🥶 KOLD: Korean Offensive Language Dataset

Younghoon Jeong[†]   **Juhyun Oh**   **Jongwon Lee**
Jaimeen Ahn[†]   **Jihyung Moon**[‡]   **Sungjoon Park**[‡]   **Alice Oh**[†]

[†]School of Computing, KAIST

[‡]SoftlyAI Research, SoftlyAI

hoon2j@gmail.com, 411juhyun@snu.ac.kr, mybizzer@gmail.com,
jaimeen01@kaist.ac.kr, {jihyung.moon, sungjoon.park}@softly.ai,
alice.oh@kaist.edu

Abstract

Warning: this paper contains content that may be offensive or upsetting.

Recent directions for offensive language detection are *hierarchical modeling*, identifying the type and the target of offensive language, and *interpretability* with offensive span annotation and prediction.

These improvements are focused on English and do not transfer well to other languages because of cultural and linguistic differences.

In this paper, we present the Korean Offensive Language Dataset (KOLD) comprising 40,429 comments, which are annotated hierarchically with the type and the target of offensive language, accompanied by annotations of the corresponding text spans. We collect the comments from NAVER news and YouTube platform and provide the titles of the articles and videos as the context information for the annotation process.

We use these annotated comments as training data for Korean BERT and RoBERTa models and find that they are effective at offensiveness detection, target classification, and target span detection while having room for improvement for target group classification and offensive span detection. We discover that the target group distribution differs drastically from the existing English datasets, and observe that providing the context information improves the model performance in offensiveness detection (+0.3), target classification (+1.5), and target group classification (+13.1). We publicly release the dataset and baseline models.

1 Introduction

Online offensive language is a growing societal problem. It propagates negative stereotypes about the targeted social groups, causing representational harm (Barocas et al., 2017). Among various research directions for offensive language detection,
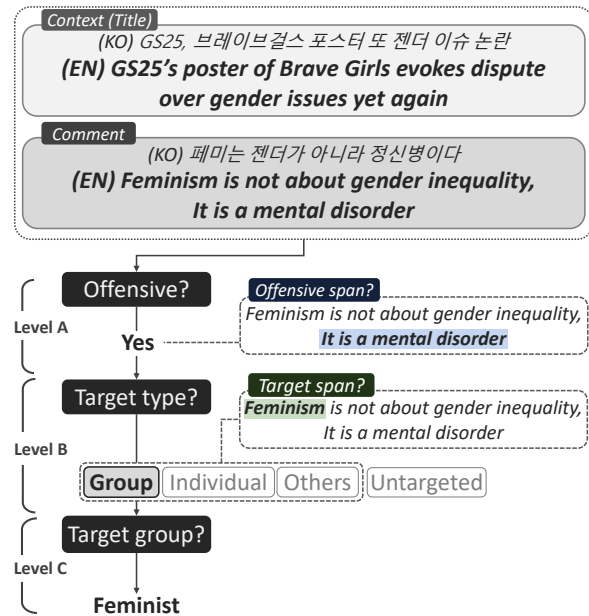

Figure 1: An illustration of the annotation process of KOLD. The title is given to provide context to the annotators. Along with the categorical labels, annotators are asked to mark spans of a sentence that justifies their decision.

Waseem et al. (2017) proposes a taxonomy to distinguish hate speech and cyberbullying based on whether the offensive language is directed toward a specific individual or entity, or toward a generalized group. Zampieri et al. (2019) integrates this taxonomy into a hierarchical annotation process to create the Offensive Language Identification Dataset (OLID), which is adopted in other languages as well (Zeinert et al., 2021; Zampieri et al., 2020).

Many problems remain in offensive language detection such as failure to generalize (Gröndahl et al., 2018; Karan and Šnajder, 2018), over-sensitivity to commonly-attacked identities (Dixon et al., 2018; Kennedy et al., 2020), and propagating bias in annotation (Sap et al., 2019; Davidson et al., 2019). Among various attempts to solve those problems, first is offensive language target detection to iden-

Context (title)	Comment	Label			
				Level C	
		Level A	**Level B**	**Target Group Attribute**	**Target Group**
심화하는 젠더갈등...SNS 때문? Gender in Conflict, deepening by SNS?	사람은 사람일뿐 People are all just people	NOT			
한국 106개 사회단체, '아프가니스탄 난민 보호책' 수립 촉구 'Protection of Afghan Refugees' by 106 Korean civil societies	정신좀 차리자 Snap out of it	OFF	UNT		
북한, 아프간 사태로 자신감 얻었나... 미국에 '인권 문제' 역공 N. Korea calls US-led Afghan war 'Human Rights Crime'	누가 김정은 인권 좀 말살 안하나? Anyone for the extermination of Kim Jong-un?	OFF	IND		
틱톡 감성 거부하다 못해 패버리는 흑인 형님 모음 Angry TikTok reactions by Black bro	틱톡 이라는 이름의 쓰레기는 제거되야되 The garbage named TikTok should be eliminated	OFF	OTH		
"이 법은 죽을 뻔한 누구를 살리기도 해요" "This law could save desperate lives who almost died"	수술받은 트랜스젠더들 을 봐도 분노가 솟아 칼을 꺼내 죽이겠다! I get mad when I see transgender people who had surgery, I'll take a knife and kill them all!	OFF	GRP	Gender & Sexual Orientation	LGBTQ+
치명적인 유혹 - 조선족이랑 결혼한 후기 Deadly Temptation - A marriage with Korean-Chinese	조선족 들 중국으로 다돌아갔으면 속이 시원하겠다 😎 It would be so refreshing if all the Korean-Chinese went back to China 😎	OFF	GRP	Race, Ethnicity & Nationality	Korean-Chinese
한국의 "이슬람교" "Islam" in Korea	뭘 믿든상관 없는데 사람을 죽이니까 무서운거지 I don't care what (they) believe, what matters is the fact that (they) kill people	OFF	GRP	Race, Ethnicity & Nationality	Muslim
GS25, 브레이브걸스 포스터 또 젠더 이슈 논란 GS25's poster of Brave Girls evokes dispute over gender issues yet again	페미 는 젠더가 아니라 정신병이다 Feminism is not about gender inequality, it is a mental disorder	OFF	GRP	Misc.	Feminists

