Korean Online Hate Speech Dataset for Multilabel Classification

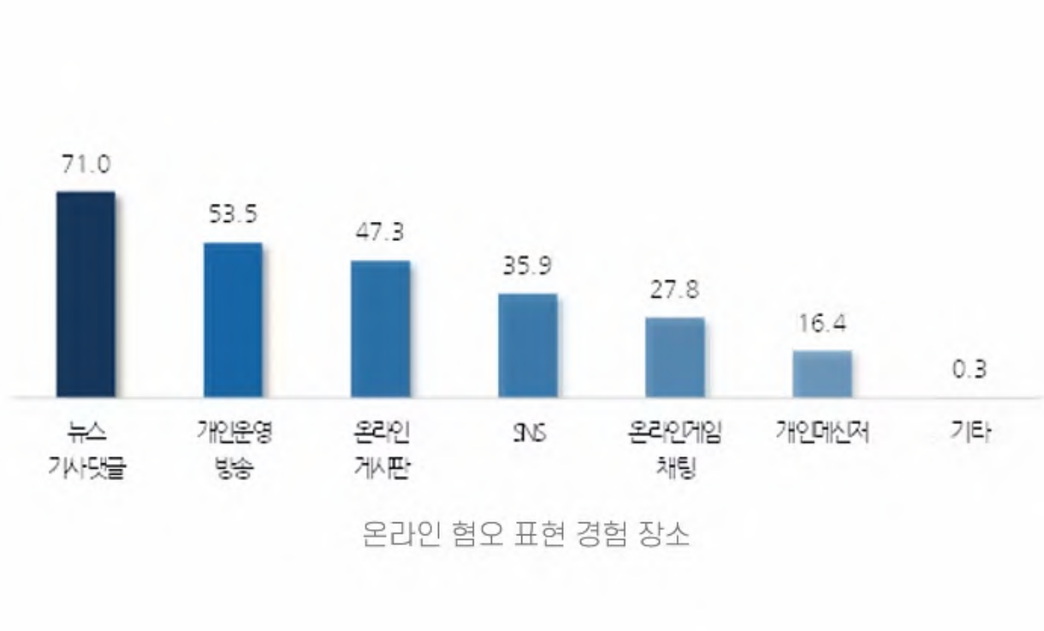
목차

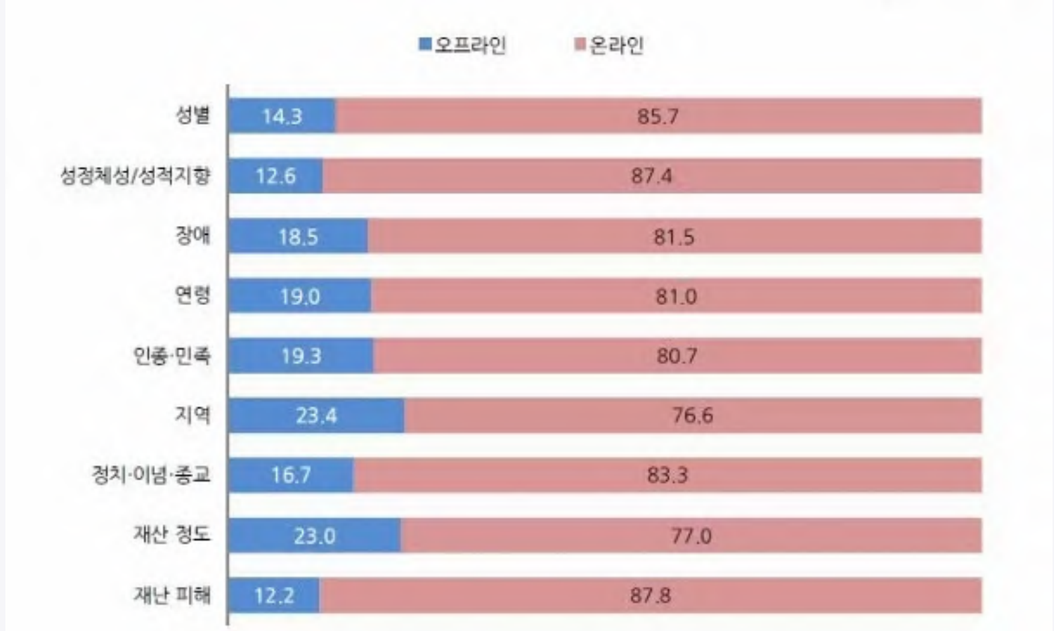
동기

많은 사람이 sns나 댓글 등 다양한 방법으로 사람들과 소통할 수 있게 되면서,

온라인 상에서 다양한 문제들이 발생하게 되었다.

그 중 다양한 온라인 경로를 통해 혐오 표현을 마주하는 경우가 다수 존재한다.





