Towards Context-Based Violence Detection: A Korean Crime Dialogue Dataset

Minju Kim[*]
Sogang University, Korea
mjmjkk0307@sogang.ac.kr

Heui-Yeen Yeen[*]
LG AI Research
heuiyeen214@lgresearch.ai

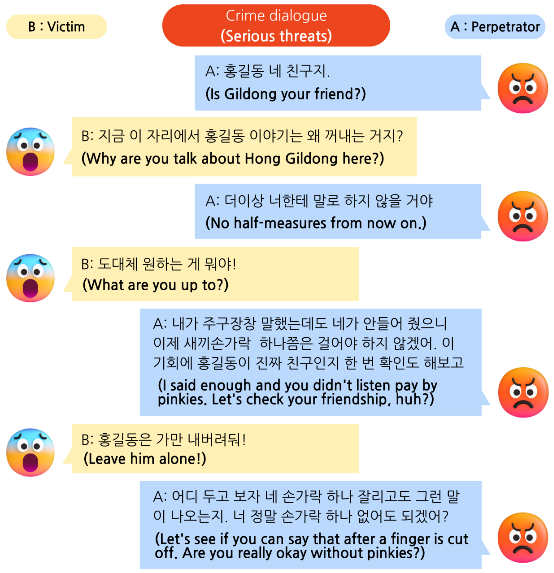
Myoung-Wan Koo†
Sogang University, Korea
mwkoo@sogang.ac.kr

Abstract

In order to enhance the security of society, there is rising interest in artificial intelligence (AI) to help detect and classify in advanced violence in daily life. The field of violence detection has introduced various datasets, yet context-based violence detection predominantly focuses on vision data, with a notable lack of NLP datasets. To overcome this, this paper presents the first Korean dialogue dataset for classifying violence that occurs in online settings: the Korean Crime Dialogue Dataset (KCDD). KCDD contains 22,249 dialogues created by crowd workers assuming offline scenarios. It has four criminal classes that meet international legal standards and one clean class (*Serious Threats, Extortion or Blackmail, Harassment in the Workplace, Other Harassment, and Clean Dialogue*). Plus, we propose a strong baseline for the proposed dataset, Relationship-Aware BERT. The model shows that understanding varying relationships among interlocutors improves the performance of crime dialogue classification. We hope that the proposed dataset will be used to detect cases of violence and aid people in danger. The KCDD dataset and corresponding baseline implementations can be found at the following link: https://sites.google.com/view/kcdd.

1 Introduction

In the pursuit of bolstering societal security, an increasingly prominent focus has emerged on harnessing the potential of artificial intelligence (AI) for the identification and categorization of sophisticated forms of aggression in everyday scenarios (Blanes i Vidal and Kirchmaier, 2017). In particular, AI is effective in discovering and preventing various forms of harm, as it can automate violence detection, allowing for early-stage awareness and prompt action (Aremu et al., 2022). However, these techniques require high-quality datasets, which are currently in short supply.



A: 홍길동 네 친구지.
(Is Gildong your friend?)

B: 지금 이 자리에서 홍길동 이야기는 왜 꺼내는 거지?
(Why are you talk about Hong Gildong here?)

A: 더이상 너한테 말로 하지 않을 거야
(No half-measures from now on.)

B: 도대체 원하는 게 뭐야!
(What are you up to?)

A: 내가 주구장창 말했는데도 네가 안들어 줬으니 이제 새끼손가락 하나쯤은 걸어야 하지 않겠어. 이 기회에 홍길동이 진짜 친구인지 한 번 확인도 해보고
(I said enough and you didn't listen pay by pinkies. Let's check your friendship, huh?)

B: 홍길동은 가만 내버려둬!
(Leave him alone!)

A: 어디 두고 보자 네 손가락 하나 잘리고도 그런 말이 나오는지. 너 정말 손가락 하나 없어도 되겠어?
(Let's see if you can say that after a finger is cut off. Are you really okay without pinkies?)

Figure 1: An example from the KCDD dataset. Our dataset was created by crowd workers, featuring conversational scenarios that could occur offline. The example data meets the criteria of the *Serious Threat* class according to the International Classification of Crime for Statistical Purposes (ICCS).

Currently, there are three main branches of application of violence detection, including surveillance of potential threats in offline situation (Mohammadi et al., 2016; Kamijo et al., 2000; Gao et al., 2016; Kooij et al., 2016; Datta et al., 2002), automatic prevention of harmful media (Vasconcelos and Lippman, 1997; Nam et al., 1998; Dai et al., 2015; Martinez et al., 2019; Singh et al., 2019; Martinez et al., 2020), and monitoring of language toxicity (Blodgett et al., 2020; Nangia et al., 2020; Wallace et al., 2019) to prevent its use in online forums or Large Language Models (LLM) (Brown et al., 2020; OpenAI, 2023; Narang and Chowdhery, 2022; Kim et al., 2021) generation. However, currently, the publicly available datasets are con-

[*] These authors contributed equally to this work.
[†] Corresponding Author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 603–623
March 17-22, 2024 ©2024 Association for Computational Linguistics

Dataset	Lang.	# Inst.	Data Source	Criteria	Context	Toxicity Labels
TCCC (AI, 2018)	Eng	310,387	Wikipedia comments	regional	No	Hate speech, Offensive
Implicit Hate (ElSherief et al., 2021)	Eng	22,584	Twitter	regional	No	Hate speech, Biased
BEEP! (Moon et al., 2020)	Kor	9,341	News comments	regional	No	Hate speech, Biased
HateScore, Unsmile (Kang et al., 2022)	Kor	31,195	News, online community comments	regional	No	Hate speech, Profanity
APEACH (Yang et al., 2022)	Kor	3,770	**Human-written**	regional	No	Offensive
KoSBI (Lee et al., 2023)	Kor	34,214	LM-generated	regional	**Yes**	Biased, Other
KCDD (Ours)	Kor	22,249	**Human-written**	**global**	**Yes**	Offensive, Biased, other

Table 1: Comparison of NLP toxicity datasets

centrated on vision datasets, and the publicly available NLP datasets rarely contain contextualized conversations, especially in offline settings. Therefore, there is a need for publicly available datasets for context-based violence detection.

We present the Korean Crime Dialogue Dataset(KCDD) to enhance violence detection. KCDD was manuscript by crowd workers, assuming potential real-world offline contexts. Figure 1 shows an example. The dataset includes 22,249 conversational scenarios of four classes of threatening situations that comply with the International Classification of Crime for Statistical Purposes (ICCS) (Bisogno et al., 2015) and one class of general conversations, enabling the detection of violence in dialogue situations. To ensure data collection and review is based on strict quality control, we provide a protocol for data gathering and control guarantees for generative datasets, which requires detailed data analysis and collaboration with legal experts. Moreover, we release Relationship-aware BERT, a robust baseline model for our dataset, which presents a methodology to enhance performance by comprehending the characteristics of conversations. Our main contributions are summarized as follows :

- We present KCDD, an NLP dataset that can be utilized in context-based violence detection. This dataset can complement areas not covered in the existing violence detection datasets and be used for international statistics as it adheres to the ICCS international standards. It consists of 22k conversations categorized into five classes.

- Rather than a simple annotation process, we propose a protocol for generating data named *Legal Expert Collaborative Data Building Process*. This protocol elaborates on the collection and legal-expert review of data.

- We also present the Relationship-Aware BERT. It is a speaker type-reflective model, which not only improves the performance on KCDD but also aids in understanding conversation-based data.

2 Related Work

This study bridges two categories of datasets: violence detection datasets and dialogue comprehension datasets. It is necessary to understand both aspects of these datasets because the primary objective of our dataset is to comprehend and detect violence in conversations. In this paper, *violence* encompasses a range of phenomena including acts of physical violence and expressions of hate.

2.1 Violence Detection Dataset

There are previous datasets designed to detect and prevent real-world violence, automatically detect harmfulness in media content, and predict toxicity in language usage. While there are image datasets and technologies for detecting anomalies like abuse in surveillance videos using CCTV data (Sultani et al., 2018; Boekhoudt et al., 2021). Additionally, for detecting harmful content, including those that annotate harmful situations or biases in image datasets or movie scripts (Edstedt et al., 2022; Singh et al., 2022). However, no publicly released language-based datasets exist for similar purposes. Also, existing datasets for detecting harmful media have not been annotated at the conversational level, reflecting the context.

Other NLP violence detection datasets are mostly publicly available to measure text toxicity in language usage (AI, 2018; ElSherief et al., 2021; Moon et al., 2020; Kang et al., 2022; Yang et al., 2022; Bourgeade et al., 2023; Lee et al., 2023). Table 1 summarizes NLP datasets related to violence detection. As shown in Table 1, existing datasets related to violence or hate speech often overlook the context. They tend to focus on identifying expressions of hate in isolated lines of text rather than in a conversational setting. Additionally,

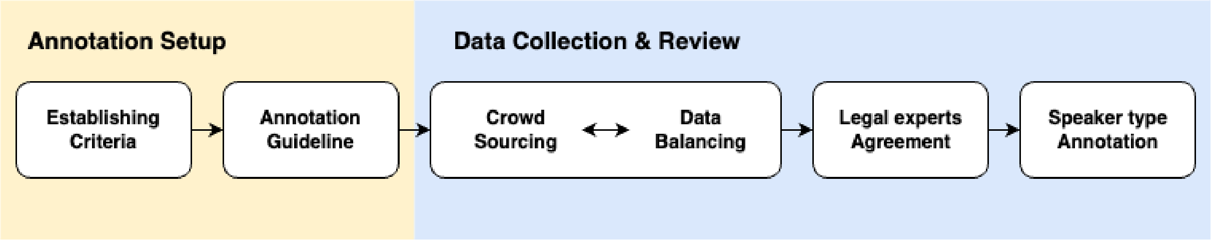


Figure 2: Diagram of the Legal Expert Collaborative Data Building Process for KCDD.

these datasets follow regional criteria and primarily concentrate on toxic situations occurring in online environments. This observation underscores the need for datasets that encompass a broader range of scenarios, including offline contexts and global perspectives. Therefore, we introduce KCDD, a dataset that meets these criteria. Our dataset is manually curated by crowd workers and legal expert, adheres to international standards, and incorporates conversational contexts, filling a significant gap in current data resources.

