뉴스 플랫폼 혐오표현 분석 비즈니스 대시보드

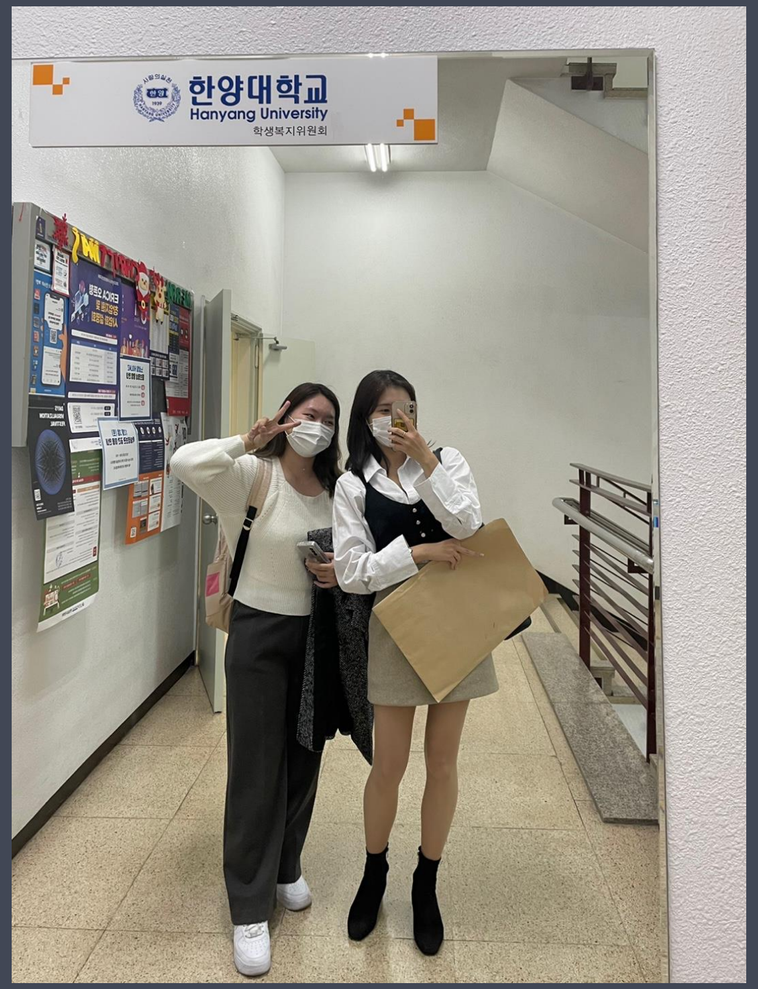
Team 은자감(Euns' Confidence)

목차

팀소개

Team 은자감(Euns' Confidence)

- 데이즈 회장, 부회장으로 구성된 팀
- 공통적으로 이름에 '은'자가 들어감



주제선정배경 및 문제상황

데이터 + 미디어

주제선정배경 및 문제상황

"STOP! Hate Speech" campaign concludes in Bosnia and ...

"Hate speech, discriminatory and inciting speech does not represent freedom of speech but is abuse and is punishable under the law. In January...

[사설]이태원 참사 트라우마 키우는 '혐오 표현' 삼가야

표현의 자유를 넘어선 혐오 표현이요 허위 사실 유포로 법적 윤리적 ... 이태원 참사의 원인을 마약이나 가스 누출로 돌리는 거짓 정보도 문제다.

"혐오와 차별을 반대한다"…사회 수업 지지하는 제주 교사들

이 교사는 학생들이 혐오표현에 대해 어떻게 인식하고 있는지 조사하고, ... 교사의 수업 내용에 문제 등에 민원을 제기해왔다.

EU Code of Conduct against online hate speech: latest ...

Today, the European Commission released the results of its seventh evaluation of the Code of Conduct on countering illegal hate speech...

[脫혐오사회]'혐오표현은 2차 가해'…네이버·카카오, 가이드라인 만든다

업계가 공동 대응에 나선 것은 온라인상 차별·혐오 표현 문제가 늘어나고 있어서다. 이번 이태원 참사 때도 마찬가지다. 희생자에게 책임을 돌리거나...

혐오표현이란?

