Improving Multi-label Malevolence Detection in Dialogues through Multi-faceted Label Correlation Enhancement

Yangjun Zhang[1], **Pengjie Ren**[2]*, **Wentao Deng**[2], **Zhumin Chen**[2], **Maarten de Rijke**[1]
[1]University of Amsterdam, [2]Shandong University
[1]{y.zhang6, m.derijke}@uva.nl, [2]{renpengjie, wentao.deng, chenzhumin}@sdu.edu.cn

Abstract

A dialogue response is malevolent if it is grounded in negative emotions, inappropriate behavior, or an unethical value basis in terms of content and dialogue acts. The detection of malevolent dialogue responses is attracting growing interest. Current research on detecting dialogue malevolence has limitations in terms of datasets and methods. First, available dialogue datasets related to malevolence are labeled with a single category, but in practice assigning a single category to each utterance may not be appropriate as some malevolent utterances belong to multiple labels. Second, current methods for detecting dialogue malevolence neglect label correlation. Therefore, we propose the task of multi-label dialogue malevolence detection and crowdsource a multi-label dataset, *multi-label dialogue malevolence detection* (MDMD) for evaluation. We also propose a multi-label malevolence detection model, *multi-faceted label correlation enhanced CRF* (MCRF), with two label correlation mechanisms, *label correlation in taxonomy* (LCT) and *label correlation in context* (LCC). Experiments on MDMD show that our method outperforms the best performing baseline by a large margin, i.e., 16.1%, 11.9%, 12.0%, and 6.1% on precision, recall, F1, and Jaccard score, respectively.

1 Introduction

Safety is an increasingly important aspect of artificial intelligence development (Amodei et al., 2016; Roegiest et al., 2019; Sun et al., 2021). When it comes to dialogue agents, taking measures to avoid risks of generating undesirable and harmful responses may have a profound positive impact on the adoption of conversational technology (Xu et al., 2020). Research on safe dialogue agents involves aspects such as inaccurate information (Gunson et al., 2021), fairness (Liu et al., 2020), and unauthorized expertise (Sun et al., 2021). Malevolence is another key aspect (Zhang et al., 2021b,a), e.g., whether the dialogue utterance contains malevolent content that is related to offensiveness (Dinan et al., 2019), toxicity (Gehman et al., 2020), ad hominem (Sheng et al., 2021), and toxicity agreement (Baheti et al., 2021), etc.

There have been several studies targeting malevolence detection (Roussinov and Robles-Flores, 2007; Saral et al., 2018; Zhang et al., 2021a,b). We build on the work of Zhang et al. (2021b) who introduce the malevolent dialogue response detection and classification task, present a hierarchical malevolent dialogue taxonomy, create a labeled multi-turn dialogue data set, and apply state-of-the-art text classification methods to the task. One important limitation of their work is that they only explore single-label dialogue malevolence detection (SDMD), i.e., they assume that each dialogue utterance corresponds to a single malevolence or non-malevolence label. However, some utterances have more than one label, e.g., in Figure 1, the utterance "f** people are disgusting"[1] belongs to both "disgust" and "negative intergroup attitude (NIA)." This is because malevolence labels are correlated with one another, which we refer to as *label correlation in taxonomy* (LCT).

Zhang et al. (2021b) propose a hierarchical malevolent dialogue taxonomy that classifies correlated malevolence labels into the same group by investigating three dimensions – negative emotions, negative psychological behavior, and unethical issues. However, the correlation of malevolence labels in different groups is not well captured. Another limitation is that the above studies neglect the impact of malevolence in dialogue contexts (i.e., previous turns) on the current utterance. Previous work concatenates the dialogue context as model input without explicitly modeling the malevolence

* Corresponding author.

[1]Words that turn a statement into a statement that may cause harm are masked in this work.

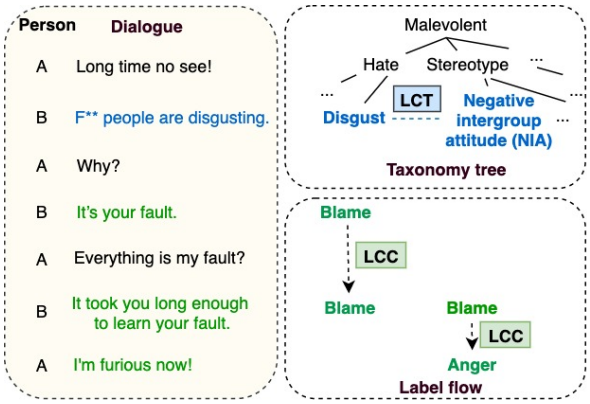


Figure 1: Label correlation in taxonomy (LCT) and label correlation in context (LCC). In terms of LCT, "negative intergroup attitude (NIA)" is correlated with "disgust", which can be reflected by the utterance in blue (LCT). In different turns, "blame" is likely to co-occur with "anger" and "blame", which can be reflected by the utterances in green (LCC).

transition. For example, in Figure 1, "blame" is likely to cause "blame" for the same person, while for different persons, "blame" is likely to cause "anger". This is due to *label correlation in context* (LCC). Zhang et al. (2021b) do not take correlation of malevolence labels in different dialogue turns into account and our label-correlation mechanisms are different from previous methods which require multi-label training sets (Kurata et al., 2016; Tsai et al., 2021).

We address the two limitations listed above. Our goal is to boost multi-label dialogue malevolence detection (MDMD) by incorporating label correlation in taxonomy and context based on a single-label dataset with re-annotated multi-label evaluation data. This goal comes with two main challenges: (1) A dataset challenge, as we only have one label per utterance in the training data, which increases the negative effect of unobserved labels during training: how to improve the single gold labels via LCT and decrease the probability of over–fitting; (2) A classification method challenge: how to capture LCC to help improve the classification.