2.2 Dialogue Comprehension Dataset

Dialogue comprehension encompasses tasks such as reading comprehension, classification, and summarization of conversation content. Due to the distinct characteristics of conversational text compared to general text, specialized datasets for performing such conversation-based tasks have been released (Sun et al., 2019; Cui et al., 2020; Zhao et al., 2022; Chen et al., 2021). As shown in these datasets, dialogue data has structural and content differences from general text, requiring consideration of speaker turns, discourse structure, common sense, and colloquial language. Therefore, additional dialogue datasets are needed, especially for PLMs, which are primarily trained on formal written text and may not understand colloquial language well. Our dataset was created in response to the need for dialogue datasets, particularly in the context of toxicity classification, and the lack of dialogue-based datasets reflecting discourse structure or conversational context in Korean.

3 The KCDD Dataset

In this section, we describe the data construction protocol named *Legal Expert Collaborative Data Building Process*. The entire process can be seen in Figure 2. Furthermore, we examine the statistics, and characteristics of the constructed data.

3.1 Legal Expert Collaborative Data Building Process

3.1.1 Critieria Estabilishment

Firstly, we define data classification criteria following ICCS, the international criteria published by the United Nations Office on Drugs and Crime (UNODC) to obtain international consistency of crime statistics. KCDD's crime-related classes adhere to the ICCS, and along with one general conversation class, comprise a total of five classes. The specific crime class definitions are as follows:

- *Serious Threats* with the ICCS code 020121 is when a person threatens someone with the intention of inflicting death or serious harm.
- *Extortion or Blackmail* with the ICCS code 02051 signifies acts that demand certain behavior through a written or verbal threat. Here, certain behavior should involve, at a minimum, deprivation of property or money and provision of services or benefits.
- *Harassment in the Workplace* with the ICCS code of 020811 means harassment by a colleague, supervisor, or other co-workers in a work environment or related to employment.
- *Other Harassment* with the ICCS code of 020819 means harassment, not in a work environment and unrelated to employment. The dataset includes a variety of harassment cases, containing physical or verbal violence, bullying, belittling of looks, personal offense, abuse of power by a customer, etc.

Among several categories of ICCS, we collected data that narrowed down to four crime categories that are relatively probable in daily life and deemed necessary for prevention, in consultation with legal experts.

3.1.2 Annotation Guidelines

As it is not a simple tagging task, but rather a complex task that requires crowd workers to create text scenarios themselves, careful efforts were made to make detailed guidelines. We provided crowd workers with class names and instructed them to write fictional conversational scenarios that could occur in offline situations, corresponding to those classes. First, We explained five class definitions that fit the ICCS criteria. For each class, more than 10 specific example situations and two example dialogues in the same format as the ones crowd workers have to write were given to help workers understand. Provided example elaborates to clarify some of the more confusing points of data creation in line with the legal standard. Appendix A gives examples of guideline for crowd workers.

3.1.3 Crowd Sourcing

We crowdsourced for the creation of our dataset, where crowd workers developed scenarios for five conversation types. Each type had an equal number of conversations written. To better manage this process, the data collection through crowdsourcing was conducted in two stages. The first stage involved university students, while the second stage was outsourced to corporations specializing in crowdsourcing. Table 2 presents statistics on crowdsourcing and figure 3 shows the demographic composition of the crowd workers. The first round targeted university students, resulting in a higher representation of individuals in their twenties, while the second round recruited a broader age range of workers. In both rounds, there were more male than female participants. Efforts were made to balance the gender ratio of crowd workers; however, some imbalance was inevitable due to the recruitment of participants who were fully aware and consented to the context of producing violent conversations. Crowd workers were compensated $1,000$ KRW, approximately equal to 1 US dollar, for creating each dialogue data. Additionally, to ensure the psychological safety of workers creating the violent conversation dataset, we limited the number of dialogues that could be created daily to 30 and established a process for psychological counseling in association with schools. The process of establishing the guidelines and crowd-sourcing the data, including the first and second rounds, took about six months.

	First round	Second round
# of workers participated	50	55
Total submitted dialogues by workers	9,749	12,500
Average # of dialogues by one worker	194.98	227.27
Max # of dialogues written by one worker	500	600

Table 2: Statics for crowdsourcing KCDD dataset.

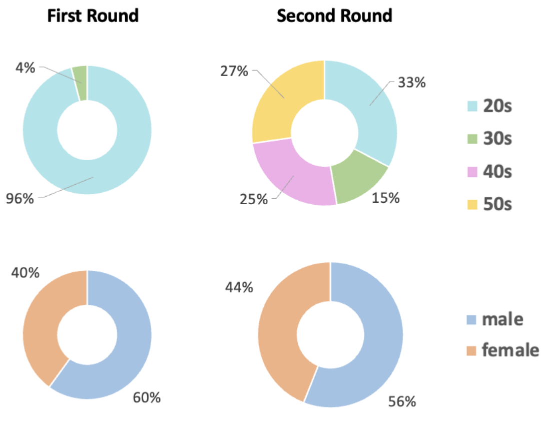


Figure 3: Demographic Composition of the Crowd Workers.

3.1.4 Data Balancing

Quantitative Balance : To balance the number of data across all classes, we instructed crowd workers to submit an even number of entries for each class from the onset. For instance, if a crowd worker created 100 pieces of data, they created 20 examples for each of class. After all data was submitted, it underwent a review process by legal experts as outlined in §Section 3.1.5, involving data review, re-annotation, and removal of irrelevant data. The resulting data statistics, as shown in the Table 3, demonstrate that the data was collected almost equally across all classes.

Qualitative Balance : We asked crowd workers to write at least 10% of adversarial data, that intentionally contains words frequently appearing in other classes. This is to prevent certain words from appearing too frequently in only a few classes. For example, "kill" in the *Serious Threats* class, property-related words in the *Extortion or Blackmail* class, and words denoting the workplace in the *Harassment in the Workplace* class appeared particularly often. In this case, the model may overfit certain words when performing the classification

Class	# of dialogue
Serious Threats	4,024
Extortion or Blackmail	4,219
Harassment in the Workplace	4,562
Other Harassment	4,566
Clean Dialogue	4,878
Total	22,249
Percentage of Std per class	1.34

Table 3: Class distribution of the dataset.

# of utterance	178,991
# of words	1,307,678
Min turns per dialogue	3
Max turns per dialogue	32
Avg turns per dialogue	8
Avg words per utterance	7,3

Table 4: Statics for the entire dataset.

task rather than the context itself. Therefore, we deliberately put dialogues like "you are killing it!" in *Clean Dialogue* so that the word "kill" can be distributed to other classes besides *Serious Threats* class. The generated adversarial data to prevent this is shown in Appendix E.

3.1.5 Legal Experts Agreement

After creation of data by crowd-socured workers, the legal team exmanied every created samples. Four legal team members reviewed each class-annotated conversation written by the crowd workers to examine if the data needed to be re-annotated, modified, or deleted. During this process, they decided final label by majority vote. Also, they removed data that could cause bias or personal information infringement based on the law. This process aimed to generate data aligned to the ICCS code and proactively review ethical issues that may arise in crowdsourcing.

3.1.6 Speaker Type Annotation

Following the completion of dialogue data creation and review, we annotated the speaker type with the goal of better reflecting the characteristics of the dialogues in our dataset. This process, conducted by the authors, involved tagging speakers as perpetrator, victim, or normal person, based on the predominance of violent situations in the dialogues. This was the final step in the data collection process, taking a total of one year, and as a result, our dataset now includes both conversation level and speaker type annotations.

# of speakers who start dialogue		
Perpetrator	Victim	Normal person
17,057	1,731	3,461
# of speakers who close dialogue		
Perpetrator	Victim	Normal person
12,297	6,237	3,715

Table 5: The number of speakers who start and close the dialogue

Class	**# of dialogue (interlocutors >2)**
Serious Threats	534
Extortion or Blackmail	409
Harassment in the Workplace	656
Other Harassment	832
Clean Dialogue	510

Table 6: The number of dialogues where the number of interlocutors is greater than two.

Class	**P&V**	**P**	**P&V&N**
Serious Threats	3,687	147	102
Extortion or Blackmail	3,967	32	42
Harassment in the Workplace	3,909	273	174
Other Harassment	3.637	479	109
Clean Dialogue	448	73	20

Table 7: The number of dialogues with relationship combinations; P is for the perpetrator, V is for the victim, and N is for the normal person.

3.2 Dataset Analysis

3.2.1 Statistics

KCDD is a dataset containing dialogues that belong to one of five classes: *Serious Threats, Extortion or Blackmail, Harassment in the Workplace, Other Harassment* and *Clean Dialogue*. The dataset consists of a total of 22,249 dialogues and train/dev/test data is split into 17,799/2,225/2,225. The distribution of data by class can be seen in Table 3. Additionally, the statistics for the entire dataset are shown in Table 4.