Table 1: Examples of the comments in KOLD, along with the annotation results. The *Title* is either the headline of the news article or the title of the video where the comment is posted on. The subject in the parentheses is omitted in the original sentence in Korean. (OFF: offensive, NOT: not offensive, UNT: untargeted, IND: individual, OTH: other, GRP: group, blue : offensive span, green : target span)

tify the individuals or groups who are the victims of offensive language (Sap et al., 2020; Mathew et al., 2021; Shvets et al., 2021). Second is offensive language span detection to provide some explanation for offensive language (Pavlopoulos et al., 2022; Mathew et al., 2021). Third is to provide context information of offensive language to improve offensive language detection (Vidgen et al., 2021; de Gibert et al., 2018; Gao and Huang, 2017).

However, these steps forward are limited to English because we lack comprehensive datasets in other languages (Mubarak et al., 2017; Fortuna et al., 2019; Chiril et al., 2020; Rizwan et al., 2020). Language-specific datasets are essential for offensive language which is very culture- and language-dependent (Reichelmann et al., 2020; Albadi et al., 2018; Ousidhoum et al., 2019). Although Korean is a comparably high-resourced language (Joshi et al., 2020), there are only a few publicly available Korean offensive language corpora (Moon et al., 2020), and they do not consider the type of the target (e.g., group, individual) to differentiate hate speech from cyberbullying, include context information such as titles of articles, nor annotate text spans for explainability.

In this paper, we describe and publicly release the Korean Offensive Language Dataset (KOLD), 40,429 comments collected from news articles and videos.[1] The unique characteristics of our dataset are as follows:

- It is the first Korean dataset with a hierarchical taxonomy of offensive language (see Figure 1). If the comment is group-targeted offensive language, we additionally annotate among the 21 target group labels tailored to Korean culture.

- The specific spans of text that are offensive or that reveal targeted communities are annotated (see Table 1 for examples). KOLD is the first publicly released dataset to provide both types of spans for offensive language in Korean.

- The comments in our dataset are annotated with the original context. We provide the titles of the articles and videos during the annotation process, which resembles the realistic setting of actual usage.

---

[1] KOLD: Korean Offensive Language Dataset is available at http://github.com/boychaboy/KOLD.

2 Annotation Task Design

We use a hierarchical annotation framework based on the multi-layer annotation schema in OLID (Zampieri et al., 2019). Additionally, we identify the specific target group of the offensive language. We also annotate the spans that support the labeling decision if the comment is offensive and/or contains a target of offensiveness. Figure 1 illustrates an overview of our annotation task, and Table 1 shows examples.

2.1 Level A: Offensive Language Detection

At level A, we determine whether the comment is offensive (OFF) or not (NOT), and which part of the comment makes it offensive (offensive span). We consider a comment offensive if it contains any form of untargeted profanity or targeted offense such as insults and threats, which can be implicit or explicit (Zampieri et al., 2019).

We define offensive span as a specific segment of text that justifies why a comment is offensive, also known as a rationale (Zaidan et al., 2007). In parallel with the definition of offensiveness, the span includes not only explicit profanity but also implicit offensive language (e.g., sarcasm or metaphor (ElSherief et al., 2021)). If the offensiveness is conveyed across multiple sentences in the comment, all of them are captured as the offensive span. Taking into account the *faithfulness* of a rationale (DeYoung et al., 2020), the offensive span is the minimal snippet of the text (i.e., sufficient) that includes all forms of expressions that convey even the slightest intensity of offense (i.e., comprehensive), such as affixes and emojis.

2.2 Level B: Target Type Categorization of Offensive Language

Level B categorizes the type of the target and highlights the supporting span of the target (target span). There are four possible categories.

- **Untargeted (UNT):** An offensive comment that does not contain a specific target.

- **Individual (IND):** An offensive comment that is targeted at a specific individual. This includes a famous person or a named/unnamed individual with specific reference in the text. Comments targeted at an individual are categorized as cyberbullying (Chen et al., 2012).

- **Group (GRP):** An offensive comment targeted at a group of people with shared pro-

tected characteristics, such as gender or religion. Offensive language in this category is generally considered as hate speech (Zhang and Luo, 2019).

- **Others (OTH):** An offensive comment whose target does not belong to the above two categories. Targeting an organization, a company, or an event.

We define a target span as a span of characters in the comment that indicates the target of the offensive language. It is collected for all types of targeted offensive speech, regardless of the target type (IND, GRP, OTH). If the term used to indicate the target is offensive, target span can overlap with the offensive span (e.g., *jjangkkae*, which corresponds to *ching-chong* in English).

2.3 Level C: Target Group Identification of Group Targeted Offensive Language

Level C identifies the specific targets of offensive language, which consists of two hierarchical levels: target group attribute and target group. The target group represents the specific social or demographic groups that share the same identity (e.g., Women, Muslim, Chinese), and the target group attribute is a superclass for the target group. We allow multi-group annotation if the target entity of the comment belongs to more than one group. For instance, *"페미녀 (feminist bitch)"*, a word that disparages a feminist woman, targets two groups: *Women* and *Feminist*. Table 10 in the Appendix contains the full set of 21 target groups. To determine the set of target groups, we begin with categorizing targets in Sap et al. (2020) and add several categories to better reflect the Korean language and culture. As the result of analyzing the targets in 1,000 initial samples, we add *Chinese*, *Korean-Chinese*, and *Indian* to the *Race, Ethnicity & Nationality* attribute, as they take up larger portions than the initial target groups (*White*, *Asian*). Group characteristics that do not belong to the four target group attributes (e.g., *Disabled*, *Feminist*) are grouped under *Miscellaneous*. Note that we are aware that feminism is a gender-related issue, but classified *Feminist* into *Miscellaneous* because feminists embody a group of people that share the same ideology rather than being a subclass of gender. We show the distribution of the top two levels (A, B) and the target group attributes (Level C) in Table 2 and the subsequent target group categories in Table 3.

