인 클린봇을 평가하기 위해 네이버 뉴스에서 작성된 댓글을 수집
- 기간: 2019년 11월 12일 ~ 2019년 12월 14일
- 키워드: '문재인', '김건모'
- 수집된 뉴스 건수: 360건
- 수집된 댓글 건수: 91,716건

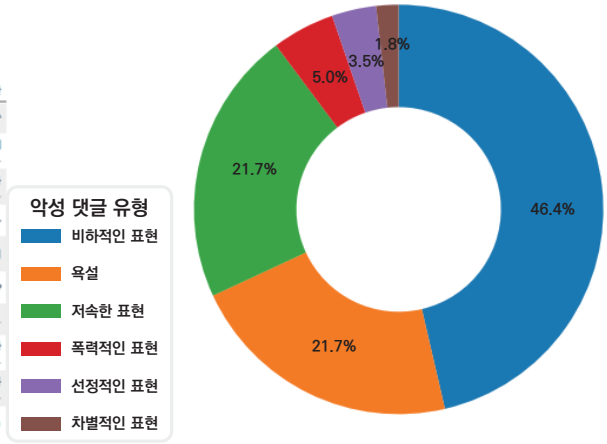
키워드	댓글 개수	차단된 댓글 개수	차단율
문재인	24,726	1,102	4.46%
김건모	66,990	1,809	2.70%
총합	91,716	2,911	3.17%

	date	keyword	link	nick	type	비고	댓글
3314	2019.12.11. 23:55	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	lijs0****	댓글	.	자체분열 기가막힌다^^
11059	2019.12.08. 03:24	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	dcli****	대댓글	.	국민들은 원자력을 발전시켜 안전한 원전이 국가 산업에 중요한 근간이라 생각합니다 국...
15633	2019.12.06. 10:47:45	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	yee2****	댓글	.	대한민국의 정의수호와 권력증의 부패,권력비리척결에 맡은바 임무를 다하고 있는 검찰청...
29905	2019.12.13. 15:21:32	김건모	https://news.naver.com/main/read.nhn?mode=LSD&...	ys92****	댓글	.	이러다 나중엔 다 사실이더라, 얼른 사과하고, 처벌 받아라~~~
88945	2019.12.09. 18:43	김건모	https://news.naver.com/main/read.nhn?mode=LSD&...	onow****	대댓글	.	접대부 있는 술집에 가은 모든 남성은 다 준 강간범이라는 논리
59185	2019.12.11. 10:32	김건모	https://news.naver.com/main/read.nhn?mode=LSD&...	yang****	댓글	.	관중 대 관중 근데 통방에 한달에 한번씩 간게 성폭행 증거는 아니잖아 ?
44412	2019.12.11. 10:08	김건모	https://news.naver.com/main/read.nhn?mode=LSD&...	wanj****	댓글	.	지금 악혼자 잘 생각하십시요. 좋을때는 금방 지나가니...
22686	2019.12.06. 09:58	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	algy****	댓글	클린봇	"거지 쓰레기같은 문재인 청와대는 거짓말을 밥먹듯이 거짓말로 포장하여 진실처럼 이야...
12018	2019.12.07. 03:49	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	thde****	댓글	클린봇	으휴 문재앙 봉ㅣ선놈 이럴줄 알았다. 달창. 전라디언들아 진짜 쉴드좀 쳐봐. 이정도...
39451	2019.12.13. 11:52	김건모	https://news.naver.com/main/read.nhn?mode=LSD&...	slad****	대댓글	클린봇	ㅈㄴ 중립기어가 가해자편이라는ㅈ개소리워내김흥국만봐도 ㅅㅂ

연구 사례 2: 인공지능 거버넌스 연구(2020)

수집된 데이터를 사람의 눈으로 확인하여 악성 댓글 유무를 판단
또한 네이버에서 정의한 6가지의 악성 댓글 유형을 기반으로 악성 댓글 유형 분류

1. 욕설: 일반적인 욕설, 모욕적인 표현 또는 남을 저주하는 표현
2. 저속한 표현: 타인에게 불쾌감을 주는 속되고, 격이 낮은 표현
3. 선정적 표현: 성적으로 자극적인 표현
4. 폭력적인 표현: 신체적 위협에 대한 표현
5. 차별적인 표현: 지역/인종/국가/종교/ 등에 기반한 차별 표현
6. 비하적인 표현: 상대방에게 모멸감과 수치심을 주는 비하 표현



모델 안정성 평가

네이버의 클린봇이 여러 상황에서 악성 댓글을 안정적으로 분류하는지 모델의 품질 평가

- 중요한 단어나 키워드를 직관적으로 시각화하는 도구인 WordCloud를 사용하여 사용자가 주로 사용하는 욕설을 확인
- "욕설"로 분류된 댓글 데이터의 WordCloud로 사용자는 주로 '미친', '지랄', 'ㅅㅂ', 'ㅂㅅ'의 욕을 사용하는 것으로 확인
- 모델 평가를 위해 '지랄'이라는 단어를 얼마나 안정적으로 분류하는지 확인
- 전체 댓글 중 '지랄'이 포함된 댓글은 총 527건, 그 중 클린봇에 의해 차단된 댓글은 67건으로 12.71%의 차단율을 확인

같은 댓글이지만 서로 다르게 판별한 경우

- 같은 작성자, 유사한 뉴스 기사, 비슷한 작성 시간 등 입력값이 유사한 상황에서 서로 다른 결과를 출력

	date	keyword	link	nick	type	비고	댓글
1141	2019.12.12. 09:35	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	acez****	댓글	클린봇	온지는 해도 지랄 안해도 지랄할 모양이구나?
1980	2019.12.10. 10:16	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	ghwl****	대댓글	클린봇	가도 지랄 안가도 지랄
4189	2019.12.11. 10:27	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	wsel****	댓글	클린봇	지랄났네 ㅋㅋㅋ 망하길 바라겠지 ㅋㅋㅋㅋㅋ
43570	2019.12.12. 22:23	김건모	https://news.naver.com/main/read.nhn?mode=LSD&...	yoog****	댓글	클린봇	저런 놈다리가 애 지랄하고있네
84686	2019.12.09. 13:00:22	김건모	https://news.naver.com/main/read.nhn?mode=LSD&...	hyun****	댓글	클린봇	이년 돈줄생기려다 말대로 안되니까 저지랄하는군ㅋㅋㅋ

"지랄"이 포함된 차단 댓글

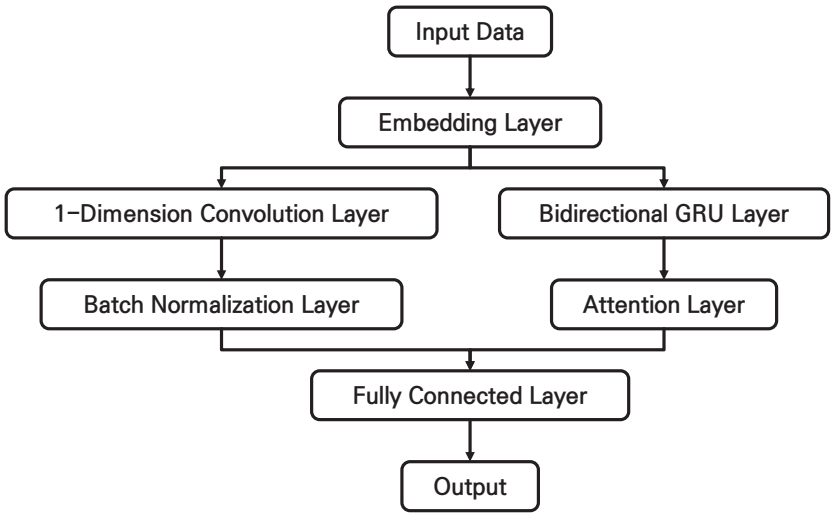
	date	keyword	link	nick	type	비고	댓글
20332	2019.11.28. 15:29	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	fran****	댓글		지랄도 풍년이다 진짜
4005	2019.12.11. 19:50	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	chlw****	대댓글		좌좀 새끼들 꼭 택트로 조지면 저 지랄들이지
9857	2019.12.07. 20:14:39	문재인	https://news.naver.com/main/read.nhn?mode=LSD&...	logi****	대댓글		문재인 대통령님 지랄합니다. 한반도 경제파탄에 애써주셔서 감사합니다
51935	2019.12.11. 06:34:27	김건모	https://news.naver.com/main/read.nhn?mode=LSD&...	seoy****	대댓글		지랄 업주잘못된거지 ㅆㅍㅇ
75139	2019.12.09. 15:12	김건모	https://news.naver.com/main/read.nhn?mode=LSD&...	ch25****	댓글		참 강용석 또 지랄한다~

"지랄"이 포함된 비차단 댓글

	클린봇 적용 전	클린봇 적용 후
연합뉴스	lmia**** ⓘ 2019.12.09. 15:03 쒸레기	lmia**** ⓘ 2019.12.09. 15:03 • 신고 쒸레기
서울신문	lmia**** ⓘ 2019.12.10. 18:50 • · 쒸레기	lmia**** ⓘ 2019.12.10. 18:50 ⓘ 클린봇이 부적절한 표현