3.2.2 Analysis of Relationships between Speakers in Dialogues

Our dataset contains conversations about criminal situations. Therefore, the dialogue features characters such as the perpetrator, the victim, or a normal person. Moreover, the relationship between these characters significantly influences the overall context of the conversation. For instance, the perpetrator leads the dialogue by uttering threats or harassing, so that the conversation openers and closers mostly come from perpetrators as described

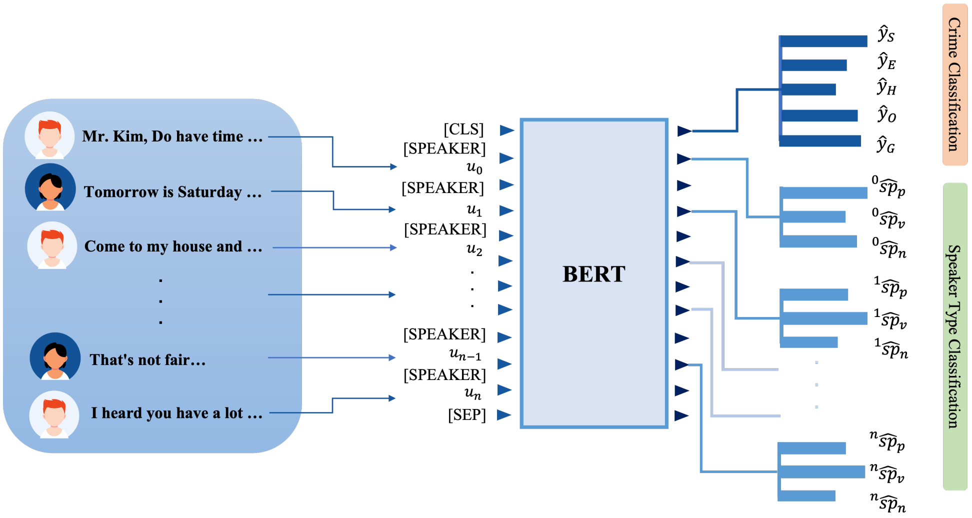


Figure 4: Relationship-Aware BERT for KCDD.

in table 5. Each class shows slightly different types of speakers, along with the relationships between them. In Table 6, there are more participants in the *Harassment in the Workplace* class and *Other Harassment* class than others. Additionally, in these classes, a more diverse combination of relationships appears compared to other classes. In other words, among the participants in the conversation, the combinations of perpetrator, victim, and normal person are more varied. (Table 7). This is because circumstances revolving around the workplace, school, or conversations between friends include more people and a greater probability of having a normal person who is not directly related.

3.2.3 Analysis of Dialogue Structure

The dialogues within our dataset are meticulously crafted to have well-structured plots, as described by (Egan, 1978). Each dialogue has a central incident corresponds to the designated class label. For instance, in the *Extortion or Blackmail* category, the narrative starts with the perpetrator intimidating the victim, followed by the victim's response, culminating in the act of extortion and the victim's subsequent loss. This well-structured plot distinctly sets KCDD apart from traditional conversational datasets and those aimed at detecting toxicity without defined context, commonly found in the NLP community. The dialogues in KCDD are characterized by their clearly articulated story arcs, revolving around pivotal incidents in each conversation. Further elaboration on this distinction is available in Appendix C.

4 Relationship-Aware BERT for KCDD

We propose the strong baseline for KCDD to classify dialogues according to the crime situation. We consider this task not simply text classification but a dialogue comprehension task that requires understanding context. Therefore, we exploit methods for models to learn the characteristics of the dialogue format.

To this end, we introduce Relationship-Aware BERT, a multi-task Transformer model (Radford et al., 2018) that is jointly trained for crime dialogue classification, as well as classifying types of interlocutors. We used KLUE-BERT (Park et al., 2021b) a model that was further pretrained in Korean using BERT as a backbone. Figure 4 shows the proposed model. We use two types of special tokens to learn two tasks jointly: $[CLS]$ token of BERT (Devlin et al., 2018) for classifying crime dialogue situation, and predefined special $[SPEAKER]$ token for classifying the type of interlocutors (perpetrator, victim, normal person).

Consider the entire dialogue data $D = \{d_0, d_1, \ldots, d_t\}$ where t represents the total number of dialogue, and each dialogue data $d = \{u_0, u_1, \ldots, u_n\}$ comprises individual utterances u. For constructing the input of the proposed model, $[CLS]$ token and $[SEP]$ token are appended at the beginning and end of each dialogue d respectively. The special token $[SPEAKER]$ is prepended to each utterance to identify the type of speaker for each utterance. Therefore, the input of Relationship-Aware BERT is as follows:

$$x = \{[CLS], [SPEAKER], u_0, [SPEAKER], u_1, \ldots, [SPEAKER], u_n, [SEP]\}$$

The number of [SPEAKER] tokens is equal to the number of utterances. To distinguish the speaker type, each $[SPEAKER]$ token goes through randomly initialized a multi-layer perceptron (MLP) layer. Next, followed by a softmax function (Goodfellow and Courville, 2016), the probability of a speaker type (perpetrator, victim, normal) ${}^{i}\widehat{sp_p}, {}^{i}\widehat{sp_v}, {}^{i}\widehat{sp_n} \in \mathbb{R}$ is predicted for each utterance.

To classify the crime situation, the $[CLS]$ token is sequentially passed through the MLP layer and softmax function. Finally, the probability of five classes *(Serious Threat, Extortion or Blackmail, Harassment in the Workplace, Other Harassment, Clean Dialogue)* $\hat{y}_S, \hat{y}_E, \hat{y}_H, \hat{y}_O, \hat{y}_C \in \mathbb{R}$ is predicted for a dialogue.

For loss of classifying the type of speaker, we employ cross-entropy loss between the predicted probability ${}^{i}\widehat{sp}$ and the ground truth ${}^{i}sp$ according to each $[SPEAKER]$ token. Adding all the values of the loss on each $[SPEAKER]$ token, the final ℓ in a dialogue is obtained.

$$\ell_{relationship} = -\sum_{i} {}^{i}sp \log {}^{i}\widehat{sp} \tag{1}$$

Similarly, the loss for crime situation classification is obtained by taking the cross-entropy loss between the predicted probability $\hat{y}$ and ground truth y on $[CLS]$ token in the data.

$$\ell_{crime} = -\sum y \log \hat{y} \tag{2}$$

Finally, the multi-task loss is composed as Equation 3. λ is a hyper-parameter, controlling the ratio of two losses.

$$\mathcal{L} = \ell_{crime} + \lambda \cdot \ell_{relationship} \tag{3}$$

Basically, λ was set to 1 so that both losses could be appropriately reflected. The effect of λ is described in Appendix **??**.

Exploiting multi-task learning, performance is improved for both tasks. This is because the classification tasks exchange signals with each other to comprehend the whole context of a dialogue during model training.

5 Experiments

Considering the characteristics of KCDD, we explored several methodologies to properly reflect the conversational context in classifying crime situations. Therefore, we compared the proposed model, Relationship-Aware BERT, with other methods.

5.1 Baselines

We compre the proposed method to five linear classification models and one multi-task classification model.

- **LSTM :** Applying a multi-layer long short-term memory RNN (Luan and Lin, 2019; Hochreiter and Schmidhuber, 1997) to an input sequence with bag-of-words vocab.

- **Dialogue TF-IDF+SVM :** A dialogue-level multi-class linear Support Vector Machine (Hearst et al., 1998) with vectorized Tf-IDF bag-of-words.

- **KLUE-BERT :** KLUE BERT base is a pre-trained BERT Model on Korean Language. The developers of KLUE BERT base developed the model in the context of the development of the Korean Language Understanding Evaluation (KLUE) Benchmark (Park et al., 2021a). Inputs are composed of sequentially concatenated all the utterances in a dialogue.

- **KLUE -BERT with Speaker embedding :** A fine-tuned KLUE-BERT model with speaker embeddings, exploiting proposed method (Gu et al., 2020). When the speaker changed in a dialogue text, the model distinguishes the speaker's turn by 0 and 1 with speaker embeddings added to model input sequences. This model, unlike the one we propose, reflects only turn changing between speakers.

- **KLUE-BERT with supervised attention :** A fine-tuned KLUE-BERT model trained by supervising the model's attention values, utilizing proposed method (Stacey et al., 2022). The methodology described involves enhancing classification performance by supervising the attention values of tokens defined as important during the training of the model. We supervised the model for higher attention value on the perpetrator's utterance.

- **AT-BMC :** A joint classification and rationale extraction model proposed by Li et al. (2022).

Crime Classification Model (Single Task)			
Method	**Metric**		
	ACC	**F1**	
LSTM	63.6	64.0	
Dialogue TF-IDF	79.6	79.6	
KLUE-BERT	84.3	82.1	
KLUE-BERT w/SE	86.3	86.2	
KLUE-BERT w/SMA	86.5	86.8	
Multi-task Learning Model			
Method	Metric		
	ACC	F1	Token F1
AT-BMC	79.7	79.7	**74.6**
Ours	**88.0**	**88.0**	**74.6**

Table 8: Results of Crime Classification Model (Single Task) and Multi-task Learning Model. In multi-task learning, accuracy and macro f1 score are adapted for the crime classification task, and speaker type classification of speaker type task is measured as token f1.

Method	Crime classification		Speaker type classification
	Acc	F1	Token F1
(a) grouping utterances by the speaker	86.8	86.8	**84.3**
(b) each utterance	**88.0**	**88.0**	74.3

Table 9: Comparison of two input methods.

It can yield accurate predictions and provide closely-related extractive rationales as potential reasons for predictions. In this experiment, the model is jointly trained to classify criminal situations and extract utterances of perpetrators as the rationale. We also adapted the same pretrained model.

5.2 Experiment Settings

Metrics We measure accuracy and the macro f1-score to compare the crime dialogue classification performances of different models. For the speaker type classification task, we measure the token f1 score. For fair comparision, we evalute all models in four different seeds and reported averaged result.

