


# [국문요약] Korean Online Hate Speech Dataset for Multilabel Classification (Kang, TaeYoung, et al., 2022)

Created By

 TaeYoung Kang

Last Edited

2022년 4월 8일 오전 4:44

Tags

Research

Property

비어 있음

## 기초 정보

### • 논문

BEEP! Korean Corpus of Online News Comments f...

Toxic comments in online platforms are an unavoidable social issue under the cloak of anonymity. Hate speech



<https://arxiv.org/abs/2005.12503>



### • 데이터셋



korean\_unsmile\_dataset

smilegate-ai



hatescore-korean-hate-speech

sgunderscore

### • 소개 영상 (Korean Unsmile Dataset)

[Smilegate AI] UnSmile 데이터셋

Smilegate AI x Underscore] UnSmile 데이터셋을 공개합니다.  
데이터셋 링크: <https://github.com/smilegate->



<https://youtu.be/XmCnlczWtQ>



## 요약

- 본 연구의 목적은 사회과학적 논의들을 바탕으로, 비서구·비영어권 문화를 소재로 한 혐오발언 데이터셋을 개발하는 것
- 총 3.5만건의 온라인 텍스트 데이터를 수집했으며 다중레이블(multi-label) 방식으로 태깅·학습함
  - SmilegateAI UnSmile Dataset 2.4만 건 (기본 데이터셋)
  - HateScore Dataset 1.1만 건 (중립 레이블 위주 추가 데이터셋)
- 데이터의 카테고리는 총 10가지
  - 혐오발언 - 여성·가족, 성소수자, 지역, 종교, 인종·국적, 연령, 남성, 기타
  - 단순악플
  - 중립 댓글
- KcBERT 기반의 분류 모델 성능(LRAP, Label Ranking Average Precision)의 경우, 기본 데이터셋 2.4만 건을 사용했을 때에는 0.89를 기록했으나 보조 데이터셋 1.1만 건을 함께 학습 시 0.92로 소폭 증가했으며 실제로도 사회적 소수자나 논쟁적인 소재를 언급만해도 혐오발언으로 오분류하던 문제를 어느 정도 해소함
- 본 연구는 스마일게이트 엔터테인먼트 산하 스마일게이트 AI의 지원을 받아 진행되었으며 논문에서 다루는 사회과학적 논의, 문헌 검토 내용, 예시 문장 사례 등은 모두 스마일게이트AI의 공식 입장과는 무관합니다.

## 연구의 배경

- 혐오발언(hate speech)은 단순히 거시적인 사회 이슈일 뿐만 아니라 챗봇과 같은 텍스트 데이터 기반 서비스를 개발하는 IT기업에게도 중요한 문제
- 기존의 악플(online profanity) 관련 데이터셋들은 일반적으로 이진 분류만을 활용함 (1=악플, 0=악플 아님)
- 혐오발언 데이터셋의 경우, 영어 텍스트들은 반유대주의(anti-semitism), 흑인 혐오(anti-black), 무슬림 혐오(anti-muslim)등의 다양한 카테고리들을 다루고 있으나 여전히 비서구·비영어권에 적합한 분류 체계 및 실제 데이터를 제공하는 사례는 많지 않았음
- 지역적인 맥락은 물론 혐오발언(hate speech)과 사회적 소수자(social minority)라는 개념에 대한 사회과학적 고민들을 반영하여 관련 데이터셋을 구축할 필요가 있음

- 따라서 본 연구의 목표는 단순히 한국어 온라인 혐오발언 데이터셋 개발을 넘어 다양한 문화권에서 어떻게 하면 각 국가 별 데이터셋을 구축할 수 있을지, 일반적인 매뉴얼을 제시하는 것
- 또한 혐오발언은 한 텍스트가 여러명의 혐오 대상을 지칭할 수 있음. 그렇기에 단순히 상호 독립적인(mutually exclusive) 레이블링을 하는 대신 다중레이블(multi-label) 방식으로 주어진 문장이 다루는 혐오를 모두 태깅한 데이터셋을 구축함