Based on Conditional Random Field (CRF), we propose a *multi-faceted label correlation enhanced CRF* (MCRF) framework to improve MDMD from single-label training data. The approach contains a *position-based label correlation in taxonomy* (PLCT)-based encoder and a multi-faceted CRF layer, which includes a LCC-based feature function and LCT-based label distribution learning. For the dataset challenge, we build a LCT-based label distribution learning module to exploit the label correlation in hierarchical taxonomy, which can alleviate the unobserved label problem. For the classification method challenge, we build an LCC-based transition function to exploit the label correlation in context.

We crowdsource a new dataset based on the previously released malevolent dialogue response detection and classifying (MDRDC) dataset, conduct experiments on this dataset, and show that MCRF with a pretrained model, i.e., BERT-MCRF, outperforms competitive baselines by a large margin. We also conduct further analyses of the LCT and LCC modules, which reveal that multi-faceted label correlation does enhance multi-label dialogue malevolence detection.

We summarize our contributions as follows: (1) We crowdsource a new dataset, i.e., MDMD, for the task of multi-label dialogue malevolence detection from single-label training data. (2) We propose multi-faceted label correlation, including LCC and LCT, which is shown to be beneficial for dialogue malevolence detection. (3) We introduce a new framework, MCRF, and compare it with competitive baseline models on the MDMD dataset and demonstrate its effectiveness.

2 Related Work

2.1 Malevolence detection taxonomies

The taxonomies for hate speech, aggressiveness, offensiveness, and condescending only contain a few categories (Waseem and Hovy, 2016; Kumar et al., 2018; Zampieri et al., 2019; Wang and Potts, 2019), which are lack of unified understanding of what constitutes malevolence. To address this gap, Sheng et al. (2021) introduce a two-level ad hominem taxonomy and Sun et al. (2021) introduce a safety taxonomy, both of which contain seven different aspects. Furthermore, Zhang et al. (2021b) define a three-level malevolence taxonomy that contains eighteen categories in total. In this work, we follow the taxonomy proposed by Zhang et al. (2021b).

2.2 Malevolence detection datasets

There are several datasets to support malevolence classification or detection research. Many of them investigate hate speech detection, e.g., Predictive Features for Hate Speech Detection (PFHSD) (Waseem and Hovy, 2016), Hate Speech Detection Dataset (HSDD) (Davidson et al., 2017), and Multilingual Detection of Hate Speech (MDHS) (Basile et al., 2019), which are all col-

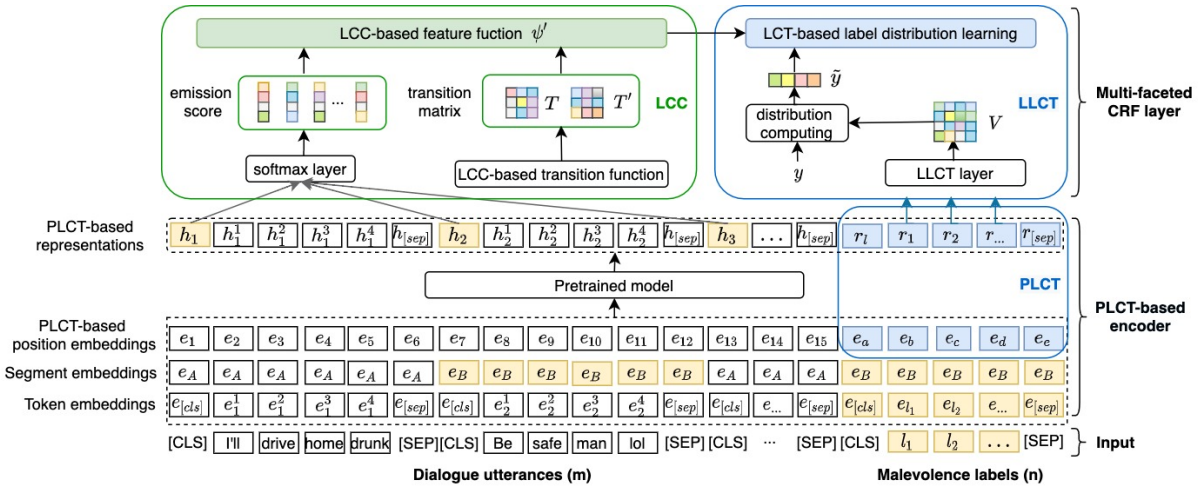


Figure 2: Framework of the proposed *multi-faceted label correlation enhanced CRF* (MCRF) model.

lected from Twitter. These datasets lack diversity, have a small data size, low inter-annotator agreement, and small lexicon size. The others are on aggressiveness, offensiveness, and condescending, e.g., Trolling, Aggression and Cyberbullying (TRAC) (Kumar et al., 2018), Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019), and TALKDOWN (Wang and Potts, 2019), which have been collected from Facebook, Reddit, and Twitter, respectively. These datasets have a larger size than those mentioned before, but problems such as low diversity and limited lexicon size affect them too. To sum up, none of these datasets is in the form of multi-turn dialogues. To address this, recent studies have released the TOXICHAT (Baheti et al., 2021), ADHOMINTWEETS (Sheng et al., 2021), MDRDC (Zhang et al., 2021b), and DIASAFETY datasets (Sun et al., 2021), for research into offensiveness, ad hominem, safety detection, etc. However, the above datasets all fall into single-label dialogue malevolence detection.