대안 모델 구현: CNN, GRU, Attention을 조합

- **빠르고 불규칙적으로 변하는** 욕설 및 비속어 댓글은 추상화 과정을 통해 비 선형적인 특징을 찾는 CNN 기법이 적합
- **욕설은 없지만 문맥을 파악하여** 악성 여부를 판단해야 할 경우엔 GRU + Attention 기법이 적합
- Attention 메커니즘은 전체 문장을 동일한 비율로 참고하지 않고 **분류 문제에 연관 있는 단어에 더 집중하여 학습**
- 성능 비교를 위해 기존의 클린봇과 그 외 세가지의 모델을 설계하여 모델의 정확도 측정
- 4개의 모델의 정확도 및 악성 댓글 유형별 평균 차단율을 비교한 결과 CNN + GRU + Attention 모델이 가장 뛰어난 성능으로 확인



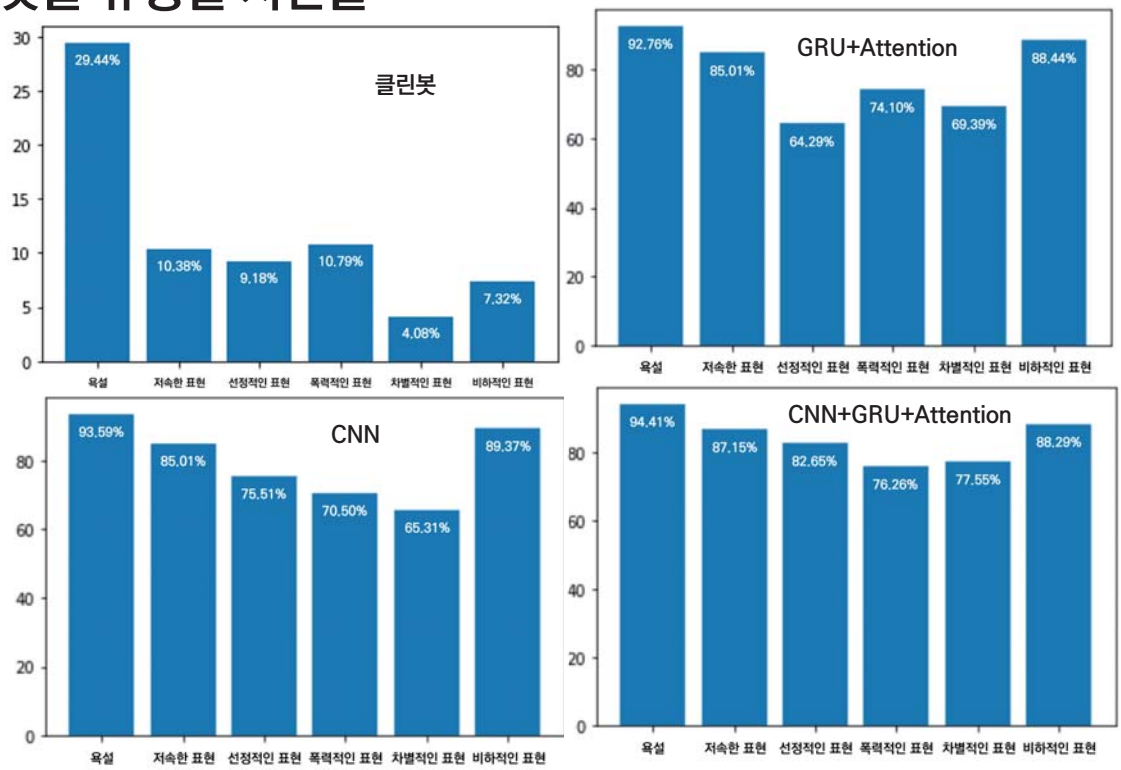
모델	정확도	악성 댓글 유형별 평균 차단율
클린봇	55.87%	11.87%
CNN	87.01%	79.88%
GRU + Attention	87.12%	79.00%
CNN + GRU + Attention	**88.21%**	**84.39%**

4개 모델의 악성 댓글 유형별 차단율

연구 사례 2: 인공지능 거버넌스 연구(2020)

악성 댓글을 6가지 항목에 따라 분류한 정보를 기반으로 각 모델의 유형별 차단율 확인

- CNN 모델은 '욕설', '비하적인 표현', '선정적인 표현'에서는 높은 성능
- 문맥을 파악해야 하는 '폭력적인 표현', '차별적인 표현'에서는 GRU+Attention 우수



네이버 클린봇 평가 및 대안 모델 연구 시사점 및 결론

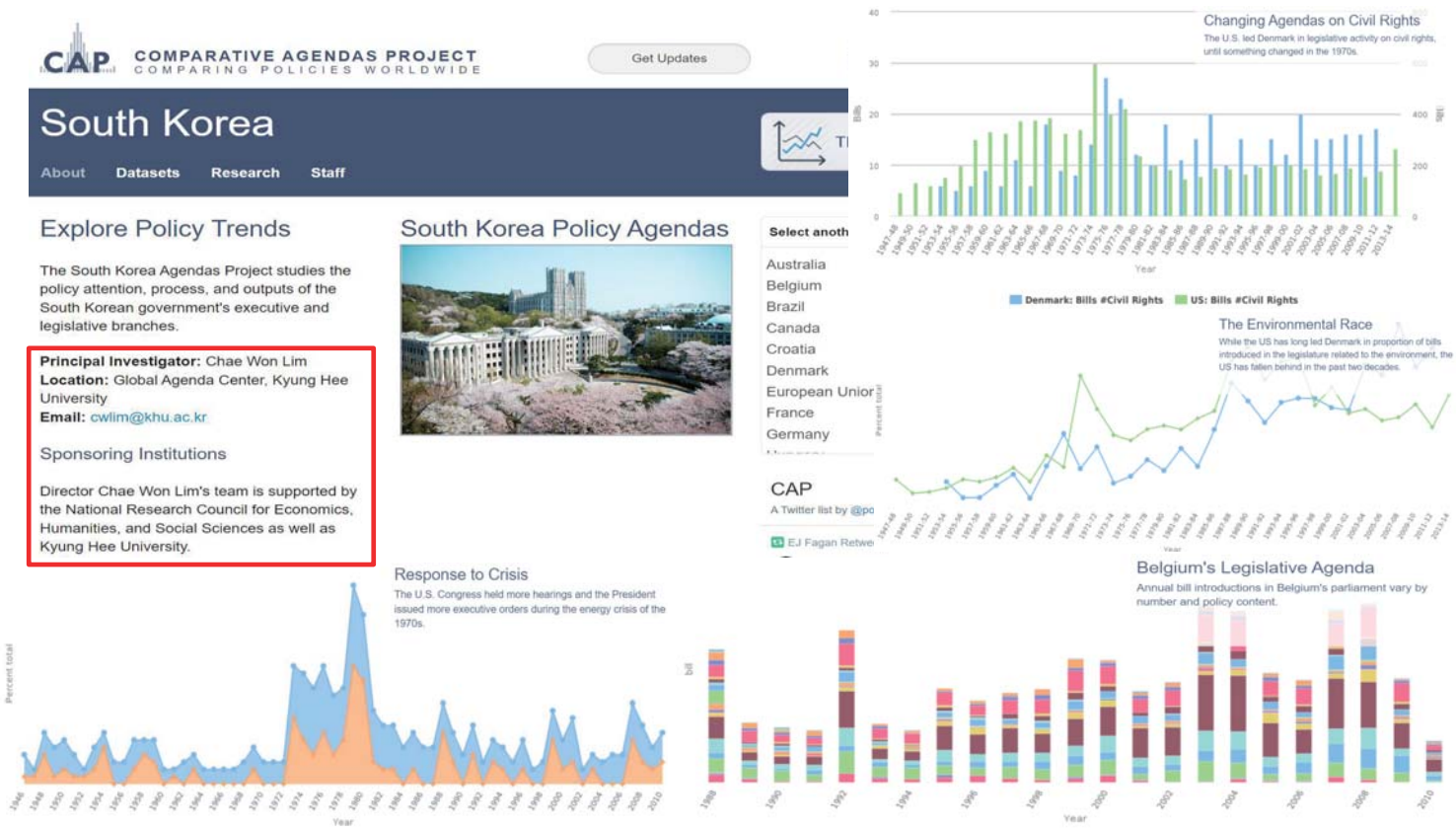
- 미디어(임원/실무진, 비밀창구) 알고리듬의 불공정한 운용에 대한 사회적 의심 검증
  - 네이버 '추미애' 검색 오류 사건
- 미디어 알고리듬의 오류로 발생하는 문제의 해결
- 미디어 사용자 관점의 오남용 및 불법적 사용의 감시
  - 예) 김경수/드루킹
  - 자동화된 댓글은 적법한가? K대 B교수 vs. 영업 방해, 선거법 위반
  - 자동화된 댓글에 대한 감시 필요
- 포털 및 미디어의 자정 노력은 한계가 있고, 자칫 여론 공론장의 폐쇄와 규제로 이어질 수 있음(예: 스포츠, 연예 댓글)
  - 빅데이터와 AI에 기반한 민간의 노력이 표현의 자유와 건전한 공론장이라는 두마리 토끼를 잡을 수 있음

불공정한 운용, 오류의 발견 및 시정, 정치권/사용자의 오남용 및 불법적 사용의 감시와 대응을 위한 AI기반 자동 모니터링 시스템과 이른바 AI NGO의 역할 필요