#### 혐오발언 카테고리, 어떻게 선정했는가?

- 좁은 의미의 혐오 발언은 **“사회적·역사적으로 차별과 억압을 받아온 소수자 집단의 정체성을 그 대상으로 하여 공격을 일삼는 표현”**(박미숙·추지현, 2017)을 의미하나, 보다 포괄적으로는 **“합리적인 근거 없이 인종, 성별, 출신지역, 장애 등을 이유로 특정 개인이나 집단에 대해 경멸적 표현을 사용하여 배척하거나 사회적 편견을 조장하는 내용”**(방송통신심의위원회, 2015)이라고 정의할 수 있음.
- 온라인 데이터는 익명으로 된 날 것의 혐오 발언을 얻을 수 있다는 강점이 있으나 이는 통계적으로 대표성이 있는 데이터가 아님. 대부분의 댓글은 소수의 유저에 의해서만 작성되고, 또 ‘가장 보편적인 웹 페이지’의 개념 역시 불명확하며 인구통계적인 정보를 알 수도 없음. 즉, 온라인 텍스트 데이터 상에서의 분포(distribution)만으로 혐오발언 카테고리를 데이터 기반으로 정하는 것은 편향된(biased) 결과를 낳을 수 있음
- 이에 UN, 국가인권위원회 등 국내외의 공공 기관들의 혐오발언·차별 관련 보고서 및 관련 사회과학 연구를 바탕으로 ①지역 ②종교, ③인종·국적, ④연령, ⑤여성·가족, ⑥성소수자, ⑦남성의 일곱 가지 유형의 혐오발언을 주요 카테고리로 선정함.
- 전문성의 부족으로 인한 레이블링의 주관성을 최소화하기 위해 사회과학 분야 석사 과정 이상의 연구자들만을 매뉴얼 구축 및 실제 레이블링에 참여시킴.



#### 각 카테고리는 어떠한 특성에 유의해서 레이블링했는가?

## • 지역

- 지역주의는 1960년대 이후 한국 사회의 핵심적인 갈등 축이며 '영호남 대립'이 그 전통의 핵심. 박정희 이후 약 40년 간 대통령을 배출한 영남 지역은 대한민국에서 패권적 지위를 얻었음. 그에 반해 80년 5월 항쟁으로 대표되는 국가 폭력은 한국의 지역갈등, 특히 호남 지역 출신자들의 원한 감정을 증폭함. 자연스레 보수당계 정당과 민주당계 정당 간의 갈등 역시 이 두 지역과 결부되어 지역 갈등이 진행됨
- 다만 최근의 온라인 커뮤니티의 경우 영호남 지역 갈등을 주제로 한 혐오표현을 넘어 모든 지역에 대해 전방위적인 조롱을 하고는 함. 가령 디시인사이드의 야구갤러리는 연고지가 있는 모든 지역에 대해 희화화와 조롱을 하고는 함.
- 최근 성별이나 세대가 새로운 갈등의 축으로 부상했으나 여전히 지역주의는 한국 사회의 차별과 적대를 다루는 중요한 축이기에 혐오 카테고리 중 하나로 포함함.

## • 종교

- 서구의 경우, 종교 관련 혐오 발언이 이민 이슈와 결부되어 무슬림에 집중되어 있으나 한국은 개신교 관련 적대적 발언이 주를 이룸.
- 물론 개신교는 국내에서 가장 신자 수가 많은 종교이기에, 이러한 적대감은 소수 종파에 대한 이질감에서 비롯되었다기 보다는 다수 종교로서의 개신교가 보여준 부정적 모습이 매체를 통해 확대·재생산되며 본격화한 것으로 판단함.