In contrast, we build a dataset for the evaluation of multi-label malevolence detection, considering an utterance may contain multiple labels.

2.3 Malevolence detection methods

Methods for malevolence detection include rule based (Roussinov and Robles-Flores, 2007), traditional machine learning based (Waseem and Hovy, 2016; Davidson et al., 2017; Saral et al., 2018; Basile et al., 2019), and deep learning based (Kumar et al., 2018; Zampieri et al., 2019; Wang and Potts, 2019; Sheng et al., 2021; Zhang et al., 2021b) approaches. Roussinov and Robles-Flores (2007) define malevolence by filtering the keywords. Saral et al. (2018) survey the machine learning-based detection methods, including KNN and SVM-based methods. The performance of these methods is not strong enough as malevolence detection requires a deep understanding of semantics. Kumar et al. (2018) apply CNNs and LSTMs for aggressiveness detection. Zampieri et al. (2019) apply CNNs and Bi-LSTMs for offensiveness detection. More recently, pretrained models, e.g., BERT and RoBERTa, have been used for ad hominem, malevolence, and safety (Sheng et al., 2021; Zhang et al., 2021b; Sun et al., 2021) , demonstrating better performance than LSTM, CNN, RCNN, and GNN based models (Zhang et al., 2021b).

Compared with previous methods, we model malevolence detection as a multi-label dialogue malevolence detection task instead of a single-label dialogue malevolence detection task. Moreover, we propose two label correlation mechanisms, i.e., label correlation in taxonomy (LCT) and label correlation in context (LCC).

3 Method

3.1 Overall

Given a dialogue that contains m utterances, $x = [x_1, x_2, \ldots, x_i, \ldots, x_m]$ and x_i is the i-th utterance in the dialogue. $y = [y_1, y_2, \ldots, y_i, \ldots, y_m]$ denotes the label sequence of one dialogue, where $y_i \in \{0, 1\}^n$ is the label for each utterance. $l = \{l_1, l_2, \ldots, l_j, \ldots, l_n\}$ denotes the label set, where l_j is the j-th label, n is the total number of label categories. *Multi-label dialogue malevolence detection* (MDMD) aims to assign the most reliable labels to each x_i. Since there is no large-scale MDMD dataset, during training, we observe one

non-malevolent label or only observe one malevolent label per utterance, while the other malevolent labels are unknown. We build a MDMD dataset for evaluation only, the details of which can be found in the experiments.

We propose a model, *multi-faceted label correlation enhanced CRF* (MCRF), for MDMD. As shown in Figure 2, MCRF consists of a PLCT-based encoder and a multi-faceted CRF layer, where the PLCT-based encoder is used to encode the utterances x and labels l, and output the representations H and R; the representations are fed into the multi-faceted CRF layer to predict the multi-labels $\hat{y}$. The PLCT-based encoder is enhanced by a taxonomy tree-based position embedding e_{pos}; the multi-faceted CRF layer is enhanced by *learning-based label correlation in taxonomy* (LLCT) (i.e., $\tilde{y}$), LCC (i.e., T and T'), and the representation output of the PLCT-based encoder (i.e., H and R). In the following subsections, we detail each component.

3.2 Utterance and label encoder

As shown in Figure 2, the utterance and label encoder takes the utterances and labels as input, and the output is the representations of utterances and labels. Following Liu and Lapata (2019), each utterance is encoded separately by inserting "[CLS]" at the start of each utterance and "[SEP]" at the end of each utterance. The labels are encoded by inserting "[CLS]" between the last utterance and labels and "[SEP]" at the end of labels. We utilize three kinds of embeddings, namely token embeddings, segment embeddings, and position embeddings. Token embeddings follow the original transformer paper (Vaswani et al., 2017). Segment embeddings distinguish each utterance, as well as the labels, by e_A or e_B, where e_A and e_B are odd or even. Position embeddings for utterances capture the position of the utterances (Wang and Chen, 2020). In order to improve the representation of labels, we change the position embeddings of labels into PLCT-based position embedding (see §3.3). We feed the three embeddings into a pretrained model (i.e., BERT) to get the representations of utterances and labels:

$$\begin{aligned} H, R &= PTM([e(x_i), e(l_j)]), \\ e &= e_{tok} + e_{seg} + e_{pos}, \end{aligned} \tag{1}$$

where PTM is the pretrained model; e_{tok}, e_{seg}, and e_{pos} are the token, segment and position embeddings, respectively. $H = \{h_1, h_2, \ldots, h_i, \ldots, h_m\}$ denotes the repsentations of the utterances with h_i (corresponding to pooler output of "[CLS]") representing the i-th utterance x_i. $R = \{r_1, r_2, \ldots, r_j, \ldots, r_n\}$ are the representations of the labels with r_j (corresponding to sequence output of labels) representing the j-th label l_j.

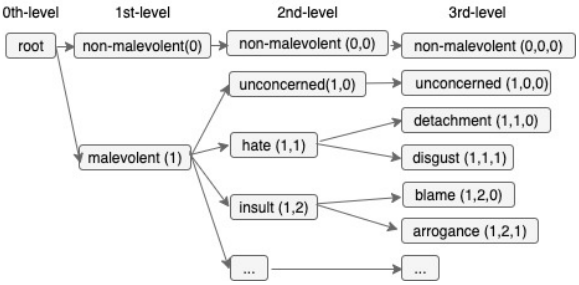


Figure 3: Demonstration of taxonomy tree of labels.