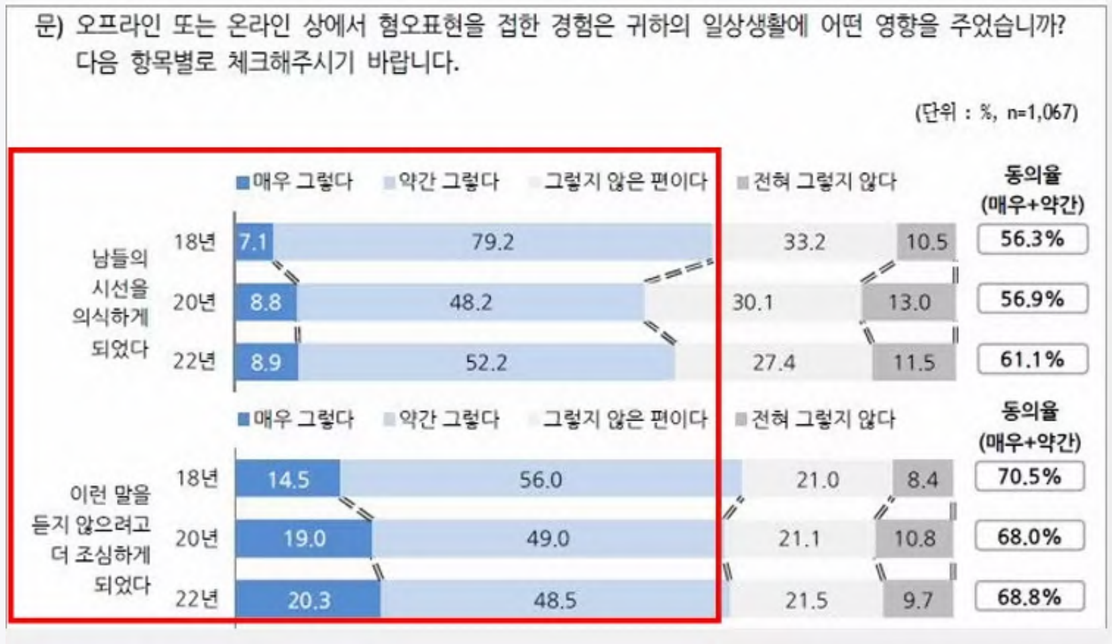
1) (문화체육관광부) 2022년 혐오표현 관련 대국민 인식조사 : https://www.korea.kr/archive/expDocView.do?docId=40270

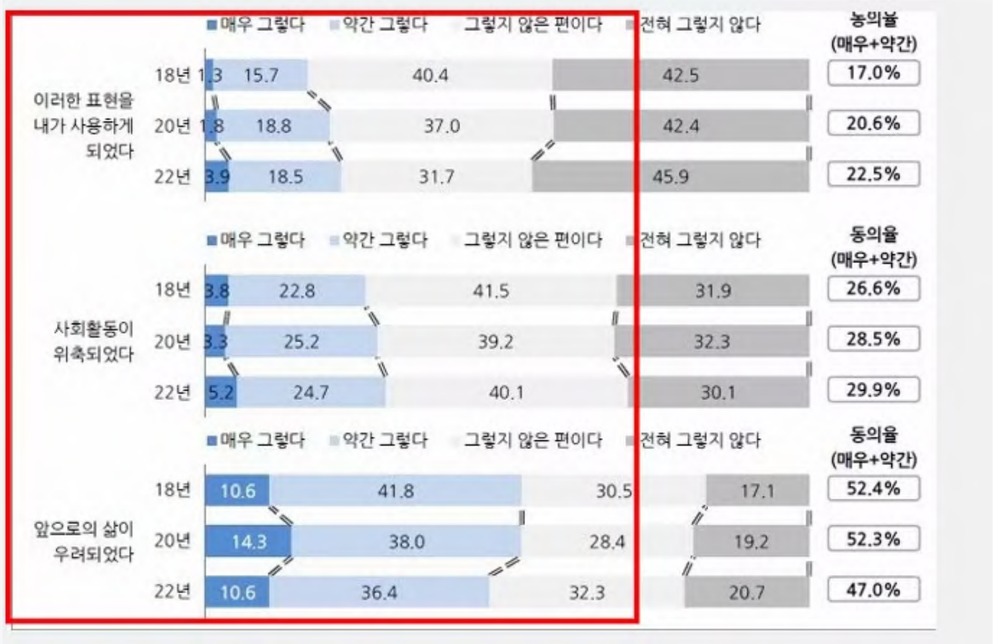
2) (KBS 뉴스) 온라인 혐오 표현은 여성·특정지역·페미니스트를 노린다 2021.09.02: https://news.kbs.co.kr/news/mobile/view/view.do?ncd=5270674

동기

온라인 혐오 표현

- 혐오 표현을 접한 경험에 대해 대부분의 사람이 부정적인 영향을 받는다고 응답
- 혐오 표현에 대한 영향을 줄이기 위해 혐오표현을 사전에 차단할 수 있는 예방책이 필요하다.