## • 인종·국적

- 인종적으로 동질적인 한국 사회의 특성 상, 서구와 달리 한국은 엄밀하게는 '국적'에 대한 차별 발언의 비중이 높으며 대개 다음의 세 가지 유형을 포함함.
- 첫 번째, 개인의 문제를 인종·국가의 문제로 환원하는 발언. "조선족들은 다 오원춘 같은 살인범임"이나 "무슬림은 솔직히 잠재적 테러리스트 아님?"과 같은 문장들이 그 예시.
- 두 번째, "너희 나라로 돌아가라", "이자스민은 한국에 무슨 꿀을 빨려고 왔냐" 등, '순수한 내국인'으로 구성된 내집단과 외국인을 완전히 분리하고 배제하는 표현.
- 세 번째, 특정 인종·국가에 대한 선입견을 강화시키고 재생산하는 표현. 이는 단순히 "필리핀 사람들은 게을러"와 같은 부정적인 통념 외에도 "와따 흑형 이두랑 빵디 크기 보소?", "스시녀들은 순종적이어서 좋아"와 같이 선망과 우호적인 의도를 함축한 고정관념까지도 포괄함.

## • 연령

- 연령 관련 혐오 표현은 세대(generation)에 관한 혐오 표현과 생물학적 연령(age)에 관한 혐오 표현의 두 가지로 구분됨.
- 빠른 사회 변화로 인해 세대 간 격차가 컸기에, 세대 개념은 점차 한국 사회의 주요한 갈등 축으로 부상하고 있으며 2000년대 이후 한국 사회에서 영향력을 높여 간 '586 세대'와 전통적으로 보수적인 경향을 지니는 '산업화 세대'가 찬사·비난·동정의 소재로서 정치적으로 호명되고는 함.
- 물론 '급식충'이나 '꼰대', '틀딱충'과 같이 세대 개념이 아닌 생물학적 연령을 소재로 하는 조롱 및 희화화 표현 역시 근래에 나타남.

## • 여성·가족

- 사회의 모습이 변화하는 동안 젠더 불평등의 모습도 변화했지만, 젠더 불평등 자체는 아직까지 사라지지 않았음. 남성성은 아직까지 사회적 표준으로 여겨지며 여성적 행동, 취향, 생각, 태도 등의 여성성은 남성성에 비해 평가 절하되고는 함.
- 노동시장에서의 차별 및 여성에 대한 디지털 성폭력 등의 적대적인 성차별(hostile sexism)은 '김치녀', '메갈쿵왕이', '피싸개' 등과 같은 어휘와 함께 온라인 공간에서도 일상적으로 등장함.
- "우리 000은 얼굴이 예쁘니까 공부 좀 못해도 괜찮아", "서른 전에 시집 갔으니 너무 늦기 전에 잘 했구나", "여자답고 조용하니까 인기 많겠다"와 같은 발언들은 직접적인 비하 발언이나 욕설이 포함된 발화는 아니나, 특정한 성 역할, 즉 여성으로서 수행해야 하는 행동과 지녀야만 하는 성격에 대한 고정관념을 전제하고 있기에 유저에게 불쾌감을 줄 수 있는 자애적인 성차별(benevolent sexism)에 해당됨.

## • 성소수자

- 동성애, 양성애, 트랜스젠더 등의 LGBTQ(Lesbian, Gay, Bisexual, Transgender, and Queer)에 대해서는 이성애 규범성(heteronormativity)에 부합하지 않고 젠더에 불순응(gender non-conformity) 한다는 이유로 차별이 존재함
- '이성애 규범성'이란 이성 관계만을 사회에서 허용되는 관념으로 여기고, 그 외의 성적 관계는 배제하는 사회적 규범을 지칭하며 '젠더 불순응'은 여성성과 남성성에 부여된 젠더 관념에서 벗어나는 것에 대한 반감을 의미함.
- 최근 국내에서는 차별금지법이 성소수자 인권 문제가 대두되고 있으나 여전히 동성혼 합법화와 같은 관련 이슈에서 보수적인 경향을 보이고 있음. 한국의 동성애 및 성소수자 차별은 개신교를 주체로 하는 경우가 많음.