3.3 Multi-faceted label correlation

Multi-faceted label correlation is the main component of MCRF, which is composed of two major modules: LCT and LCC. The former is meant to decrease the probability of over-fitting caused by single-label annotated data, while the latter is meant to leverage the influence of the previous label on the next label of the utterances from the same user and the other user.

Label correlation in taxonomy. The LCT module contains two parts: PLCT and LLCT. First, the PLCT module captures label correlation in the taxonomy tree. The input of the module is the taxonomy tree, the output is the label position, and the module is used for improving the encoder. PLCT is defined by the taxonomy tree-based position of each label, which is formulated by its path from the root in the taxonomy tree (Wang et al., 2021). The taxonomy of malevolence consists of a root and three levels of labels. We use the 1st-level, 2nd-level, and 3rd-level of labels to get the coordinate for the 3rd-level labels. For instance, in Figure 3, the taxonomy tree-based positional label embedding for "blame" is $(1, 2, 0)$. We use label position output of PLCT to improve e_{pos} in Eq. 1, and the encoder is improved as *PLCT-based encoder*.

Second, the LLCT module captures label correlation by learning a correlation matrix $V^{n\times n}$. Each element of the matrix corresponds to the correlation of two labels accordingly as follows:

$$V = \frac{1}{2}(\hat{V}_{j,j'} + V'_{j,j'}), \tag{2}$$

where $\hat{V}$ is the learned LCT correlation matrix by representations of labels, $\hat{V}_{j,j'} = d(r_j, r_{j'})$; V' is the fixed LCT correlation matrix, $V'_{j,j'} = d(c_j, c_{j'})$; d is the correlation function and we use the Cosine similarity; r_j and r'_j are the representations of the j-th and j'-th label by PLCT-based encoder with taxonomy tree position, i.e., R from Eq. 1; c_j and c'_j are the n-gram bag-of-words vectors of the utterances belong to the j-th and j'-th label, respectively. The label correlation matrix V is used for hierarchical label distribution learning later in §3.4.

Label correlation in context. The LCC module captures the label correlation between the labels of different utterance turns. We use two kinds of LCC correlation functions, i.e., label correlation functions between utterance turns from different users (t) and the same user (t'), which are defined as follows:

$$\begin{aligned} t(y_{i-1} = l_j, y_i = l_{j'}) &= T_{(l_j, l_{j'})}, \\ t'(y_{i-2} = l_j, y_i = l_{j'}) &= T'_{(l_j, l_{j'})}, \end{aligned} \tag{3}$$

where l_j and $l_{j'}$ denote the j-th and j'-th labels. T and T' are two $n \times n$ matrices initialized randomly and trained by LCC-based label distribution learning, which is introduced next.

3.4 Multi-faceted CRF layer

Given a sequence of utterances, a linear chain CRF can be used to predict the label of an utterance:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_i \psi(x_i, y_i)\right), \tag{4}$$

where Z is a normalization function, and

$$\psi(x, y) = \sum_i s(y_i, x) + \sum_i t(y_{i-1}, y_i), \tag{5}$$

where t is defined in Eq. 3. s is the emission function. Next, we introduce the components of our multi-faceted CRF layer, including the LCC-based feature function and the LCT-based label distribution learning.

LCC-based feature function. The LCC-based feature function contains two parts: the emission function and the LCC-based transition function. First, the *emission function* s is defined as follows:

$$s(y_i, x) = \text{softmax}(h_i), \tag{6}$$

where h_i is the representation of each utterance x_i. Second, the *LCC-based feature function* is defined as follows:

$$\begin{aligned} \psi'(x, y) = &\frac{1}{2}\Bigg(\psi(x, y) + \sum_i s(y_i, x) \\ &+ \sum_i t'(y_{i-2}, y_i)\Bigg), \end{aligned} \tag{7}$$

where t', ψ and s and are defined in Eq. 3, 5 and 6, respectively.

LCT-based label distribution learning. We get the estimated gold label distribution $\tilde{y}$ for CRF label distribution learning. We calculate the estimated distribution $\tilde{y}_i$ from the original distribution y_i of the i-th utterance as follows:

$$\tilde{y}_i = \lambda V y_i + y_i, \tag{8}$$

where λ denotes how much the original one-hot distribution is redefined and V is the matrix that estimates the LCT in Eq. 2.

Our training objective is the KL-divergence loss except that we replace gold label y with estimated gold label $\tilde{y}$:

$$\mathcal{L} = \sum_y q(y|x) \log \frac{q(y|x)}{p(y|x)}, \tag{9}$$

where $q(y|x)$ is the target distribution to learn, we use the probability of $\tilde{y}$ given x for $q(y|x)$; $p(y|x)$ is the predicted distribution.

The KL loss can be transformed into the following function by expanding and marginalizing $p(y|x)$ (Liu and Hockenmaier, 2020):

$$\begin{aligned} \mathcal{L} = &\sum_i \sum_{y_i} \{q(y_i|x) \log q(y_i|x)\} \\ &- \sum_y \{q(y|x)\psi'(y, x)\} + \log Z(x), \end{aligned} \tag{10}$$

where q is the target distribution, ψ' is the feature function, Z is the normalization function.

4 Experimental Setup

We conduct experiments to answer the following research questions: (RQ1) How does BERT-MCRF compare to baselines on the MDMD test set? (RQ2) What is the impact of the number of labels on the performance of BERT-MCRF? (RQ3) What is the influence of different LCT and LCC settings? (RQ4) What do the components of BERT-MCRF contribute to its overall performance?

4.1 Dataset

We conduct experiments on an extension of the MDRDC dataset released by Zhang et al. (2021b). The original MDRDC dataset is for single-label dialogue malevolence detection; it contains 6,000 dialogues (with 10,299 malevolent utterances and 21,081 non-malevolent utterances) annotated by Amazon MTurk workers.