3 Collecting Annotations

3.1 Source Corpora Collection

We choose two social media platforms, NAVER and YouTube, as our source of data, which are two of the top three mobile apps used in Korea in 2021.[2] In particular, we collect titles and comments on NAVER news articles and YouTube videos distributed from March 2020 to March 2022.

Due to the scarcity of offensive comments, we collect articles and posts by using predefined keywords, which is a commonly used method in hate speech dataset construction (Waseem and Hovy, 2016). Every keyword is potentially highly correlated with articles or videos that may have abusive comments. Keywords are listed in Appendix A.

To ensure we do not reveal users' personally identifiable information, we do not collect user ids. We replace mentions of a username with <user> tokens, URLs with <url> tokens, and emails with <email> tokens to conceal private information.

3.2 Annotation Procedure

The steps we took for high-quality of annotations include providing a detailed guideline, selecting the annotators deliberately, and managing the annotation process carefully. In the guideline, we resolve predictable difficulties during the process. For example, we provide rules for delimiting morphological boundaries specific to Korean to collect consistent text spans, and provide guidance of implicit hate speech based on the taxonomy proposed in ElSherief et al. (2021). To ensure overall annotation quality, we only allow annotators who pass a qualification test to participate in the main annotation process. We make it clear that annotators refer to the title of the article or video as context to reduce ambiguity. As hate speech annotation can also be influenced by the bias of the annotators (Davidson et al., 2019; Sap et al., 2022; Al Kuwatly et al., 2020), we include annotators of diverse demographic backgrounds by limiting the maximum amount of annotation per worker to 1% of the data. A total of 3,124 annotators participate in creating the final dataset. To ensure each label is genuine, we embed questions with clear known answers and remove users who answer those wrong. We use SelectStar, a crowdsourcing platform in Korea, to collect annotations. The full task is shown in the appendix (Figure 2).

Level A	Level B	Level C	Count (%)
NOT	-	-	20,119 (49.8)
OFF	UNT	-	2,596 (6.4)
	IND	-	3,899 (9.6)
	OTH	-	1,402 (3.5)
	GRP	Gender & Sexual Orientation	2,986 (7.4)
		Race, Ethnicity & Nationality	2,970 (7.3)
		Political Affiliation	1,939 (4.8)
		Religion	1,699 (4.2)
		Miscellaneous	2,819 (7.0)
Total			40,429 (100)

Table 2: Statistics of labels in KOLD. Here, the lowest level of data granularity is the target group attribute in Level C.

To decide on the gold label, we apply majority voting among the three annotations for the categorical labels, and take character offsets that more than two out of three annotators highlighted for text spans. When the gold label cannot be determined by majority voting as there are more than two choices, inspectors[3] resolve the disagreement to determine the gold label.

3.3 Annotation Result

Overall, the average Krippendorff's α for inter-annotator agreement of each annotation level is 0.55. The label distribution of the collected data is shown in Table 2.

Level A: Offensive Language Detection The dataset contains 40,429 comments, of which 20,130 comments are classified as offensive language. Offensive comments targeted at group characteristics (also known as hate speech) comprise 30.7% of the whole data. Specifically, Krippendorff's α for the inter-annotator agreement is 0.55 for classifying offensiveness.

Level B: Target Type of Categorization of Offensive Language Among 20,130 offensive comments, 2,596 comments are classified as untargeted offense (UNT). These comments include comments with non-targeted profanity and swearing. Group (GRP) is the most common type of target, taking up 70.1% of the three types of targeted offenses, followed by individual (IND) (22.0%) and others

---

Target Group Attribute	Target Group	Count
Gender & Sexual Orientation (24.63%)	LGBTQ+	1,369
	Women	1,129
	Men	591
	Others	108
Race, Ethnicity & Nationality (23.93%)	Chinese	475
	Korean-Chinese	460
	Black	175
	Others	1,605
Political Affiliation (15.55%)	Progressive	1,052
	Conservative	316
	Others	651
Religion (13.42%)	Muslim	1,059
	Christian	518
	Catholic	57
	Others	91
Miscellaneous (20.32%)	Feminist	1,483
	Socio-economic Status	88
	Agism	62
	Others	971
Total		12,981

Table 3: Breakdown of target group attributes of group-targeted offensive comments (Level C). We present the three most frequent target groups for each target group attribute. Multi-targeted groups are split into single groups when counting.

(OTH) (7.9%). Krippendorff's α for agreement on deciding the type of target is 0.45 .

Level C: Target Group Identification of Group Targeted Offensive Language Among 12,413 group-targeted offensive comments, the most common attribute is *Gender & Sexual orientation* taking up 7.4 of the whole dataset. *Political Affiliation* and *Religion* both appear in less than 5% of the data. Table 3 shows a breakdown of the target group attributes within the top three most frequent as well as the group *Others*, which is tagged at the target group attribute level but does not belong to the target group choices we provide. In *Race, Ethnicity & Nationality*, *Others* take up the largest portion with 1,605 comments, since it includes small but various origins ranging from Afghans and Americans to North Koreans and North Korean defectors. The three most frequently targeted group characteristics in the whole dataset are *Feminist*, *LGBTQ+*, and *Women*, which amount to 11.42%, 10.55%, and 8.7% of the group-targeted offensive language, respectively. Krippendorff's α is 0.65 for specifying the target group of the offensive language.