- 남성

- 2015-2016년의 메갈리아-워마드 현상의 유행과 함께 '미러링'이라는 이름으로 남성에 대한 조롱과 희화화를 포함한 악플의 유형이 새롭게 등장하기 시작.
  - 여성에 대한 차별이 현재까지 계속되고 있고, 또 젠더 관련 악플들 역시 2000년대 초반 '김치녀', '된장녀' 등의 표현으로 '남성→여성'의 방향으로 시작되었기에 '여성→남성' 방향의 악플을 동일한 "젠더 혐오"로 개념화해야 하는지는 다소 논쟁적.
  - 다만 "특정한 집단에 대한 적대감을 일반화하고 낙인 찍는 표현" (Howard, 2019)이라는 보다 포괄적인 정의에 따르자면, 온라인 공간에서 빠르게 성장 중인 특정 유형의 악플을 완전히 제외할 수는 없음. 이에 해당 유형의 발언을 악플·혐오발언 카테고리에 포함함.
  - 남성 대상 조롱·희화화 은어가 자주 사용되는 온라인 커뮤니티가 성전환자에 대해 배제적인 급진 페미니즘(TERF, Trans-Exclusionary Radical Feminism)의 입장을 보이는 특성 상, 성소수자 혐오 표현과의 다중 레이블링에 유의함.
- 위의 7가지 유형에 속하지는 않지만 외모, 직업군, 장애, 소득 등을 소재로 조롱·희화화·차별하는 발언의 경우 **기타 혐오**로, 욕설 댓글 및 음담패설은 **단순 악플**로, 아무런 문제가 없는 텍스트는 **중립**으로 레이블링함.



#### 어떤 데이터를 수집했고 어떻게 레이블링했는가?

- 2019년 상반기부터 2021년 상반기까지의 뉴스 데이터 ([한국언론진흥재단 BigKinds 뉴스DB](#) 활용)를 수집 후 네이버·다음 포털 뉴스와 매칭 후 댓글 수집
- 온라인 커뮤니티 (일간베스트저장소, 디시인사이드, 워마드, 오늘의유머) 최신글 댓글

- 위 웹사이트들에서 아래의 쿼리를 기준으로 기사·게시물들을 필터링함

주제	검색 쿼리
여성·가족	페미니즘, 젠더, 성차별, 여성 인권, 미혼모, 맘충, 메갈리아, 워마드, PC주의, PC충
성소수자	퀴어, 퀴어퍼레이드, 동성애, 게이, 트랜스젠더, 레즈비언, LGBT, 성소수자, 차별금지
남성	미투 운동, N번방, 성폭행, 성매매, 이남자, 20대 남자, 한남충
인종·국적	예멘 난민, 다문화, 조선족, 탈북자, 반일, 반중, 이자스민, 이민자, 무슬림, 이주민, 이민
연령	틀딱, 틀니딱딱, 급식충, 잼민이, 노인 혐오, 아동 혐오
지역	지역주의, 지역 차별, 지역 혐오, 전라도, 호남, 머구, 쌍도, 라도
종교	무슬림, 이슬람, 기독교, 목사, 개독교, 개슬람

- 이 중 **2.4만건의 데이터**에 대해 3명의 연구자들이 레이블링에 참여해 다수결로 최종 레이블을 결정함. 레이블의 일관성을 측정하는 지표인 **Krippendorff's Alpha 값은 0.713**으로, 10가지의 다중 레이블이라는 방식을 감안할 때 상당히 우수한 편
- 다만 위 데이터만 활용 시 특정 어휘가 포함되어있기만 해도 혐오발언으로 오분류하는 한계 (ex. 저 사람은 중국인이다, 너는 페미니즘에 대해 어떻게 생각하니 등)가 존재해 **1.1만건의 '일반 댓글' 레이블 위주의 데이터셋**을 추가 개발함. 이는 위의 검색 쿼리를 다루는 위키피디아 문서들에서 추출한 문장 2.2천 건, 2021년 하반기 데이터를 추가 수집해 Human-in-the-Loop 방식으로 레이블링한 1.7천 건, 규칙 기반으로 생성한 중립 문장 7.1천 건으로 구성됨.