To conduct the evaluation for multi-label dialogue malevolence detection, we re-annotate the validation and test set of the MDRDC dataset using Amazon MTurk following the annotation protocols in (Zhang et al., 2021b). We select workers with a test score of at least 90, 500 approved human intelligence tasks (HITs) and 98% HIT approval rate and the location is limited to countries where English is one of the official languages. The workers are also asked to consider dialogue context and implicit words. Before the annotation, we warn the crowd workers that the task may contain malevolent content. The crowd workers are asked to annotate each utterance of the dialogue with 18 3rd-level labels in the taxonomy of Zhang et al. (2021b). We ask three workers to annotate the data. Cohen's multi-Kappa value of the three workers is 0.701 for the re-annotated data, which is considered substantial (McHugh, 2012).

	Malevolent		Non-malevolent		Total
	Valid.	Test	Valid.	Test	
1-label	413	733	2,088	4,276	7,510
2-label	264	574	–	–	838
3-label	22	85	–	–	107
4-label	2	5	–	–	7
Total	701	1,397	2,088	4,276	8,462

Table 1: Statistics of the validation and test sets of MDMD.

The MDMD dataset statistics are shown in Table 1. We have re-annotated 8,462 utterances in total, with 2,098 malevolent and 6,364 non-malevolent utterances. There are 7,510 (88.7%), 838 (9.9%), 107 (1.3%) and 7 (0.1%) utterances for 1-label, 2-label, 3-label and 4-label group separately. For all the collected data, 952 (11.3%) of 8,462 utterances have 2–4 labels. For the malevolent utterances, 952 (45.4%) of 2,098 utterances have 2–4 labels, which indicates the importance of MDMD task considering the percentage of multi-label utterances. We use the training, validation, and test splits provided in (Zhang et al., 2021b), which has a ratio of 7:1:2.

4.2 Baselines

We compare BERT-MCRF against BERT and BERT-CRF. The two baselines are competitive since BERT with a softmax classifier performs well in a previous SDMD task (Zhang et al., 2021b), and BERT-CRF with modified encoder for separate sentences is the state-of-the-art model for sequence labeling task (Cohan et al., 2019).

4.3 Implementation details

We use the 'bert-base-uncased' version of BERT as the pretrained model with a vocabulary size of 30,522. The max sequence length is set to 512. For BERT-MCRF, we first do BERT fine-tuning with learning rate 2e-5, and BERT is fine-tuned with 2 epochs. Then, we train the multi-faceted CRF layer and fine-tune BERT together, with multi-faceted CRF layer learning rate 7e-4 and BERT-encoder learning rate 5e-7, we train 10 epochs together. The batch size is 8 for training, validation, and test. The dropout ratio is 0.1. More runtime and parameter details are provided in Appendix B. All the neural models are trained on GeForce GTX TitanX GPUs.

4.4 Evaluation metrics

We use the precision, recall, F1 score, and Jaccard score as our evaluation metrics (Manning et al., 2008). We report the macro scores since the data is imbalanced in terms of labels (Zhang et al., 2021b).

5 Results and Analysis

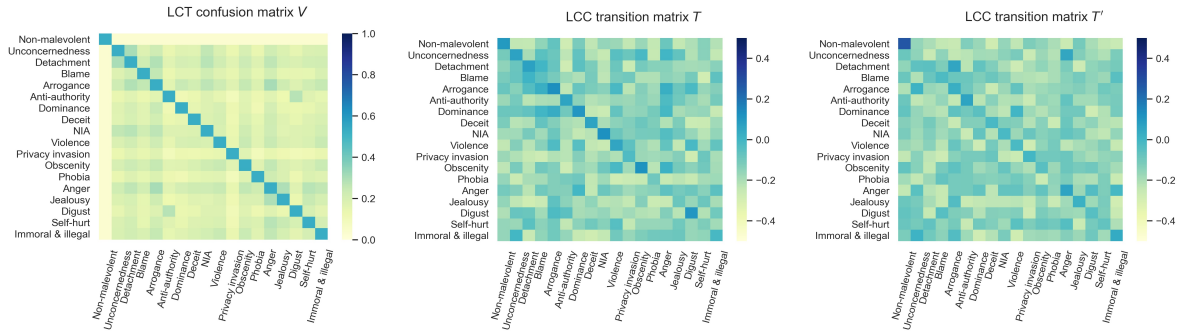
5.1 RQ1: Comparison with baselines

To determine how MCRF compares to baseline models on the MDMD task, we report the results in terms of precision, recall, F1, and Jaccard score in Table 2. In terms of overall performance, adding

Model	Precision	Recall	F1	Jaccard
BERT	67.73	33.59	42.32	37.25
BERT-CRF	69.62	33.57	43.30	40.83
BERT-MCRF	**82.99**	**38.12**	**49.20**	**43.46**

Table 2: Main results of MCRF on the MDMD test set.

LCT and LCC improves the performance of dialogue malevolence detection. In general, the performance of BERT-MCRF is better than BERT and BERT-CRF. The precision, recall, F1, and Jaccard score of BERT-MCRF outperform the second-best model (i.e., BERT-CRF) by 16.1%, 11.9%, 12.0%,



(a) LCT confusion matrix V. (b) LCC transition matrix T. (c) LCC transition matrix T'.

Figure 4: Visualization of LCT and LCC.

and 6.1%, respectively. The results in terms of precision and recall indicate that incorporating LCT and LCC provides benefits to both precision and recall, and more benefits to precision than recall.