어떤 개인 혹은 집단에 대해 그들이 사회적 소수자의 속성을 가졌다는 이유로 차별·혐오하거나 차별·적의·포격을 선동하는 표현

주제선정배경

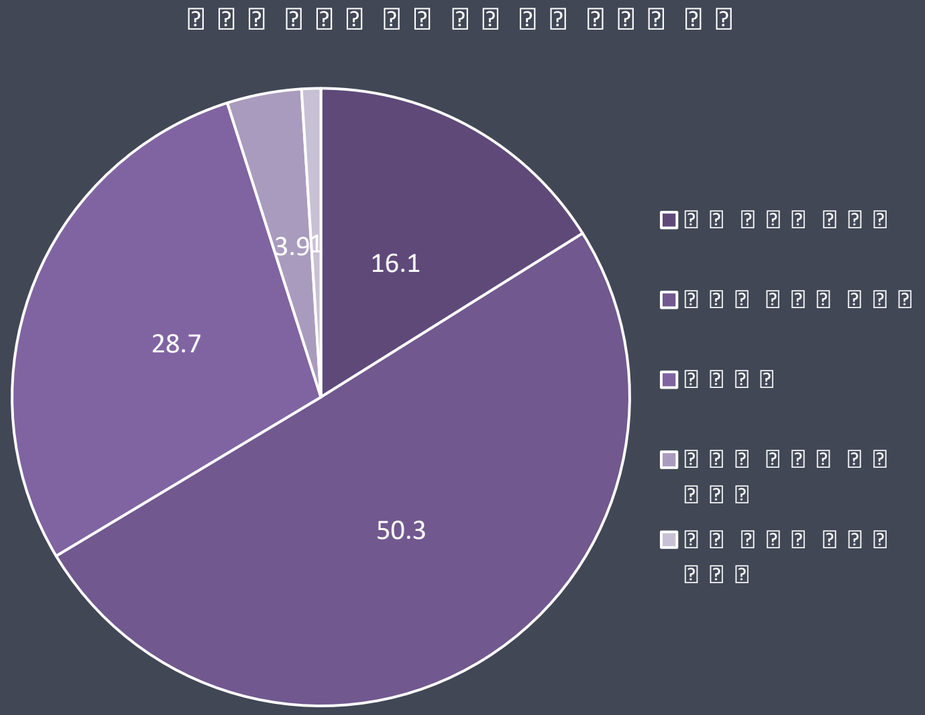
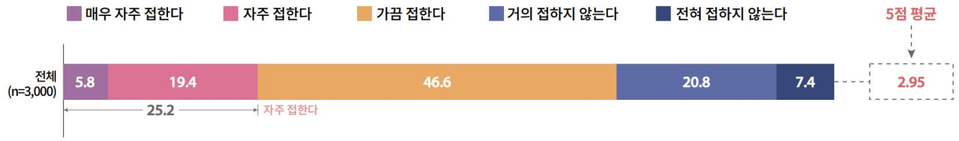
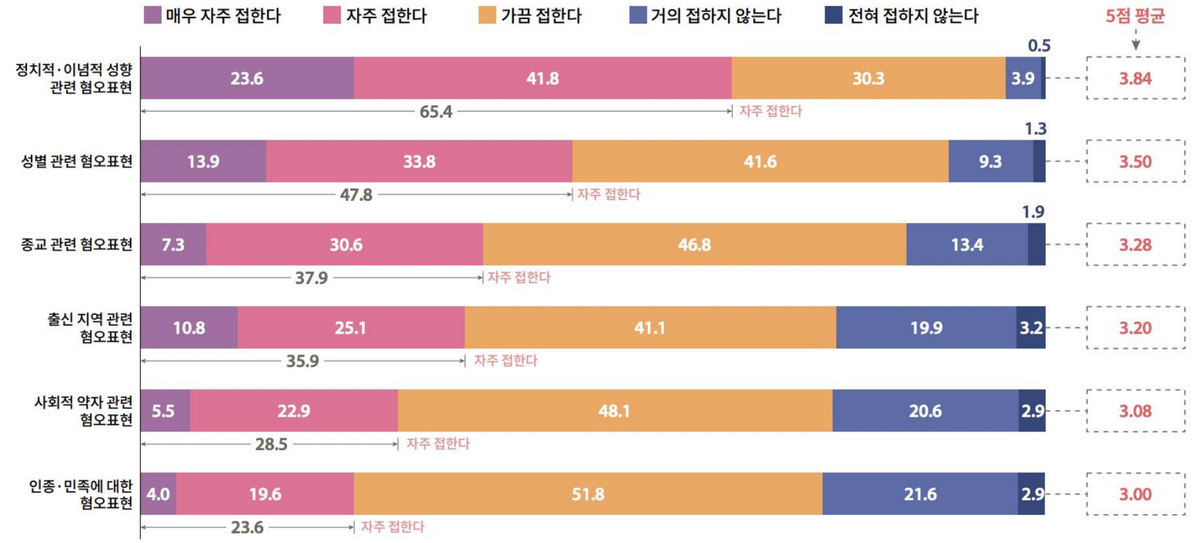


그림 1-4 소셜미디어 혐오 표현 접촉 빈도 (단위: %)



* 표본 – 전체(n=3,000)
N6 귀하께서는 소셜미디어에서 혐오표현을 얼마나 자주 접하십니까?

그림 4-7 유형별 혐오 표현 접촉 빈도 (단위: %)



* 표본 – 소셜미디어에서 혐오 표현 '가끔 접한다' 이상 응답자(n=2,154)
N7 귀하께서는 소셜미디어에서 혐오 표현을 접한다고 응답하셨습니다. 다음 각 유형의 혐오 표현을 얼마나 자주 접하시는지 선택해 주십시오.

주제선정배경 및 문제상황

전세계적으로 혐오 표현을 줄이기 위한 움직임

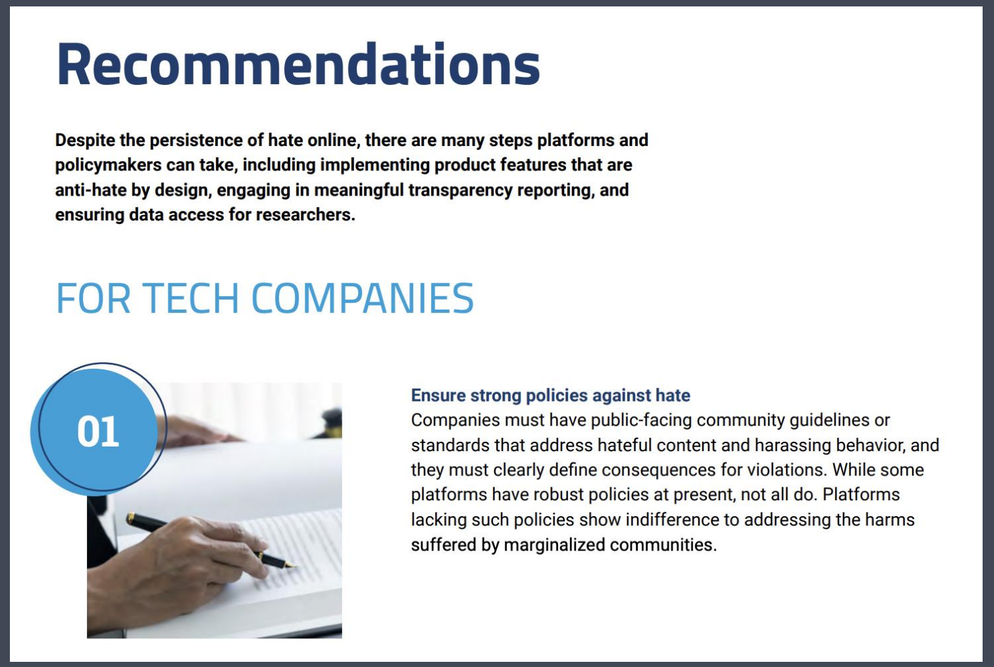
“자율규제 + 인권경영”

페이스북 : 투명성 보고서(Transparency Report)를 통해
혐오표현 금지정책 위반 현황을 발표

유튜브 : 커뮤니티 가이드라인(Community Guidelines)에서
‘증오성 콘텐츠’(Hate Speech)를 자사 정책을 위반하는
게시물의 유형 중 하나로 규정

카카오 : 2021년 1월 국내 기업 최초로 ‘증오발언 근절을 위한 카카오의 원칙’을 제시

ADL(세계 최고 증오 방지 조직) : 테크 기업들에게
혐오 표현 관련 권고 사항을 담은 문헌 제작



Recommendations

Despite the persistence of hate online, there are many steps platforms and policymakers can take, including implementing product features that are anti-hate by design, engaging in meaningful transparency reporting, and ensuring data access for researchers.

