A Large-scale Comprehensive Abusiveness Detection Dataset with Multifaceted Labels from Reddit

Hoyun Song[*] **Soo Hyun Ryu**[*†] **Huije Lee** **Jong C. Park**[‡]
School of Computing
Korea Advanced Institute of Science and Technology
{hysong,shryu,jae4258,park}@nlp.kaist.ac.kr

Abstract

As users in online communities suffer from severe side effects of abusive language, many researchers attempted to detect abusive texts from social media, presenting several datasets for such detection. However, none of them contain both comprehensive labels and contextual information, which are essential for thoroughly detecting all kinds of abusiveness from texts, since datasets with such fine-grained features demand a significant amount of annotations, leading to much increased complexity. In this paper, we propose a Comprehensive Abusiveness Detection Dataset (CADD), collected from the English Reddit posts, with multifaceted labels and contexts. Our dataset is annotated hierarchically for an efficient annotation through crowdsourcing on a large-scale. We also empirically explore the characteristics of our dataset and provide a detailed analysis for novel insights. The results of our experiments with strong pre-trained natural language understanding models on our dataset show that our dataset gives rise to meaningful performance, assuring its practicality for abusive language detection.

1 Introduction

While enhancing the freedom of expression, online discussion has also brought massive harm inflicted by abusive language. To address this problem, numerous studies have attempted to automatically identify abusive expressions in social media (Nobata et al., 2016; Kumar et al., 2018; Wiegand et al., 2018; Zampieri et al., 2019; Pedersen, 2019). However, the definition of abusive language varies to those aspects the researchers have considered important (Nobata et al., 2016; Price et al., 2020). These aspects include lexical profanity (Pedersen, 2019; Koufakou et al., 2020) or implicit abuse (Kumar et al., 2018; Caselli et al., 2020; Wiegand et al., 2021). Other researchers put their importance on distinguishing targeted attacks from profanity (Poletto et al., 2017; Zampieri et al., 2019) or identifying the targets' demographic characteristics (Kumar et al., 2018; Davidson et al., 2019). Hence, detecting abusive language is quite challenging as it is inherently a high-dimensional phenomenon as to how it is expressed or what the speakers intend to convey.

Due to the high-dimensionality, devising a model for identifying the abusiveness[1] from social media texts calls for a dataset covering comprehensive aspects. However, to the best of our knowledge, no study has yet provided a dataset with such fine-grained features for the abusiveness detection. Constructing such a dataset with multifaceted and detailed labels demands a significant amount of annotations, leading to much increased complexity for dataset construction. To address this challenge, we propose a new annotation scheme that exploits the hierarchical structure of interrelated features. Our hierarchical annotation scheme makes use of a series of easy-to-answer questions that mitigate the complexity from high-dimensional information, enhancing annotation efficiency in crowdsourcing.

As an outcome of our annotation scheme, we present a large-scale *Comprehensive Abusiveness Detection Dataset (CADD*[2]*)*. The CADD contains diverse linguistic information with a broad range of aspects of abusive language, detailed in Section 3.3. We perform various analyses to confirm that our annotation scheme leads to a high quality and diverse dataset. We make empirical studies of our dataset to look into its characteristics for insights, and assess the performance of strong pre-trained Natural Language Understanding (NLU) models

[*] Equal contribution

[†] Present Address: Department of Psychology, University of Michigan, soohyunr@umich.edu.

[‡] Corresponding author

[1] We use the term '*abusiveness*' to cover a broad range of aspects of abusive language.

[2] https://github.com/nlpcl-lab/CADD_dataset

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 552–561
November 10–11, 2021. ©2021 Association for Computational Linguistics

	# Instances	Type	Target	D.C.	Implicit	Profanity	Context
Waseem and Hovy (2016)	17K	✗	✗	✓	✗	✗	✗
Poletto et al. (2017)	2K	✗	✓	✗	✓	✓	✗
Wiegand et al. (2018)	9K	✓	✗	✗	✓	✗	✗
de Gibert et al. (2018)	10K	✗	✗	✗	✗	✗	✓
Kumar et al. (2018)	21K	✗	✓	✓	✓	✗	✗
Basile et al. (2019)	19K	✗	✓	✗	✗	✓	✗
Zampieri et al. (2019)	14K	✓	✓	✗	✗	✗	✗
Caselli et al. (2020)	13K	✓	✗	✗	✓	✗	✗
CADD (ours)	24K	✓	✓	✓	✓	✓	✓

Table 1: An overview of the related corpora for abusive language detection. Each column indicates the features of abusive language. *D.C.* indicates demographic characteristics of a target of an abusive content.

trained on our dataset to show its practicality.