5.2 RQ2: Performance of different label groups

We divide the samples in the MDMD test set into different groups according to the number of labels. We report the Jaccard scores of different label groups in Table 3.

Model	1-label	2-label	3-label	4-label
BERT	40.16	11.84	11.48	8.00
BERT-CRF	44.02	13.06	11.89	**11.33**
BERT-MCRF	**46.39**	**15.23**	**12.88**	10.00

Table 3: Jaccard scores of different label groups.

First, the results suggest that BERT-MCRF has better performance with regard to different label groups. BERT-MCRF's Jaccard scores for the 1-label, 2-label, and 3-label are 5.4%, 16.6%, 8.3% higher than the second best performing approach. An exception is that for the 4-label group, the result of BERT-MCRF is lower than BERT-CRF. The reason is that the size of 4-label utterances is small for the test set and the performance of 4-label changes dramatically when we evaluate at different epochs. Second, the results show that the MDMD task becomes more challenging as the number of labels increases. The Jaccard score results for all the models in Table 3 decrease as the number of labels increases.

5.3 RQ3: Influence of the LCT and LCC settings

First, we study the influence of the hyperparameter λ of LCT in Eq. 8, as shown in the upper part of Table 4. As λ increases, the performance increases and then decreases. The reason is that as with overly large λ, the original one-hot distribution is redefined too much as to make the learning target deviate from the real target. We visualize the LCT confusion matrix V (Eq. 8) in Figure 4(a). Yellow or blue suggests the correlation is low or high, separately. The variation of correlation value suggests our model can capture the label correlation in taxonomy, which contributes to final results.

Settings	Precision	Recall	F1	Jaccard
LCT ($\lambda = 0$)	83.60	36.78	47.96	42.75
LCT ($\lambda = 1/2$)	84.58	37.04	48.50	42.85
LCT ($\lambda = 1$)	**82.99**	**38.12**	**49.20**	**43.46**
LCT ($\lambda = 2$)	82.28	38.09	49.10	42.98
LCC (T)	84.37	37.08	48.58	43.43
LCC (T')	84.43	35.99	47.10	42.62
LCC (T+T')	**82.99**	**38.19**	**49.20**	**43.46**

Table 4: BERT-MCRF performance w.r.t. different LCT and LCC settings. λ is the hyperparameter in Eq. 8, T and T' are the transition matrices by Eq. 3.

Second, we study the influence of different transition function matrices of LCC, i.e., T is LCC between the same user, T' is LCC between different users, as shown in the bottom part of Table 4. For the three LCC settings, T has better recall thus improving the final performance compared with T'; T' has the better precision than the other two groups, but he overall performance is the lowest; BERT-MCRF with both T and T' combine the advantages to achieve the best performance. We visualize the LCC confusion matrices T in Figure 4(b) and T' in Figure 4(c); yellow and blue suggests a negative and positive correlation, respectively. First, LCC captured by transition matrices can be both positive and negative, e.g., for T', "non-malevolent" is likely to transit to "non-malevolent" and not-likely to transit to "immoral & illegal";

second, the LCC captured by T and T' is different.

5.4 RQ4: Ablation study

We perform an ablation study on BERT-MCRF by removing LCT or LCC. The results are reported in Table 5. The results suggest that both LCC and LCT are important for BERT-MCRF.

First, removing LCC decreases the performance of BERT-MCRF by 2.9%, 1.3%, and 0.1% for recall, F1 and Jaccard, respectively, while the precision increase by 1.7%. LCC has a positive influence since it considers both the LCC from the same user and different users, while BERT-CRF only contains the label correlation from different users, as explained in §5.3.

Second, removing LLCT decreases the performance of recall, F1 and Jaccard score by 3.7%, 2.5%, and 1.6%; LLCT has a positive influence since it predicts estimated gold labels to improve model learning. An exception is that the precision increases by 0.7%, which is not significant, and the reason might be that BERT-MCRF tends to predict more labels, which results in a much higher recall but decreases precision a bit.

Model	Precision	Recall	F1	Jaccard
BERT-MCRF	82.99	38.19	49.20	43.46
−LCC	84.37	37.08	48.58	43.43
−LLCT	83.60	36.78	47.96	42.75
−PLCT	69.34	33.79	43.27	40.86
−LCT	69.87	33.16	42.62	40.83

Table 5: Ablation study results. Note that LCC of different users T is already captured by BERT-CRF, therefore the ablation of LCC keeps T but deletes T'.

Third, removing PLCT decreases the performance of precision, recall, F1, and Jaccard by 16.4%, 11.5%, 12.1%, and 6.0%. The performance suggests that PLCT has a positive influence on the results. The fixed correlation between the 3rd-level labels with the same node based on the taxonomy tree is captured well by the position embedding.

Fourth, removing both LLCT and PLCT decreases the performance of recall, F1, and Jaccard score by 15.8%, 13.2%, 13.4%, and 6.1%. Compared with the results with LLCT ablation and PLCT ablation, both LLCT and PLCT have a positive influence on the BERT-CRF model. Previously, some methods have utilized label correlation in training data to improve multi-label classification, i.e., label co-occurrence (Zhang et al., 2018). However, for MDMD, there is no label co-occurrence information; our results suggest that LCT is able to increase the MDMD; the reason might be that the LCT reduces overfitting of single-label training data.

5.5 Case study

We randomly select two examples from the test set to illustrate the performance of BERT, BERT-CRF, and BERT-MCRF (see Table 7 in Appendix A.2).