4 Dataset Analysis

4.1 Target Group Distribution

Our novel finding is that target groups are defined based on the specific language and culture to embrace ongoing social phenomena and reflect them to the dataset. Shown in Table 4, the distribution of the target groups in KOLD largely differs from the English HateXplain dataset (Mathew et al., 2021). We observe that groups such as *Jewish, Arabs* and *Hispanic* which commonly appear in English datasets (e.g., Sap et al. (2020); Ousidhoum et al. (2019)), do not frequently appear in KOLD. *Africans*, the target group that appears most frequently in HateXplain, is not included in the top ten ranked groups in our dataset, and the reverse is true for *Feminist*, the first-ranked target group in our dataset. While *Women, LGBTQ+* (which includes Homosexual) and *Men* are common target groups in both datasets, other identity groups such as *Chinese, Korean-Chinese, Progressive* and *Conservative* only appear in our dataset.

Specifically, we observe that within the Korean language, the Asian race as a target of offensive language should be more finely partitioned. While *Asians* appear as a frequent target in English datasets without race or ethnic division (An et al., 2021; Ousidhoum et al., 2019; Hartvigsen et al., 2022), in KOLD, it is further separated into fine-grained targets grouped by nationality or ethnicity (e.g., *Chinese, Indian, Southeast Asian*). Moreover, our dataset demonstrates that a single Asian race should be divided into separate target groups. A large portion of *Chinese* and *Korean-Chinese* targeted offensive comments in KOLD highlights the uniqueness of offensive language in Korean and reflects the cultural differences between Korean speakers and English speakers. This demonstrates the prevalence of social bias among Asians even though they share similar cultural values and phenotype (Lee et al., 2017).

4.2 The Role of Title for Target Group Identification

In KOLD, 55.1% of group-identified offensive comments have no target span marked in the comment, which implies that in the majority of the cases, titles contain information about the targets. For

---

[4]We calculate the statistics of the HateXplain dataset ourselves as Mathew et al. (2021) do not provide the specific distribution of the target groups. We changed the target group *Islam* to *Muslim*, to match with our term.

Rank	Ours		HateXplain	
	Target group	*%*	**Target group**	*%*
1	Feminist	11.42	African	21.58
2	LGBTQ+	10.55	Jewish	13.04
3	Women	8.7	Muslim	12.4
4	Muslim	8.16	Homosexual	10.22
5	Progressive	8.1	Women	9.5
6	Men	4.55	Arab	6.17
7	Christian	3.99	Refugee	3.46
8	Chinese	3.66	Caucasian	3.11
9	Korean-Chinese	3.54	Hispanic	2.74
10	Conservative	2.43	Men	2.13

Table 4: Comparison of target group distribution with existing dataset in English (HateXplain). We list the top 10 target groups of each dataset based on the frequency.[4] Groups in brown only appear in KOLD, while groups in mint only appear in HateXplain.

example, given the title of the article *"'Islam' in Korea / Yonhap News"*, it is easy to find a comment without explicitly mentioning the target such as *"I don't care what (they) believe, what matters is the fact that (they) kill people"* (penultimate row of Table 1).

5 Experiments and Results

We experiment using three different model architectures: (1) sequence classification model for predicting offensiveness, target type, and target group categories, (2) token classification model for predicting the offensive and target span, and (3) multi-task model for predicting both category and span at once. We report the score of single-task models using various sizes of Korean BERT and RoBERTa (Park et al., 2021), and compare the result against the multi-task model.

We further conduct an ablation study by excluding the title from the input to discover how much it impacts the prediction performance of the model. We also compare the results of our model with translated versions of English data, and a multilingual span prediction model.

For the experiments, we use an 80-10-10 split for each task, and report the best performances based on the F_1 score of the test set result with the tuned hyperparameters. Training details are reported in the Appendix B.

5.1 Category Prediction

For each level of annotation, we fine-tune the pretrained models to predict the label given the ti-

Model		Metric		
		P	**R**	**F1**
Level A: Offensive	BERT$_{base}$	75.8	78.8	77.2
	RoBERTa$_{base}$	76.8	77.6	77.2
	RoBERTa$_{large}$	76.6	81.1	**78.8**
Level B: Target Type	BERT$_{base}$	58.5	58.5	57.6
	RoBERTa$_{base}$	60.8	58.7	59.4
	RoBERTa$_{large}$	63.9	62.4	**62.7**
Level C: Target Group	BERT$_{base}$	56.9	57.4	55.9
	RoBERTa$_{base}$	58.6	58.5	57.3
	RoBERTa$_{large}$	59.1	59.6	**57.5**

Table 5: Evaluation results of category prediction models of each task. Binary-F1 is used to evaluate *offensive* task and macro-F1 is used for the other tasks whose number of labels is more than two. **Bold** indicates the best performance across the models.

Model		Metric		
		P	**R**	**F1**
Offensive Span	BERT$_{base}$	47.7	43.4	41.9
	RoBERTa$_{base}$	46.4	41.8	40.6
	RoBERTa$_{large}$	50.8	47.8	**45.4**
Target Span	BERT$_{base}$	60.2	67.3	61.9
	RoBERTa$_{base}$	60.6	68.2	62.4
	RoBERTa$_{large}$	60.8	67.8	**62.5**

Table 6: Evaluation results of offensive/target span prediction models. **Bold** indicates the best performance across the models.

tle and comment, and then evaluate the model using precision, recall, and F_1 scores of the positive class.

In Table 5, we observe that the more the categories are fine-grained, the task becomes more difficult, and the larger model shows better performance.

5.2 Span Prediction

We convert each span in the comment to BIO-tags to formulate the span prediction task as a token classification task and fine-tune the pre-trained models to predict BIO-tags assigned to each token. To evaluate the model, we follow the work of Da San Martino et al. (2019) by computing the F_1 score of the predicted character offsets with the ground truth. If the ground truth is empty, a perfect score ($F_1 = 1$) is assigned whereas if the predicted set of offsets is empty, a score of zero ($F_1 = 0$) is assigned.