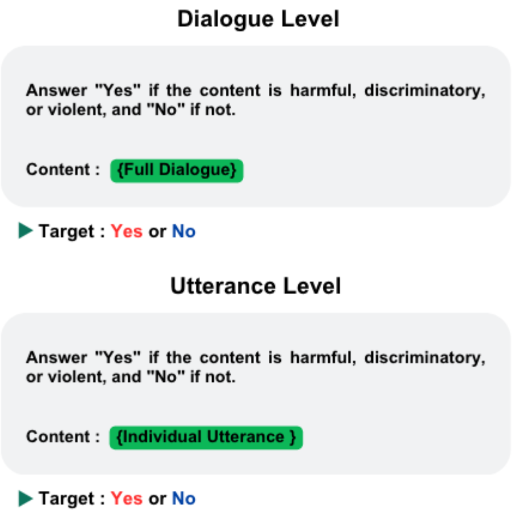
Hyper-parameters We used PyTorch (Paszke et al., 2019) for the model implementation. We set the AdamW optimizer (Kingma and Ba, 2014) as the optimizer, 32 as the train batch size, 5e-5 as the learning rate, and 256 as the max sequence length. The GPU used for training is a single NVIDIA RTX A5000 24G.

Results Table 8 shows the performance of Relationship-Aware BERT and other baseline models. Relationship-Aware BERT scored the best in the crime dialogue classification task. The result represents that understanding relationships among interlocutors helps detecting and classifying criminal situations. Comparing among models only leant crime classification, adding speaker embedding improves the model performance compared to the vanilla KLUE-BERT model. Also, supervising the model to get a higher attention value on the perpetrator's utterance contributes better to improving performance than simply distinguishing the speaker. AT-BMC can solve two tasks simultaneously but has decreased performance. For crime classification, it seems that detecting the perpetrator's utterance on just a token is not very useful. In contrast, Relationship-Aware BERT, which classifies the speaker's type, has the highest score. It represents identifying speaker type based on an utterance helps to increase performance on crime classification.

6 Discussion

Influence of Input Format for Learning Speaker Relationships We experimented with various input formats to find the most efficient way to predict the relationship between speakers. We compared two methods: (a) grouping utterances by the speaker and adding [$SPEAKER$] tokens in front of the group so that tokens appear equal to the number of speakers. (b) adding [$SPEAKER$] in front of each utterance. Appendix F gives examples of the input (a) and (b) and the results of the crime classification task. Table 9 shows a higher score with method (b). When utterances are grouped by the speaker as method (a), the story structure in dialogue is broken, resulting in performance degradation. However, since utterance is concatenated for each speaker, the speaker type classification becomes easier, and speaker classification performance is improved. In summary, since the entire context is considered during the multi-task learning, method (b) seems to have been learned more effectively. Thus, the Relationship-Aware BERT ultimately reported its performance using method (b).

Analysis of LLM's Violence Detection Ability on Contextual Data We experimented with having LLM classify whether a conversation in our dataset is violent or not. Then, we sample 50 dialogues that LLM misclassified and analyzed them. 50 sample



Dialogue Level

Answer "Yes" if the content is harmful, discriminatory, or violent, and "No" if not.

Content : {Full Dialogue}

▶ Target : Yes or No

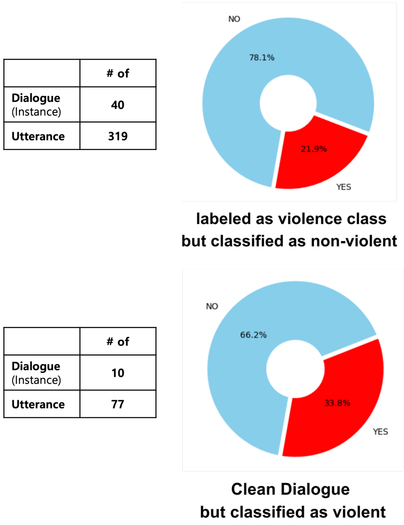
Utterance Level

Answer "Yes" if the content is harmful, discriminatory, or violent, and "No" if not.

Content : {Individual Utterance }

▶ Target : Yes or No

Figure 5: Examples of Prompts for LLM's Violence Detection Ability Experiment. The part in bold is the template for the prompt, the part highlighted in green is the respective input, and the output is either yes or no.



	# of
Dialogue (Instance)	40
Utterance	319

labeled as violence class but classified as non-violent

	# of
Dialogue (Instance)	10
Utterance	77

Clean Dialogue but classified as violent

Figure 6: Pie chart of violence detection at the utterance level for a dialogue labeled as violence class but classified as non-violent (left) and a *Clean Dialogue* but classified as violent (right).

dialogues consists of 40 dialogues labeled as crime class but LLM classified as non-violent and 10 dialogues labeled as a *Clean Dialogue* class but classified as violent. The analysis involved assessing violence detection utterance level using OpenAI's `GPT-3.5-turbo`[1]. We construct prompts accordingly using the Entailment-oriented Instruction approach mentioned in the (Lou et al., 2023; Yin et al., 2019). The prompts used to guide LLM in classifying violence are presented in Figure 5. Also, as shown in Figure 6, the distribution of violent utterances in both dialogues, which are violence label and *Clean Dialogue* class, was similar. These findings imply that while the LLM excels at detecting overt harm within individual utterances (Dixon et al., 2018; Gehman et al., 2020; Zhang et al., 2022; Li et al., 2023; Hartvigsen et al., 2022), but it demonstrates limitations in capturing harm that is context-dependent. We hope future research will address violence classification considering factors like the relationship between participants, offline violence, and situation-based violence.

7 Conclusion

In this paper, we introduced the Korean Crime Dialogue Dataset (KCDD), comprising 22,249 dialogues adhering to the International Classification of Crime for Statistical Purposes (ICCS). We also developed the *Legal Expert Collaborative Data Building Process* for crowd-sourced data creation, ensuring quality through expert collaboration. Moreover, we proposed the Relationship-Aware BERT, demonstrating superior performance on KCDD dataset. We hope that our dataset can be utilized for various context-based violence detection studies.

8 Limitations

International Criteria-based Classification of Violence This dataset is built to classify crimes in the real world according to the International Classification of Crime for Statistical Purposes (ICCS) code. However, it does not encompass all types of crimes that exist in practice. Legal experts we collaborated with selected the five most frequent classes in real life. While the current dataset is limited to these classes, we believe there is potential for expansion using methodologies involving Large Language Models (LLMs). Utilizing LLMs to augment the dataset with examples from other classification codes presents an exciting area of exploration. Therefore, we consider researching methodologies to expand beyond the current limited classes as an intriguing future research topic. We hope future research and datasets will extend to cases that follow other ICCS codes, potentially leveraging LLM capabilities for this expansion.

User Diversity The collected dataset was created by Korean worker and written in Korean, so it has the limitation of potentially reflecting the social culture of Korea more prominently. However, since it

[1] `https://platform.openai.com/docs/models/gpt-3-5`

was built based on the definition of ICCS codes, we anticipate it can be similarly expanded in diverse countries.

Annotation Complexity The ambiguity in the data was partially addressed through the Legal Experts Agreement process. Specifically, cases that either encapsulate all four predefined violence classes or contain violent elements outside these classes were generally excluded. However, it's important to note that instances might still be included if there is a consensus among the majority by legal experts. Consequently, this approach may introduce limitations in interpretation, varying depending on individual legal expert perspectives. This highlights the inherent complexity in annotating data that straddles multiple violence classes or ambiguous situations.

9 Ethics

Managing the Potential Violence in the Dataset Our legal team rigorously reviewed all datasets to identify and rectify any biases. The dataset has been constructed using hypothetical scenarios, ensuring there is no risk of compromising anonymity or leaking personal information. However, it's possible that some discriminatory language remains undiscovered; we are committed to continuously updating and refining our dataset to eliminate such content upon its detection. Note that, due to the inclusion of violent scenarios in the dataset, its use is strictly limited to research purposes related to violence detection and is strictly prohibited for any other application. The KCDD is available for non-commercial use under the custom license CC-BY-NC 4.0.[2]

Managing the Psychological Safety of Crowd Workers We collected our dataset through crowd-sourcing, which involved crowd workers creating the dataset directly, including writing scenarios involving violent situations. Recognizing the potential psychological stress this could cause, we implemented safety measures to manage it. Firstly, we limited the submission to a maximum of 30 dialogues per day to prevent excessive psychological stress. Since our research was conducted by a university research team, we established a process in conjunction with the university's psychological counseling center to provide support for crowd workers in case of any issues. Lastly, we ensured that only those who had received a thorough explanation of the dataset creation and consented to participate were engaged, and we allowed crowd workers to discontinue their participation at any time if they chose to do so. By implementing these measures, we aimed to safeguard the psychological well-being of the crowd workers. We hope that such safety protocols will be considered in future research involving violent situations.

10 Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00621, Development of artificial intelligence technology that provides dialog-based multi-modal explainability)

References

Jigsaw/Conversation AI. 2018. Toxic comment classification challenge identify and classify toxic online comments.

Toluwani Aremu, Li Zhiyuan, Reem Alameeri, Moayad Aloqaily, and Mohsen Guizani. 2022. Towards smart city security: Violence and weaponized violence detection using dcnn. *arXiv preprint arXiv:2207.12850*.

Enrico Bisogno, Jenna Dawson-Faber, and Michael Jandl. 2015. The international classification of crime for statistical purposes: A new instrument to improve comparative criminological research. *European Journal of Criminology*, 12(5):535–550.

Jordi Blanes i Vidal and Tom Kirchmaier. 2017. The Effect of Police Response Time on Crime Clearance Rates. *The Review of Economic Studies*, 85(2):855–891.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Kayleigh Boekhoudt, Alina Matei, Maya Aghaei, and Estefanía Talavera. 2021. Hr-crime: Human-related anomaly detection in surveillance videos. In *Computer Analysis of Images and Patterns: 19th International Conference, CAIP 2021, Virtual Event, September 28–30, 2021, Proceedings, Part II 19*, pages 164–174. Springer.

[2] https://creativecommons.org/licenses/by-nc/4.0/

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696, Dubrovnik, Croatia. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. *arXiv preprint arXiv:2004.04494*.

