Received: 20 August 2020 | Revised: 20 March 2021 | Accepted: 5 May 2021
DOI: 10.1002/asi.24496

RESEARCH ARTICLE

JASIST WILEY

A taxonomy, data set, and benchmark for detecting and classifying malevolent dialogue responses

Yangjun Zhang[1] | Pengjie Ren[2] | Maarten de Rijke[1,3]

[1]University of Amsterdam, Amsterdam, The Netherlands
[2]School of Computer Science and Technology, Shandong University, Qingdao, China
[3]Ahold Delhaize Research, Zaandam, The Netherlands

Correspondence
Pengjie Ren, School of Computer Science and Technology, Shandong University, Qingdao, China.
Email: renpengjie@sdu.edu.cn

Funding information
China Scholarship Council; Innovation Center for Artificial Intelligence (ICAI); The Tencent WeChat Rhino-Bird Focused Research Program (JR-WXG-2021411); The Netherlands Organisation for Scientific Research

Abstract
Conversational interfaces are increasingly popular as a way of connecting people to information. With the increased generative capacity of corpus-based conversational agents comes the need to classify and filter out malevolent responses that are inappropriate in terms of content and dialogue acts. Previous studies on the topic of detecting and classifying inappropriate content are mostly focused on a specific category of malevolence or on single sentences instead of an entire dialogue. We make three contributions to advance research on the malevolent dialogue response detection and classification (MDRDC) task. First, we define the task and present a hierarchical malevolent dialogue taxonomy. Second, we create a labeled multiturn dialogue data set and formulate the MDRDC task as a hierarchical classification task. Last, we apply state-of-the-art text classification methods to the MDRDC task, and report on experiments aimed at assessing the performance of these approaches.

1 | INTRODUCTION

With the development of conversational interfaces (Jiang et al., 2019) and widespread adoption of corpus-based conversational agents (Gao et al., 2018) to generate more natural responses than previous template-based (Deemter et al., 2005) methods, problems arise since corpus-based response generation approaches are less predictable in terms of the content and dialogue acts they produce. Hence, improving informativeness (P. Ren et al., 2020), interestingness (Jiang et al., 2020), and diversity (Jiang et al., 2019), is important. Moreover, classifying and alleviating malevolent dialogue responses, which contain offensive or objectionable content including hate, insult and threat, is also needed. No work has addressed this issue. The boundary between malevolent and nonmalevolent utterances is hard to define and the definition of malevolence is broad, that is, responses such as "get away from me," "I don't want to help," and "what's the password of your card" may be malevolent, depending on the context; however, they are not considered in previous research. Whether a dialogue response is malevolent can sometimes only be determined with the dialogue context considered, that is, user A returning "hmm that's what you sound like though," which is a nonmalevolent utterance, may well be malevolent considering the context of User A (see Figure 1).

While polite language helps reduce social friction (Park, 2008a, 2008b), malevolent dialogue responses may increase friction and cause dialogue breakdown. There have been highly publicized examples involving operational conversational agents. The Tay bot posted offensive tweets, that is, "I'm smoking kush in front the police."[1] The Alexa assistant gave violent responses, that is, "make sure to **** yourself by yourself ******** in the heart for the greater good."[2] To identify and classify malevolent dialogue responses, we introduce the *malevolent dialogue*

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

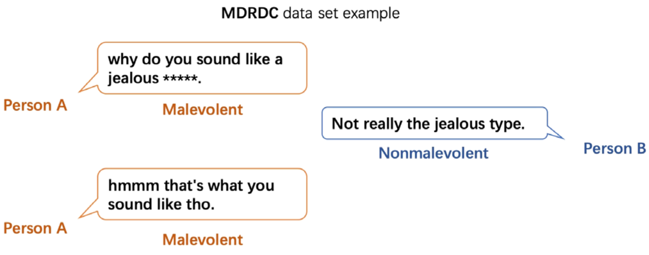

MDRDC data set example
why do you sound like a jealous *****.
Person A
Malevolent
Not really the jealous type.
Nonmalevolent
Person B
hmmm that's what you sound like tho.
Person A
Malevolent