As demonstrated in Table 6, the best character-level F1 score is 45.4 for offensive span and 62.5 for target span. The pattern of higher score with a

System	Offensive			Offensive Span			Target Type			Target Span		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Seq	75.8	78.7	**77.2**	-	-	-	58.5	58.5	**57.6**	-	-	-
Span	-	-	-	47.7	43.4	41.9	-	-	-	60.2	67.3	61.9
Seq + Span	75.0	78.1	76.5	56.3	55.4	**51.7**	56.8	57.0	56.5	72.3	71.9	**71.6**

Table 7: Evaluation results of the multi-task baseline models compared against the single-task baseline models. "Seq" refers to the sequence classification model, "Span" refers to the span prediction model, and "Seq+Span" refers to the multi-task model. All models have fine-tuned the same pre-trained language model, BERT-base.

larger model is consistent with the results of the category prediction.

5.3 Category and Span Prediction

We employ a multi-task learning approach to train a model capable of classifying the category and predicting the span at the same time. In the multi-task model, the sequence classifier and the token classifier share the neural representation of the pre-trained model and only differ in the output layers for each task. The representation of the first token ([CLS]) is fed into an output layer for sequence classification, and the other representations are fed into the layer for token classification. The model jointly learns the global information of a given input sequence and span information. We train two types of multi-task models. First, we train a multi-task model with the binary label of offensiveness and the corresponding offensive span. Second, using the data labeled as offensive, we train a multi-task model to predict the target type (Level B) and the corresponding target span.

As shown in Table 7, multi-task models outperform single-task models in span prediction by 10% in the F1 score, mainly due to joint learning of both types of information. However, the performance of sequence classification drops in both models.

Examples of model predictions and ground truths are illustrated in Table 11 and Table 12 in the Appendix.

5.4 Title Ablation on Classification Tasks

To find out how much the context information (titles of the articles and the videos) contributes to the offensiveness and target classifications, we conduct an ablation study by excluding the titles from the input.

As can be seen in Table 8, if only the comments are given, accuracy and the f1 scores drop significantly compared to the setting where titles and comments are given together. This phenomenon becomes more significant as the granularity of the

Level	Comment	Comment+Title	Δ F1
A	76.9	77.2	+ 0.3 (0.4%)
B	56.1	57.6	+ 1.5 (2.6%)
C	42.8	55.9	+ 13.1 (30.6%)

Table 8: Category prediction results on the offensiveness (Level A), target type (Level B), and target group categories (Level C) with and without the title. Binary-F1 is used to evaluate the offensive task and macro-F1 is used for the other tasks. ΔF1 refers to the absolute and relative performance gap. For all models, we use the same BERT-base model.

	System	F1
Offensive	$KOLD_{Seq}$	**76.9**
	$OLID_{EN \rightarrow KO}$	63.2
Offensive Span	$KOLD_{Span}$	**40.6**
	MUDES	12.8

Table 9: $OLID_{EN \rightarrow KO}$ is an offensiveness classification model trained with the English OLID dataset translated into Korean. MUDES is a multilingual offensive span prediction model. We compare the results against our baseline performance.

label increases. When predicting the fine-grained target group, the f1 score dropped by more than 30%. We conclude that providing the context with the comments helps the model predict the target groups more precisely, as the comments may not contain sufficient information.

5.5 Translated Data and Multilingual Models

To see how much translation and multilingual model are effective at distinguishing offensiveness, we compare our baselines against (1) sequence classification model trained on translated dataset, and (2) multilingual offensive span detection model. For the translation experiment, we translate the OLID to Korean via google translate api[5] and use the dataset for training the same sequence classifier described in Section 5.1. For the multilin-

[5] Requested in April 2022.

gual experiment, we adopt the multilingual token classification model (MUDES) (Ranasinghe and Zampieri, 2021) trained on English toxic span dataset (Pavlopoulos et al., 2021). For all tasks, evaluation is done with the KOLD test set. We report the results in Table 9.

Overall, both translation and multilingual approaches are not more effective than our baselines. For the offensive category prediction, our model is 13.7 higher, and for the span prediction, our model is 27.8 higher. Although MUDES scores high on English (61.6), the performance drops significantly in Korean (12.8).

6 Related Work

6.1 Offensiveness & Hate Speech Detection

Most datasets created for the detection of offensive language have dealt with the subtypes of offensive language such as hate speech, cyberbullying, and profanity as a flat multi-level classification task (Waseem and Hovy, 2016; Davidson et al., 2017; Wiegand et al., 2018; Mollas et al., 2022). Waseem et al. (2017) and Zampieri et al. (2019) have proposed a hierarchical taxonomy of offensive speech, emphasizing the need for annotating specific dimensions of offensive language, such as the content's explicitness and the type of targets. Rosenthal et al. (2021) further expands the size of the dataset using the OLID dataset proposed by Zampieri et al. (2019) with semi-supervising method. The hierarchical annotation has also made possible systematic expansion to subtypes of hate speech in the following works, such as misogyny (Zeinert et al., 2021). Our work also builds upon the taxonomy proposed by Zampieri et al. (2019), further identifying the targeted social group of offensive languages.

Recent papers focus on more diverse aspects, such as interpretability and context information. To train a human-interpretable classification models, Sap et al. (2020) collect social bias implicated about the targeted group in a free-text format. In a similar spirit, Pavlopoulos et al. (2022) and Mathew et al. (2021) create datasets annotated with particular span of the text that makes the post toxic (Zaidan et al., 2007). As most text in the real world appears in context (Seaver, 2015), considering context is important for the development of practical models. Recent work on offensive language detection incorporates the context of the post (Vidgen et al., 2021; de Gibert et al., 2018; Gao and Huang, 2017), albeit the benefits of the context are controversial

(Pavlopoulos et al., 2020; Xenos et al., 2021). Using hierarchical annotation, KOLD dataset systematically classifies multiple depths of contextualized offensiveness, and collects textual spans to justify such classification at the same time.