In summary, our major contributions are as follows: 1) We introduce a large-scale CADD, designed to contain multifaceted labels, annotated hierarchically; 2) we present an effective scheme, assessing the quality and diversity of the presented dataset; and 3) we present observations found through empirical studies and experiments on the dataset which may be helpful for future studies.

2 Related Work

Given that there is little consensus on the clear definition of abusive language (Waseem et al., 2017; Castelle, 2018), each study used a different definition of the abusive language and narrowed down its scope accordingly. For instance, some studies focused on lexicon-based profanity (Saleem et al., 2017; Pedersen, 2019; Koufakou et al., 2020), while others focused rather on hate speech, which is understood to attack a specific group by mentioning the typical aspects of its members such as age, gender, or sexual orientation (de Gibert et al., 2018; Fortuna and Nunes, 2018). Other researchers focused on the derogatory language that attacks an individual or a group without such identity aspects (Nobata et al., 2016; Davidson et al., 2019). Motivated by these types of abusive language (hate speech, derogatory, and profanity), researchers have also worked on identifying information about targets (Poletto et al., 2017; Zampieri et al., 2019) or demographic characteristics (Park and Fung, 2017; Kumar et al., 2018; Davidson et al., 2019), because these are the key factors to discriminate the types. Some studies have emphasized that contexts surrounding abusive language should also be taken into account (Castelle, 2018; Qian et al., 2019), or whether it is an implicit abuse such as sarcasm, rhetorical questions, or satire (Poletto et al., 2017; Kumar et al., 2018; Caselli et al., 2020; Wiegand et al., 2021). We note that some of these aspects share closely interrelated features, so that they would be best considered together.

With the necessity of detecting a broad range of aspects of abusive language, we first sought to include its diverse types. Waseem et al. (2017) and Ross et al. (2017) speculated that abusive language annotation could never be complete or made reliable with a single binary value (e.g., “abusive" or “non-abusive"). It is thus not surprising that several studies introduced corpora with novel typologies of abusive language. Table 1 shows various features that are included in each corpus.

However, Table 1 also shows that current open datasets do not cover all features and contextual information at the same time. Specifically, Wiegand et al. (2018) employed 4-way classification, annotating multiple subcategories (Abuse, Insult, Profanity, or Other). Waseem and Hovy (2016) studied demographic characteristics for abusive remarks such as racism or sexism. Poletto et al. (2017) covered information about target, action, aggressiveness, offensiveness, irony, and stereotype. de Gibert et al. (2018) emphasized that contexts surrounding abusive language should also be taken into account. Basile et al. (2019) and Zampieri et al. (2019) looked into the target of abusive remarks, such as whether the offensive message has a target or not and whether the target of the offensive message is an individual, a group, or other. Kumar et al. (2018) presented a corpus that has the most enriched range of labels about subcategories of abusive language but they dismissed the importance of context. Caselli et al. (2020) proposed to include implicit abuse for abusiveness classification. While

Example	L1	L2	L3	L4	L5	L6
Context: I don't like black people, and never associate with them. **Text**: We hate n****r, but what we hate more are low effort trolls.	**HS**	A	Y	Y:2	N	Y
Context: I have but one silver to give. **Text**: If you freak me out, I want to get the f**k away from you.	**D**	A	Y	N	N	Y
Context: Reminder: Stay home if you're sick. **Text**: It's f**king ridiculous that this post is needed.	**P**	A	N	-	N	Y
Context: Dear Americans, what is the most forgotten state in US? **Text**: Absolutely Montana. Nobody is even mentioning it.	**N**	N	-	-	N	N

Table 2: Examples of Reddit posts from our dataset with corresponding labels. *L2* to *L4* show a hierarchical structure where some of the labels do not need to be annotated (marked as '-' symbol). The first label (*L1*) also does not need to be annotated since it can be determined automatically according to other labels (from *L2* to *L4*).

we find that all of these approaches address important aspects of abusive language, we see that they do not offer a broad enough range of semantic and linguistic information on a large-scale.