First, for the first example, BERT-MCRF predicts the right labels "violence" and "self-hurt". The LCT correlation value between label "violence" and "self-hurt" is 0.1923, and suggests that LCT may help predict the two labels together. Second, in the second example, BERT-MCRF predicts a sequence of labels for different dialogue turns more accurately than BERT and BERT-CRF. We found that the LCC value between "non-malevolent" and "non-malevolent" is 0.2725, while the LCC value between "non-malevolent" and "immoral & illegal" is -0.1183, which implies that it helps BERT-MCRF predict the right label "non-malevolent" for the third utterance considering the label of the first utterance. In summary, LCC is able to boost the performance of BERT-MCRF. In addition, there are also cases where BERT-MCRF fails. An example is the label with implicit expression, i.e., "deceit", which leaves room for further improvement by considering implicit meaning.

6 Conclusion and Future Work

We have studied multi-label dialogue malevolence detection and built a dataset MDMD. The dataset statistics suggest that the dataset quality is substantial and that it is essential to do multi-label dialogue malevolence detection as almost 12% of the utterances have more than one malevolent label. We have proposed BERT-MCRF by considering label correlation in taxonomy (LCT) and label correlation in context (LCC). Experimental results suggest that BERT-MCRF outperforms competitive baselines. Further analyses have demonstrated the effectiveness of LCT and LCC.

A limitation of BERT-MCRF is that it is not good at detecting implicitly malevolent utterances, e.g., "deceit." As to future work, we plan to address this type of utterance and investigate how to enhance BERT-MCRF in terms of implicit multi-label dialogue malevolence detection by semi-supervised learning as there are large-scale unlabeled datasets.

7 Ethical Considerations

The data collection process for the re-annotated MDMD dataset follows the regulations of Twitter. The data is anonymized so the data can not be linked to a particular user. The crowd workers are fairly compensated with a minimum wage per hour (using the minimum wage from a Western European country). The data collection process has been approved by the ethics committee of the authors' university. The data will be made available to researchers that agree to the ethical regulations of the ethics committee. Characteristics and quality control of the re-annotated dataset are described in Section 5.

The claims in the paper match the results and the model can be generalized to multi-label dialogue safety detection tasks. This paper can be used for the deployment of dialogue systems, hoping to improve the ability of dialogue systems to detect malevolent human natural language. Multi-label classification has a positive impact on the application of dialogue systems. Detecting and filtering dialogue responses that are not malevolent may decrease the diversity of the dialogue. For the deployment of non-malevolent dialogue systems, it is better to consider the extent of malevolence according to malevolence label counts of each utterance or the perception of different labels.

This paper does not involve identity characteristics nor does it categorize people.

Acknowledgements

This research was partially supported by the Natural Science Foundation of China (62102234, 61972234, 61902219, 62072279), the Natural Science Foundation of Shandong Province (ZR2021QF129), the National Key R&D Program of China with grant No. 2020YFB1406704, the Key Scientific and Technological Innovation Program of Shandong Province (2019JZZY010129), the China Scholarship Council and the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, 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 *SemEval*, pages 54–63.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. 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 3693–3699.

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

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Nancie Gunson, Weronika Sieińska, Yanchao Yu, Daniel Hernandez Garcia, Jose L Part, Christian Dondrup, and Oliver Lemon. 2021. Coronabot: A conversational ai system for tackling misinformation. In *Proceedings of the Conference on Information Technology for Social Good*, pages 265–270.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. 2016. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–526.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416.

Jiacheng Liu and Julia Hockenmaier. 2020. Phrase grounding by soft-label chain conditional random field. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 5112–5122. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740.

Christopher D. Manning, Hinrich Schütze, and Prabhakar Raghavan. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

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

Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, Anubrata Das, Asia Biega, Bart Voorn, Claudia Hauff, Damiano Spina, David Lewis, Douglas W. Oard, Emine Yilmaz, Faegheh Hasibi, Gabriella Kazai, Graham McDonald, Hinda Haned, Iadh Ounis, Ilse van der Linden, Jean Garcia-Gathright, Joris Baan, Kamuela N. Lau, Krisztian Balog, Maarten de Rijke, Mahmoud Sayed, Maria Panteli, Mark Sanderson, Matthew Lease, Michael D. Ekstrand, Preethi Lahoti, and Toshihiro Kamishima. 2019. FACTS-IR: Fairness, accountability, confidentiality, transparency, and safety in information retrieval. *SIGIR Forum*, 53(2):20–43.

Dmitri Roussinov and José A Robles-Flores. 2007. Applying question answering technology to locating malevolent online content. *Decision Support Systems*, 43(4):1404–1418.

Shubham M Saral, Rahul R Sawarkar, and Priyanka A Jalan. 2018. A survey paper on malevolent word detection and hazy vicious imaging. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 723–728. IEEE.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "nice try, kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2021. On the safety of conversational models: Taxonomy, dataset, and benchmark. *arXiv preprint arXiv:2110.08466*.

Shang-Chi Tsai, Chao-Wei Huang, and Yun-Nung Chen. 2021. Modeling diagnostic label correlation for automatic icd coding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4043–4052.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yu-An Wang and Yun-Nung Chen. 2020. What do position embeddings learn? an empirical study of pre-trained language model positional encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6840–6849.

Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2021. Tuta: Tree-based transformers for generally structured table pre-training. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1780–1790.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. 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 3711–3719.

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

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

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

Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2021a. A human-machine collaborative framework

for evaluating malevolence in dialogues. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5612–5623.

Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2021b. A taxonomy, data set, and benchmark for detecting and classifying malevolent dialogue responses. *Journal of the Association for Information Science and Technology*, 72:1477–1497.