6.2 Non-English Datasets

There is relatively little work done on developing offensive language datasets in languages other than English. Simple translation of English datasets is not enough as there are some well-known issues in using automatically translated English datasets in NLP, such as *translationese* (Koppel and Ordan, 2011) and over-representation of source language's culture (Hu et al., 2020). Several papers have emphasized the need of high-quality monolingual data (Hu et al., 2021; Park et al., 2021). This is also true in offensive language datasets. The focus of hatred differs by culture and country (Reichelmann et al., 2020). Ousidhoum et al. (2019) observe that there are significant differences in terms of target attributes and target groups in the three languages (English, French, Arabic) of which they constructed hate speech datasets. Moreover, Nozza (2021) shows that zero-shot, cross-lingual transfer learning of English hate speech has limitations. Some datasets for detection of toxicity or abuse exist in other languages (e.g., Zeinert et al. (2021) for Danish, Fortuna et al. (2019) for Portuguese, Mubarak et al. (2021) for Arabic, and Çöltekin (2020) for Turkish). For Korean, Moon et al. (2020) have paved the way for hate speech detection, but they are relatively small in size and lack focus on the target of offensiveness. In comparison, KOLD is built upon an extensive taxonomy that can handle a broad range of offensive language with clearly annotated target groups of 21 categories and textual spans.

7 Conclusion

We present KOLD, a dataset of 40,429 comments of news articles and video clips, annotated within context. It is the first to introduce a hierarchical taxonomy of offensive language in Korean with textual spans of the offensiveness and the targets. We establish baseline performance for multi-task model that both detects the categories and the spans that support the classification. Through analysis and experiments, we show that target terms are often omitted in offensive comments, and title information helps models predict the target of the offense. This

finding can be applied to other syntactically null-subject languages other than Korean (e.g., Arabic, Chinese, Modern Greek) as well. By comparing the distribution of target groups with existing English data and showing the inadequacy of multilingual models, we demonstrate that offensive language corpus customized for the language and its corresponding culture is necessary. We acknowledge that our dataset does not cover all communities of Korean social media whose offensive language patterns may differ from each other. Despite this limitation, KOLD will serve as a stepping stone to developing more accurate and adaptive offensive language detection models in Korean.

8 Ethical Considerations

This study has been approved by the KAIST Institutional Review Board (#KH2021-177). During the annotation process, we informed the annotators that the content might be offensive or upsetting and limited the amount that each worker could provide. Annotators were also paid above the minimum wage. We are aware of the risk of releasing a dataset containing offensive language. This dataset must not be used as training data to automatically generate and publish offensive language online, but by publicly releasing it, we cannot prevent all malicious use. We will explicitly state that we do not condone any malicious use. We urge researchers and practitioners to use it in beneficial ways (e.g., to filter out hate speech). Another consideration is that the names of political figures and popular entertainers mentioned in the comments remain in our dataset. This is because offensive language detection becomes difficult without those mentions. This is consistent with the common practice in other offensive language datasets, and as a community, we need to deliberate and discuss the potential implications.

9 Limitations

We discuss two limitations of our work in this section. First, our annotation method requires a high annotation cost and a lot of time. Guiding the annotators to familiarize them with our annotation process takes much time since the guideline is complicated, and they are updated whenever ambiguous comments are reported during the process, to give annotators clear direction. Furthermore, as we collect three annotations per comment, we need a large number of annotators (3,124) and spent a significant amount of annotation cost to pay them above the minimum wage.

In most cases, there is a trade-off between quantity and quality. For example, if one plans to build a large-scale dataset with limited amount of resources, he/she should sacrifice the complexity of the annotations by reducing the amount of work for each annotator, or the accuracy of the labels by decreasing the number of annotators for each sample. This is the reason why it is challenging to build a large dataset with accurate and rich annotations. Recently, there has been an approach to make a large-scale machine-generated hate speech detection dataset (Hartvigsen et al., 2022). This might be an alternative to overcome such limitations. By collaborating with such models, we can obtain large-scale datasets with accurate labels while reducing annotation costs and time.

Second, detecting patterns of offensive language changing over time requires constant update. For example, hateful comments related to COVID-19 emerged recently, and offensive language toward political figures or celebrities also changes constantly. It is difficult to train a model that captures such changes well with a dataset within a limited time period. A model trained on our dataset might not perform so well in detecting hateful comments that emerge in the future. To overcome this limitation, a continuous update of datasets as well as methods to efficiently update models (Qian et al., 2021), is needed.

Acknowledgments

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra.

2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Jisun An, Haewoon Kwak, Claire Seungeun Lee, Bogang Jun, and Yong-Yeol Ahn. 2021. Predicting anti-Asian hateful users on Twitter during COVID-19. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4655–4666, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

Çağrı Çöltekin. 2020. A corpus of turkish offensive language on social media. In *Proceedings of the 12th language resources and evaluation conference*, pages 6174–6184.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.

Hai Hu, He Zhou, Zuoyu Tian, Yiwen Zhang, Yina Patterson, Yanting Li, Yixin Nie, and Kyle Richardson. 2021. Investigating transfer learning in multilingual pre-trained language models through Chinese natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3770–3785, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1318–1326.

Jenny Lee, Jae-Eun Jon, and Kiyong Byun. 2017. Neo-racism and neo-nationalism within east asia: The experiences of international students in south korea. *Journal of Studies in International Education*, 21(2):136–155.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex Intelligent Systems*.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. Lifelong learning of hate speech classification on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2304–2314, Online. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual detection of offensive spans. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 144–152, Online. Association for Computational Linguistics.