Qi Dai, Rui-Wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, and Yu-Gang Jiang. 2015. Fudanhuawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning. In *MediaEval*, volume 1436.

A. Datta, M. Shah, and N. Da Vitoria Lobo. 2002. Person-on-person violence detection in video data. In *2002 International Conference on Pattern Recognition*, volume 1, pages 433–438 vol.1.

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*.

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.

Johan Edstedt, Amanda Berg, Michael Felsberg, Johan Karlsson, Francisca Benavente, Anette Novak, and Gustav Grund Pihlgren. 2022. Vidharm: A clip based dataset for harmful content detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1543–1549. IEEE.

Kieran Egan. 1978. What is a plot? *New Literary History*, 9(3):455–473.

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.

Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. 2016. Violence detection using oriented violent flows. *Image and Vision Computing*, 48-49:37–41.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Y.; Goodfellow, I.; Bengio and A Courville. 2016. Deep learning.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

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. *arXiv preprint arXiv:2203.09509*.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi. 2000. Traffic monitoring and accident detection at intersections. *IEEE Transactions on Intelligent Transportation Systems*, 1(2):108–118.

TaeYoung Kang, Eunrang Kwon, Junbum Lee, Youngeun Nam, Junmo Song, and JeongKyu Suh. 2022. Korean online hate speech dataset for multilabel classification: How can social science aid developing better hate speech dataset? *arXiv preprint arXiv:2204.03262*.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. 2021. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. *arXiv preprint arXiv:2109.04650*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

J.F.P. Kooij, M.C. Liem, J.D. Krijnders, T.C. Andringa, and D.M. Gavrila. 2016. Multi-modal human aggression detection. *Computer Vision and Image Understanding*, 144:106–120. Individual and Group Activities in Video Event Analysis.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-Woo Ha. 2023. Kosbi: A dataset for mitigating social bias risks towards safer large language model application. *arXiv preprint arXiv:2305.17701*.

Dongfang Li, Baotian Hu, Qingcai Chen, Tujie Xu, Jingcong Tao, and Yunan Zhang. 2022. Unifying model explainability and robustness for joint text classification and rationale extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10947–10955.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. " hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2023. Is prompt all you need? no. a comprehensive and broader view of instruction learning. *arXiv preprint arXiv:2303.10475*.

Yuandong Luan and Shaofu Lin. 2019. Research on text classification based on cnn and lstm. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)*, pages 352–355. IEEE.

Victor Martinez, Krishna Somandepalli, Yalda Tehranian-Uhls, and Shrikanth Narayanan. 2020. Joint estimation and analysis of risk behavior ratings in movie scripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4780–4790, Online. Association for Computational Linguistics.

Victor R Martinez, Krishna Somandepalli, Karan Singla, Anil Ramakrishna, Yalda T Uhls, and Shrikanth Narayanan. 2019. Violence rating prediction from movie scripts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 671–678.

Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, and Vittorio Murino. 2016. Angry crowds: Detecting violent events in videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 3–18. Springer.

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.

J. Nam, M. Alghoniemy, and A.H. Tewfik. 1998. Audio-visual content-based violent scene characterization. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, volume 1, pages 353–357 vol.1.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Sharan Narang and Aakanksha Chowdhery. 2022. Pathways language model (palm): Scaling to 540 billion parameters for breakthrough performance.

OpenAI. 2023. Gpt-4 technical report.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook 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, Jungwoo Ha, and Kyunghyun Cho. 2021a. Klue: Korean language understanding evaluation.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021b. Klue: Korean language understanding evaluation. arXiv:2105.09680.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Roshni Ramnani, Anutosh Maitra, Shubhashis Sengupta, et al. 2022. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues. *arXiv preprint arXiv:2205.15951*.

Shubham Singh, Rishabh Kaushal, Arun Balaji Buduru, and Ponnurangam Kumaraguru. 2019. Kidsguard: Fine grained approach for child unsafe video representation and detection. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 2104–2111, New York, NY, USA. Association for Computing Machinery.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11349–11357.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

N. Vasconcelos and A. Lippman. 1997. Towards semantically meaningful feature spaces for the characterization of video content. In *Proceedings of International Conference on Image Processing*, volume 1, pages 25–28 vol.1.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

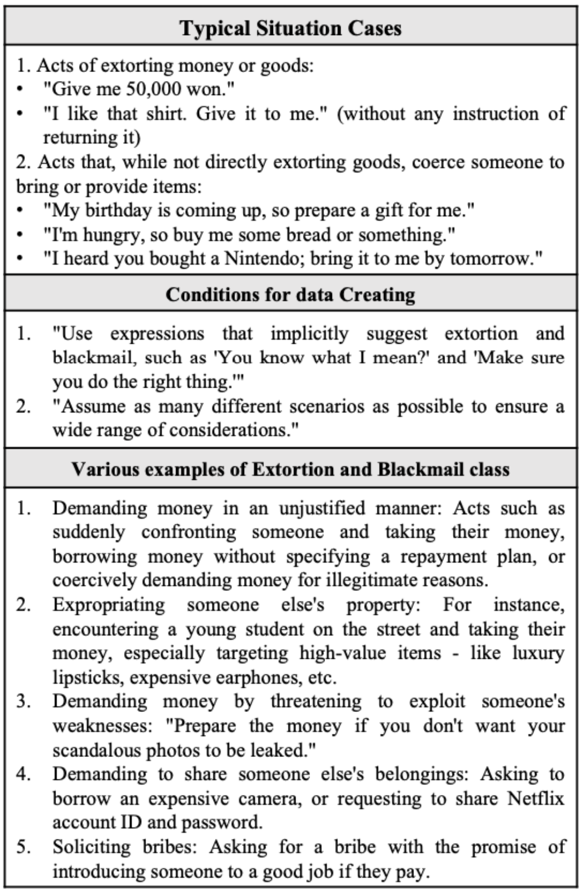
Kichang Yang, Wonjun Jang, and Won Ik Cho. 2022. APEACH: Attacking pejorative expressions with analysis on crowd-generated hate speech evaluation datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7076–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Chao Zhao, Wenlin Yao, Dian Yu, Kaiqiang Song, Dong Yu, and Jianshu Chen. 2022. Learning-by-narrating: Narrative pre-training for zero-shot dialogue comprehension. *arXiv preprint arXiv:2203.10249*.

A Example of Guidelines



Typical Situation Cases

1. Acts of extorting money or goods:
- "Give me 50,000 won."
- "I like that shirt. Give it to me." (without any instruction of returning it)

2. Acts that, while not directly extorting goods, coerce someone to bring or provide items:
- "My birthday is coming up, so prepare a gift for me."
- "I'm hungry, so buy me some bread or something."
- "I heard you bought a Nintendo; bring it to me by tomorrow."

Conditions for data Creating

1. "Use expressions that implicitly suggest extortion and blackmail, such as 'You know what I mean?' and 'Make sure you do the right thing.'"
2. "Assume as many different scenarios as possible to ensure a wide range of considerations."

Various examples of Extortion and Blackmail class

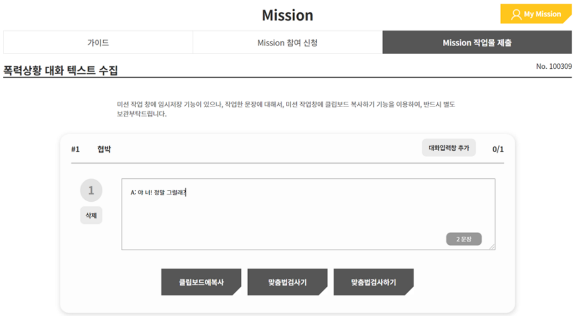
1. Demanding money in an unjustified manner: Acts such as suddenly confronting someone and taking their money, borrowing money without specifying a repayment plan, or coercively demanding money for illegitimate reasons.
2. Expropriating someone else's property: For instance, encountering a young student on the street and taking their money, especially targeting high-value items - like luxury lipsticks, expensive earphones, etc.
3. Demanding money by threatening to exploit someone's weaknesses: "Prepare the money if you don't want your scandalous photos to be leaked."
4. Demanding to share someone else's belongings: Asking to borrow an expensive camera, or requesting to share Netflix account ID and password.
5. Soliciting bribes: Asking for a bribe with the promise of introducing someone to a good job if they pay.

Figure 7: An example of guideline for the *Extortion and Blackmail* class.

Figure 7 represents a guideline for the *Extortion and Blackmail* class that was offered to crowd workers. The guideline includes representative cases and examples of criminal situations according to the class. Also, we provide conditions for data creation. Referring to various examples in the guidelines, crowd workers created virtual criminal situation dialogue data.

B Crowdsourcing Statistics and Data Annotating Tools UI/UX

Figure 8 gives screenshot of the data annotation tool given to crowd workers. The first round of crowd workers were university students, and the second round was outsourced to a crowdsourcing company so that individuals of all genders and ages could complete the data. We selected crowdsourcing company[3] with convenient UI/UX data annotation tools, because it is a crucial factor affecting data quality.



폭력상황 대화 텍스트 수집

미션 작업 창에 임시저장 기능이 있으나, 작업한 문장에 대해서, 미션 작업창에 클립보드 복사하기 기능을 이용하여, 반드시 별도 보관부탁드립니다.

A: 야 너! 정말 그럴래?

Figure 8: UI/UX of Data Authoring Tools

The process of creating data using this annotation tool by crowd workers is as follows. 1) Checking class name of the data. 2) Writing dialogue data according to the class, assuming a criminal situation or a clean conversation. The data is created in accordance with the format (i.e. A: utterance 1, B: utterance 2, A: utterance 3. . .), assigning different alphabets to each speaker. 3) After finishing writing a dialogue, workers checked the number of sentences, so that the data was not too short or too long. 4) Through the spelling checker, it was possible to correct the spelling error. 5) When data was submitted, it was automatically changed to excel format so that it could be provided to the examinee.