Yinyuan Zhang, Ricardo Henao, Zhe Gan, Yitong Li, and Lawrence Carin. 2018. Multi-label learning from medical plain text with convolutional residual models. In *Machine Learning for Healthcare Conference*, pages 280–294. PMLR.

APPENDICES

We present additional details on our experimental in the appendices below. We include the validation performance for the main result (Appendix A.1); the case study (Appendix A.2); our source code (Appendix A.3); the average runtime of each module and detailed information about the parameters (Appendix B); further details about the newly created dataset that we release with this paper (Appendix C); and ethical considerations of this work (Appendix 7).

A Experimental Results

A.1 Performance of BERT-MCRF on the validation set

In terms of validation performance, we report the validation performance of comparing BERT-MCRF to BERT-classifier and BERT-CRF, as shown in Table 6. BERT-MCRF surpasses both BERT and BERT-CRF in dialogue malevolence detection. Compared to the second-best model, the precision, recall, F1, and Jaccard score of BERT-MCRF increase by 9.1%, 8.8%, 11.4%, and 4.2%, respectively. This suggests that for the validation set, BERT-MCRF is better than BERT and BERT-CRF for dialogue malevolence detection. Moreover, the results of the validation set and the test set are similar. Compared to results of the test set, F1 of the validation set increase by 2.8%, while Jaccard decrease by 1.8%. This suggests the test result is consistent with validation. 1, but lower Jaccard, as shown in Table 2 (presented in Section 5) and Table 6.

Metric	Precision	Recall	F1	Jaccard
BERT	64.14	36.33	43.97	37.02
BERT-CRF	73.05	35.84	44.81	40.89
BERT-MCRF	**80.39**	**39.82**	**50.58**	**42.68**

Table 6: Main results of BERT-MCRF on the multi-label validation set.

A.2 Case study examples

We show two examples for case study that explains how LCT and LCC work, as shown in Table 7; the description is in §5 of the main paper.

A.3 Code

Our code is uploaded to https://github.com/repozhang/MCRF with the paper.

User A: S** my head into a brick building while having a steal toe boot r** up my a** would hurt less.
BERT: non-malevolent. **BERT-CRF**: non-malevolent. **BERT-MCRF**: violence, self-hurt. **Gold**: violence, self-hurt.
User A: Mom: I can't believe you haven't seen birdman , Edward Norton is in it ! n Me: I know she gets me. User B: Hope Gasols forgive me when I marry him. User A: Invite me so i can get drunk and be inappropriate.
BERT: non-malevolent; non-malevolent; immoral & illegal. **BERT-CRF**: non-malevolent; non-malevolent; immoral & illegal. **BERT-MCRF**: non-malevolent; non-malevolent; non-malevolent. **Gold**: non-malevolent; non-malevolent; non-malevolent.

Table 7: Case study. Upper: utterances and labels of example 1; bottom: utterances and labels of example 2.

B Runtime and Parameters

In terms of average runtime, the time cost for our BERT-MCRF model is acceptable. The time costs for BERT-MCRF is 2 hours. The run time of BERT-CRF is the same as BERT-MCRF and the run-time for BERT is less than 1 hour.

In terms of parameters, BERT-MCRF has 109,496,802 parameters, BERT has 109,496,118 parameters, BERT-CRF has 109,496,478 parameters. As described in §4.3, in terms of the BERT-MCRF model, we first fine-tune BERT. We choose the best result of learning rate 2e-5 and training epochs 2. Second, we train multi-faceted CRF layer with BERT together, where BERT is not completely frozen but has a relatively small learning rate. In this step, the learning rate for BERT is 5e-7 and for the multi-faceted CRF layer is 7e-4. The reason that the BERT learning rate is small during the joint training is that we have fine-tuned BERT for 2 epochs before feeding the representations to multi-faceted CRF Layer. We train BERT-MCRF for 10 epochs and choose the best result based on the validation set results. For the λ parameter in Eq. 8, we use the value range $[0, 0.5, 1, 2]$ and select the best result. In terms V' in Eq. 2, we use n-gram settings of $[1, 2, 3, 4]$, and select 2 for the

final estimation of V' based on the best result. In terms of the BERT classifier, the learning rate is 2e-5, the epoch number is 2. In terms of BERT-CRF, the parameter selection process is similar to BERT-MCRF, the BERT fine-tuning parameters for the first step same to BERT-MCRF; and for the second step that trains both BERT and CRF, the final learning rate is 5e-7 for BERT and 3e-4 for CRF layer.

C Dataset

Our data is uploaded to https://github.com/repozhang/malevolent_dialogue with the paper. The statistics and splits are described in §4.1. The language of the dataset is in English. For data prepossessing, we use all the data from the dataset. In terms of the data collection process, we follow the previous research (Zhang et al., 2021b), except that the workers are asked to choose multiple choices from the labels. The label taxonomy is grounded in negative emotion, negative psychological behavior, and unethical issues. It includes three levels of labels, with two, eleven, and eighteen labels in 1st-level, 2nd-level, and 3rd-level labels. The third level labels, as shown in Figure 4, includes 'non-malevolent', 'unconcernedness', 'detachment', 'blame', 'arrogance', 'anti-authority', 'dominance', 'deceit', 'negative intergroup attitude (NIA)', 'violence', 'privacy invasion', 'obscenity', 'phobia', 'anger', 'jealousy', 'disgust', 'self-hurt', 'Immoral and illegal'. For the 2nd-level categories, the taxonomy put the set of 3rd-level categories that have correlations in linguistic characteristics with each other into the same group (Zhang et al., 2021b).