Ashley Reichelmann, James Hawdon, Matt Costello, John Ryan, Catherine Blaya, Vicente Llorent, Atte Oksanen, Pekka Räsänen, and Izabela Zych. 2020. Hate knows no boundaries: Online hate in six nations. *Deviant Behavior*, 42:1–12.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online. Association for Computational Linguistics.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Nick Seaver. 2015. The nice thing about context is that everyone has it. *Media, Culture & Society*, 37(7):1101–1109.

Alexander Shvets, Paula Fortuna, Juan Soler, and Leo Wanner. 2021. Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190, Online. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. 2021. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

Appendix

A Data Collection Keywords

- Gender : 백래시 (backlash), 여성 단체 (female organization), 여성 혐오 (misogyny), 젠더 (gender), 페미니즘 (feminism)

- Sexual orientation : 동성혼 (homosexual marriage), 성소수자 (sexual minority), 차별금지법 (anti-discrimination legislation), 퀴어 (queer), 퀴어 활동가 (queer activist)

- Race & Ethnicity & Nationality : 난민 (refugee), 동남아 (Southeast Asia), 백인 (White), 외국인 근로자 (foreign worker), 이민자 (immigrant), 인도 (India), 조선족 (Korean-Chinese), 중국 동포 (ethnic Korean from China), 탈북민 (North Korean defectors), 흑인 (Black)

- Religion : 이슬람 (Islam), 탈레반 (Taliban), 기독교 (Christianity), 카톨릭 (Catholic), 교회 (Church), 목사 (minister)

B Training Details

All experiments are conducted using the Transformers library via HuggingFace[6]. For the experiments, we searched for a learning rate out of {1e-5, 2e-5, 3e-5, 5e-5} and the number of epochs out of {1, 2, 3, 4, 5}. We kept the batch size fixed at 32 for training, 64 for validation. We report a mean score of 5 runs. The experiments are conducted on GeForce RTX 2080 Ti 10GB, with 10.2 CUDA version. The single experiment takes from an hour to 4 hours. The number of parameters for BERT base model is 110 million, 123 million for RoBERTa base and 354 million for RoBERTa large.

---

[6]https://github.com/huggingface/transformers

Target Group Attribute	Target Group
Gender & Sexual Orientation	LGBTQ+, Men, Women
Race & Ethnicity & Nationality	Asian, Black, Chinese, Indian, Korean-Chinese, Southeast Asian, White
Political Affiliation	Conservative, Progressive
Religion	Buddhism, Catholic, Christian, Islam
Miscellaneous	Agism, Disabled, Diseased, Feminist, Physical Appearance, Socio-economic Status

Table 10: Target group attributes and target groups in the KOLD dataset.

Gold	Pred	Title	Comment
OFF	OFF	미국서 백인 여성 한국계 부부에 욕설 White female curses to Korean-American couples in the U.S	미국은 소수의 똑똑한사람들이 전체를 먹여살리는거임 나머지는 걍 개돼지 아이큐수준 In the US, few elite people feed the whole country The rests are just the great unwashed with low-level IQ
OFF	OFF	차별금지법을 반대하는 교회들에게 To churches who oppose anti-discrimination laws	요즘 개독들이 신천보다 더함 개독은 종교도 않임 These days Jesus freaks are way worse than Shincheonji. Halle-fxxkin-lujah is not a even religion.
NOT	NOT	캐나다 이민 | 다시 한국으로 돌아가고 싶어요. Immigration to Canada | I want to go back to Korea.	자연의 위치..저수지 에 사는·물고기가 바다에.살면.. 적응하기 어려운. 환경에.. The law of nature... When a fish living in reservoirs goes to the sea it has to have many difficulties to adapt
OFF	NOT	탈동성애자가 밝히는 동성애의 실체 The reality of homosexuality revealed by ex-gay	난 이 재미난 동성애에 빠져버렸어 ~ ~ ~ I'm into this funny homosexuality ~ ~ ~
OFF	NOT	이준석, 차별금지법, 기독교 Lee Jun-seok, Anti-discrimination law, Christianity	이준석 ... 차별금지법을 옹호한다고 ??? 이준석은 OUT Lee Jun-seok ... advocate the anti-discrimination law ??? Lee Jun-seok is OUT
OFF	NOT	미국이 만든 난민은 스스로 책임져야 US has to take responsibility for the refugees they created	어중떠중천하 A bunch of ragtag

Table 11: Examples of Level A (Offensiveness) predictions for labels and spans by a multi-task model. Top 3 examples show that the multi-task model predicts correct labels and spans, however, below 3 examples show its errors in labels and spans. (OFF: offensive, NOT: not offensive, blue : gold span, **underline**: predicted span, red : incorrect label)

Gold	Pred	Title	Comment
GRP	GRP	조선족조폭의 강남 장악기 가리봉 잔혹사 Korean-Chinese Gang's Domination in Gangnam the cruel history of Garibong	정부가 무능해서 그렇지 한심한 정치인 Because the government is incompetent. Pathetic **politician**
GRP	GRP	[정치와 젠더 세미나 2회차] '젠더와 군대' [Politics and Gender Seminar 2nd] 'Gender and the Army'	정의당 에 정의는 어디? Where is justice in **Justice Party** ?
IND	IND	"미국으로 가자!" 과테말라 국경 뚫은 이민자들 "Let's go to America!" Immigrants Breaking the Guatemala Border	또라이 바이든 Freak **Biden**
GRP	GRP	"코로나보다 무서운 건 혐오의 시선"... 코로나 시대의 중국동포 "Eyes of Hatred scarier than COVID-19"... The age of COVID-19, Compatriots in China	동포는 뭔 개소리냐 Compatriots? What the h**l are you talking about?
GRP	GRP	[한국교회 CPR] 삯꾼 목사 청소는 얼마면 돼요? [Korean Church CPR] How much is it to clean up all the wage-earner pastors?	맘몬 이 강단을 장악했네ㅜㅜ 합동을 떠나야 하나..안에서 싸워야 하나 ㅜㅜ **Mum**-Roach took control of podium ;(Should I leave the general meeting or fight against ;(
GRP	UNT	탈영병, 중국서 조선족 4명 살해 Deserters Killed 4 Korean-Chinese in China	착한 살인 인정합니다 I admit that it's a good murder

Table 12: Examples of Level B (Target) predictions for labels and spans by a multi-task model. Top 3 examples show that the multi-task model predicts correct labels and spans, however, below 3 examples show its errors in labels and spans. (UNT: untargeted, IND: individual, GRP: group, OTH: other, green : gold span, **underline**: predicted span, red : incorrect label)

댓글 읽고 혐오 표현 찾기

카테고리 선택 후 라벨을 생성해주세요

혐오 표현 혐오 대상

여자들 뚫린입이라고 막뱉네

=================

난민은 어디로 가나 / 연합뉴스

● 혐오 표현
● 혐오 대상

혐오 대상을 체크해 주세요.