FOR TECH COMPANIES

01

Ensure strong policies against hate
Companies must have public-facing community guidelines or standards that address hateful content and harassing behavior, and they must clearly define consequences for violations. While some platforms have robust policies at present, not all do. Platforms lacking such policies show indifference to addressing the harms suffered by marginalized communities.

비즈니스 시나리오

선정 기업 : KISO(한국인터넷자율정책기구)

설립목적

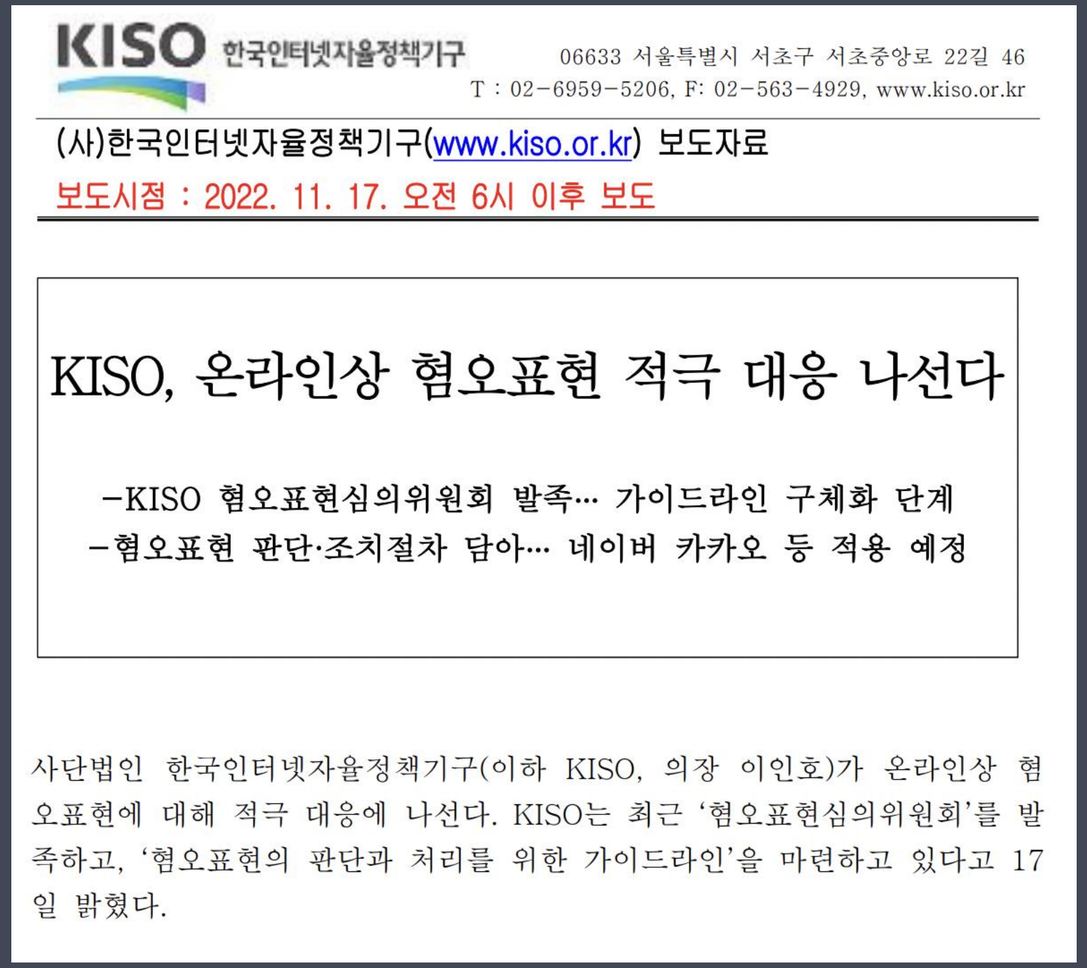
KISO는 인터넷 사업자들이 이용자들의 표현의 자유를 신장하는 동시에 이용자들의 책임을 제고해 인터넷이 신뢰받는 정보 소통의 장이 될 수 있도록 하고, 인터넷 사업자들이 이용자 보호에 최선의 노력을 기울이는 등 사회적 책무를 다하기 위해 설립 되었습니다.
또, 공정하고 투명한 정책을 통해 인터넷 공간의 질서와 사회적 합의를 도출하기 위해 노력하고 있습니다.

주요사업

KISO는 기구의 목적에 맞게 아래의 사업을 주요 사업 내용으로 합니다.

- 기구 강령 및 가이드라인 수립
- 회원사 등으로부터 요청 받은 인터넷 게시물 등의 여러 정책에 관한 사항
- 국제 자율규제기구와의 교류협력 및 국제기구 활동 참여
- 기타 기구 목적에 부합되는 사업