#### 분류 모델의 성능은 어떠한가?

- 다중레이블 분류 모델 평가에 활용하는 지표인 LRAP(Label Ranking Average Precision) 기준으로 분류 모델의 성능은 다음과 같음. 모델 테스트에는 KcBERT와 KcELECTRA를 활용함.

모델명	UnSmile	Unsmile + HateScore
KcBERT-base	.886	.914
KcBERT-large	.892	.919
KcELECTRA-large	.884	.912

- 아래의 예시에서 볼 수 있듯 중립 레이블을 다수 포함한 추가 데이터셋을 함께 학습 시 문장 오분류 확률이 소폭 감소하는 것을 확인할 수 있음. (예시는 KcBERT-base 기준이며 괄호 안의 수치는 혐오발언 분류 확률을 의미)

예시 문장	UnSmile	Unsmile + HateScore
저 사람 중국인이네	<b>0.867</b>	0.196
너 페미니스트니?	0.028	0.006
동성혼은 논쟁적이지	0.347	0.008
무슬림을 다 죽인다고?	<b>0.835</b>	<b>0.761</b>

- KcBERT base 모델 및 Unsmile+HateScore 데이터셋 기준으로 문장 분류 예시는 아래와 같음

여자는 집에서 애나 봐라	<b>0.86</b>	0.01	0.03	0.03	0.01	0.01	0.01
좃족은 21세기의 홍어다	0.03	0.02	0.03	<b>0.68</b>	<b>0.89</b>	0.04	0.03
너는 전라도 사람이니?	0.00	0.00	0.00	0.00	0.01	0.00	0.00
상폐 한남들 다 재기하라고	0.09	0.02	<b>0.88</b>	0.05	0.05	0.04	<b>0.55</b>
도심에서 변태성욕 축제라니 말세	0.06	<b>0.79</b>	0.02	0.01	0.13	0.01	0.01
원내나는 태극기들 틀니 압수	0.07	0.03	0.06	0.09	0.03	0.05	<b>0.94</b>
개독이나 짱깨나 거기서 거기	0.05	0.03	0.02	<b>0.84</b>	0.09	<b>0.92</b>	0.05
저 친구는 필리핀 출신이다	0.00	0.00	0.00	0.01	0.00	0.00	0.00
콩깡이들도 필리핀 그지는 싫지?	<b>0.74</b>	0.01	0.04	<b>0.71</b>	0.02	0.01	0.01



#### 데이터셋 및 모델의 한계는 어떻게 되는가?

- 본 데이터셋은 온라인 댓글 데이터만을 다룸. 그렇기에 웹 커뮤니티 제목 텍스트에 모델을 적용할 경우, 혐오발언 여부를 오분류할 가능성이 높음.



- 중립 문장 오분류 문제 역시 사실 인위적인 중립 레이블 데이터셋 추가보다는 모델 단에서 해당 문제를 해결하는 것이 보다 본질적인 해결책일 것. 물론 본 연구는 데이터셋 페이지지 모델 성능의 SOTA 달성에 집중하지는 않기에 이는 후속 연구의 영역일 것.



## 연구 참여 인력

- **협오발언 유형 설정, 레이블링 매뉴얼 작성, 모델 개발**
  - 강태영 (KAIST 경영공학 석사)
  - 권은낭 (연세대학교 사회학 박사과정)
  - 김학준 (서울대학교 사회학 석사)
  - 남영은 (Ph.D. candidate in Sociology at Purdue University)
  - 서정규 (Ph.D. candidate in Political Science at University of Houston)
  - 송준모 (연세대학교 사회학 박사과정)
  - 이준범 (서울대학교 데이터사이언스학 석사)
- **레이블링 보조참여자**
  - 권혜윤 (서울대학교 인류학 석사)
  - 박형준 (싱가포르국립대학교 심리학 석사)
  - 이보미 (서강대학교 정치외교학 석사)
  - 이성일 (서강대학교 사회학 석사)
  - 지소연 (서강대학교 사회학 석사)
  - 홍수민 (연세대학교 정치외교학 석사과정)
  - 황지영 (서강대학교 사회학 석사과정)
- **문의**  
master@underscore.kr