집단-성 정체성-여성

집단-성 정체성-남성

제출하기

건너뛰기

Discovering Hate Speech in Comments

Please first select the category and then create a label

Offensive Target

Women talk like shit

=================

Where do refugees go / Yonhap News

● Offensive
● Target

Please check the target of hatred
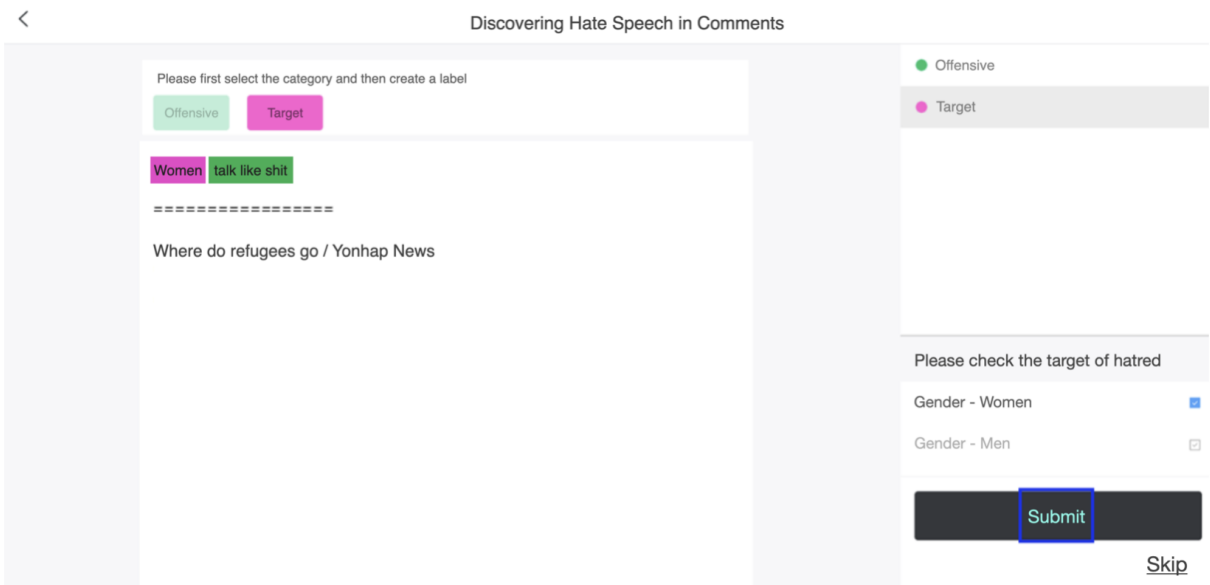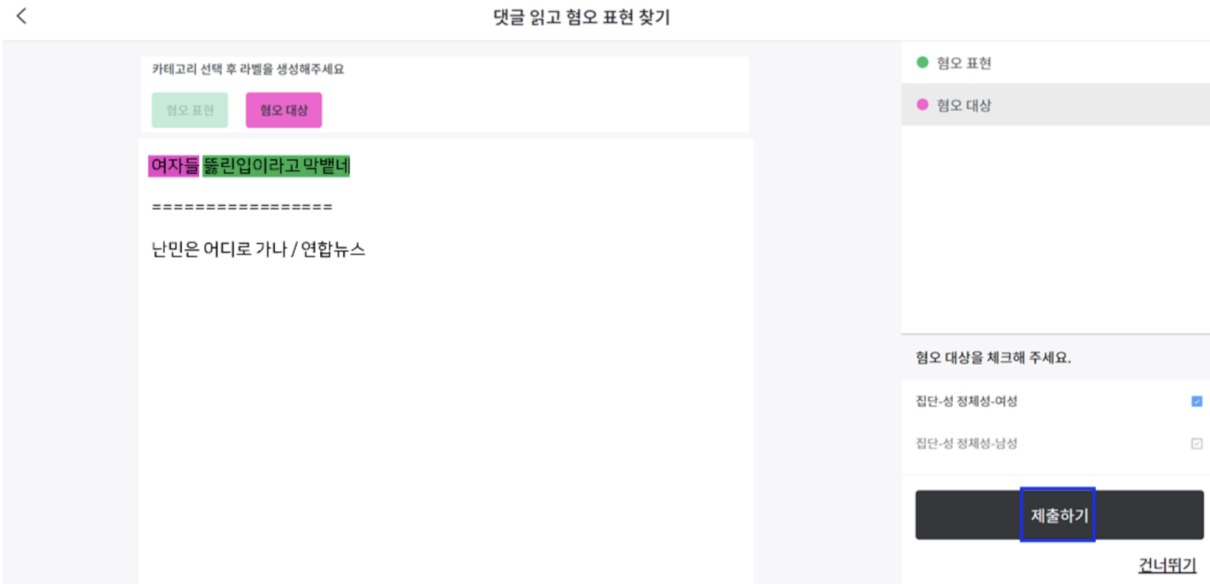
Gender - Women

Gender - Men

Submit

Skip

Figure 2: Annotation interface used to collect label and span annotations of KOLD.