비즈니스 시나리오



KISO 한국인터넷자율정책기구

06633 서울특별시 서초구 서초중앙로 22길 46
T : 02-6959-5206, F: 02-563-4929, www.kiso.or.kr

(사)한국인터넷자율정책기구(www.kiso.or.kr) 보도자료

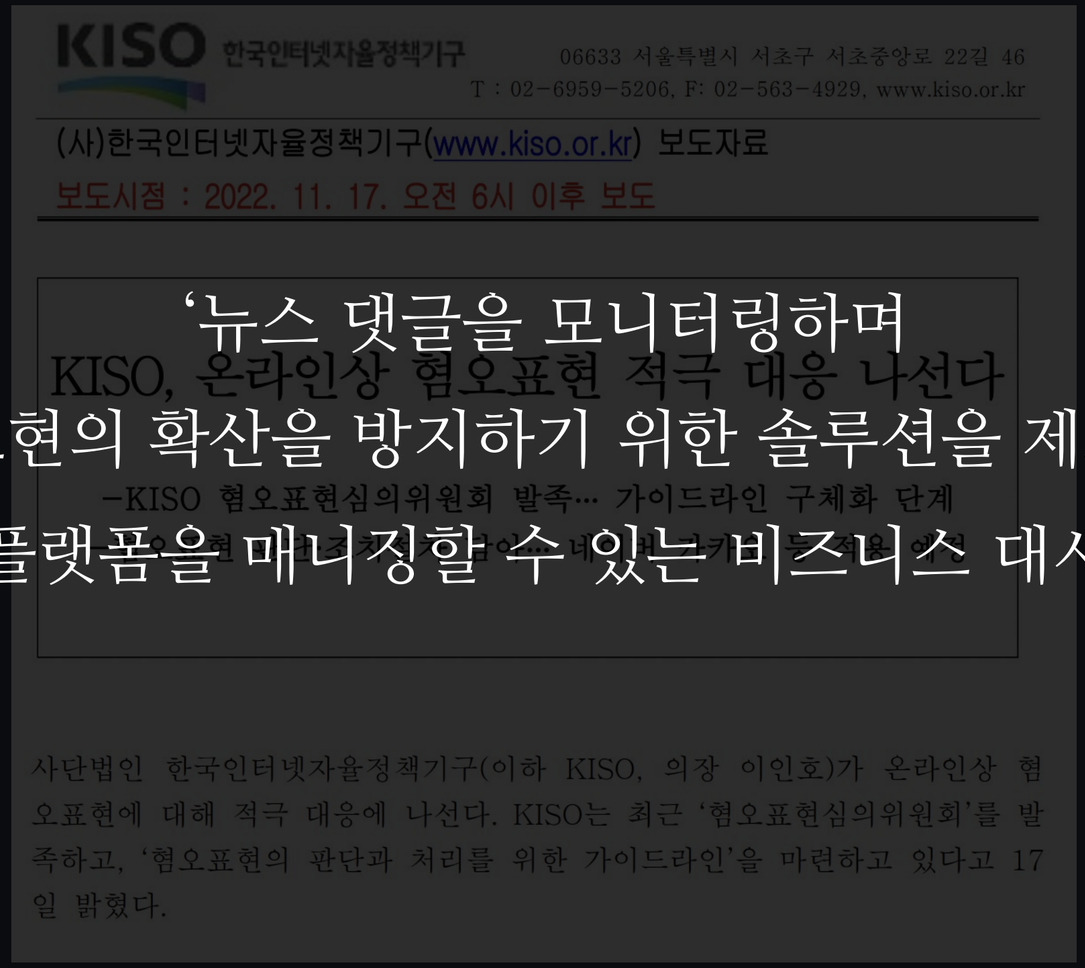
보도시점 : 2022. 11. 17. 오전 6시 이후 보도

KISO, 온라인상 혐오표현 적극 대응 나선다

-KISO 혐오표현심의위원회 발족… 가이드라인 구체화 단계
-혐오표현 판단·조치절차 담아… 네이버 카카오 등 적용 예정

사단법인 한국인터넷자율정책기구(이하 KISO, 의장 이인호)가 온라인상 혐오표현에 대해 적극 대응에 나선다. KISO는 최근 '혐오표현심의위원회'를 발족하고, '혐오표현의 판단과 처리를 위한 가이드라인'을 마련하고 있다고 17일 밝혔다.

비즈니스 시나리오



KISO 한국인터넷자율정책기구

06633 서울특별시 서초구 서초중앙로 22길 46
T : 02-6959-5206, F: 02-563-4929, www.kiso.or.kr

(사)한국인터넷자율정책기구(www.kiso.or.kr) 보도자료

보도시점 : 2022. 11. 17. 오전 6시 이후 보도

KISO, 온라인상 혐오표현 적극 대응 나선다

-KISO 혐오표현심의위원회 발족… 가이드라인 구체화 단계

사단법인 한국인터넷자율정책기구(이하 KISO, 의장 이인호)가 온라인상 혐오표현에 대해 적극 대응에 나선다. KISO는 최근 '혐오표현심의위원회'를 발족하고, '혐오표현의 판단과 처리를 위한 가이드라인'을 마련하고 있다고 17일 밝혔다.

'뉴스 댓글을 모니터링하며
혐오표현의 확산을 방지하기 위한 솔루션을 제고하고
뉴스 플랫폼을 매니징할 수 있는 비즈니스 대시보드'

데이터 수집

1. Kocohub 한국어 혐오표현 말뭉치 데이터셋

Korean HateSpeech Dataset

We provide the first human-annotated Korean corpus for toxic speech detection and the large unlabeled corpus.
The data is comments from the Korean entertainment news aggregation platform.

	news_title	comments	gender_bias	others_bias	none_bias	hate	offensive	none_hate	labels
0	"나혜미, 예뻑 아내→본업 컴백…연기 갈증 풀까 [MD픽]"	곧 이혼 뉴스 뜬다.	0	0	1	0	1	0	[0.0, 0.0, 1.0, 0.0, 1.0, 0.0]
1	"[단독] 조병규, '나혼자산다' 출연 확정…드라마 이어 예능까지 접수"	그냥 폐지해라 프로그램 오래해먹었잖아 관찰예능지겹다.법써 섭외봐라 범죄자들이랑 연결...	0	1	0	0	1	0	[0.0, 1.0, 0.0, 0.0, 1.0, 0.0]
2	한예슬 '코 피어싱' 본 BTS 정국 반응에 '갑론을박'	아니 코 피어싱 했다고 아팠겠다고 형들한테 얘기해주고 있잖아 ㅡ	0	0	1	0	0	1	[0.0, 0.0, 1.0, 0.0, 0.0, 1.0]
3	"[종합]""탐관오리에게는 죽음을""…첫방 '녹두꽃' 최무성, 먼란 일으켰다"	탐관오리에게는 죽음을~ 문재앙 정부 저격? 하긴; 강원도 산불났는데 시장이란넘은 ...	0	0	1	1	0	0	[0.0, 0.0, 1.0, 1.0, 0.0, 0.0]
4	"[천일야화] '슈돌' 손동이→까불이 윌리엄, 장난기 이 정도였어?"	윌리엄 그동안 스킵했었는데 오늘 이모습하고 서준이에서 눈여겨볼 아이로 성장했음	0	0	1	0	0	1	[0.0, 0.0, 1.0, 0.0, 0.0, 1.0]