C Comparison with Dialogue Data and Online Toxic Data

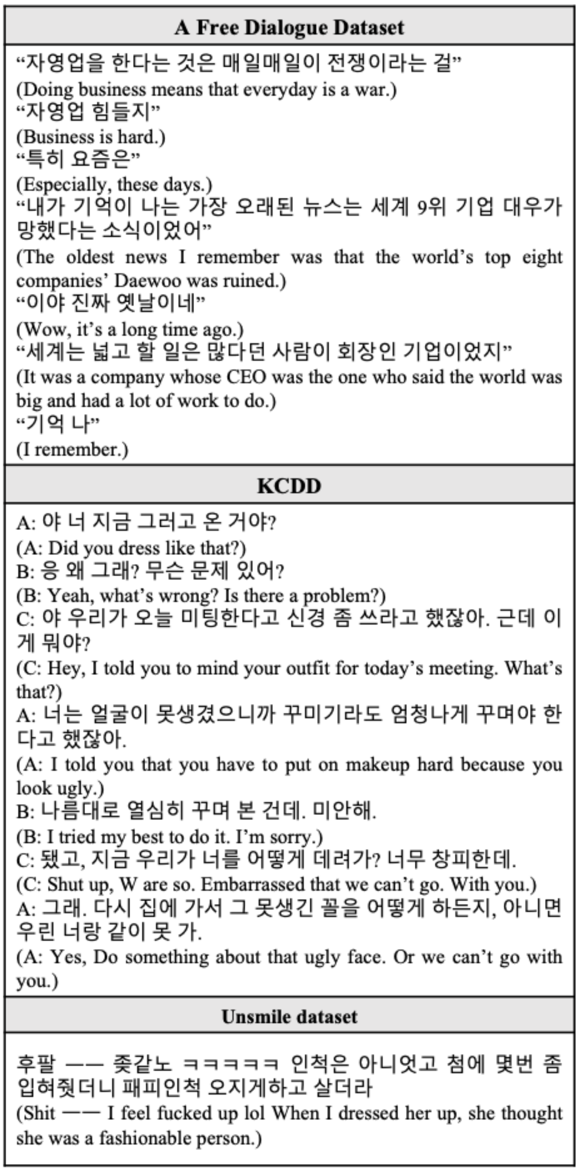
KCDD has a face-to-face dialogic structure and semantically contains a toxic situation that may occur in an offline situation. To demonstrate these characteristics, we compare our dataset with Korean dialogue data and online toxic data. For comparison, we choose a free conversation voice dataset[4] published by AI Hub and Korean Unsmile dataset.[5] A free conversation voice dataset published by AI Hub consists of conversations between two speakers given a topic. The dataset also gives text transcription of spoken dialogue, which we used for this comparison. The Korean Unsmile dataset published by Smile-Gate is built to detect toxicity in online interactions consisting of ten toxic classes and one clean class.

Table 9 shows samples of each dataset. The Korean Unsmile dataset has a format of online comments (i.e. vowels only), and contains verbal abuse

[3] https://metworks.co.kr/home/main/

[4] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=109

[5] https://github.com/smilegate-ai/korean_unsmile_dataset



A Free Dialogue Dataset

"자영업을 한다는 것은 매일매일이 전쟁이라는 걸"
(Doing business means that everyday is a war.)
"자영업 힘들지"
(Business is hard.)
"특히 요즘은"
(Especially, these days.)
"내가 기억이 나는 가장 오래된 뉴스는 세계 9위 기업 대우가 망했다는 소식이었어"
(The oldest news I remember was that the world's top eight companies' Daewoo was ruined.)
"이야 진짜 옛날이네"
(Wow, it's a long time ago.)
"세계는 넓고 할 일은 많다던 사람이 회장인 기업이었지"
(It was a company whose CEO was the one who said the world was big and had a lot of work to do.)
"기억 나"
(I remember.)

KCDD

A: 야 너 지금 그러고 온 거야?
(A: Did you dress like that?)
B: 응 왜 그래? 무슨 문제 있어?
(B: Yeah, what's wrong? Is there a problem?)
C: 야 우리가 오늘 미팅한다고 신경 좀 쓰라고 했잖아. 근데 이게 뭐야?
(C: Hey, I told you to mind your outfit for today's meeting. What's that?)
A: 너는 얼굴이 못생겼으니까 꾸미기라도 엄청나게 꾸며야 한다고 했잖아.
(A: I told you that you have to put on makeup hard because you look ugly.)
B: 나름대로 열심히 꾸며 본 건데. 미안해.
(B: I tried my best to do it. I'm sorry.)
C: 됐고, 지금 우리가 너를 어떻게 데려가? 너무 창피한데.
(C: Shut up, W are so. Embarrassed that we can't go. With you.)
A: 그래. 다시 집에 가서 그 못생긴 꼴을 어떻게 하든지, 아니면 우린 너랑 같이 못 가.
(A: Yes, Do something about that ugly face. Or we can't go with you.)

Unsmile dataset

후팔 —— 좆같노 ㅋㅋㅋㅋㅋ 인척은 아니엇고 첨에 몇번 좀 입혀줬더니 패피인척 오지게하고 살더라
(Shit —— I feel fucked up lol When I dressed her up, she thought she was a fashionable person.)

Figure 9: The comparison with a free dialogue dataset, Unsmile dataset, and KCDD.

which correspond to *Other Harassment* class of ICCS. A free dialogue dataset has a dialogic structure that would be in a face-to-face situation and includes general dialogue content. An example of KCDD corresponding to *Other Harassment* class has the same structure of free dialogue data which is in form of dialogue. However, the content contains the toxicity of bullying same as Unsmile data.

Figure 10 visually shows the BERT embeddings of three datasets. After fine-tuning the KLUE-BERT model on KCDD dataset, 768-dimensionnal embedding vector were reduced to 2-dimension with t-SNE for visualization. We took the [CLS] token embedding of last layer as the representative embedding value of data. Since the model trained with our dataset, embedding vectors of our datset are well classified for the five classes. In addition, free conversation data is located close to the *Clean Dialogue* in a vector space and the Unsmile data is located close to *Other Harassment* class. This represents that semantically the Unsmile data is close to *Other Harassment* and free conversation is closer to *Clean Dialogue*. On the other hand, data on *Serious Threats, Extortion or Blackmail*, and *Harassment in the Workplace* is located relatively distant from the two other data in a vector space, because they contain toxicity in offline situations, which is not covered in previous online toxic data or dialogue data.

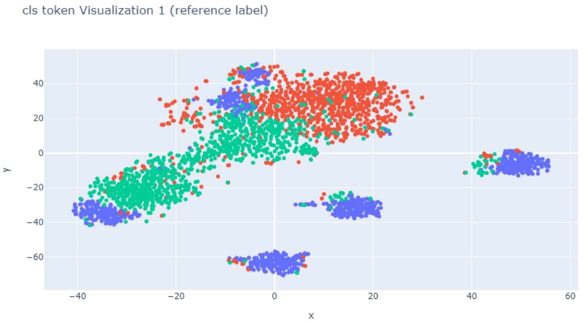


Figure 10: Visualization of BERT embedding of three datasets, blue for KCDD, red for a free speech dialogue, green for unsmile dataset. Because we fine-tuned BERT model on KCDD, the embedding of KCDD are well divided to five classes. For the KCDD embedding, *Clean Dialogue*is at the top, then *Other Harassment, Extortion or Blackmail, Serious Threats, Harassment in the Workplace* in a counterclockwise direction.

D Analyzing the results of generating adversarial data

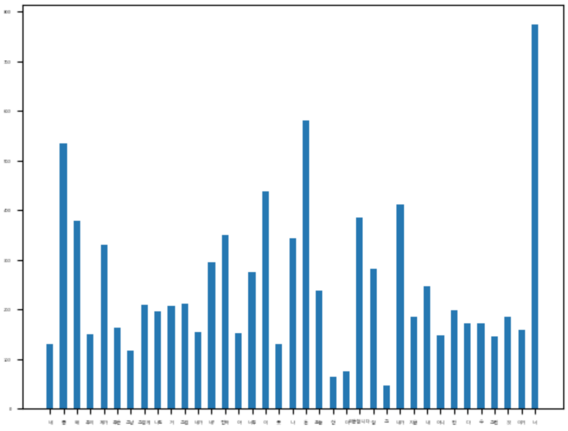
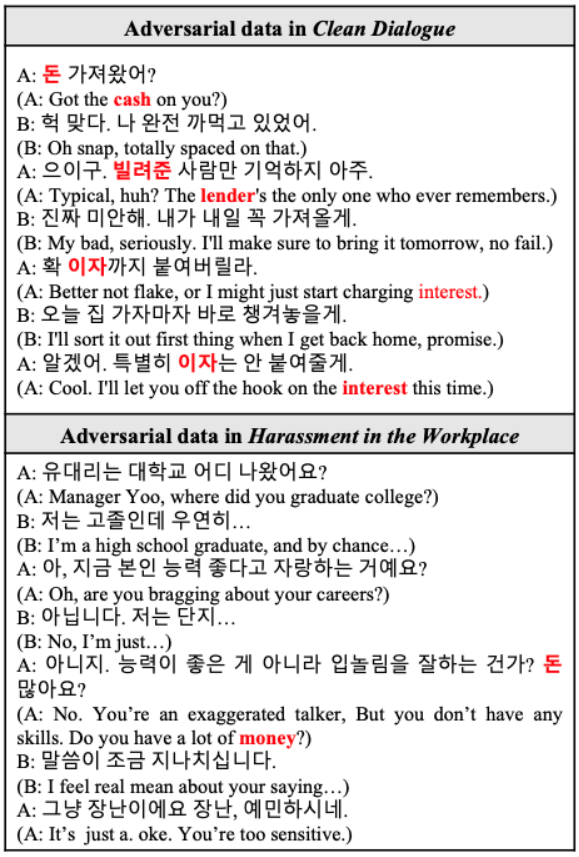


Figure 11: Normalized variance of the top 20 most frequent words in each class

In order to reduce the number of words that appear frequently in a particular class, we constructed a candidate set of key words for each class and generated adversarial data from crowd workers. As a result, we constructed a candidate set of the top 20 most frequent words for each class (the sum of 37 words), and confirmed that the mean of the variance was below 300. Therefore, we could confirm that there were no words that appeared exceptionally frequently in a particular class. The results can be seen in Figue 11.