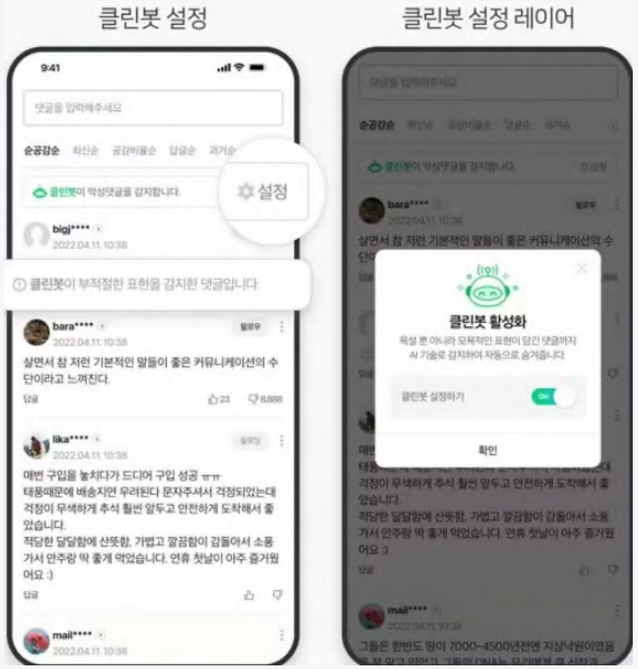


1) (문화체육관광부) 2022년 혐오표현 관련 대국민 인식조사 : https://www.korea.kr/archive/expDocView.do?docId=40270

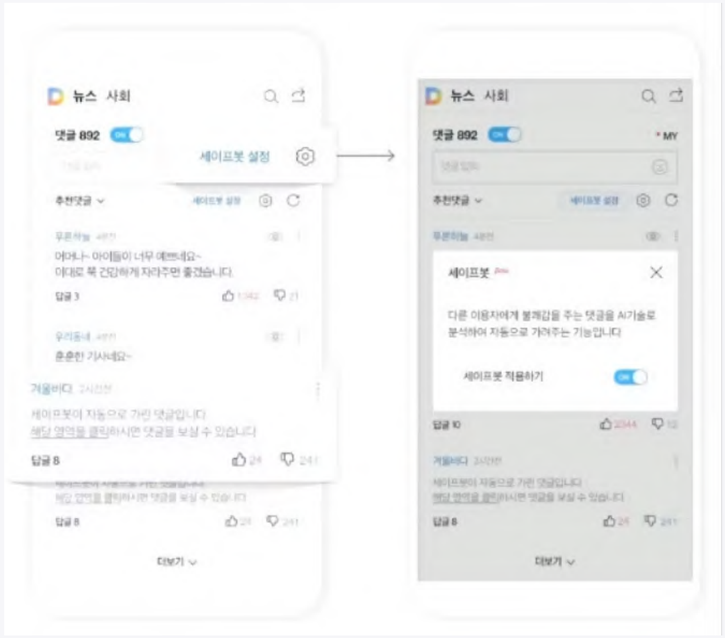
동기

온라인 혐오 표현을 탐지하고 방지하는 기술

- 네이버 클린봇: 악성댓글 탐지 인공지능으로 주석자가 레이블링한 35만 개의 데이터를 사용한다.
- 카카오 세이프봇: AI 기반 댓글 필터링 기능으로 데이터 라벨링에는 Auto Labeling 기법을 사용하고, 일부 데이터는 사람이 직접 검증한다.

