K-MHAS : A Multi-label Hate Speech Detection Dataset in Korean Online News Comment

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han

The University of Sydney, The University of Western Australia, BigWave AI, Keimyung University, and National Information Society Agency

Introduction

Background

- **Growth of online content**
  - (e.g.) social media, news comments, Wikipedia, and in-game chat.
  - challenges in detecting hate speech.
- **Rise in popularity of Korean TV, movies, and music**
  - (e.g.) Squid Game, Parasite, and BTS.
  - could result in exposure to harmful content and hate speech in Korean.

Problems

(1) **Language:** Extremely limited resources other than English.
(2) **Single Label** classification of particular aspects: commonly used → difficult to explain the subjectivity of hate speech.

Contributions

✓ **A large size Korean multi-label hate speech detection dataset.**
  - representing *Korean language patterns effectively.*

✓ **A multi-label hate speech annotation scheme.**
  - handling the *subjectivity* and the *intersectionality.*

✓ **Strong baseline experiments** on our dataset.
  - using four Korean-BERT-based language models;
  - with six different metrics.

Korean Multi-label Hate Speech Detection Dataset (K-MHaS)

Data Collection

- Unlabelled Korean online news comments (Kaggle and Github).
- Period: Between Jan. 2018 and Jun. 2020.

Multi-label Annotation

(a) **Binary classification**: Hate Speech (HS) or Not Hate Speech;
(b) **Fine-grained classification:** 8 labels (***Politics, Origin, Physical, Age, Gender, Religion, Race, and Profanity***) or Not HS.

- *Non-exclusive concepts*: accounting for the overlapping shades of given categories.
- Selection of 8 labels: in order to *reflect the social and historical context* in Korean (e.g. '*Politics*').
- Annotation: by five native speakers manually (*IAA: 0.892*).

K-MHaS dataset

- Total **109,692 utterances.**
- Providing multi-label classification from 1(one) to 4(four) labels.

Label Types		Count (%)
Total Utterances		**109,692 (100%)**
Multi-label (Hate Speech)	1 label (Single)	36,470 (33.2%)
	2 labels	12,073 (11.0%)
	3 labels	1,440 (1.3%)
	4 labels	94 (0.1%)
Not Hate Speech		59,615 (54.3%)

Table 1: **Dataset Statistics.** The total is the combination of all '*hate speech*' and '*not hate speech*' label. Together the '*hate speech*' label makes up 45.7% of the data.

Publication	Language	Source	Data size	Labels	M-label
Waseem and Hovy (2016)	English	Twitter	16.2k	Sexism, Racism, Neither	N
Davidson et al. (2017)	English	Twitter	24.8k	Hate Speech, Offensive, Neither	N
Wulczyn et al. (2017)	English	Wikipedia comments	115k	Toxic, Severe Toxic, Obscene, Threat, Insult, Identity Hate, Neutral	Y
Ibrohim and Budi (2019)	Indonesian	Twitter	11k	(a) Individual, Group (b) Religion, Race, Pysical, Gender, Other (c) Weak, Moderate, Strong Hate Speech	P
Fortuna et al. (2019)	Portuguese	Twitter	5.6k	(a) Hate Speech, Not Hate Speech (b) Sexism, Body, Origin, Homophobia, Racism, Ideology, Religion, Health, Other-Lifestyle	P
Ousidhoum et al. (2019)	English French Arabic	Twitter	6k (EN) 4k (FR) 3k (AR)	Labels for five different aspects (a) Directness, (b) Hostility, (c) Target, (d) Group, and (e) Annotator	P
Moon et al. (2020)	Korean	News comments	9k	(a) Hate Speech, Offensive, None (b) Gender, Others, None	N
Ours	**Korean**	**News comments**	**109k**	**(a) Hate Speech, Not Hate Speech** **(b) Politics, Origin, Physical, Age, Gender Religion, Race, Profanity, Not Hate Speech**	**Y**

Table 2: **Comparison of datasets.** A "M-label" indicates a multi-label annotation scheme that allows overlapping labels for intersectionality (P = partially applied). The (a) - (e) indicates a layer containing a single label from each aspect.

Dataset Analysis

Label Distribution in Single(-s) and Multi-label(-m)

- **'*Religion*' (5.1%-s, 1.8%-m) and '*Race*' (0.4%-s, 0.6%-m)** classes.
  - the *smallest* portions in both distributions.
  - indicating **cultural aspect** that Korea is a highly homogenous monoculture (단일민족국가).
- **'*Gender*' (9.2%-s, 16.3%-m)** class.
  - at almost *twice* the frequency in a multi-label distribution.
  - indicating **combined aspects** used in gender-based HS.

Keyword Analysis (Lexical Aspects)

- One-word tokens **used in their stem form.**
  - to *modify* the meanings of other words.
  - (e.g.) "denture"[teulni] (틀니) → [teul] (틀) : to the elderly.
- One-word tokens **combined with other neutral words.**
  - to create a *new* offensive term as a prefix or a suffix.
  - (e.g.) "dog"[gae] (개), "insect"[chung] (충).

Class	Count - Single (%)	Count - Multi (%)
Politics	6,931 (19.0%)	4,961 (17.2%)
Origin	5,739 (15.7%)	4,458 (15.5%)
Physical	5,443 (14.9%)	3,364 (11.7%)
Age	4,192 (11.5%)	3,178 (11.0%)
Gender	3,348 (9.2%)	4,696 (16.3%)
Religion	1,862 (5.1%)	513 (1.8%)
Race	160 (0.4%)	163 (0.6%)
Profanity	8,795 (24.1%)	7,509 (26.0%)

Table 3: **Fine-grained label distributions** on hate speech labels. A 'not hate speech' label is not included.