연구 사례 3: 정책 분류 자동화를 위한 AI 모형 개발에 관한 연구

- 비교 아젠다 프로젝트(Comparative Agendas Project, CAP) – 에딘버러대학 주도
  - 정책변동 및 정책 아젠다에 대한 글로벌 연구로서 비교 아젠다 프로젝트가 진행되어 왔음
  - 정책 공통기준으로 분류, 공공의제 변동 특성, 글로벌 공공정책 비교를 통해 협력체계 구축
  - 참여국은 거시 경제, 교육, 보건 등 21의제로 범주화(213 Subtopics): 질적, 계량분석 활용
  - 구축된 데이터 통해 정책변동 특성과 미래 글로벌 의제의 등장 예상하고 정책적 시사점 도출
  - 미, 영 등 국가는 1900년부터 데이터 수집: 정책 전문가들이 모두 수작업으로 정책을 분류
  - 한국: 1987-2019 미디어, 입법, 행정부 의제 분석과 구축을 정책 분석가의 수작업으로 진행

- 수작업 정책 분류는 정책 전문가에 따라 다르게 해석될 수 있으며, 상황에 따라 휴먼에러 발생할 수 있기에 정책 분류작업의 일관적 분류 및 생산성 높이는 AI 알고리듬 필요
- 수작업 구축 DB 기반으로 AI 알고리듬 활용하여 정책 분류 자동화 모델 세계 최초 개발
- 정책 분류 자동화를 위한 인공지능 모형을 개발하여 알고리듬을 구현한 Policy Perceptron을 제안: 정책 분류의 첫 알고리듬으로서의 의미
- *이경전, 황보유정, 정백, 유지웅, 배성원, 임채원, Policy Perceptron: 정책 분류 자동화를 위한 인공지능 모형 개발, 2022 한국행정학회 춘계학술대회, 여수.*



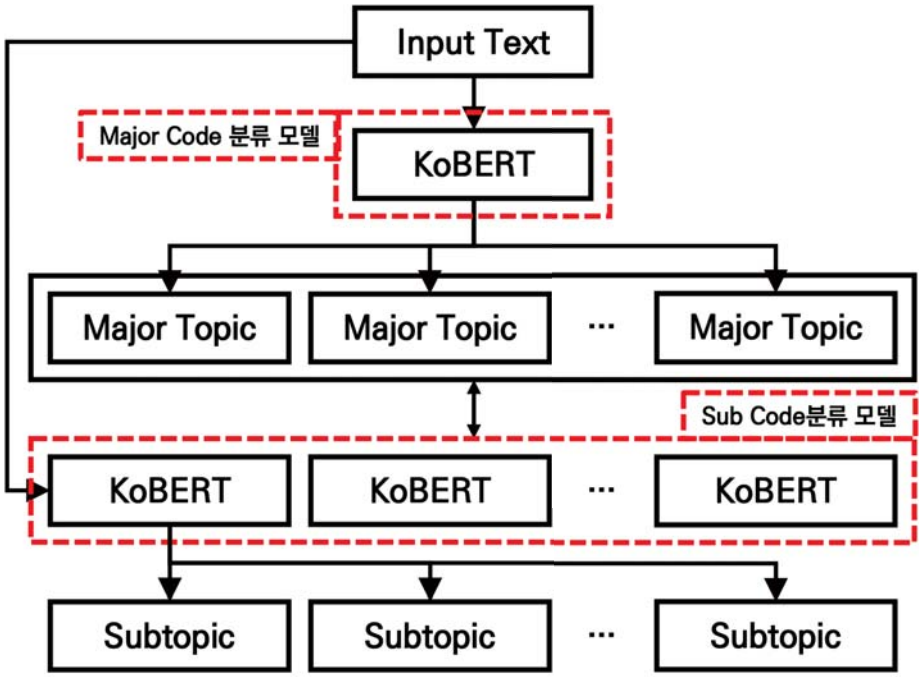
한국판 코딩북 예시:
한국판 정책범주 및 하위범주 코드

Major Topic		Subtopic
1	거시 경제	100. 국내 거시 경제 관련 정책
		101. 인플레이션, 가격, 이자율
		103. 실업률
		104. 통화정책, 한국은행
		105. 국가예산 및 국가부채
		107. 조세, 조세정책, 조세개혁
		108. 산업정책
		110. 물가관리 및 안정
		119. 그 외
2	인권, 소수자 문제, 시민의 자유, 주권	200. 인권, 소수자 문제, 시민권, 주권 관련 일반
		201. 사회적 소수자에 대한 차별
		202. 성차별, 성차에 따른 차별
		204. 연령 차별
		205. 장애인 및 특정 질병 보유자에 대한 차별
		206. 선거권
		207. 표현, 종교, 결사의 자유
		208. 사생활에 대한 권리 및 정부 정보에 대한 접근
		209. 반정부활동
		210. 종교단체에 대한 규제
		230. 이민 및 난민
		299. 그 외

Major Topic		Subtopic
3	보 건	300. 보건 일반
		301. 보건 의료 체계 개혁
		302. 국민 건강 보험
		321. 제약, 의료 기기, 의원에 대한 규제
		322. 의료 관련 시설의 건설, 규제 및 보조
		323. 의료인 및 보험 단체
		324. 의료 사고, 의료인의 일탈 행위 및 사기 행위
		325. 의료인의 육성 및 관리
		331. 질병의 예방 및 관리와 건강 검진
		332. 영유아
		333. 정신건강
		334. 장기 요양 치료, 자택 진료, 불치병 환자 케어, 재활 서비스
		335. 처방전 필요 의약품의 지정과 국가 부담의 범위
		336. 그 외 및 복수의 의료 혜택 관련
		341. 담배 흡연, 흡연자 치료, 흡연 예방 및 금연 교육
		342. 주류 과다 소비에 대한 정책, 음주 예방 및 금주 교육
		343. 마약 및 금지 약품 규제, 예방, 교육
		398. 의술 및 의약품 관련 연구 개발(R&D)
		399. 그 외
4	농 업	400. 농업 일반
		401. 농수산물 교역
		402. 농가 및 어가에 대한 보조금 지급 및 재해 피해 대책
		403. 농수산물 안전 및 검역
		404. 농수산물 판촉
		405. 가축 및 농작물의 질병 및 해충 관리
		406. 가축의 건강과 돌봄
		407. 농수산물 생산과 관련된 환경 문제
		408. 어업
		498. 농업 관련 연구 개발(R&D)
		499. 그 외
...	...	...

연구 사례 3: 정책 분류 자동화 AI

Policy Perceptron 구조도



한국 정책 아젠다 데이터베이스 구축

- 한국 비교아젠다연구팀은 2018년도 부터 비교 아젠다 프로젝트에 참여 23 Major Topic 및 213개의 Subtopic 지정하여 각 기준에 기반한 입법, 행정, 미디어에 대한 데이터베이스 구축
- **행정 데이터:** 1988-2018년 대통령 연설문을 문장 단위로 분해하여 구축. 행정 데이터는 대통령의 연설문을 문장 단위로 쪼개어 각 Major Topic 및 Subtopic의 코드로 레이블링 하였으며 만약 해당 문장이 둘 이상의 정책 내용을 포함하면 추가 분해 작업을 통해 세분화하였다. 또한 문장이 정책 내용을 포함하지 않을 경우 Major Topic 및 Subtopic을 9999(그 외)로 레이블링.
- **입법 데이터:** 한국의 주요 통과 법안에 대한 내용을 수집한 데이터베이스로 1987년부터 2018년 까지의 데이터가 수집. 각 법안의 발의 날짜와 발의한 상임위원회, 법안명 등의 데이터 수집.
- **미디어 데이터:** 미디어 데이터는 각 시기에 미디어의 제 1면에 기재된 뉴스 기사들을 수집한 데이터베이스. 조선일보 신문사의 기사를 수집하였으며 1988년부터 2020년까지의 기사 데이터베이스를 구축. 해당 기사의 발행 날짜, 기사 제목, 기사 요약 문장 수집.
- **최종 데이터베이스:** 데이터의 수가 10개 미만인 Subtopic은 제외 => Major Topic은 23개, Subtopic은 184개로 확정.

데이터 추가 수집, 전처리 & 모델 설계

연구 사례 3: 정책 분류 자동화 AI

- 입법 데이터: 국회 의안정보시스템 홈페이지에서 웹 크롤링 과정을 통해 해당 법률에 대한 세부 내용을 수집. 불필요한 텍스트 또는 특수문자 제거 및 한자->한글 변환.
- 미디어 데이터: 조선일보 신문 제 1면에 보도된 뉴스 기사를 기반으로 구축. 조선 뉴스 라이브러리 100 홈페이지에서 웹 크롤링 과정을 통해 뉴스의 본문 내용을 수집. 한자 -> 한글 변환.
- 데이터가 상대적으로 긴 입법과 미디어 데이터를 각기 다른 방식으로 데이터 압축
  - TF-IDF(Term Frequency-Inverse Document Frequency)를 데이터 전부에 적용 불용어 제거.
  - 카카오 브레인 Pororo(Platform of Neural Models for Natural Language Processing)를 입법/미디어 데이터에 적용, 각기 다른 두 개 데이터를 재생산하여 공통된 딥러닝 모델에 적용 성능 비교 통해 최종 기법 선택.
- Major Topic 및 Subtopic 코드는 각각 23개, 184개: 184개 중 하나 맞추는 문제
- 2중 분류 모델 설계: 먼저 Major Topic을 분류하고 다음으로 Subtopic을 분류
- 1차 딥러닝 모델은 전체 데이터에 Major Topic을 레이블링으로 학습하고 2차 딥러닝 모델은 전체 데이터를 Major Topic에 따라 분류하여 각각 학습
- Major Topic 분류 모델 1 & 각 Major Topic의 Subtopic 분류 모델 23: 총 24개 딥러닝 모델
- 정답 범주를 예측할 경우 하나의 범주를 예측하여 맞추는 경우(Top-1)와 3개의 범주를 예측하여 정답을 맞추는 경우(Top-3)를 비교하여 성능을 확인.