3 Data Design

3.1 Broad definition of abusive language

Abusiveness is an inherently fuzzy term that can be defined in many different ways, and therefore researchers needed to determine what is abusive to make their studies clear. In the present study, we assume that abusive language can fall into four types, following Nobata et al. (2016):

Hate Speech (HS) A language that attacks people with a particular identity with respect to properties such as race, religion, gender, and age;

Derogatory (D) A language that attacks a group or an individual, but is not considered hate speech;

Profanity (P) A language that contains any sexual remarks or slur; or

Non-abusive (N) A language that is not in any categories of abusive language.

The reason behind this definition is that it covers both the target information about the attack and the demographic characteristics of the target, also addressed by previous studies.

3.2 A hierarchical annotation scheme

We introduce a hierarchical annotation scheme for annotating multifaceted labels. There are two key ideas behind this scheme. First, the aforementioned types of abusive language can be structured hierarchically by combining interrelated features (Park and Fung, 2017; Zampieri et al., 2019). In our case, there are three levels of hierarchy: i) abusiveness, ii) target, and iii) demographic characteristics, where (i) through (iii) indicate the levels of the hierarchy (i > ii > iii). The first level (i) discriminates all types of abusive language from clean texts, and the second level (ii) distinguishes the targeted attack (i.e., hate speech and derogatory) and profanity. The third level (iii) distinguishes the hate speech and derogatory remarks.

Second, annotation schemes should provide intuitive guidelines for crowdsourcing because the performance of crowdworkers tends to become less reliable when they are given more complicated tasks. We assume that a branch to consecutive questions is much more intuitive than that of a bunch of multiple choices (Hellinga and Menkovski, 2019). We designed our scheme hierarchically so as to make questions less complicated and help annotators follow the guidelines more easily.

3.3 Detailed labels of interest

To address diverse aspects of abusive language, we introduce six labels to the dataset (from *L1* to *L6*, respectively): *Type*, *Abusiveness*, *Target*, *Demographic Characteristics*, *Implicitness*, and *Profanity*. Table 2 lists entries that exemplify[3] the four types of abusive language with their corresponding labels.

L1: Type Following the classification of Nobata et al. (2016), we use four types of abusive language, described in Section 3.1. This label does not need to be annotated since it is automatically determined with regard to other labels. The use of this label allows our dataset to train multi-class classification models, which is more complicated than classifying binary labels. The values are **Hate Speech** (*L1:HS*); **Derogatory** (*L1:D*); **Profanity** (*L1:P*); or **Non-abusive** (*L1:N*).

[3] The presented examples are shortened from the actual data because of privacy and ethical issues.

L2: Abusiveness It shows whether a given text works as abusive or not. This label is at the first level (i) of the hierarchical structure of our annotation scheme. The values are **Abusive** (*L2:A*), containing any form of abusiveness; or **Non-abusive** (*L2:N*), that does not contain any profane words and does not convey the intention to attack.

L3: Target This label tells the presence of targets to be attacked (either implicitly or explicitly), the second level (ii) of our hierarchical structure. The *L3* label is the key feature of discriminating a non-targeted profanity (*L1:P*), against other targeted attacks (*L1:HS* and *L1:D*). The values are **Targeted** (*L3:Y*), containing a language that attacks a group or an individual; or **Non-targeted** (*L3:N*).

L4: Demographic Characteristics It is the last level (iii) of the hierarchy, categorizing whether the attack is at one or more of the identities of the target. If posts contain a targeted attack towards specific identities, *L4* would be marked as **Yes** (*L4:Y:1-8*); otherwise, **No** (*L4:N:0*). The following number from 1 to 8 indicates the related properties; *gender*, *sexual orientation*, *race*, *religion*, *disability*, *age*, *others*, and *unclear*, respectively, where *unclear* is not agreed upon among the annotators. If *L4* is annotated as one of *L4:Y:1-8*s, the post will be automatically classified as *L1:HS*.

L5: Implicitness Posts can express abusiveness, whether implicitly or not. Given that different strategies can be implemented to detect or mitigate abusiveness depending on the way abusiveness is expressed, we included the implicitness of abusive comment in our dataset (*L5:Y* or *L5:N*).

L6: Profanity If posts contain any words expressing abusiveness in an explicit way, we have *L6:Y*; otherwise *L6:N*.

3.4 Contextual information

Even for the same sentences, the delivered message could differ significantly depending on the context in which they are uttered (Kamp and Partee, 2004). In this regard, whether a remark is abusive or not may not be determined without considering the relevant context. Given the importance of such contextual information in abusiveness determination, the present dataset was constructed and is released with full context information.