Rank	Politics	Origin	Physical	Age
1	재앙 (1427)	짱깨 (615)	얼굴 (962)	틀 (1918)
2	문재인 (951)	전라도 (596)	돼지 (772)	나이 (599)
3	좌파 (464)	중국 (539)	여자 (294)	노인 (139)
4	좌빨 (402)	쪽 (448)	성형 (216)	충 (112)
5	빨갱이 (367)	짱 (446)	관상 (183)	놈 (106)
Rank	**Gender**	**Religion**	**Race**	**Profanity**
1	여자 (1704)	개독 (526)	흑인 (44)	새끼 (1103)
2	남자 (990)	신천지 (460)	백인 (32)	년 (1014)
3	페미 (172)	사이비 (409)	양키 (32)	지랄 (564)
4	맘충 (138)	종교 (305)	깜둥이 (19)	개 (459)
5	여성 (134)	예수 (227)	놈 (13)	놈 (404)

Table 4: **Top 5 keywords(token count)** associated with each fine-grained label.

Experiments

Setup

- Train/valid/test sets: 72%/8%/20% of samples.
- Baselines: **Multilingual-BERT, KoELECTRA, KoBERT, KR-BERT-c** (character-level) and **KR-BERT-s** (sub-character-level tokenizer).
- Evaluation Metrics: F1-[macro, micro, weighted], Exact Match(E.M), AUC and Hamming Loss(H.L).

Evaluation for All Labels

- **KoELECTRA**: overall the best or second best among six metrics.
  - its corpus: **modern slang and buzzwords**.
  - indicating the effects of the pre-training data source.

Model	F1 (macro)	F1 (micro)	F1 (weighted)	E.M.	AUC	H.L. (↓)
BERT	0.6912	0.8139	0.8119	0.7579	0.8878	0.0464
KoELECTRA	0.7245	0.8493	**0.8480**	**0.7994**	**0.9122**	0.0380
KoBERT	**0.7651**	0.8413	0.8424	0.7926	0.9083	0.0401
KR-BERT-c	0.7444	**0.8500**	0.8470	0.7901	0.9028	**0.0368**
KR-BERT-s	0.7245	0.8445	0.8437	0.7825	0.9076	0.0390

Table 5: **Overall multi-label classification performance on K-MHaS** at epoch 4.

Evaluation for Multi-labels

- **KR-BERT-c** (using a character tokenizer): the best for a single label.
- **KR-BERT-s** (using a sub-character): overall the best for multi-labels.
  - decomposing Korean syllables into sub-characters.
  - providing greater granularity in detecting HS.
  - handling the bottom consonant (받침) or initial consonant (초성).
  - (e.g.) 개빠ㄹ갱이년 = 개("dog"–'*Profanity*') + 빠ㄹ갱이 ("communist"–'*Politics*') + 년 ("bitch"–'*Gender*').
  - (e.g.) "gold-digger" [kko#t#baem] 꼬#ㅊ#뱀.

# Labels	Model	F1 (Macro)	F1 (Micro)	F1 (Weighted)	E.M.	AUC	H.L. (↓)
1	BERT	0.6666	0.8190	0.8202	0.7919	0.9011	0.0406
	KoELECTRA	0.6953	0.8490	0.8508	**0.8263**	**0.9213**	0.0341
	KoBERT	0.7321	0.8320	0.8370	0.8142	0.9110	0.0379
	KR-BERT(w. char)	**0.7336**	**0.8553**	**0.8543**	0.8239	0.9145	**0.0318**
	KR-BERT(w. sub)	0.6985	0.8392	0.8419	0.8062	0.9123	0.0360
2	BERT	0.6389	0.8043	0.8174	0.5580	0.8524	0.0788
	KoELECTRA	0.6777	0.8612	0.8700	0.6511	0.8934	0.0577
	KoBERT	**0.7249**	**0.8854**	**0.8911**	**0.6794**	**0.9112**	**0.0482**
	KR-BERT(w. char)	0.6748	0.8405	0.8451	0.5912	0.8735	0.0642
	KR-BERT(w. sub)	0.6718	0.8703	0.8723	0.6535	0.9000	0.0542
3	BERT	0.5784	0.7517	0.7522	0.2448	0.8040	0.1402
	KoELECTRA	0.6146	0.7987	0.7953	0.3310	0.8362	0.1169
	KoBERT	**0.6523**	0.8290	0.8251	**0.3759**	0.8589	0.1019
	KR-BERT(w. char)	0.5828	0.7827	0.7732	0.2828	0.8239	0.1230
	KR-BERT(w. sub)	0.6164	**0.8329**	**0.8263**	0.3586	**0.8615**	**0.0996**
4	BERT	0.4776	0.7093	0.7029	**0.1200**	0.7610	0.2222
	KoELECTRA	0.4511	0.7044	0.6639	0.0000	0.7680	0.2089
	KoBERT	0.4177	0.6832	0.6460	0.0400	0.7510	0.2267
	KR-BERT(w. char)	0.4837	0.7439	0.7226	**0.1200**	0.7930	0.1867
	KR-BERT(w. sub)	**0.5068**	**0.7771**	**0.7618**	**0.1200**	**0.8120**	**0.1733**

Table 6: **A breakdown of multi-label classification performance** from 1 to 4 labels on K-MHaS at epoch4.

The research is conducted by the AdNLP (Australia Deep learning NLP) Team