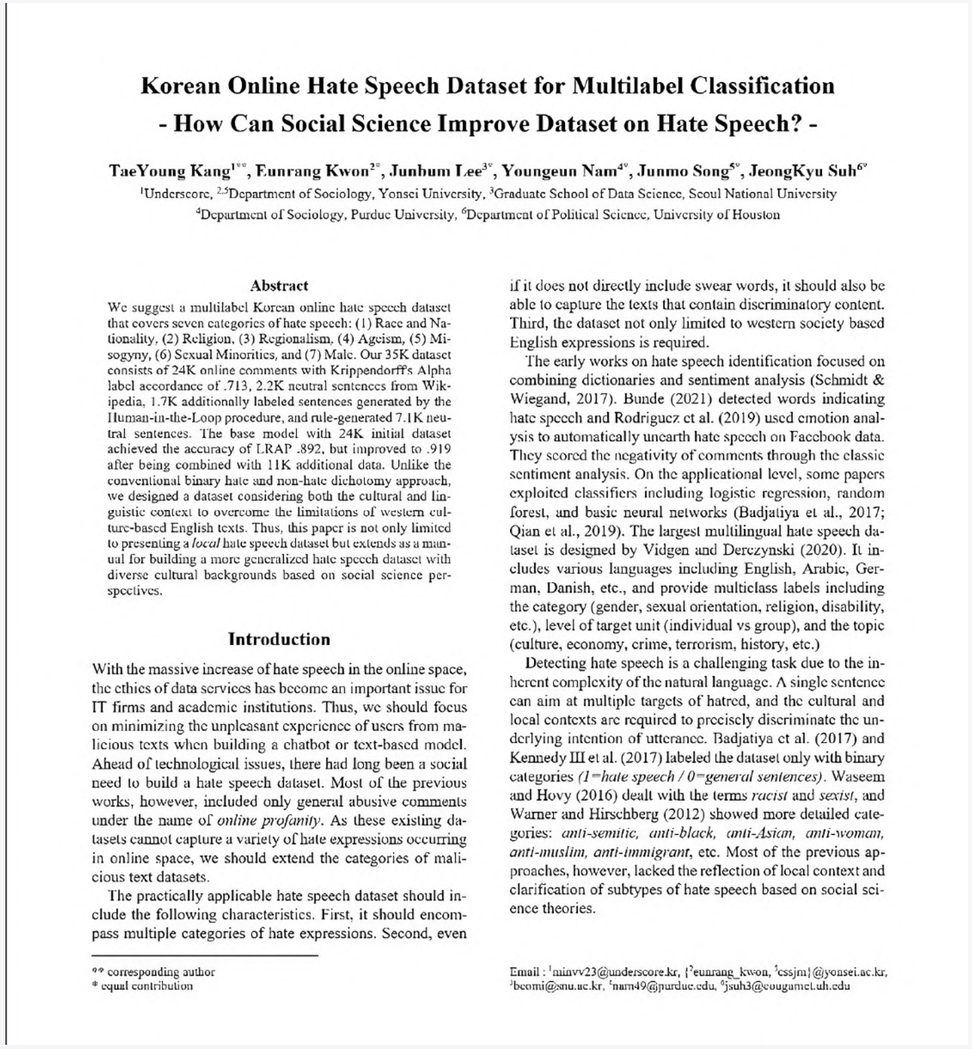


네이버 클린봇



카카오 세이프봇

관련 연구



Korean Online Hate Speech Dataset for Multilabel Classification
- How Can Social Science Improve Dataset on Hate Speech? -

TaeYoung Kang[1*†], Eunrang Kwon[2*], Junbum Lee[3*], Youngeun Nam[4*], Junmo Song[5*], JeongKyu Suh[6†]
[1]Underscore, [2,5]Department of Sociology, Yonsei University, [3]Graduate School of Data Science, Seoul National University
[4]Department of Sociology, Purdue University, [6]Department of Political Science, University of Houston

Abstract

We suggest a multilabel Korean online hate speech dataset that covers seven categories of hate speech: (1) Race and Nationality, (2) Religion, (3) Regionalism, (4) Ageism, (5) Misogyny, (6) Sexual Minorities, and (7) Male. Our 35K dataset consists of 24K online comments with Krippendorff's Alpha label accordance of .713, 2.2K neutral sentences from Wikipedia, 1.7K additionally labeled sentences generated by the Human-in-the-Loop procedure, and rule-generated 7.1K neutral sentences. The base model with 24K initial dataset achieved the accuracy of LRAP .892, but improved to .919 after being combined with 11K additional data. Unlike the conventional binary hate and non-hate dichotomy approach, we designed a dataset considering both the cultural and linguistic context to overcome the limitations of western culture-based English texts. Thus, this paper is not only limited to presenting a *local* hate speech dataset but extends as a manual for building a more generalized hate speech dataset with diverse cultural backgrounds based on social science perspectives.

Introduction

With the massive increase of hate speech in the online space, the ethics of data services has become an important issue for IT firms and academic institutions. Thus, we should focus on minimizing the unpleasant experience of users from malicious texts when building a chatbot or text-based model. Ahead of technological issues, there had long been a social need to build a hate speech dataset. Most of the previous works, however, included only general abusive comments under the name of *online profanity*. As these existing datasets cannot capture a variety of hate expressions occurring in online space, we should extend the categories of malicious text datasets.

The practically applicable hate speech dataset should include the following characteristics. First, it should encompass multiple categories of hate expressions. Second, even if it does not directly include swear words, it should also be able to capture the texts that contain discriminatory content. Third, the dataset not only limited to western society based English expressions is required.