4 Corpus Construction

We constructed a new dataset to support the tasks of identifying a broad range of aspects of abusiveness from social media texts. The Comprehensive Abusiveness Detection Dataset (CADD) was built with posts extracted from the English Reddit dataset[4].

For annotation on Amazon Mechanical Turk[5] (AMT), we asked crowdworkers to provide values for multifaceted labels on each data. Although we were aware of the trade-off between relying on experts with assured quality and crowdsourcing with scalable quantity (Tekiroglu et al., 2020), we eventually chose to crowdsource because it could magnify productivity and provide perspectives of non-experts, amplifying the representativeness of online users. Specifically, given that the purpose of the present study lies more in collecting the large-scale data that best represent online users rather than maximizing the reliability and consistency, we concluded that crowdsourcing is more appropriate.

We conducted two steps of annotation tasks: First, we asked the workers to determine if the comments contain abusiveness, to reduce costs at the second step by maximizing the number of abusive comments; second, we carried out a very detailed annotation using our hierarchical annotation scheme. Figure 1 shows an overview of our annotation process. A more detailed explanation for each step is given below.

4.1 Data collection and preprocessing

We sampled posts from Reddit, which is one of the largest online communities. Each post consists of a title and a body, together with a comment. We choose a comment as a target of abusive language detection, and a title and a body as contextual information. Even though the discussion across multiple comments also provides important contextual information, we just included a single top-level comment for each post to keep the uniformity of the length. In order to prevent our data from being skewed toward non-abusive comments, we sought to balance the numbers of abusive and non-abusive comments. To this end, we crawled two categories of comments as follows:

Possibly Abusive Comments We employed an offensive/profane word list[6] to collect abusive comments. The word list contains not only profanity

[4] https://files.pushshift.io/reddit/

[5] https://www.mturk.com/

[6] https://www.cs.cmu.edu/~biglou/resources/

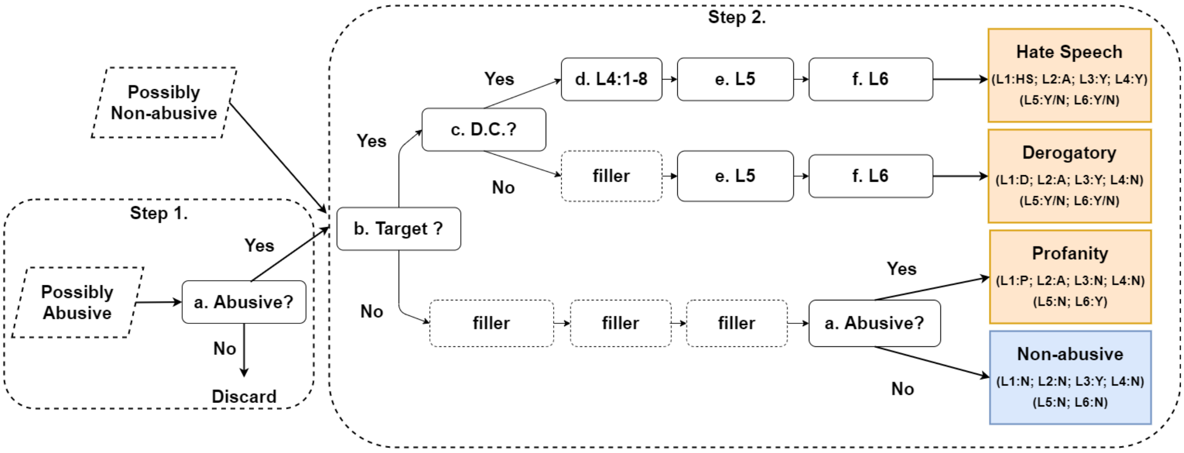
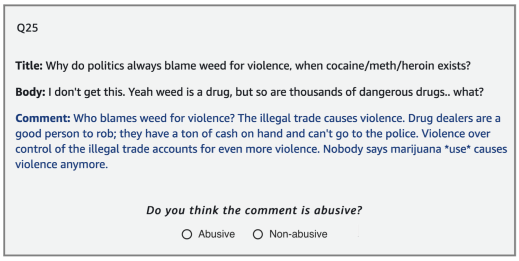


Figure 1: An overview of the hierarchical annotation scheme for multifaceted labels. ‘*filler*’ is a question to balance the number of questions for each of the four cases.



Q25

Title: Why do politics always blame weed for violence, when cocaine/meth/heroin exists?

Body: I don't get this. Yeah weed is a drug, but so are thousands of dangerous drugs.. what?