Policy Perceptron 모형 성능

- Major Topic 분류 모형
  - 문장 입력 시 Major Topic 23개의 범주를 분류하는 모형
  - Top-1일 경우 23개 중 1개를 예측하여 정확도를 확인한 결과 69.2%
  - Top-3에서는 23개 중 3개를 예측하였을때 87.6%의 정확도
- Subtopic 분류 모형
  - Subtopic 분류 모형의 성능을 나타낸 것으로, 각 Major Topic에 해당하는 모형의 성능과 평균을 Top-1과 Top-3의 정확도를 확인
  - 정확도는 각 23개 Subtopic 모델의 정확도를 평균낸 값으로 Top-1의 경우 74.2%, Top-3의 경우 91.4%의 정확도
- Policy Perceptron 최종 분류 모형
  - Major Topic 분류 모형과 Subtopic 분류 모형을 결합한 Policy Perceptron 모형의 최종 모형 성능
  - Top-3의 경우 Major Topic 분류 모형에서 도출되는 3개의 결과 확률과 이에 해당하는 Subtopic 분류 모형의 9개 도출 결과 확률을 곱하여 가장 높은 확률을 가지는 3개의 결과를 토대로 정확도를 출력
  - Top-1의 경우 정확도가 62.4%로 나타났으며, Top-3의 경우 71.6%의 정확도

연구 사례 3: 정책 분류 자동화 AI

정책적 활용 내용 & 기대효과

- 비교 아젠다 정책 분류를 인공지능 모형 Policy Perceptron이 정책 전문가를 보조하게 되는 경우, 정책 전문가의 생산성이 향상되어 다양한 영역에서의 정책 연구를 진행 가능
- 중앙 정부 정책 외에도 지방자치 정책 등 많은 수의 정책을 분류하고 비교·평가함으로써 국정운영의 방향을 모색하는데 기여.
- 정책 분류 자동화의 인공지능 모형 개발은 동일한 기준으로 정책을 분류 가능
- Policy Perceptron 모델은 향후 사람의 정책 분류 의사결정을 돕는 DSS로 활용 가능
  - AI와 빅데이터 분석, 인간의 시너지 창출:AI는 자동 분류하고 빅데이터 기법은 시사점 도출 및 시각화 – 인간은 해석 및 AI의 오류 검증
- 향후 연구: Policy Perceptron과 사람이 함께 최적의 선택을 할 수 있도록, 사람과 기계의 지속적 상호 작용을 다룬 휴먼인더루프(Human In The Loop) 시스템을 구현하여 Policy Perceptron의 잘못된 판단을 수정하여 점점 더 정확도를 높이는 시스템을 구현
- 본 연구에서 세계 최초로 개발된 정책 분류 자동화 인공지능 모형은 한국어를 사용하여 개발하였지만, 향후 글로벌 정책을 비교를 위하여 영어 등 다른 언어로 개발하게 된다면 글로벌 협력체계를 구축하는 비교 아젠다 프로젝트에 큰 기여를 하게 될 것으로 기대
- 정책 전문가와 AI 전문팀의 원활한 협조로 신속한 Pilot개발
  - 데이터 검증 및 모델 검증

연구 사례 4: 썸트렌드 비즈를 활용한 빅데이터 활용 교육 사례

발표: 박아름 교수 (용인예술과학대학교 빅데이터경영과)

빅데이터경영과의 커리큘럼 무엇이 문제인가

K대, D대, Y대, 한국데이터산업진흥원 빅데이터인재양성 프로그램 등 학부생들을 대상으로 코딩교육을 한 결과, 학생들은 어렵고, 재미없으나 해야만 하는 과목정도로 인식

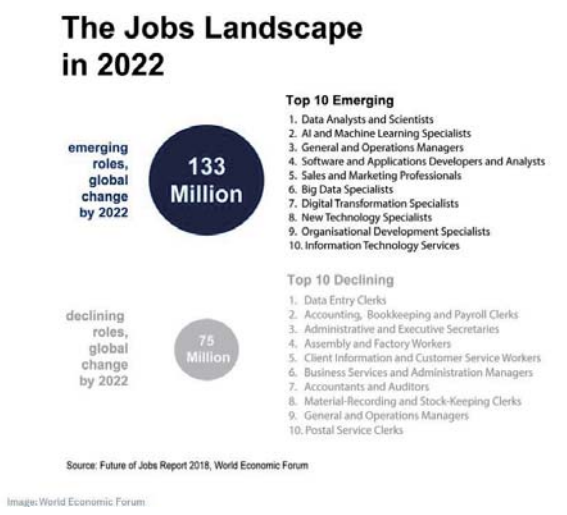
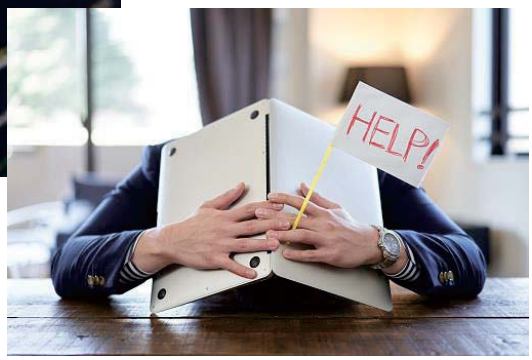


Image: World Economic Forum



경영직군을 위한 빅데이터분석 교육커리큘럼

〈K대 커리큘럼〉

Phase01 빅데이터 분석을 위한 기초학습	데이터베이스 입문 + Big DATA 이해 + DB 운영 관리	데이터처리운영 + DB성능관리 + SQL tuning + 고 가용성을 위한 서버 구축	운영체제 Internal + Unix / Linux + Storage 관리 기법 + OS shell programming
Phase02 빅데이터 분석도구 학습	R빅데이터 기본분석 + R Studio기본실습 + R과 선형대수 + 데이터 시각화 및 특성분석 + 기술통계 및 통계분석	파이썬 기본분석 + 파이썬 개발환경 구축 및 기본분석 + Numpy / pandas 학습 + 데이터수집 및 가공 + 시각화+시계열분석	통계분석 실습 + 공공데이터 활용한 실습 + 빅데이터 분석 결과 시각화 + pandas + matplot 활용 + SciPy
Phase03 머신러닝+딥러닝 인공신경망	통계기반 심화학습 + 지도학습(상관,분류분석) + 비지도학습 (군집분석, 연관분석) + 시계열 분석(예측도식화) + 기술통계분석	머신러닝 (Machine learning) + Clustering (고객 데이터 분석, 고객 분류) + Association Rule (장바구니 분석, 상품 추천) + Decision tree(결정 트리) + 회귀분석 - 연관성 분석	Deep Learning + 인공신경망 + 에이다부스트 알고리즘 + Tensorflow 구현 + Tensorflow에서 NN/DNN + CNN - RNN
Phase04 빅데이터 프로젝트	빅데이터 설계 분석 프로젝트 + Data 분석 + Data 분석 설계 + 데이터마이닝 기법 적용 + 시각화 표현	분석 도구 활용 + R + Python + 분석 알고리즘	프로젝트 시연 + 프로젝트 발표자료 및 기술서 작성 + 발표 및 시연

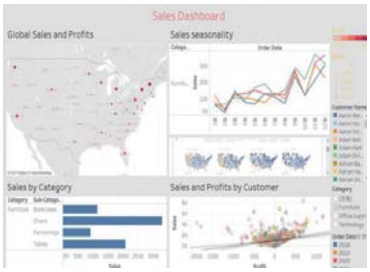
〈국내외 데이터과학 커리큘럼〉

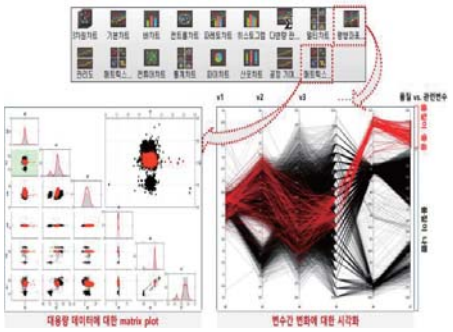
역량 구분	역량 강화를 위한 표준 커리큘럼	교육 기관별 표준 커리큘럼 커버리지 비율				
		해외 A 대학	해외 B 사설교육	국내 A 대학	국내 B 사설교육	국내 C 사설교육
기반 역량 (Foundation)	산업 별 빅데이터 활용 사례, 빅데이터와 Creative Thinking, 빅데이터 보안 분석, 데이터 과학자의 역할 등	40 %	0 %	25 %	20 %	20 %
기술 역량 (Platform Technique)	하둡 Core 및 Eco System의 이해, HDFS와 MapReduce의 활용, NoSQL(Monggo DB, Cassandra 등)	100 %	16.7 %	50 %	83.5 %	16.7 %
분석 역량 (Analysis Technique)	분석 모형의 이해, R분석 및 Visualization, 상용 Tool 활용법, 데이터 마이닝 프로세스, 텍스트 마이닝, Social Network Analysis 등	100 %	100 %	12 %	20 %	100 %
사업 역량 (Business Analytics)	산업 별(제조, 유통, 통신, 금융, 공공, 소매 등) 핵심 업무의 이해, 산업/업무별(Risk, Social, CRM 등) Analytics 방법 및 적용 등	62.5 %	37.5 %	13 %	12.5 %	12.5 %