- 한글로 된 뉴스 댓글과, 뉴스 댓글 내 혐오표현의 정도를 측정할 수 있는 데이터

- 뉴스 제목과 댓글 정보가 있으며, 혐오표현의 유형은
  Gender Bias, Other Bias, None Bias, Hate, Offensive, None Hate 총 6개로 구성됨

- 해당 댓글이 각 혐오표현에 해당하는지의 여부를 0과 1로 구분해놓은 데이터

2. Real World Fake Data

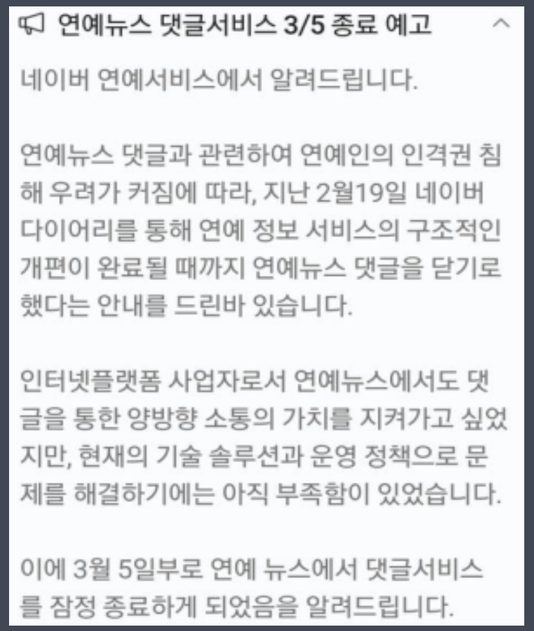
- 실제 비즈니스 상황에서 뉴스 데이터를 분석할 때 필요한 '뉴스 게시 일자', '뉴스 공감 수', '댓글 좋아요 수' 등의 정보를 추가하기 위해, data.world에서 제공하는 비즈니스 분석을 위한 가상의 데이터, RWFD를 사용함

출처1: https://github.com/kocohub/korean-hate-speech
출처2: https://data.world/markbradbourne/rwfd-real-world-fake-data/workspace/file?filename=SocialMedia.csv

데이터 수집

데이터 관련 추가 설명

- 실제 네이버 포털 연예뉴스 댓글은 2020년 폐지



우리의 대시보드 구축 목적

이번 프로젝트의 비즈니스 대시보드가 향후 타 SNS 게시글 또는 정치·경제 분야 뉴스 댓글로 확장,
실제로 다양한 미디어 비즈니스 상황에서 활용될 수 있을지에 대한 가능성을 제시하는 데에 초점을 두고 있음

→ 따라서, 본 프로젝트에서 **목표하는 비즈니스 상황을 구축하기 위해, 일부 가상의 데이터를 사용**하고 있음

데이터 분석 개요

분석 환경

- Google Colab (하드웨어 가속기: GPU)

- Programming Language: 파이썬 (Python)
- Pytorch: 딥러닝을 구현을 위한 파이썬 기반의 오픈소스 머신러닝 라이브러리

- Library: Pandas (파이썬의 데이터프레임 관리 라이브러리),
  Scikit-Learn (데이터 분석을 위한 머신러닝 라이브러리),
  Transformer (딥러닝 기반 자연어 처리 라이브러리)

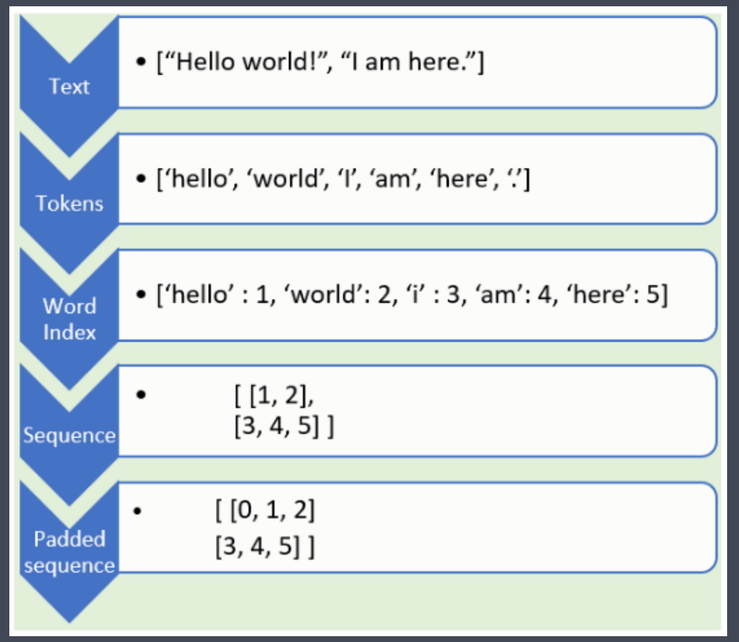
데이터 분석 과정

[딥러닝 기반 자연어 처리] 다중분류 (Multi Label Classification)를 이용한 감성분석

특정 댓글이 6개의 혐오표현 유형 각각에 속해 있을 확률을 분류해 예측하는 것이므로, 다중 분류에 해당됨

데이터 분석 파이프라인 (Tokenizer ~ Classification Model)

1) Auto Tokenizer: 뉴스 댓글 토크나이징, 패딩 진행
특정 뉴스 댓글의 혐오표현 점수를 예측하는 딥러닝 모델을 훈련시키기 위해,
뉴스 댓글 텍스트를 형태소 단위로 자르고, 텍스트를 정수형으로 변환



2) LRAP(Label Ranking Average Precision)
원 데이터셋에 주어진 '혐오표현 해당 여부'를 기반으로, 각 댓글의 혐오유형별 점수를 계산
(주로 다중분류 분석에서 사용되는 계산식)

데이터 분석 과정 및 결과

3) Bert For Sequence Classification

BERT란? 자연어 분류를 위해 '사전에 훈련된 언어 모델'

본 프로젝트에서는 Pytorch의 Transformer 라이브러리를 제작한 'Hugging Face'에서 제공하는 사전 훈련 모델을 사용해 모델 훈련을 진행함

* 성능 향상을 위한 모델 훈련 옵션 (batch size:64, epoch:5)