The early works on hate speech identification focused on combining dictionaries and sentiment analysis (Schmidt & Wiegand, 2017). Bunde (2021) detected words indicating hate speech and Rodriguez et al. (2019) used emotion analysis to automatically uncarth hate speech on Facebook data. They scored the negativity of comments through the classic sentiment analysis. On the applicational level, some papers exploited classifiers including logistic regression, random forest, and basic neural networks (Badjatiya et al., 2017; Qian et al., 2019). The largest multilingual hate speech dataset is designed by Vidgen and Derczynski (2020). It includes various languages including English, Arabic, German, Danish, etc., and provide multiclass labels including the category (gender, sexual orientation, religion, disability, etc.), level of target unit (individual vs group), and the topic (culture, economy, crime, terrorism, history, etc.)

Detecting hate speech is a challenging task due to the inherent complexity of the natural language. A single sentence can aim at multiple targets of hatred, and the cultural and local contexts are required to precisely discriminate the underlying intention of utterance. Badjatiya et al. (2017) and Kennedy III et al. (2017) labeled the dataset only with binary categories *(1=hate speech / 0=general sentences)*. Waseem and Hovy (2016) dealt with the terms *racist* and *sexist*, and Warner and Hirschberg (2012) showed more detailed categories: *anti-semitic, anti-black, anti-Asian, anti-woman, anti-muslim, anti-immigrant*, etc. Most of the previous approaches, however, lacked the reflection of local context and clarification of subtypes of hate speech based on social science theories.

†* corresponding author
* equal contribution

Email : [1]minvv23@underscore.kr, [1]*eunrang_kwon, [2]cssjm1@yonsei.ac.kr, [3]beomi@snu.ac.kr, [4]nam49@purdue.edu, [5]jsuh3@cougarnet.uh.edu

HateScore : Human-in-the-Loop and Neutral Korean Multi-label Online Hate Speech Dataset

- (Kang, Tayoung, et al., 2022) 논문에서 활용한 데이터셋 중 보조 데이터셋 1.1만건에 대한 라벨링 기준을 다룸.
- 데이터셋의 크기는 약 1.1만 건으로, base model을 활용해 HITL(Human-in-the-Loop) 방식, 위키피디아에서 수집한 혐오 이슈 관련 중립 문장, 규칙 기반으로 생성된 중립 문장 세 가지로 구성되며 중립 문장 오분류 방지를 주 목적으로 개발되었음.

Model Performance : LRAP (Label Ranking Average Precision)

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

Base Model 기준 비교 예제 (표 안의 값은 혐오발언 분류 확률)

혐오발언 분류 확률	Unsmile	Unsmile+HateScore
저 사람 중국인이네	**0.87**	0.20
너 페미니스트니?	0.03	0.01
동성혼은 논쟁적이지	0.35	0.01
무슬림을 다 죽인다고?*	**0.84**	**0.76**

*두 모델 모두 오분류한 사례

→ 혐오 데이터 필터링 시 혐오 데이터 뿐만이 아니라 문맥에 따라 중립적일수도 있는 문장 등은 해당 데이터셋을 포함하여 학습한 경우가 더 우수했음

Try Pitch https://github.com/sgunderscore/hatescore-korean-hate-speech/tree/main

MULTI LABEL의 필요성

01

쿵쾅이들도 필리핀 그지는 싫지?

여성 + 인종 혐오

02

개신교나 중국인이나 똑같다

종교 + 인종 혐오

03

상폐 한남들 재기해라

세대 + 남성 혐오

기존의 이진분류와 비교해서, 한국어 온라인 커뮤니티의 악성 댓글 문제 해결을 위해 정확한 다중 라벨 분류 모델 필요

HITL 효과

Example	basic model	HITL model
That Guy is Chinese 저 사람 중국인이네	.867 (race)	.196 (race)
Are you feminist? 너 페미니스트니?	.028 (women)	.006 (women)
Same-sex marriage is controversial. 동성혼은 논쟁적이지	.347 (LGBT)	.008 (LGBT)
You're going to kill all Muslims? 무슬림을 다 죽인다고?	.835 (religion)	.761 (religion)

인종, 성소수자 관련 키워드 오류 교정

Example	prob>.5
Hey girls. Go back home and just care your child. 여자는 집에서 애나 봐라	.858 (women)
Korean codgers, just die. 상폐 한남들 다 재기하라고	.551 (ageism) .877 (male)
Wow, a pervert festival in the city center. 도심에서 변태성욕 축제라니	.810 (LGBT)
Don't you fat feminazis also hate Filipino beggars? 쿵쾅이들도 필리핀 그지는 싫지?	.740 (women) .706 (race)
Christians are just like Chinese. 개신교인이나 중국인이나 거기서 거기	.576 (religion) .727 (race)

복잡 혐오 (멀티레이블) 정확도 개선

데이터셋

Smilegate AI에서 공개하는 한국어 혐오표현 "☹ UnSmile" 데이터셋과 Hatescore 데이터셋을 사용을 함께 사용한다.