E Example of Adversarial Data



Adversarial data in *Clean Dialogue*

A: **돈** 가져왔어?
(A: Got the **cash** on you?)
B: 헉 맞다. 나 완전 까먹고 있었어.
(B: Oh snap, totally spaced on that.)
A: 으이구. **빌려준** 사람만 기억하지 아주.
(A: Typical, huh? The **lender**'s the only one who ever remembers.)
B: 진짜 미안해. 내가 내일 꼭 가져올게.
(B: My bad, seriously. I'll make sure to bring it tomorrow, no fail.)
A: 확 **이자**까지 붙여버릴라.
(A: Better not flake, or I might just start charging interest.)
B: 오늘 집 가자마자 바로 챙겨놓을게.
(B: I'll sort it out first thing when I get back home, promise.)
A: 알겠어. 특별히 **이자**는 안 붙여줄게.
(A: Cool. I'll let you off the hook on the **interest** this time.)

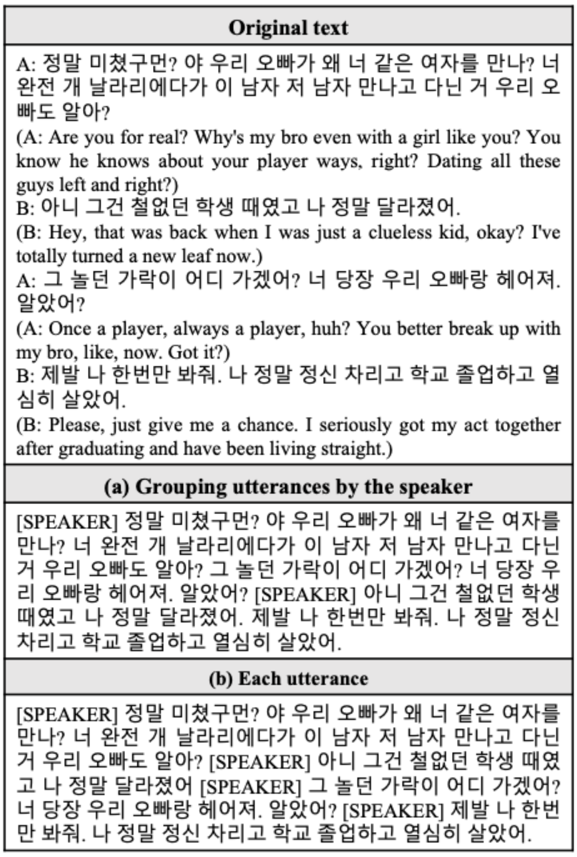
Adversarial data in *Harassment in the Workplace*

A: 유대리는 대학교 어디 나왔어요?
(A: Manager Yoo, where did you graduate college?)
B: 저는 고졸인데 우연히…
(B: I'm a high school graduate, and by chance…)
A: 아, 지금 본인 능력 좋다고 자랑하는 거예요?
(A: Oh, are you bragging about your careers?)
B: 아닙니다. 저는 단지…
(B: No, I'm just…)
A: 아니지. 능력이 좋은 게 아니라 입놀림을 잘하는 건가? **돈** 많아요?
(A: No. You're an exaggerated talker, But you don't have any skills. Do you have a lot of **money**?)
B: 말씀이 조금 지나치십니다.
(B: I feel real mean about your saying…)
A: 그냥 장난이에요 장난, 예민하시네.
(A: It's just a. oke. You're too sensitive.)

Figure 12: The Examples of Adversarial data for the *Extortion and Blackmail* class. They include words that appear frequently in *Extortion and Blackmail* class (highlighted in red).

To prevent model to overfit on certain words which frequently appeared in a certain class, we also collected adversarial data as described in § Section 3.1.4. Figure 12 shows examples of adversarial data for *Extortion or Blackmail* class. To prevent overfitting words related to money, monetary, and private property, that frequently appear in *Extortion and Blackmail* class, we collected data including the keyword related to money, monetary, and private property, but belonging to other classes. Through this process, we ensured that word expressions could be well distributed across several classes.

F Example of [SPEAKER] Token Input Style



Original text

A: 정말 미쳤구먼? 야 우리 오빠가 왜 너 같은 여자를 만나? 너 완전 개 날라리에다가 이 남자 저 남자 만나고 다닌 거 우리 오빠도 알아?
(A: Are you for real? Why's my bro even with a girl like you? You know he knows about your player ways, right? Dating all these guys left and right?)
B: 아니 그건 철없던 학생 때였고 나 정말 달라졌어.
(B: Hey, that was back when I was just a clueless kid, okay? I've totally turned a new leaf now.)
A: 그 놀던 가락이 어디 가겠어? 너 당장 우리 오빠랑 헤어져. 알았어?
(A: Once a player, always a player, huh? You better break up with my bro, like, now. Got it?)
B: 제발 나 한번만 봐줘. 나 정말 정신 차리고 학교 졸업하고 열심히 살았어.
(B: Please, just give me a chance. I seriously got my act together after graduating and have been living straight.)

(a) Grouping utterances by the speaker

[SPEAKER] 정말 미쳤구먼? 야 우리 오빠가 왜 너 같은 여자를 만나? 너 완전 개 날라리에다가 이 남자 저 남자 만나고 다닌 거 우리 오빠도 알아? 그 놀던 가락이 어디 가겠어? 너 당장 우리 오빠랑 헤어져. 알았어? [SPEAKER] 아니 그건 철없던 학생 때였고 나 정말 달라졌어. 제발 나 한번만 봐줘. 나 정말 정신 차리고 학교 졸업하고 열심히 살았어.

(b) Each utterance

[SPEAKER] 정말 미쳤구먼? 야 우리 오빠가 왜 너 같은 여자를 만나? 너 완전 개 날라리에다가 이 남자 저 남자 만나고 다닌 거 우리 오빠도 알아? [SPEAKER] 아니 그건 철없던 학생 때였고 나 정말 달라졌어 [SPEAKER] 그 놀던 가락이 어디 가겠어? 너 당장 우리 오빠랑 헤어져. 알았어? [SPEAKER] 제발 나 한번만 봐줘. 나 정말 정신 차리고 학교 졸업하고 열심히 살았어.

Figure 13: Examples of the Realationship-Aware BERT with input style. A methods of (a) grouping utterances by the speaker and adding [SPEAKER] tokens in front of the group, so that tokens appear equal to the number of speakers. And (b) adding [SPEAKER] to each utterance. We made example in English for helping to understad the example.

Figure 13 is an example of the different input styles.

G Legal Expert Group that We Collaborated with

We worked with law school professors and students to establish data guidelines and conduct data quality checks. We will be able to release more details on this once it is accepted.

H Dataset Card

1. **Motivation**

   (a) **For what purpose was the dataset created?**
   This dataset was built with the purpose of creating a high-quality dataset for creating models that can perform context-based violence detection and classification tasks. Previously, datasets for violence detection in real world or harmful media classification were mostly focused on vision data, and NLP datasets for violence detection did not consider context. Therefore, this dataset was built to fill this gap. In addition, the dataset was built in accordance with ICCS legal standards to be widely used through global criteria.

   (b) **Who created the dataset and on behalf of which entity?**
   The dataset design, guidelines, crowdsourcing management, and data quality checks were conducted by the authors of this paper and a team of legal experts, including law school professors and students. This was done to ensure that ethical issues were taken into account as the dataset deals with violent situations and to ensure that the dataset was aligned with the ICCS standards. Our data is human-written created by crowd workers. The first round of crowd workers were university students, and the second round was outsourced to a crowdsourcing company to ensure that the data was compiled by individuals of different genders and ages.

   (c) **Who funded the creation of the dataset?**
   During the first and second rounds and the entire crowdsourcing process, crowd workers were paid $23 million for data production. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00621,Development of artificial intelligence technology that provides dialog-based multi-modal explainability).

2. **Composition**

   (a) **What do the instances that comprise the dataset represent?**
   Our dataset consists of text in the form of conversations. Each conversation unit is annotated as belonging to the following classes: *Serious Threats, Extortion or Blackmail, Harassment in the Workplace, Clean Dialogue*. Each utterance in each conversation is also annotated as to whether the speaker is the perpetrator, the victim, or a normal person.

   (b) **How many instances are there in total?**
   It contains a total of 22,249 conversations.

   (c) **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
   Our dataset consists of four classes of violent conversations, which are part of the ICCS taxonomy. These four classes were selected by a team of legal experts as they are most likely to be encountered in the neighborhood and are expected to be of high utility. There is room for extension to conversations that fall under other crime classifications.

   (d) **What data does each instance consist of? "Raw" data or features?**
   The KCDD dataset is a human-written senario dataset created by crowd workers.

   (e) **Is there a label or target associated with each instance?**
   Annotations were made according to the international standardized crime classification system called ICCS.

   (f) **Is any information missing from individual instances?**
   No.

   (g) **Are relationships between individual instances made explicit?**
   No.

   (h) **Are there recommended data splits?**
   Our dataset is split into 17,799/2,255/2,225 for train/dev/test. We categorized them for model training, validation, and evaluation.