Comment: Who blames weed for violence? The illegal trade causes violence. Drug dealers are a good person to rob; they have a ton of cash on hand and can't go to the police. Violence over control of the illegal trade accounts for even more violence. Nobody says marijuana *use* causes violence anymore.

Do you think the comment is abusive?

○ Abusive ○ Non-abusive

Figure 2: An example of an annotation task, shown to the annotators at the first step.

words but also more general terms that can often be accompanied with offensiveness, such as ‘*asian*’, ‘*black*’, or ‘*remains*’. This is intended to satisfy the purpose of the present study, which is to collect diverse types of abusive comments rather than focusing only on the comments with profanity.

Possibly Non-Abusive Comments For the comparison of abusive comments, we randomly collected comments as possibly non-abusive comments. Then, we randomly selected 10,000 posts out of the collected comments to be employed in Section 4.2.2.

We performed preprocessing by discarding posts that contain URLs or Emojis. We also discarded posts that are shorter than 3 or longer than 512 words (i.e., tokens). We only retained posts in English; otherwise, they are discarded.

4.2 Data annotation

4.2.1 Step 1. Determination of abusiveness

To make sure that we include enough abusive comments in the next step, we first asked annotators to determine whether each comment is abusive or not and extracted only those comments that were considered abusive, as shown in Figure 1.

To this end, we recruited 2,000 participants through AMT with Master status, whose IP addresses were restricted to those of the English-speaking countries. We provided 50 Reddit posts to each participant to determine whether each comment is considered abusive or not. Every post has been annotated by two participants; a total of 50,000 posts were annotated. The average time spent by participants to complete the task was 7 minutes, and the participants were rewarded with $ 0.5. An example question provided to participants is shown in Figure 2. We accepted posts as abusive if they are fully agreed upon by two participants. Among these accepted posts, we randomly selected 20,000 posts to be included in the next step. Step 1 performs the first level (i) of hierarchical annotation, described in Section 3.2.

4.2.2 Step 2. Annotation of detailed features

Through Step 1, we obtained 20,000 posts whose comments are likely to be abusive, and 10,000 posts whose comments are considered as non-abusive, with a total of 30,000 posts. In the second step, we also had the comments annotated with labels through AMT with 6,000 participants, with the same restrictions as in Step 1. The average time spent by participants to complete the task was 13 minutes, and the participants were rewarded with $ 1. Each participant was given 15 Reddit posts per task.

For each post, participants were required to annotate the values of multi-labels based on the scheme explained in Section 3.2. Figure 1 shows the details of our decision-tree shaped scheme, where

Type	Train	Val.	Test	Total
Hate speech	2,515	388	772	3,675
Derogatory	1,632	241	494	2,367
Profanity	4,595	631	1,339	6,565
Non-abusive	8,412	1,190	2,297	11,899
All	17,154	2,450	4,902	24,506

Table 3: The size of our corpus, which is divided into training, validation (Val.), and test sets.

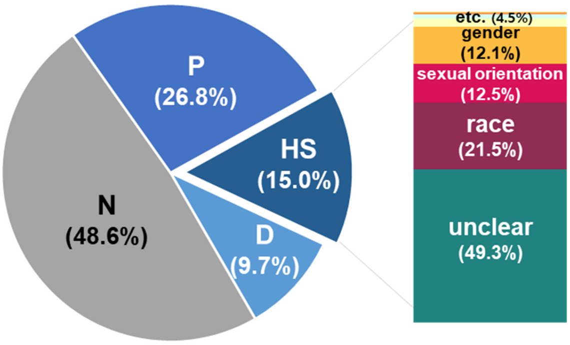


Figure 3: A distribution of abusive types and demographic characteristics of our corpus. The *etc.* includes religion, disability, age, and others.

each question is indexed in alphabetical order from *a* to *f*. Each question from *a* to *c* is for determining the types of abusive language (*L1*), where questions *b* and *c* are about the second and third levels (ii, iii) of our hierarchical annotation, respectively. Even though question *a* has already been checked in Step 1, it is double-checked in Step 2 to determine the final result of annotation. Question *d* is for specifying the demographic categories, and *e* and *f* are for determining *L5* and *L6*, respectively.

Going through our annotation scheme, every input is determined to have an abusive type out of four, as well as fully annotated with six labels. As a result, we obtained 30,000 annotated posts, each of which was annotated by three participants.