HateScore

Unsmile Dataset

본 테이터셋에서의 혐오 표현은 “특정 사회적 (소수자) 집단에 대한 적대적 발언, 조롱, 희화화, 편견을 재생산하는 표현”으로 정의

- 혐오의 대상이 속한 집단을 명확히 지칭하는 비하·차별발언
- 대상에 대한 고정관념
- 대상의 특성이나 성향을 특정한 통념에 고착시키는 발언
- 화자 스스로를 자조적으로 표현하는 경우는 혐오 발언이 아님

단일 데이터는 [혐오표현, 악플/욕설, clean]으로 분류될 수 있으며, 혐오 표현은 다중 레이블(multi-label)로 전문가 집단을 통해 레이블링되었다.

항목	문장 수
혐오표현	10,139
악플/욕설	3,929
Clean	4,674
Total	18,742

데이터셋

Smilegate AI에서 공개하는 한국어 혐오표현 "☹ UnSmile" 데이터셋과 Hatescore 데이터셋을 사용을 함께 사용한다.

HateScore

- Korean UnSmile Dataset의 base model을 활용해 HITL(Human-in-the-Loop) 방식으로 태깅된 데이터, 위키피디아에서 수집한 혐오 이슈 관련 중립 문장 데이터, 규칙 기반으로 생성된 중립 문장 데이터로 구성된다. 데이터셋의 크기는 약 1.1만 건이다.
- Human-in-the-Loop 방식 태깅된 방식은 '모델의 분류 확률'과 '연구원 한 명의 의견'의 두 가지 값을 활용하였다.
- 태깅에 사용된 모델은 UnSmilie 데이터로 학습된 모델을 사용하였다.

LABELS : 여성/가족,남성,성소수자,인종/국적,연령,지역,종교,기타혐오,악플/욕설

Metric

$$LRAP(y, \hat{f}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \frac{1}{||y_i||_0} \sum_{j:y_{ij}=1} \frac{|\mathcal{L}_{ij}|}{rank_{ij}}$$

Metric (LRAP(Lable ranking average precision))

- 라벨 지표 (multi-label classification) 문제에서 모델의 예측 결과와 실제 라벨 간의 순위 관계를 평가하는 지표
- 순위 관계를 통해 라벨의 중요도를 판단하고, 이를 반영하여 precision을 계산
- score 값은 0~10이며, score가 높을 수록 성능이 높다고 판단

Binary indicator matrix of the ground truth labels $y \in \{0, 1\}^{n_{samples} \times n_{labels}}$

Score associated with each label $\hat{f} \in \mathbb{R}^{n_{samples} \times n_{labels}}$

실험 목표

1	데이터셋을 이용한 한국 사회 맥락에 맞는 7개 혐오 카테고리에 따른 멀티레이블 분류 태스크 실험
2	HateScore 데이터셋을 통한 HLTL(주석자 개입) + 중립 데이터 확장의 효과 검증
3	추가적인 실험
4	웹사이트 데모

Baseline Model & Hyperparameter Setting

Model

Pretrained Korean Large-Language Models(PLM)

BERT, ELECTRA, RoBERTa

→ 추후에 적용한 방법론과 비교해보기 위해 논문에서 사용한 KcBERT base, large와 다른 모델을 이용

Hyperparameter Tuning

Epoch = 3 ~ 6

Learning Rate = 1e-5 ~ 5e-5

Batch Size = 16, 32

Metrics

- F1-Score
- LRAP(Lable ranking average precision)

대규모 크롤링 대신 augmentation 해보자

Problem Definition: Existing Studies

Bengali toxic comment 연구는

- **이진 분류(Binary)** → 독성 여부
- **다중 라벨 분류(Multi-label)** → 독성 유형 6종
- 두 단계를 결합한 **hierarchical pipeline** 제시

  → 한국 온라인 혐오 표현의 맥락 기반 Multi Label 분류 모델을 설계

	Existing Approach (Binary/Simple)	Our Approach (Context-Aware)
Single Label	toxic / non toxic	hate / offensive/ clean
Multi Label	toxic 이면 6개 라벨	hate 7개 라벨
category	VULGAR, HATE, Religious, Threat, Troll, insult	gender, LGBT, age, region, race, religion, …
증강	데이터 증강 사용 x	데이터 증강 0