→ 경영직군의 데이터 분석가를 양성하는데 있어, 사업역량과 분석역량의 비중이 더 중요하며, 분석역량을 위해 R, Python 중심의 교육은 학생들이 흥미를 잃을 수 있고 포기할 수 있음

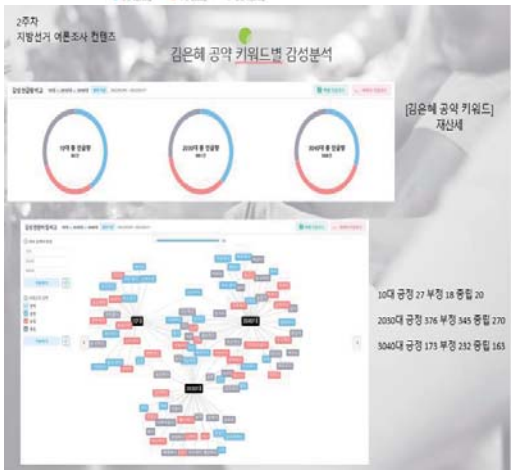
경영직군을 위한 빅데이터분석 교육커리큘럼

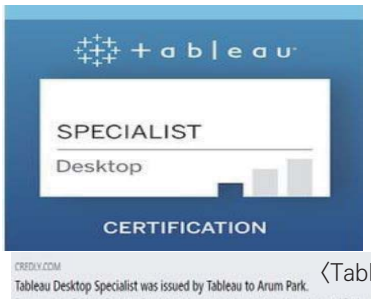
〈Sometrend〉



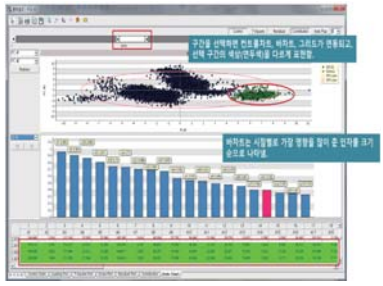


〈ECMiner〉





〈Tableau와 자격증〉



→저학년 학부생 대상으로 데이터분석과 시각화를 쉽게 할 수 있는 Sometrend ,ECminer, Tableau, 엑셀 등을 활용한 교육진행 결과 학생들의 참여도가 높아짐

→ 특히, Z세대에게 롤플레이 기반의 교육이 효과적이며 이러한 툴 활용시 롤플레이 강의가 원활하게 진행될 수 있음

썸트렌드 비즈를 활용한 빅데이터 활용 교육: 지방선거 결과 예측

- 과목: 디지털마케팅 전략, 마케팅관리
- 프로젝트 주제: '빅데이터 분석에 기반한 선거결과 예측 및 정책별 감성분석을 통한 선거전략 제안' 을 각 팀별 인스타그램계정에 공유
- 기간: 4.25~6.15
- 주차별 팀프로젝트
  - 1주차: 후보자별 추이분석과 연관어 분석
  - 2주차~4주차: 각 후보자의 공약별 감성분석과 타겟분석 이슈 도출(5.19일 발표)
  - 5주차: 최종 결과 예측
  - 6주차: 예측과 결과 비교 및 토론

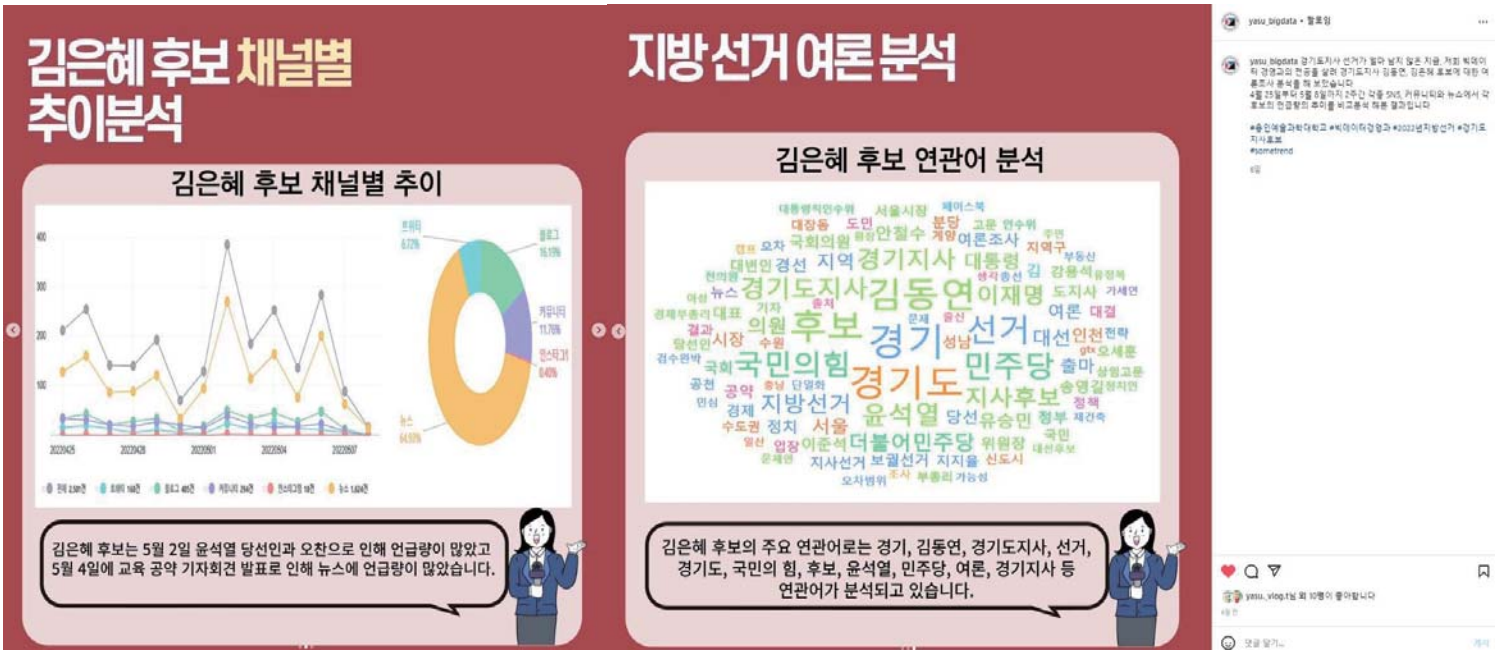
썸트렌드 비즈를 활용한 빅데이터 활용 교육_1주차 과제사례

1주차: 후보자별 추이분석과 연관어 분석

- 세종시장: 이춘희 vs 최민호
- 서울시장: 오세훈 vs 송영길
- 부산시장: 박형준 vs 변성완
- 경기도지사: 김동연 vs 김은혜
- 용인시장: 백군기 vs 이상일
- 강원도지사: 이광재 vs 김진태
- 인천시장: 박남춘 vs 유정복