4.3 Ethical consideration

Our annotation task was approved by *Korea Advanced Institute of Science and Technology* Institutional Review Board (IRB)[7], and the informed consent was read and acknowledged by participants (annotators) prior to their tasks[8]. Since all the possible privacy concerns of the data should be respected (Šuster et al., 2017; Benton et al., 2017), our dataset is fully anonymized and will be made available to researchers who are informed of, and agree to ethical guidelines.

5 Corpus analysis

5.1 Data quality and statistics

In order to ensure the quality of our annotated dataset obtained through crowdsourcing, we filtered out some of the annotation results, which are possible mistakes or spamming (Hovy et al., 2013; Yang et al., 2019; Smith et al., 2020). To filter out possible spamming, we discarded some posts from unreliable workers who spent too little time to complete the task compared to the average or who submitted the same answer for all questions. After removing annotation outliers, we used the majority vote for the remaining data to obtain the ground truth labels.

By going through crowdsourcing, we constructed a large-scale annotated dataset containing 12,607 comments with abusive contents and 11,899 clean comments (in total, 24,506). The average length of each comment is 26.7 tokens (median 19), while that of contextual information (i.e., Title and Body) is 47.4 tokens (median 42). Table 3 presents the detailed figures of our corpus with respect to their multi-class labels. As shown, the number of *L1* labels for our corpus is quite balanced. The CADD dataset is randomly split into training (70%), validation (10%), and test (20%) sets. We also present detailed statistics of the proportion of their abusive types and the proportion of target demographic identities (Figure 3).

We computed the Inter-Annotator Agreement (IAA) to assess the validity and efficiency of our annotation. We measured Fleiss' kappa (κ) for the three annotators on each label of our dataset. The values were $\kappa = 0.56$ for 4-way multi-class categories (*L1*), and $\kappa = 0.69$ for binary categories (*L2*). We achieved not only 'substantial' agreement on binary annotations but also 'moderate' agreement on multiple annotations (McHugh, 2012). The values for the other labels are $\kappa = 0.47$ (*L3*), $\kappa = 0.43$ (*L4*), $\kappa = 0.54$ (*L5*), and $\kappa = 0.63$ (*L6*). The result shows that, even though they are crowdsourced, whose quality is known to be more difficult to assure than in-house annotations, our dataset achieves a quality level higher than moderate for all labels. It suggests that our hierarchical annotation scheme successfully lowers the complexity of crowdsourc-

[7] Approval number: KH2020-076

[8] We gave participants advance notice of possible exposure to harmful contents during experiments and encouraged them not to participate if they have any concerns related to mental and/or physical health.

No.	Example	Feature
1	**Context**: Who do you think is the greatest black man on the earth? **Text**: The African warlord that killed the most blacks.	Semantics
2	**Context**: What are your thoughts on team killing? **Text**: I only TK if someone calls me a f****t or a n****r, or variant thereof.	Semantics
3	**Context**: White people react positively to black people calling them with n-word. **Text**: Only ninja can call another ninja ninja.	Context
4	**Context**: What do you think the campaign could be about? **Text**: The a**h*** white guy. The strong female lead. The token black guy. The buff Latino nerd. And probably who is half robot, to include AI diversity.	Sarcasm
5	**Context**: Firecrackers and fireworks should be illegal. **Text**: You can just fok on outta here.	Neologism

Table 4: Examples of Reddit posts with respect to their related features

ing to a manageable degree even with multifaceted labels.

5.2 Empirical studies

To ensure the usefulness of our dataset and provide a hint of future researches, we empirically studied our dataset. We present some qualitative examples and their related features in Table 4.

The first and second example posts show semantic diversity and emphasize the need for a considerable understanding of their precise meanings. The first post implies *hate speech* without any profane words. In contrast, the second post is *profanity*, even though it has hate-based words, assaulting a specific identity, since it does not work as a targeted attack. They suggest that our dataset enables fine-grained studies of high-dimensional semantic tasks, such as intent classification, categorizing demographic properties, and identifying a target.

The third post presents a case where the degree of abusiveness can be changed depending on its context. Given a text alone, it is often quite tricky to pin down abusiveness because the information of abusive language may not be limited to just one sentence but sometimes spread over multiple sentences. This suggests that our dataset challenges models that are trained without considering the context.

We note that our dataset also retains traditional challenges of natural language processing, such as detecting sarcasm or neologisms. Thus, we anticipate that our dataset can be analyzed further to address such challenges.