Hybrid Lexicon Construction Pipeline

Source A (Manual):
수동사전

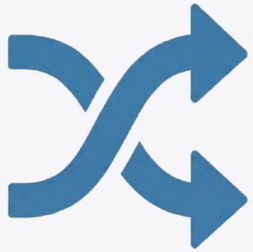
기존 정의된 욕설, 카테고리 정보

예 : 시발, 개새끼 ,한남, 짱깨

Source B (Auto-mined):
Auto Seed JSON

데이터 기반 빈도 추출 → 필터링

예 : 개노답, 좆족, 개슬람



Dedup (중복제거) +
Filtering (숫자 / 일반명사 제거)

Enhanced Lexicon

기존 사전에 없던 변형 욕설 포착

예 : 좇족, 개슬람, 냄져

Robustness-oriented Augmentation Rules

Original
시발

Morphological Variation
(형태론적 변형)

- 문자 반복 (시바아알)
- 자모 분리 (ㅅ ㅣ ㅂ ㅏ ㄹ)

Semantic/Phonetic
(의미/음운 치환)

- 숫자 치환 (시 8)
- 접두/접미 (개시발)

Model & Hyperparameter Setting

Model

Pretrained Korean Large-Language Models(PLM)

BERT, ELECTRA, RoBERTa

Hyperparameter Tuning

Epoch = 3 ~ 6

Learning Rate = 1e-5 ~ 5e-5

Batch Size = 16, 32

Metrics

- F1-Score
- LRAP(Lable ranking average precision)

모델 결과

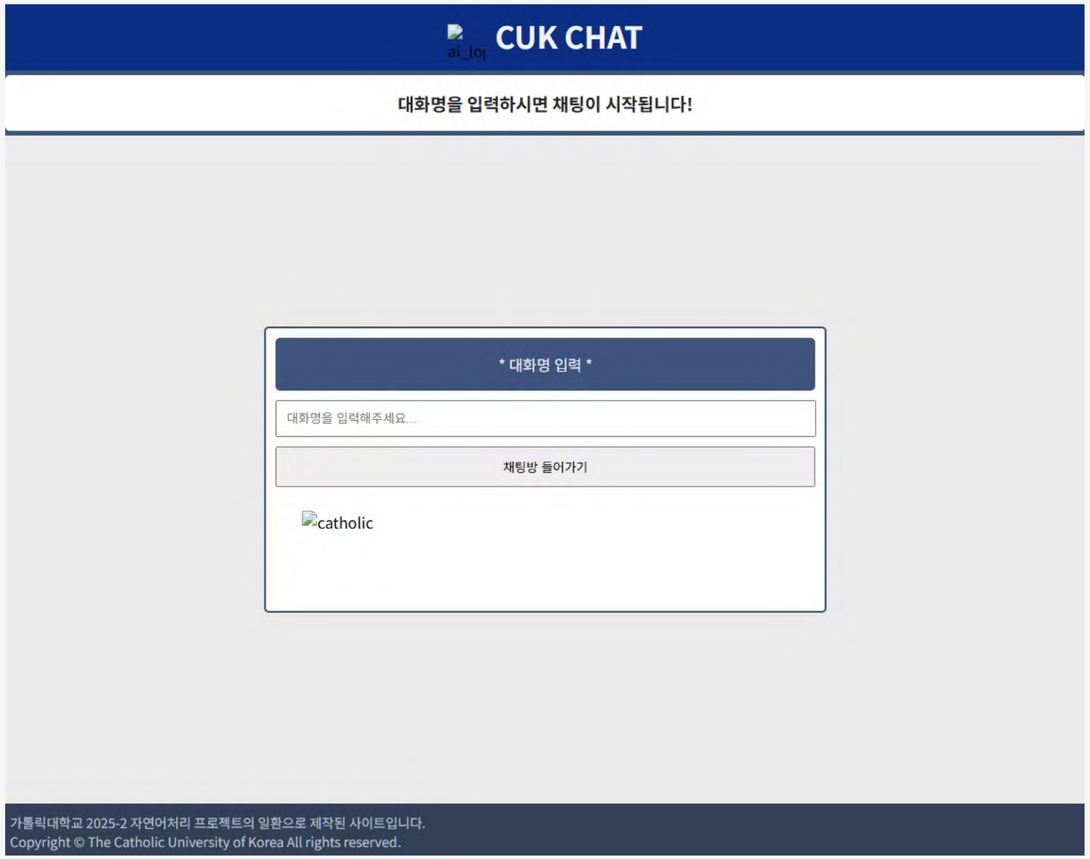
Setting(Baseline)	BERT	ELECTRA	RoBERTa
Unsmile	0.854	0.857	0.857
Unsmile+HateScore	0.854	0.850	0.857

