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, 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-
In this paper, we describe and publicly release the Korean Offensive Language Dataset (KOLD), 40,429 comments collected from news articles and videos. 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 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)

### 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.

<table>
<thead>
<tr>
<th>Context (title)</th>
<th>Comment</th>
<th>Level A</th>
<th>Level B</th>
<th>Target Group Attribute</th>
<th>Target Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>심화하는 젠더갈등...SNS 때문? Gender in Conflict, deepening by SNS?</td>
<td>사람들은 사람들을</td>
<td>NOT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>그래, 레터는 이롭다. 포기하는 게 하/datatables</td>
<td>The garbage named TikTok should be eliminated</td>
<td>OFF</td>
<td>GRP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“이 법은 희망을 털어 떨어진 누구를 살리기도 해요” “This law could save desperate lives who almost died”</td>
<td>수많은 트랜스젠더들을 보도</td>
<td>OFF</td>
<td>GRP</td>
<td>Gender &amp; Sexual Orientation</td>
<td>LGBTQ+</td>
</tr>
<tr>
<td>&quot;People are all just people NOT anyone for surgery, I don’t care what (they) believe, I get mad when I see transgender people who had surgery, I’ll take a knife and kill them all!&quot;</td>
<td>It would be so refreshing if all the Korean-Chinese went back to China.</td>
<td>OFF</td>
<td>GRP</td>
<td>Race, Ethnicity &amp; Nationality</td>
<td></td>
</tr>
<tr>
<td>&quot;How much is the weight of the people who suffer from this cruel treatment?&quot;</td>
<td>I don’t care what they believe, what matters is the fact that they kill people</td>
<td>OFF</td>
<td>GRP</td>
<td>Race, Ethnicity &amp; Nationality</td>
<td>Muslim</td>
</tr>
<tr>
<td>“I want the people who believe in ‘Islam’ in Korea”</td>
<td>Feminism is not about gender inequality, it is a mental disorder</td>
<td>OFF</td>
<td>GRP</td>
<td>Misc. Feminists</td>
<td></td>
</tr>
</tbody>
</table>

In this paper, we describe and publicly release the Korean Offensive Language Dataset (KOLD), 40,429 comments collected from news articles and videos. 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. 

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 protected 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., jiangkkae, 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, “미안한/IP '[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. 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. 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.

3.3 Annotation Result

Overall, the average Krippendorff’s \(\alpha\) for inter-annotator agreement of each annotation level is 0.55. The label distribution of the collected data is shown in Table 2.

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

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

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.

**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 \(\alpha\) 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.

\(^2\)Using big data analysis to chart a new course, The Korea Herald

\(^3\)Inspectors are selected workers who passed the qualification test with the highest scores.
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.

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

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

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
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 pre-trained models to predict the label given the title 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
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 pretrained 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 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 and use the dataset for training the same sequence classifier described in Section 5.1. For the multilin-

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.

<table>
<thead>
<tr>
<th>System</th>
<th>Offensive</th>
<th>Offensive Span</th>
<th>Target Type</th>
<th>Target Span</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>Seq</td>
<td>75.8</td>
<td>78.7</td>
<td>77.2</td>
<td>-</td>
</tr>
<tr>
<td>Span</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>47.7</td>
</tr>
<tr>
<td>Seq + Span</td>
<td>75.0</td>
<td>78.1</td>
<td>76.5</td>
<td>56.3</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Level</th>
<th>Comment</th>
<th>Comment+Title</th>
<th>Δ F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>76.9</td>
<td>77.2</td>
<td>+ 0.3 (0.4%)</td>
</tr>
<tr>
<td>B</td>
<td>56.1</td>
<td>57.6</td>
<td>+ 1.5 (2.6%)</td>
</tr>
<tr>
<td>C</td>
<td>42.8</td>
<td>55.9</td>
<td>+ 13.1 (30.6%)</td>
</tr>
</tbody>
</table>

Table 9: OLID\textsubscript{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.