6 Experiment

6.1 Experimental setup

In this section, we report the performance of natural language understanding models on our dataset (CADD). We trained each model on the CADD training set, optimized it on the CADD validation set, and evaluated it on the CADD test set. We randomly shuffled the training data at the beginning of each training epoch. In order to detect abusiveness, we conducted two tasks, using the two most prominent labels out of the multiple labels. The details of each task are as follows:

a. Task 1 is binary classification to determine whether a text is abusive or not (*L2*).

b. Task 2 is multi-class classification to choose one out of four abusive language types (*L1*).

As for the baselines, we implemented two dictionary-based classifiers, support vector machine (SVM) and random forest (RF), and three pre-trained transformer models. We experimented with them on the CADD dataset. We fine-tuned SVM with linear kernel and C=10, and RF where max depth is set to 100. We employed a BERT's vocabulary to train dictionary-based models. We fine-tuned transformer models employing the default settings from the Huggingface library (Wolf et al., 2019):

a. BERT (Devlin et al., 2018) is designed to pretrain bidirectional representations using masked language models. We fine-tuned the *bert-base-cased* model.

b. ALBERT (Lan et al., 2019) has significantly fewer parameters than a traditional BERT by two parameter reduction techniques. We fine-tuned the *albert-base-v2* model.

Model	Task 1				Task 2			
	wPre.	wRec.	wF1 (SD)	MF1	wPre.	wRec.	wF1 (SD)	MF1
SVM	0.800	0.799	0.799 (0.001)	0.799	0.589	0.600	0.592 (0.001)	0.481
RF	0.845	0.845	0.845 (0.002)	0.844	0.657	0.665	0.601 (0.003)	0.446
BERT	0.884	0.884	0.884 (0.003)	0.882	0.735	0.713	0.721 (0.004)	0.605
ALBERT	0.884	0.884	0.884 (0.003)	0.883	0.743	0.710	0.721 (0.004)	0.590
RoBERTa	0.886	0.886	0.886 (0.001)	0.885	0.739	0.716	0.723 (0.002)	0.612
BERT†	0.886	0.886	0.886 (0.003)	0.885	**0.755**	0.723	**0.735 (0.005)**	0.612
ALBERT†	0.887	0.887	0.887 (0.001)	0.887	0.752	0.718	0.729 (0.004)	0.602
RoBERTa†	**0.891**	**0.890**	**0.890 (0.001)**	**0.890**	0.750	**0.725**	0.733 (0.003)	**0.626**

Table 5: Results for two tasks on our dataset. We report weighted averaged precision (wPre.), recall (wRec.), wF1, and macro-F1 (MF1) for each model on two tasks. SD indicates the standard deviation. Model names with † are models using contextual information.

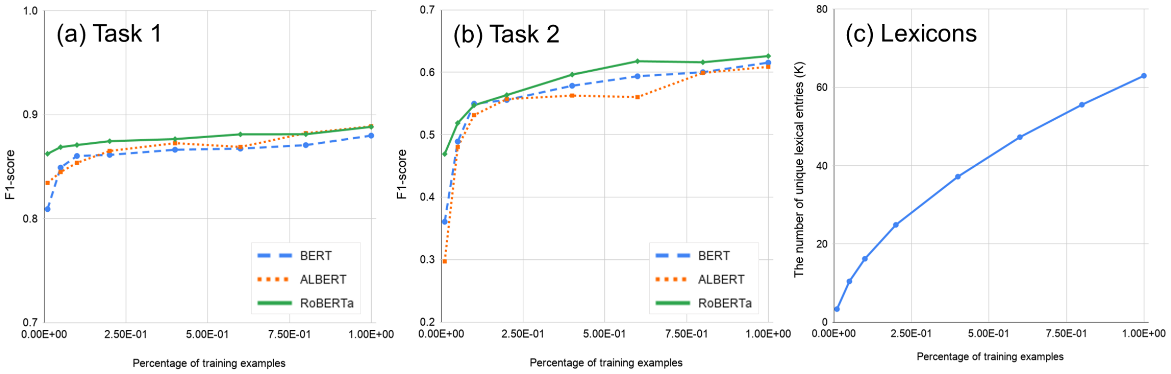


Figure 4: The F1-score ((a) Task 1 and (b) Task 2) and the number of unique lexical entries (c) by the percentage of the training dataset. We tested on a total of eight proportions (1e-02, 5e-02, 1e-01, 2e-01, 4e-01, 6e-01, 8e-01, and 1e+00). All models are fine-tuned on the same steps.

c. RoBERTa (Liu et al., 2019) is a robustly optimized BERT, through pre-training on larger data and careful validation of hyperparameters. We fine-tuned the *roberta-base* model.