썸트렌드 비즈를 활용한 빅데이터 활용 교육_1주차 이슈도출 과제사례

후보자별 추이분석과 연관어 분석 경기도 지사 예시: 단순 추이분석과 연관어 분석

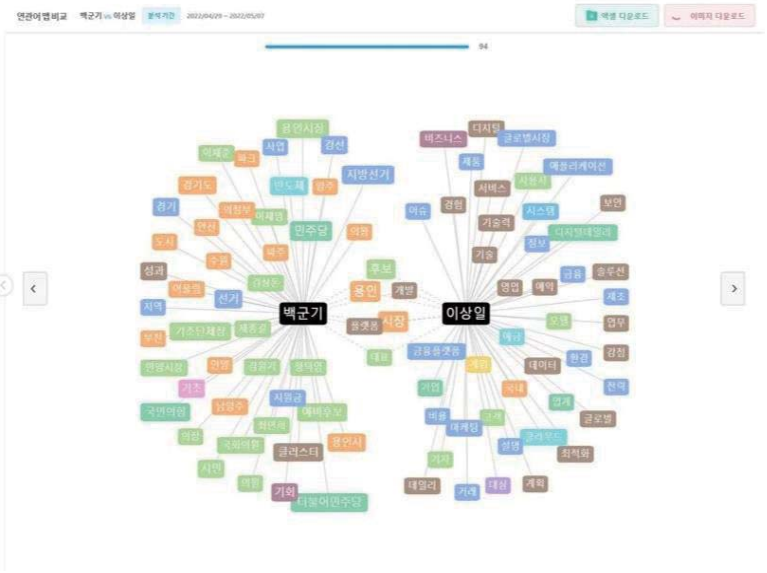


썸트렌드 비즈를 활용한 빅데이터 활용 교육_1주차 이슈도출 과제사례

후보자별 추이분석과 연관어 분석 인천 시장 예시: 공약과 추이분석을 통해 주요 정책 이슈 도출

썸트렌드 비즈를 활용한 빅데이터 활용 교육_1주차 이슈도출 과제사례

용인 시장 언급량 연관어 맵 비교분석 예시:언급량과 지지율 비교

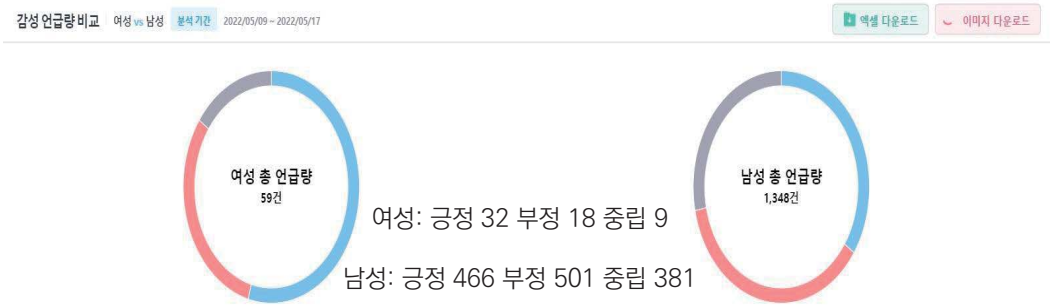


썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

각 후보자의 공약별 감성분석과 타겟분석을 통한 여론분석 경기도 지사 예시

썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

각 후보자의 공약별 감성분석과 타겟분석을 통한 여론분석 경기도 지사 예시



여성: 긍정 32 부정 18 중립 9

남성: 긍정 466 부정 501 중립 381

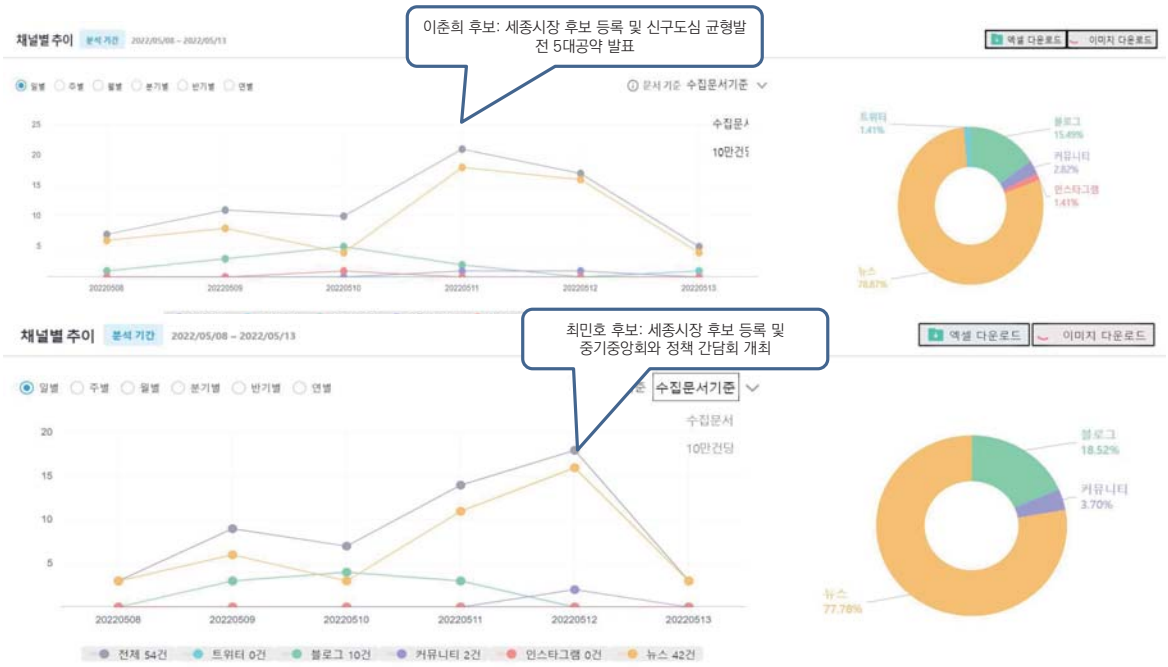
[김은혜 공약 육아 키워드 감성분석 결과]

[육아 공약에 대한 긍정적 의견]

- 김은혜 국민의힘 경기도지사 후보가 '여성의 건강한 성장과 각종 폭력으로부터 안전한 경기도'를 주요 내용으로 하는 여성 공약을 발표하고 본격적인 정책 선거에 돌입
- 여성안심귀갓길 1000개소 구축'을 통해 주거침입이나 성범죄 등 범죄가 빈번하고 1인 가구가 밀집

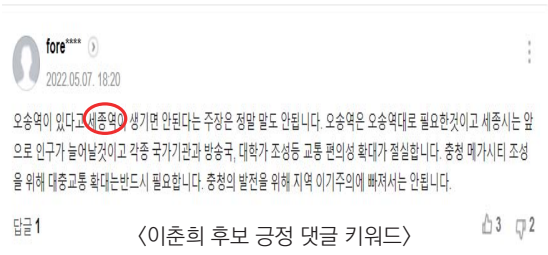
썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

각 후보자의 공약별 감성분석과 이슈 도출 세종시장 예시

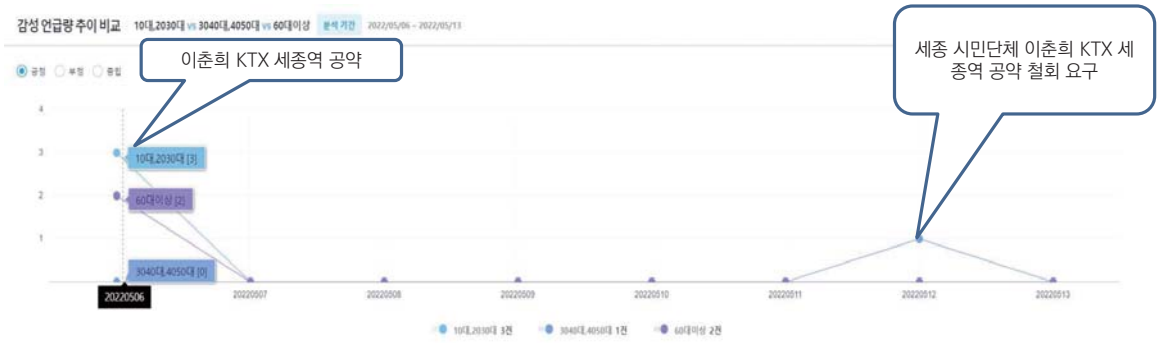


썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

각 후보자의 공약별 타겟감성분석과 이슈 도출 세종시장 예시

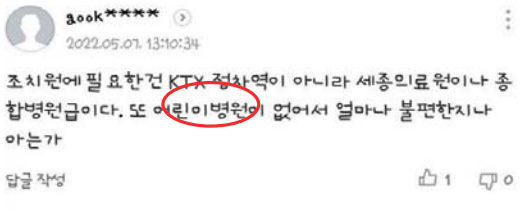


〈이춘희 후보 긍정 댓글 키워드〉

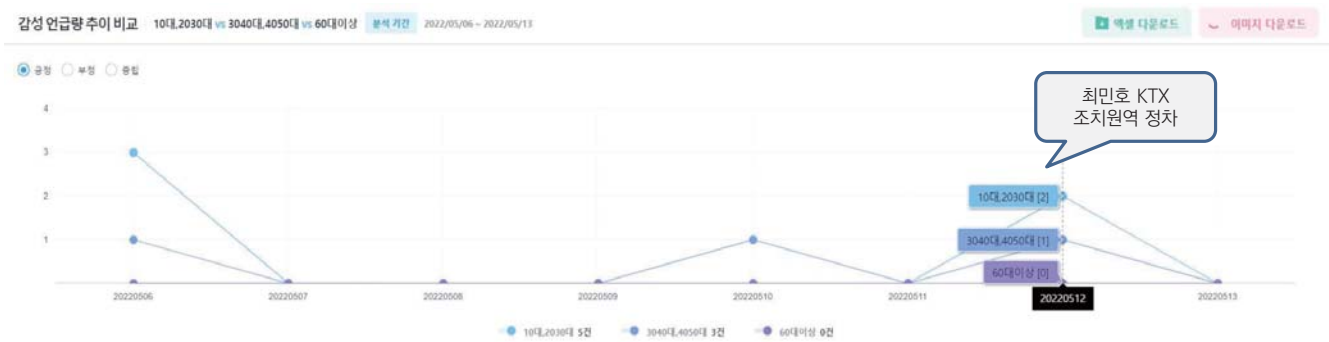


썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

각 후보자의 공약별 감성분석과 이슈 도출 세종시장 예시



〈최민호 후보 부정 댓글 키워드〉

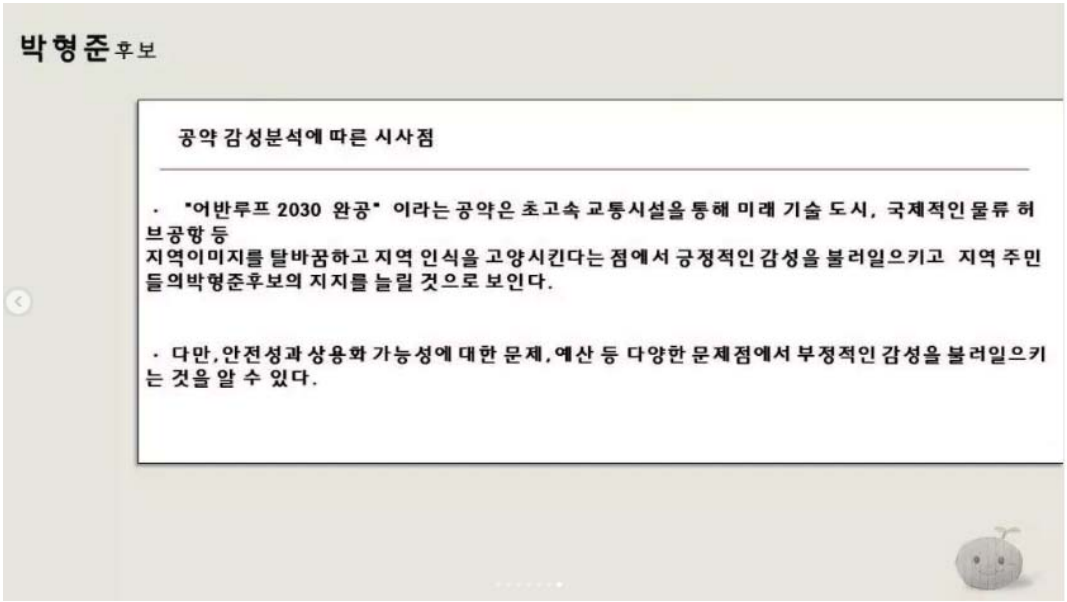


썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

후보자의 공약별 감성분석과 이슈 도출 - 부산시장 예
연령대별 공약에 대한 상반된 의견(박형준 후보)