<table>
<thead>
<tr>
<th>System</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offensive</td>
<td>76.9</td>
</tr>
<tr>
<td>OLID\textsubscript{Seq}</td>
<td>63.2</td>
</tr>
<tr>
<td>Offensive</td>
<td>40.6</td>
</tr>
<tr>
<td>Span</td>
<td>MUDES</td>
</tr>
</tbody>
</table>

\textsuperscript{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 (Pavlopulos 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 Ordoñez, 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

This project was funded by the KAIST-NAVER Hypercreative AI Center. Alice Oh is funded by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics). SelectStar provided a crowdsourcing platform for the annotation of the data.

References


Nuha Albadi, Maram Kurdi, and Shivakant Mishra.


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

Appendix

A Data Collection Keywords

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

• Sexual orientation : 동성혼 (homosexual marriage), 성소수자 (sexual minority), 차별금지법 (anti-discrimination legislation), 퀘어 (queer), 퀘어 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), 혼인 (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. 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.

---

6https://github.com/huggingface/transformers
<table>
<thead>
<tr>
<th>Gender &amp; Sexual Orientation</th>
<th>Target Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGBTQ+, Men, Women</td>
<td></td>
</tr>
<tr>
<td>Race &amp; Ethnicity &amp; Nationality</td>
<td>Asian, Black, Chinese, Indian, Korean-Chinese, Southeast Asian, White</td>
</tr>
<tr>
<td>Political Affiliation</td>
<td>Conservative, Progressive</td>
</tr>
<tr>
<td>Religion</td>
<td>Buddhism, Catholic, Christian, Islam</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>Agism, Disabled, Diseased, Feminist, Physical Appearance, Socio-economic Status</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Gold</th>
<th>Pred</th>
<th>Title</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFF</td>
<td>OFF</td>
<td>미국에서 배인 여성 한국계 부부의 육성</td>
<td>The rest are just the great unwashed with low-level IQ</td>
</tr>
<tr>
<td>OFF</td>
<td>NOT</td>
<td>지방군정법을 반대하는 교회들에게</td>
<td>These days Jesus freaks are way worse than Shincheonji.</td>
</tr>
<tr>
<td>NOT</td>
<td>NOT</td>
<td>한국인 이민 1, 2, 3로 한국으로 돌아가고 싶어요</td>
<td>The law of nature... When a fish living in reservoirs goes to the sea it has to have many difficulties to adapt</td>
</tr>
<tr>
<td>OFF</td>
<td>NOT</td>
<td>차별금지법을 용호한다고 ???</td>
<td>Lee Jun-seok... advocate the anti-discrimination law ???</td>
</tr>
<tr>
<td>OFF</td>
<td>NOT</td>
<td>미국이 만든 난민은 스스로 책임져야</td>
<td>A bunch of ragtag</td>
</tr>
</tbody>
</table>

**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)

<table>
<thead>
<tr>
<th>Gold</th>
<th>Pred</th>
<th>Title</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRP</td>
<td>GRP</td>
<td>조선축조직의 강남 장악기 가려웅 항혹사</td>
<td>The cruel history of Garibong</td>
</tr>
<tr>
<td>GRP</td>
<td>GRP</td>
<td>[정치와 펜터 셰이나 2회기] “펜터와 군대”</td>
<td>Where is justice in Justice Party?</td>
</tr>
<tr>
<td>IND</td>
<td>IND</td>
<td>&quot;미국으로 가자!&quot; 과대발라 과정 좋은 이민자들</td>
<td>Trump is a hypocrite</td>
</tr>
<tr>
<td>GRP</td>
<td>GRP</td>
<td>&quot;코로나바타 무서운 건 헤모의 시선&quot;...</td>
<td>What are you talking about?</td>
</tr>
<tr>
<td>GRP</td>
<td>UNT</td>
<td>[한국교회 CPR] 항목 목표 창설은 얼마나 때요?</td>
<td>Should I leave the general meeting or fight against?</td>
</tr>
<tr>
<td>GRP</td>
<td>UNT</td>
<td>망명병, 중국계 조선족 4명 살해</td>
<td>I admit that it’s a good murder</td>
</tr>
</tbody>
</table>

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