모델 훈련 결과, 댓글 혐오표현 점수 예측 성능

Precision 82%, Recall 61%, F1 Score 66, Support 942

→ 약 82%의 성능으로 혐오지수 예측 가능

출처: https://ebbnflow.tistory.com/151
https://velog.io/@raqoon886/KorBertSum-SummaryBot

데이터 시각화_대시보드 스케치

KISO 입장에서 얻고자 하는 정보가 무엇일지 정리,
다양한 정보를 효율적으로 제공할 수 있는 대시보드를 기획
동시에, 대시보드 제작을 위한 레퍼런스 탐색을 진행

#1 뉴스 댓글 혐오표현지수 검색 대시보드

대시보드 선택패널

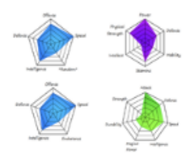
뉴스 게시일자 범위 | 조회수 범위 검색 필터 | 언론사 검색 필터

뉴스 제목 워드클라우드

뉴스 제목 목록

뉴스별 KPI 지표
뉴스 댓글 수, 반응 수

해당 기사 댓글들의
gender_bias **평균**점수..
offensive **평균**점수..



해당 뉴스의
혐오 종합지수

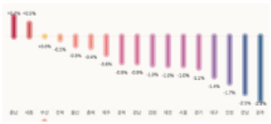
뉴스 댓글 목록
답글 수, 좋아요 수,
이 댓글의 gender bias, offensive 등.. 혐오유형
점수

#2 혐오지수 기반 언론사 매니징 대시보드

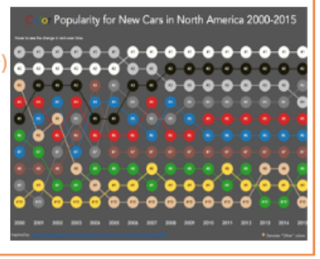
대시보드 선택패널

혐오유형별 뉴스 기사 수
추이

언론사별
종합 혐오지수,
6개 혐오유형 점수 비교



2022년
월간 언론사 순위
(혐오기사 수 내림차순)



언론사 선택

혐오유형 7개 아이콘

선택한 언론사, 혐오유형의
기사 순위

뉴스댓글 혐오지수 서치 대시보드 기능 설명

뉴스댓글 혐오지수 서치 대시보드 기능 설명

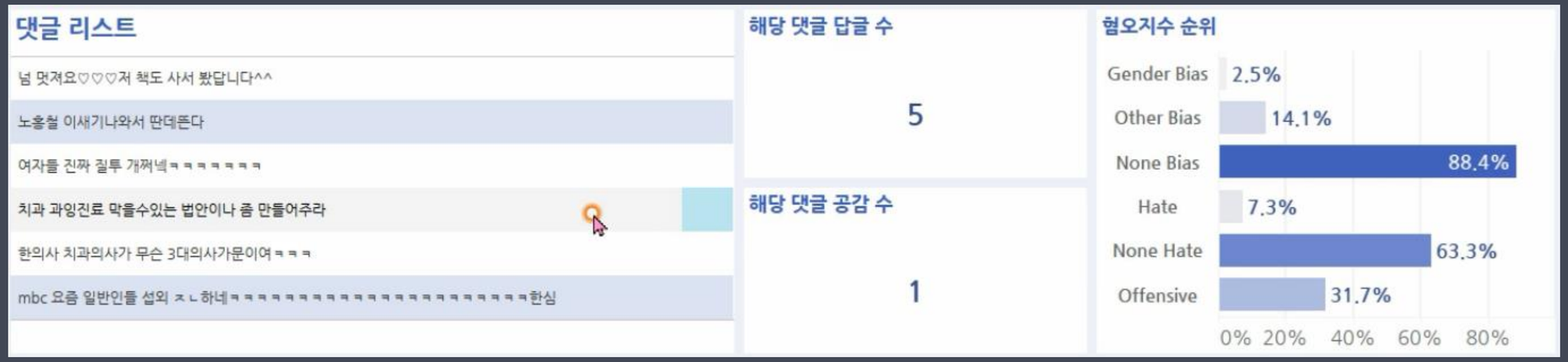
- 뉴스 날짜별 조회,
- '좋아요' '응원해요' 등의 반응 수에 따른 뉴스 목록 탐색

뉴스댓글 혐오지수 서치 대시보드 기능 설명

- 뉴스 타이틀, 댓글 선택
- 뉴스별, 뉴스 댓글별 혐오지수 확인

뉴스댓글 혐오지수 서치 대시보드 기능 설명

인사이트 예시)