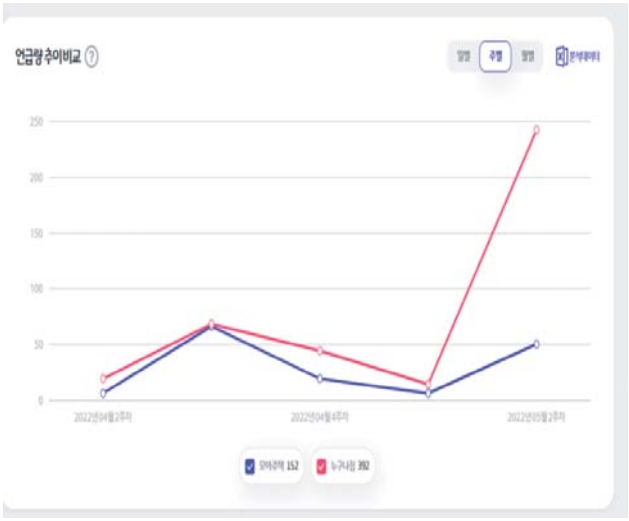
썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

각 후보자의 공약별 감성분석과 이슈 도출 부산시장 예시: 부정적 감성에 대한 이슈 도출

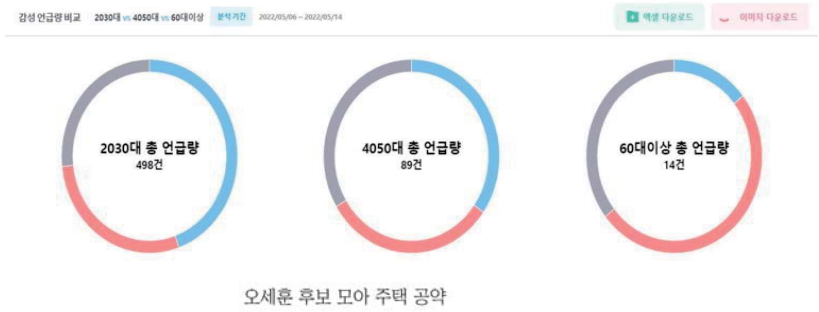


썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

각 후보자의 공약별 감성분석과 이슈 도출 서울시장 예시



오세훈 후보, 송영길 후보 공약별 긍.부정 키워드 추이

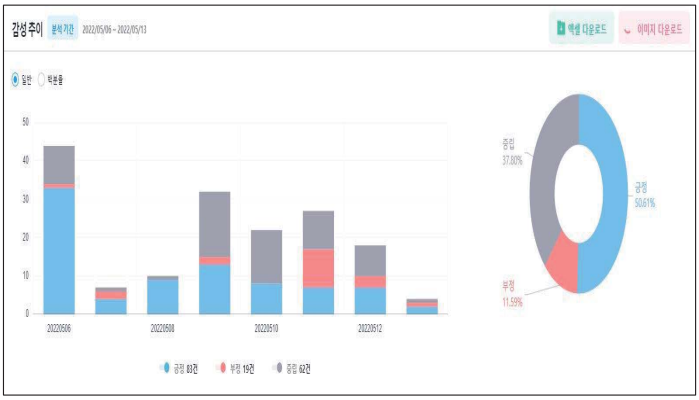


오세훈 후보 모아 주택 공약

송영길 후보 누구나 집 공약

썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

각 후보자의 공약별 감성분석과 이슈 도출 용인시장 예시

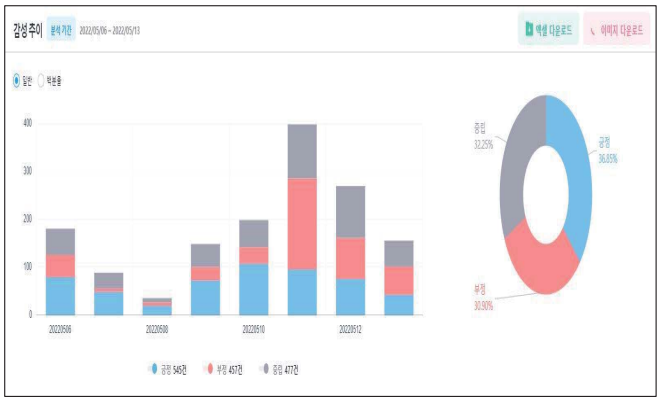


백군기 후보 주요 공약 : 개발이익 시민환원

- SK하이닉스 유치로 매년 늘어나는 세수 1조 5000억원 중 10%를 시민기금으로 적립하고, 운영수익 전액을 시민 제안사업에 투자 하는 것
- 이 공약에서 유의해야 할 점은 원금상환 이외에 민간사업자들이 추가로 얻는 수익을 어떻게 지역사회로 환원해야 할 것인지, 세수가 감소하는 경우에는 어떻게 적정선을 정해야 할 것 인지에 대한 이슈가 있음

썸트렌드 비즈를 활용한 빅데이터 활용 교육_2주차 공약별 감성분석 과제

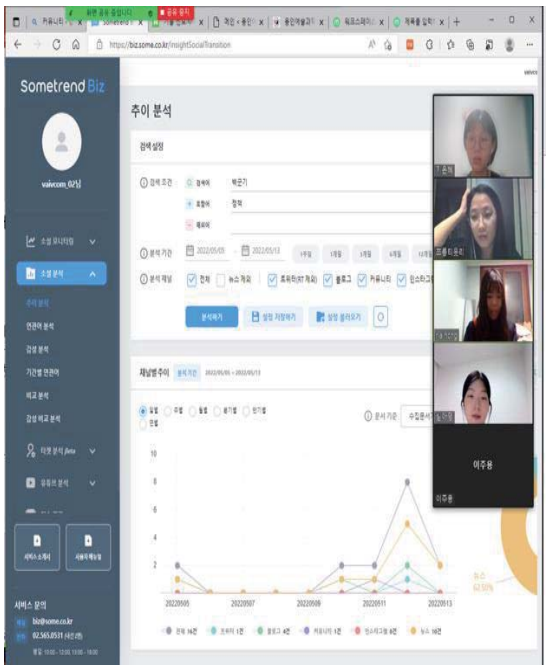
각 후보자의 공약별 감성분석과 이슈 도출 용인시장 예시



이상일 후보 주요 공약 : 재산세 감면

- 과세표준 3억원(공시가격 5억원)이하 1가구 1주택의 재산세 100% 감면해준다는 것
- 조세법률주의에 의거한 국회 입법사항으로 도지사나 특례시장권한 밖의 사항이며 포퓰리즘 공약이라는 것. 이를 활용하기 위해선 국회와 지방정부, 지방의회 차원에서 구체적이고 종합적인 대책마련을 하여 공평과세의 원칙을 철저히 지켜야 해당 공약을 활용할 수 있을 것이라는 이슈가 있음