(i) **Are there any errors, sources of noise, or redundancies in the dataset?**
All data was created by crowdsourced workers and then reviewed to ensure it met the right standards and was re-annotated, corrected, or removed to avoid ethical issues. The datasets we've released have been reviewed. However, it may contain some unidentified errors, labels may need to be corrected, or conversation text may need to be revised. If any are found, we will take immediate action.

(j) **Is the dataset self-contained, or does it link to or otherwise rely on external resources?**
KCDD is a self-contained dataset that contains no external links.

(k) **Does the dataset contain data that might be considered confidential?**
Our dataset is a fictitious creation by crowd workers of conversational texts that fit the labeling of violent situations, so it does not contain any real-world personal information.

(l) **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
Our dataset is for violence detection and includes toxicity. It contains offensive content that appears in the context of a conversation between two or more speakers. Therefore, we prohibit misuse of this dataset and release it with a general prohibition on the use of the dataset for malicious purposes other than research. We also release it under the CC-BY-NC 4.0 license to prevent it from being maliciously edited for other purposes.

(m) **Does the dataset identify any subpopulations?**
In our conversational text, each speaker is represented by an anonymized alphabet from A to D, but the context of the conversation allows us to infer subgroups such as gender and age. The workplace harassment class includes harassment that occurs in workplace relationships, so we estimate higher and lower age ranges for different job titles. The *Other harassment* class contains school bullying situations, so in this case the age can be inferred from the context to be teenagers.

(n) **Is it possible to identify individuals, either directly or indirectly ?**
Our dataset is created as a fictionalized scenario and does not specify or identify any individual or group. However, some celebrity case conversations have been adapted and redacted in a legal expert agreement process where specificity to a particular individual or group is a concern.

(o) **Does the dataset contain data that might be considered sensitive in any way?**
Our dataset is intended to facilitate research on context-based categorization of violence, bias, and toxicity, so we consider violent conversations, criminal contexts, and harassment contexts to include socially discriminatory statements. Because we recognize this risk, our collaborative review process with legal experts included modifications to avoid including too much bias against specific social groups. For example, we worked to flip datasets where foreign workers were often characterized as perpetrators of violence and Koreans as victims.

3. **Collection Process**

(a) **How was the data associated with each instance acquired?**
1) When conversations are created: Our dataset is generative, meaning that it was created by the crowd workers themselves. We provided them with class descriptions and example conversation data as guidelines, and asked them to create conversations that could fall into each class. 2) Speaker type annotation: When annotating perpetrator, victim, and normal person by utterance, we showed the entire dialog context to the crowd workers and asked them to annotate the speaker type of each of the speakers A to D.

(b) **What mechanisms or procedures were used to collect the data?**
We presented a protocol for human-created datasets and quality control

through the Legal Expert Collaborative Data Building Process. We collaborated with legal experts to provide criteria and guidelines, and the dataset was manually built by crowd workers. The data was then reviewed through a process of data balancing and legal expert agreement. Later, we also checked the speaker type through speaker type annotation. The UI and UX screens used for crowdsourcing can be found in appendix B. More details about the data collection process can be found in the main text of the paper in Section 3.1 Legal Expert Collaborative Data Building Process.

(c) **If the dataset is a sample from a larger set, what was the sampling strategy?**
N/A. Our dataset was created by crowd workers manually, not imported as part of a raw dataset.

(d) **Who was involved in the data collection process and how were they compensated?**
The data was compiled by the authors of this paper and a team of legal experts. They are a team of law school professors and students. Crowdsourcing was divided into two rounds, with university students creating the data in the first round, and crowdsourcing companies collecting the data in the second round. Crowd workers were paid 1,000 KRW to create one piece of conversation data. The authors personally attempted to write dialogues prior to crowdsourcing and found that it took approximately 5 minutes to compose one dialogue. Taking this into account, crowd workers could produce about 12 dialogues per hour, which means they could earn roughly 12,000 KRW per hour. Considering that the hourly minimum wage in South Korea in 2023 was 9,620 KRW, this payment was set at a level higher than the minimum wage.

(e) **Over what timeframe was the data collected?**
Our dataset was crowdsourced over a six-month period in the second half of 2021. It then went through a data vetting process, including a Legal Expert Agreement process, during the first half of 2022.

(f) **Were any ethical review processes conducted?**
We went through the process of having legal experts agree on whether there were any ethical issues at the agreement stage. Given that the dataset was created for violence detection, violence was included, but we tried to ensure that it was evenly distributed by including only negative perceptions of certain social groups and not the other way around. We also included steps to edit or remove data if it was clear that the scenarios were targeted at specific celebrities, even though they were fictionalized.

(g) **Did you collect the data directly from the individuals in question, or obtain it via third parties or other sources?**
The crowdsourcing process consisted of two rounds. The first round was conducted by directly recruiting university students as crowd workers as individuals, and the second round was conducted through a specialized crowdsourcing company. More details on this are mentioned in appendix B.

(h) **Were the individuals in question notified about the data collection?**
Because this dataset is not just an annotation task, but a data creation task, we provided more detailed guidelines for the crowd workers. Appendix **??** shows some of the guidelines, and appendix B contains the website screens that the crowd workers worked on.

(i) **Did the individuals in question consent to the collection and use of their data?**
During the crowd worker recruitment process, the purpose of data collection and utilization plan were clearly stated, and only those who agreed with the plan participated in crowdsourcing. In addition, the guidelines specifically stated that adversarial data creation, data balancing, etc. should be considered for AI model training.

4. **Preprocessing, Cleaning and Labeling**

(a) **Was any preprocess-**

ing/cleaning/labeling of the data done?
This data has been collected, reviewed, and labeled through the Legal Expert Collaborative Data Building Process. Crowdworkers created raw data for the five classes according to the ICCS codes. Then, a final label was determined through a major vote by four legal experts. Throughout this process, data with ethical concerns (including personal information and bias) were excluded.

(b) **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?**
The data before undergoing the refinement process will not be disclosed. The original data generated by the crowdworkers may contain some ethical concerns, and the reliability of the labels is also vulnerable.

(c) **Is the software that was used to preprocess/clean/label the data available?**
To preprocess the data into the appropriate input format for training the benchmark model(Relationship-Aware BERT), please refer to the code at https://sites.google.com/view/kcdd.

5. **Uses**

(a) **Has the dataset been used for any tasks already?**
The current dataset has been constructed for the purpose of classifying into five categories: *Serious Threats, Extortion or Blackmail, Harassment in the Workplace, Other Harassment*, and *Clean Dialogue*. This aims to contribute to pre-crime prevention. Additionally, since the speaker type for each utterance is annotated, it can also be used for the task of classifying the speaker type (perpetrator, victim, and normal person) participating in the conversation.

(b) **Is there a repository that links to any or all papers or systems that use the dataset?**
For the review stage, we are concurrently releasing the dataset and benchmark code on https://sites.google.com/view/kcdd for efficiency purposes. In the future, we plan to maintain a separate repository on GitHub for efficient maintenance. In the camera-ready version, we will provide the respective links for each.

(c) **What (other) tasks could the dataset be used for?**
We hope future research will address violence classification considering factors like the relationship between participants, offline violence, and situation-based violence.

(d) **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
The dataset was created by Korean national crowd workers and underwent scrutiny by legal experts of Korean nationality. Therefore, the dataset may have a focus on Korean culture. When using the dataset through translation or post-processing, it is necessary to consider linguistic and cultural differences. However, since it adheres to international standards and conventions, it can be used for data collection in a consistent manner. Although the scenarios are designed in a fictional format, they are based on situations that can frequently occur in offline environments. As there is a risk of imitation, this dataset is made available for research purposes only and should be used strictly for non-commercial purposes.

(e) **Are there tasks for which the dataset should not be used?**
This dataset was developed to overcome the limitations of violence and harmful content detection datasets. Therefore, it is designed for detecting and classifying violent situations from voice and text data coming from smartwatches, IoT devices, and other sources, with the purpose of pre-crime prevention. Consequently, any use of this dataset for purposes other than research related to its intended goals is strictly prohibited.

6. **Distribution**

(a) **Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?**
This dataset encourages contextualized violence classification research through openness, so any third party is welcome to download and use the data for research purposes.

(b) **How will the dataset will be distributed?**
Currently in the review phase, we are releasing the dataset and code on the same website, but in the camera-ready version, we will release their respective DOIs, website, and GitHub addresses.

(c) **When will the dataset be distributed?**
When the research paper on this dataset and benchmark is accepted and published, it will be made publicly available on the same date.

(d) **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use(ToU)?**
This dataset is licensed under CC-BY-NC 4.0. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, provided they give attribution to the author for non-commercial purposes only. For more information, see the corresponding footnotes.

(e) **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
No.

(f) **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
No.

7. **Maintenance**

(a) **Who will be supporting/hosting/maintaining the dataset?**
The authors of this paper actively maintain the dataset on a regular basis. They utilize the issue page on GitHub to address users' questions and requests, and handle other inquiries through a designated contact email. Any updates or important announcements that users need to be aware of will be consistently managed and communicated through the GitHub repository.

(b) **How can the owner/curator/manager of the dataset be contacted?**
We'll be releasing a representative email on GitHub to respond to user inquiries.

(c) **Is there an erratum? If so, please provide a link or other access point.**
All datasets have been built over the course of about a year of collection and thorough review. However, we will respond quickly to any errors you may find in your use. Please contact us via the GitHub issues page or our main email.

(d) **Will the dataset be updated?**
We do not plan to add new data, but we will announce when we do. We will also respond quickly to user requests to correct errors. Data checks will be conducted by the authors on a quarterly basis.

(e) **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?**
No.

(f) **Will older versions of the dataset continue to be supported/hosted/maintained?**
When data is updated, the dataset is named differently for each version, and both versions of the dataset are maintained.

(g) **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
We welcome all extend/augment/build on/contribute to the dataset. If someone would like to participate in any of these contributions, feel free to email the main email listed on GitHub, and you will be listed as a contributor on GitHub after your contribution.