악성댓글 X → None Bias 88.4%

악성댓글 O → Gender Bias 64.7% Hate 75%

혐오지수 기반 언론사 매니징 대시보드 기능 설명

– 각 언론사의 뉴스 평균 혐오지수 파악
혐오지수 높은 언론사 도출 → 언론사 문제점 파악 및 관리

혐오지수 기반 언론사 매니징 대시보드 기능 설명

언론사별 혐오종합지수 도출 계산식

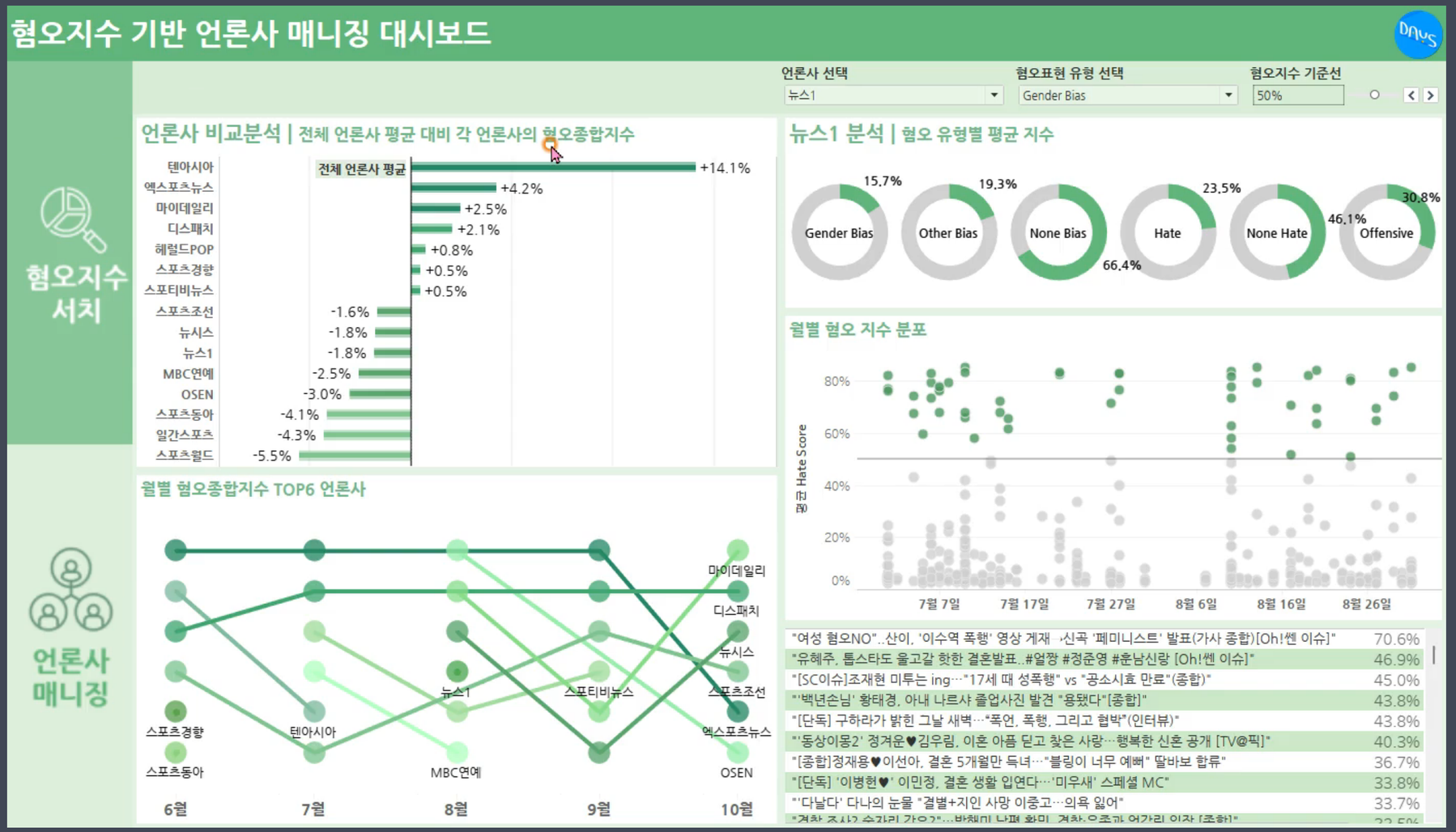
해당 언론사 뉴스별 평균 혐오표현 점수
Total Hate Score

$$\rightarrow \frac{\text{해당 언론사 뉴스별 평균 Gender Bias x Other Bias x Hate x Offensive}}{\text{해당 언론사 뉴스별 평균 None Bias x None Hate}}$$

해당 언론사의 혐오종합지수

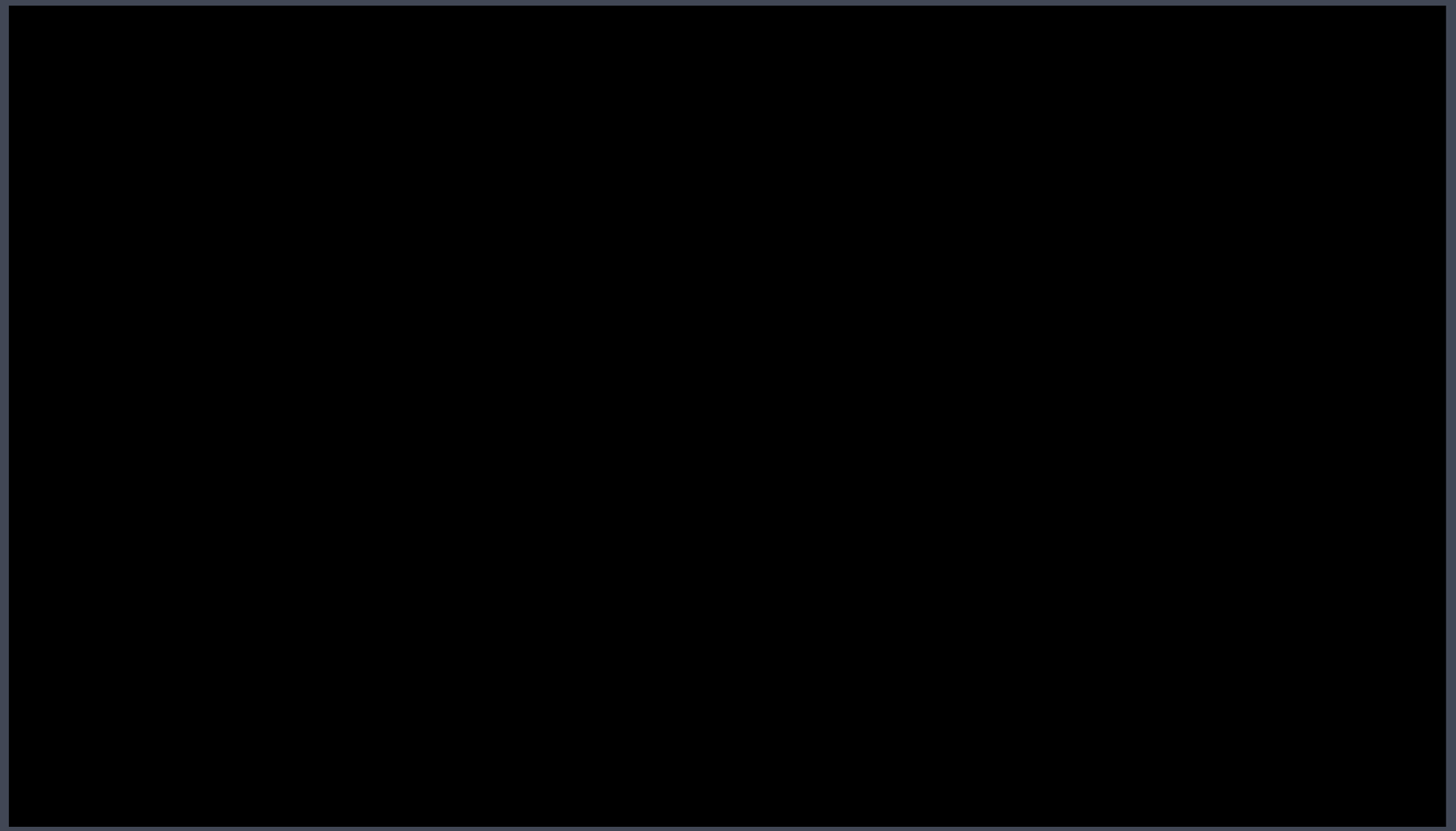
$$\rightarrow \frac{\text{Total Hate Score}}{\text{해당 언론사의 뉴스 개수}} \times 100$$