썸트렌드 비즈를 활용한 빅데이터 활용 교육: 팀 프로젝트 사진 및 소감

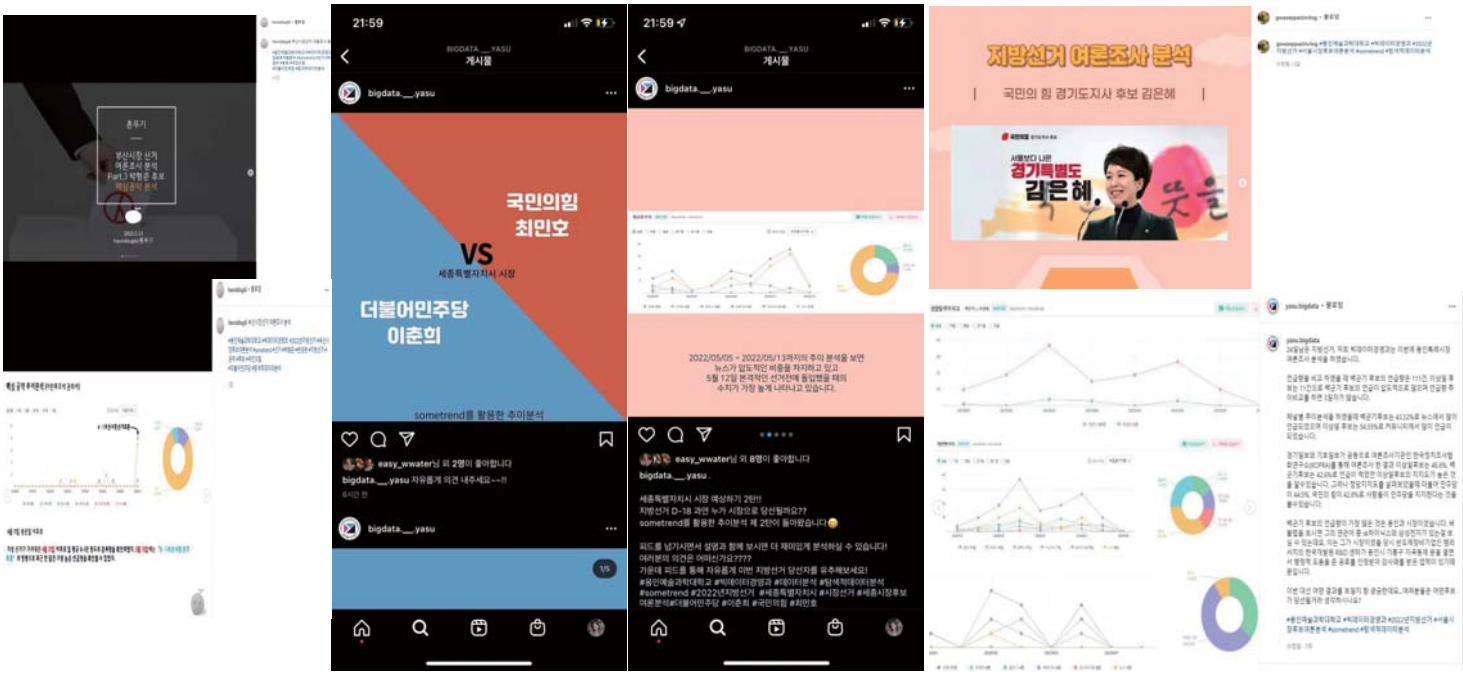


〈썸트렌드를 활용함 팀플 줌회의〉

〈이지수 학생〉

〈서주영 학생〉

썸트렌드 비즈를 활용한 빅데이터 활용 교육: 감성분석 인스타그램 공유



- 팀별 인스타그램에 아래 태그와 함께 빅데이터분석 결과 공유
  #용인예술과학대학교 #빅데이터경영과 #2022년지방선거 #sometrend #탐색적데이터분석

썸트렌드 비즈를 활용한 빅데이터 활용 교육: 썸트렌드의 기능별 의견

1. 추이분석/연관어분석:
   - 인터넷에 단어가 언급되는 게시글 원본을 확인하기 쉽고 시각화가 잘 되어 있어 그래프를 이해하기 쉽다
   - 트위터, 블로그, 뉴스 등 주요 채널별 원문 찾기에도 쉽다.
   - 알아보고 싶은 키워드의 일별 언급량 및 채널의 비율을 한눈에 볼 수 있는게 좋았고 전체적으로 정확한 량이 채널별로 잘 나뉘어져 있다.
   - 관심있는 단어를 검색하면 각 미디어별로 언급량을 한 눈에 확인 할 수 있고, 어떤 미디어에서 가장 많이 언급이 되었는지도 알 수 있다.
   - 일별, 주별, 월별, 분기별, 반기별, 연별로 간편하게 볼 수 있다.
   - 원하는 기간 내에 있던 정보를 알아서 분석 할 수 있다.
   - 간단하게 많은 양의 데이터를 확인하고 활용할 수 있어 사용하는데 용이하며 분석 단어가 얼마나 많이 언급되고 있는지 파악할 수 있고 특정 단어의 화제성 또는 이슈성을 확인하고 싶을 때 사용할 수 있다.

썸트렌드 비즈를 활용한 빅데이터 활용 교육: 썸트렌드의 기능별 의견

2.감성분석:

- 속성별 기간별 긍정 부정 단어 변화 추이를 분석하여 분석 단어에 대한 인식과 평판을 파악할 수 있다.
- 다양한 색깔들로 표시가 되어 보기 편했고 여러가지 연관어들도 편리하게 볼 수 있다.
- 키워드에 대한 사람들의 긍정,부정, 중립적인 면이 나오니까, 분석하기 편하다.

3.타겟분석:

- 각 연령별, 키워드별 단어를 빠르게 선택하여 분석하는 시간이 단축되었고, 일일히 검색하지 않아도 언급량과 연령별 등 필요한 자료들을 쉽게 얻을 수 있어서 좋았다.
- 관심있는 단어와 그 단어를 언급한 타겟을 직접 정해서 그에 대해 긍정, 부정, 중립인지의 의견을 확인할 수 있고, 그 타겟의 관심사를 유추할 수 있다. 날짜를 지정했을 때 어느 날에 가장 높았는지도 알 수 있다.
- 타겟을 정하기 쉽게 되어있고 분석을 하였을 때 자세히 분석이 되어 조사할 때 좋았다.
- 키워드에 대한 연령대별로 타켓을 분석해서 보기에 편하고, 어떤 연령대가 더 검색을 많이 하는지 빠르게 보기 편하다.

썸트렌드 비즈를 활용한 빅데이터 활용 교육: 썸트렌드에 대한 제안

1.추이분석:

- 해당 인물을 검색하였을 때 유명한 연예인들이 함께 언급되어 수정하는데 불편함이 있었다.
- 검색 하고싶은 사람이 동일인물 일 때 조건을 추가 하지 않으면 내가 원하는 사람을 검색하기 어려웠다.
- 여러 채널의 데이터를 제공하나, 채널별 주 사용연령층 등 특징을 같이 제공한다면 더 좋을 것 같다.

2.타겟분석:

-제시된 키워드의 양이 적어서 검색하는데 약간의 불편함이 있다. 다른 분석들처럼 필수 키워드 입력란이 있으면 좀 더 정확하게 타겟분석이 가능할 것 같다.

-타겟과 날짜를 지정해서 검색했을 때, 가장 높거나 낮게 언급이 된 이유를 확인할 수 없다.

-타겟을 검색 했을 때 아무것도 뜨지 않는 경우가 있다.

-타겟을 특정으로 정하고 하기 때문에 특정 타겟만 분석할 수 있다.

-설정한 타겟에 대한 분석 결과가 아주 적거나 아예 없을 수도 있어 분석하기 전 조금 더 꼼꼼하게 조사 할 필요가 있다.

-여성, 남성이 아닌 혼성으로 분석 되어 타겟을 확실하게 알고싶은데 그렇지못해 아쉽다.(타겟이 불분명해진다)

- 온라인 매체를 거의 사용하지 않는 고령 사용자들의 타겟분석은 정보와 신뢰성이 떨어지는 것 같다.

썸트렌드 비즈를 활용한 빅데이터 활용 교육 효과

6.1 지방선거 결과 예측을 위한 썸트렌드 활용결과

– 썸트렌드를 활용하여 탐색적 데이터 분석을 비교적 쉽게 이행

1) 추이분석과 연관어 분석 기능

추이분석을 통한 이벤트 시점과 이슈 발견–〉 연관어 분석을 통해 키워드 도출→공약과 키워드 비교/주차별 여론 추이 확인

2) 감성분석 기능: 후보의 정책별 대략적 여론 파악가능

: R이나 python을 활용한 감성분석은 코딩을 하는데 익숙치 않은 저학년 학부생들이 흥미를 잃을 수 있음. 썸트렌드를 활용함으로써 감성분석 결과를 빠르게 도출하고 원문으로의 쉬운 접근으로 감성분석 결과에 대한 검증이 가능

3) 타겟분석 기능: 정책에 대한 조건별 선호도 파악 가능

: 연령대별/성별/관심사별 정책에 대한 감성분석 결과로 향후 선거방향에 제언 가능

결론 및 향후 필요 연구: 썸트렌드 비즈를 활용한 빅데이터 활용 교육효과

- 빅데이터활용 교육을 위한 썸트렌드 활용 의의
  - 데이터분석가로서 분석역량은 상당히 중요. 그러나 R이나Python을 기반으로 한 코딩부터 접근할 경우, 일부 학생을 제외한 대부분은 흥미를 잃고 포기
  - 따라서, 흥미 유지와 문제 및 이슈를 탐색하기 위한 단계로서 분석결과를 빠르게 도출할 수 있는 툴을 활용하는 것이 탐색적 사고를 향상시키는데 효율적
  - 탐색적 데이터 분석 과정을 빠르게 경험하고 분석결과를 기반으로 한 시사점을 도출하는 과정을 통해 비판적 사고와 논리적 사고를 기를 수 있으며 학생들의 참여도를 높일 수 있음
  - 향후, 롤플레이 방식 수업을 진행시 유용한 툴임
- '빅데이터 분석 도구를 활용 교육이 학습역량을 높이는데 미치는 영향'에 대한 후속 연구 진행
  - ✓ Sometrend 활용 학생들을 대상으로 서베이 진행 예정
  - ✓ 가설
    - 빅데이터 분석 도구를 통한 교육이 직업에 대한 태도에 긍정적인 영향을 미친다.
    - 빅데이터 분석 도구를 통한 교육이 학습태도 및 학습효과에 긍정적인 영향을 미친다.
    - 빅데이터 분석 도구를 통한 교육이 자기효능감에 긍정적인 영향을 미친다.
    - 빅데이터 분석 도구를 통한 교육이 분석 분야에 대한 이해와 태도에 긍정적인 영향을 미친다: 빅데이터 분석 교육 뿐만 아니라 분석 대상 분야 교육(예: 정치)에서 빅데이터 분석 도구 활용 교육의 필요성 인식

빅데이터 기술 교류 세미나

빅데이터와 여론조사

한국경영정보학회
The Korea Society of Management Information Systems
서울시 용산구 한강대로 115, 702호 I office@kmis.or.kr