The batch size of all models is 32 and fine-tuned for three epochs. We truncated each post at 512 tokens for all models and used a special [SEP] token to concatenate a comment and a context. For each model and task, we manually fine-tuned the learning rates, choosing one out of {2e-5, 3e-5, 4e-5, 5e-5} that shows the best performance of weighted averaged F1-score on the CADD validation set. We fine-tuned our models on two 32GB Nvidia V100 GPUs, taking about 20 minutes for three epochs. We report the median result over five randomly-initialized runs on the CADD test set from the same pre-trained checkpoint.

6.2 Experimental results and analysis

The experimental results of all baseline models for all tasks are shown in Table 5. We experimented with six baselines, with or without context for each of the three models, respectively. Generally, all baseline models achieved quite stable performance, to endorse that our dataset is adequate for abusive language detection on both tasks.

The result shows that three transformer models outperform the dictionary-based models. It suggests that our dataset is not lexically biased, requiring a more complex model to solve its problem than simple dictionary-based models. The result also shows that the models with contextual information marginally outperform the models without it. It implies that the context may have affected crowdworkers during annotation (Pavlopoulos et al., 2020). It also suggests that the contexts of our dataset provide informative clues to abusiveness, which is also explored in Section 5.2. The performance gain due to context is more striking when conducted on Task 2. The fine-grained detection task may require more abundant information such as context, suggesting that the performance gain may become more significant if more powerful detection models are engaged.

We also investigated how the model performance is increased with different percentages of the training dataset (Figures 4a and 4b). It shows that using 20% of our training data leads to a steep improvement, which gains a marginal increment afterwards. It implies that our dataset includes a sufficient number of samples for training models. In addition, even though it is a marginal increment, the performance is improved continuously, especially for Task 2 (from 0.56 to 0.62 for the BERT case). We assume that the increased lexical diversity, shown in Figure 4c, might be one of the possible reasons for such improvement[9].

It is noted that our experiments involve just two tasks, so that if we build a model where all of the labels are combined effectively, it would help many downstream tasks on our dataset. Designing such a model, however, requires a more rigorous examination of each label. We leave it for future work.

7 Conclusion

In this paper, we presented the *Comprehensive Abusiveness Detection Dataset*, covering a broad range of aspects of abusive language. To this end, we designed a hierarchical annotation scheme to allow crowdworkers to annotate multifaceted labels, achieving reliable annotation results. We used the dataset against two tasks discriminating the binary or multiple labels on strong pre-trained NLU models, achieving comparable performance as a baseline, and assuring the practicality of our dataset.

Acknowledgements

This work was supported by Institute for Information and communications Technology Promotion (IITP) (No. 20180005820041001, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection), and the National Research Foundation of Korea (NRF) (No. 2020R1A2C101075912, A system for offensive language detection and automatic feedback with correction) grant funded by the Korea government.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6193–6202.

Michael Castelle. 2018. The linguistic ideologies of deep abusive language classification. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 160–170.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Niels Hellinga and Vlado Menkovski. 2019. Hierarchical annotation of images with two-alternative-forced-choice metric learning. *arXiv preprint arXiv:1905.09523*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Hans Kamp and Barbara Hall Partee. 2004. Context-dependence in the analysis of linguistic meaning.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43.

[9] We used spaCy (https://spacy.io/) to count the number of unique lexical entries.

Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*.

Ted Pedersen. 2019. Duluth at SemEval-2019 Task 6: Lexical Approaches to Identify and Categorize Offensive Tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 593–599.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, volume 2006, pages 1–6.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversation. *arXiv preprint arXiv:2010.07410*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. *arXiv preprint arXiv:1703.10090*.

Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. *arXiv preprint arXiv:2004.04216*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

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, Maja Geulig, and Josef Ruppenhofer. 2021. Implicitly abusive comparisons–a new dataset and linguistic analysis. In *Proceedings of The 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 358–368.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natrual Language Processing (KONVENS 2018)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, pages arXiv–1910.

Wonsuk Yang, Ada Carpenter, and Jong C Park. 2019. Nonsense!: Quality control via two-step reason selection for annotating local acceptability and related attributes in news editorials. 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 2947–2956.

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. *arXiv preprint arXiv:1902.09666*.