혐오지수 기반 언론사 매니징 대시보드 기능 설명



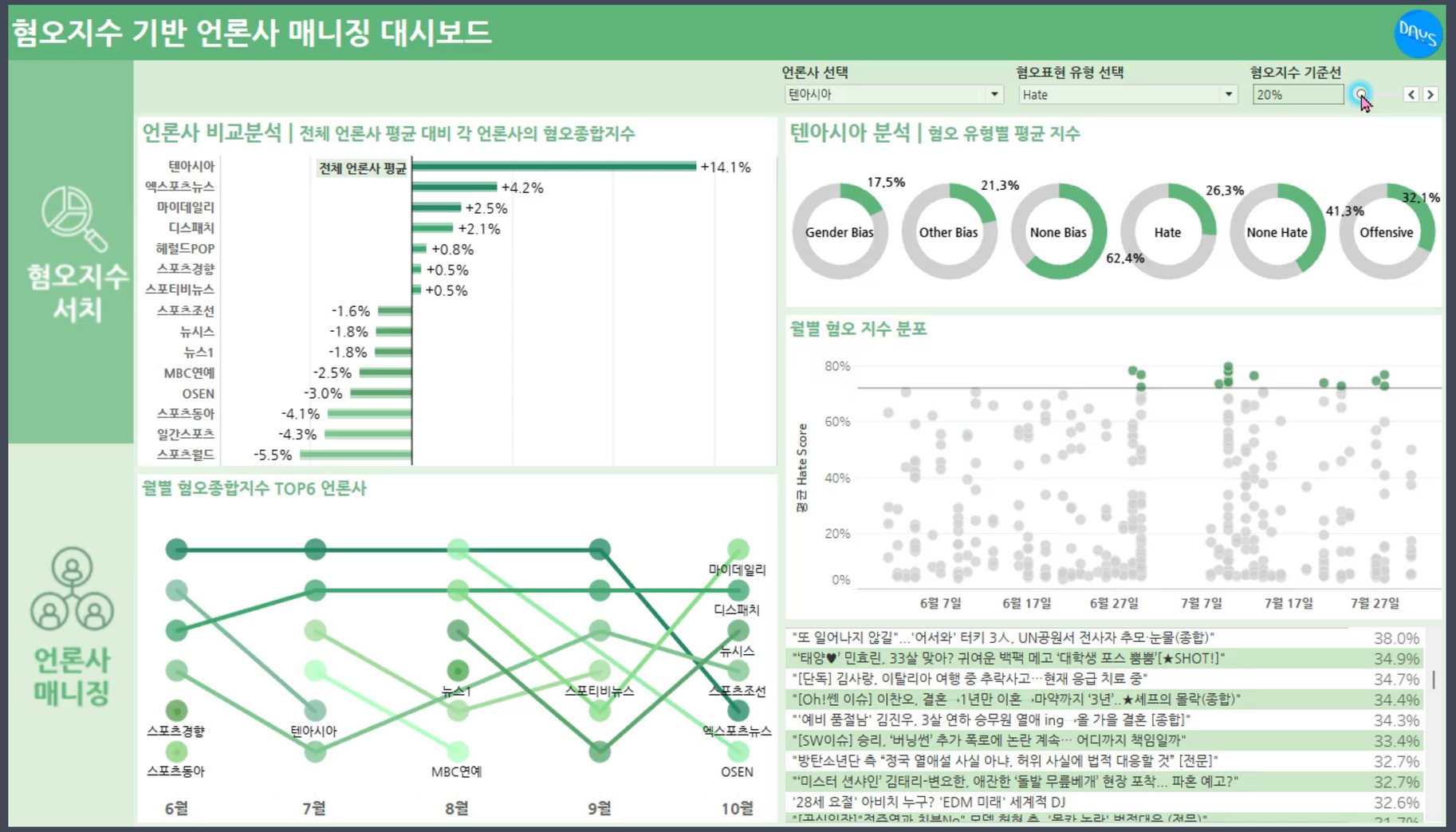
- 전체 15개 언론사 평균 혐오종합지수 17.5%
- 전체 평균 대비 각 언론사의 혐오종합지수 비교, 월별 언론사 순위 확인

혐오지수 기반 언론사 매니징 대시보드 기능 설명



특정 언론사 집중 분석 | 혐오지수 고위험군에 속하는 뉴스 추출

혐오지수 기반 언론사 매니징 대시보드 기능 설명



뉴스 제목 클릭 → 해당 뉴스의 혐오지수 분포 위치 파악

ETL 파이프라인 구축: 대시보드 자동화

1) KISO DB에 새로운 데이터 업데이트

2) 파이썬 데이터 모델링 → 아웃풋 추출

3) 태블로에 데이터 연결

4) 대시보드 업데이트

예시 이미지

프로젝트 결과 및 의의

가상의 비즈니스 시나리오를 설계하고, 이에 맞는 대시보드를 개발
고객이 원하는 정보를 제공하기 위해, 자체적으로 지수를 개발하고, 이를 대시보드에 활용함

딥러닝을 기반으로 한 NLP 분석 결과를 시각화하여, 뉴스 댓글의 혐오표현 지수를 탐색함
→ 혐오지수 높은 댓글 탐색 및 제거 가능

혐오표현지수가 높은 댓글이 반복적으로, 심한 강도로 재생산되어 배포되는 언론사는 인터넷 언론 문화를 저해함,
해당 대시보드를 통해 혐오댓글을 유발하는 자극적인 기사를 생산하는 언론사 추출 및 감시 효과 가능
→ 뉴스플랫폼 (네이버, 카카오 등)에 해당 리스트를 제공하여 건강한 언론문화를 조성하도록 유도
뉴스와 댓글에 관련된 사람들의 영향 연구에 활용 가능