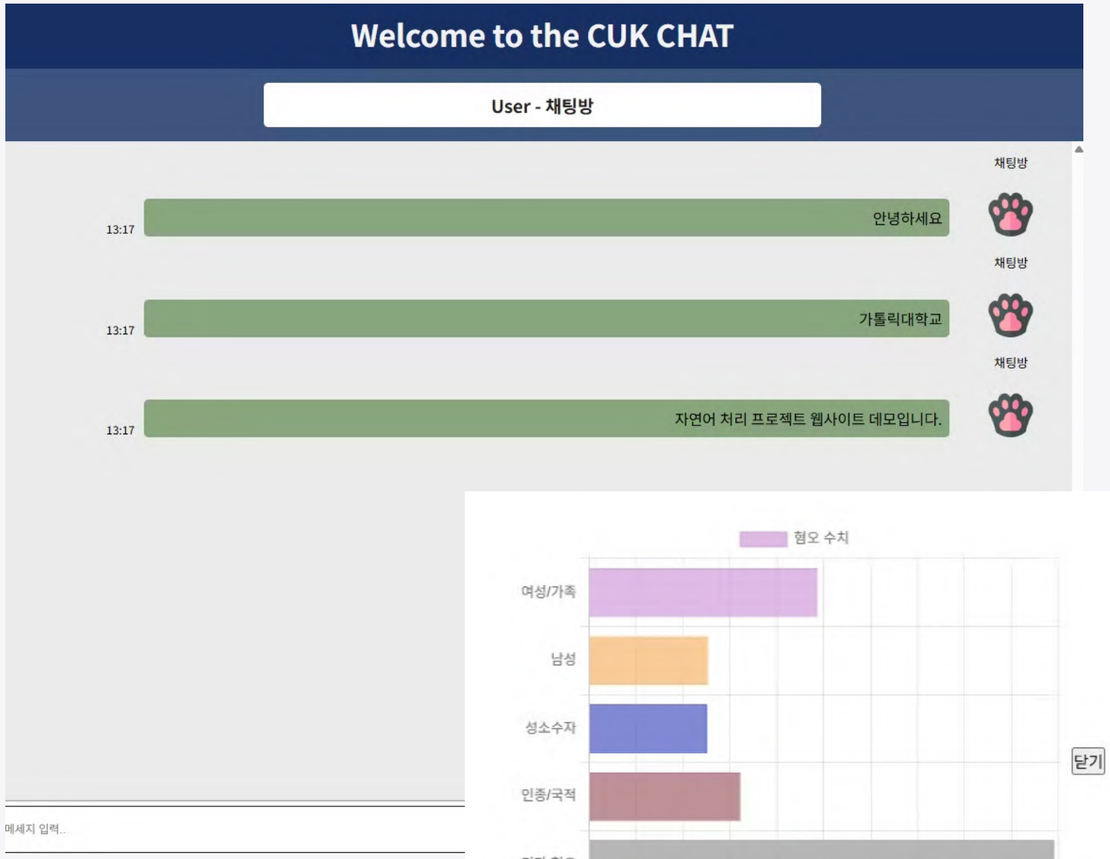
	Coarse macro f1	Coarse micro f1	Coarse Lrap	Fine macro	Fine micro	Fine LRAP
hier	0.7373	0.7092	0.8576	0.7258	0.6552	0.9560
hier(aug o)	**0.7415**	**0.7163**	**0.8577**	**0.7292**	**0.6610**	**0.9569**

Dataset : Unsmile+HateScore

Klue bert BEST 값

웹사이트 데모 시연 - 욕설 채팅 필터링





맥락 기반

계수	의미	값커지면	값작아지면
lamda_fine	final loss 비중	fine 분류 , overfitting ↑	단일 라벨 중심 안정성
lamda_hier	coarse_fine 불일치 패널티	계층 일관성 강해짐	flat과 비슷하게

	Coarse macro f1	Coarse micro f1	Coarse Lrap	Fine macro	Fine micro	Fine LRAP
flat	0.7395	0.7188	0.8584	0.7073	0.6915	0.9603
hier	0.7373	0.7092	0.8576	0.7258	0.6552	0.9560
flat (aug o)	0.7299	0.7094	0.8519	0.7085	0.6976	
hier(aug o)	**0.7415**	**0.7163**	**0.8577**	**0.7292**	**0.6610**	**0.9569**

LABEL :

여성/가족

남성

성소수자

인종/국적

기타혐오(연령, 지역, 종교, 기타혐오)

악플/욕설

→ 데이터셋을 합산하는 과정에서 라벨을 재정의

연구배경

혐오 vs 비혐오

원래 이진 분류

한국어 온라인 커뮤니티의 악성 댓글 문제 해결을 위해 **정확한 다중 라벨 분류 모델** 필요 **새롭게 7개의 label**

Pitch

Want to make a presentation like this one?

Start with a fully customizable template, create a beautiful deck in minutes, then easily share it with anyone.

Create a presentation (It's free)