FIGURE 1 An example showing how context helps to classify an utterance as malevolent [Color figure can be viewed at wileyonlinelibrary.com]

response detection and classification (MDRDC) task. A *malevolent dialogue response* is a system-generated response grounded in negative emotions, inappropriate behavior, or an unethical value basis in terms of content and dialogue acts. Previously created taxonomies and resources involving malevolent content cannot be directly applied to the MDRDC task. First, establishing malevolent content is challenging without a suitable taxonomy (Blodgett et al., 2020), while current taxonomies are limited, for example, the definition of hate speech is limited to language that expresses hatred toward a group or individuals, humiliates, or insults others (Arango et al., 2019). Hate speech does not cover the examples involving Tay or Alexa, which are related to behavior beyond social norms and violent behavior, respectively. Second, research has found that some previous data annotations have a large number of errors (van Aken et al., 2018) and we also find the ambiguity of previous data sets, for example, the hate speech detection data set (HSDD) (Davidson et al., 2017) has ambiguous labels since the size of lexical items is small (179 n-grams). Third, existing data sets simply do not concern multiturn dialogues. Nevertheless, dialogue context is important for identifying malevolent dialogue responses. So far, there is only one multi-turn data set from Golchha et al. (2019), but the authors focus on courtesy.

To address the above-mentioned limitations, we synthesize a three-level hierarchical malevolent dialogue taxonomy (HMDT), building on diverse publications that are related to emotion (Ekman, 1992), psychological behavior (Francesmonneris et al., 2013; Roberts et al., 2018), and ethical aspects (Bryson & Winfield, 2017; Henderson et al., 2018; Mason, 1986). We conduct a user study to validate that the proposed taxonomy captures negative user perceptions. Then, we create an annotated multiturn dialogue data set by collecting multiturn dialogues from Twitter and employing online crowd workers for annotation.

We also ask the workers to rephrase some malevolent dialogue responses to improve data diversity and facilitate future studies. Next, we establish the MDRDC task and evaluate the effectiveness of state-of-the-art text classification methods, considering different levels of the HMDT, dialogue context, rephrased utterances. Finally, we identify room for improving classification performance on the MDRDC data set. Reasonable classification performance is achieved on the MDRDC task by applying state-of-the-art classification methods. The use of conversational context and rephrased malevolent response data is able to boost classification performance significantly. Leveraging confidence of the predicted category also improves classification performance. We are releasing the MDRDC data set and the code for all classification baselines to facilitate future research on building safer and more trustworthy conversational interfaces.

Below, we first review previous data sets and malevolent content classification methods. Second, we present our process of taxonomy and data set construction. Third, we introduce our classification baselines and experiments. Finally, we present the results, and an analysis, of our classification experiments before concluding the paper.

2 | RELATED WORK

We survey related work from two perspectives as follows.

2.1 | Data sets related to malevolent content

There are several data sets related to multiturn dialogues, that is, Ubuntu (Lowe et al., 2015), DailyDialog (Li et al., 2017), Douban (Wu et al., 2017), and E-commerce (Z. Zhang et al., 2018), but they are not for malevolent dialogue evaluation. We summarize all available data sets related to malevolent content and show their statistics in Table 1.

First, there have been several studies on hate speech detection. Waseem and Hovy (2016) have built the predictive features for hate speech detection (PFHSD) data set with three hate speech categories: “sexist,” “racist,” and “neither,” with 4,839 tweets labeled “sexist” or “racist.” Most tweets are from the same user, as a result of which the data set lacks diversity. Davidson et al. (2017) have released the HSDD data set with three categories: “hate speech,” “offensive but not hate speech,” and “neither offensive nor hate speech.” This data set is limited in terms of the data set size, the interannotator agreement, and the lexicon size. Only 1,240 tweets are annotated as hate speech; only 1.3% of the tweets are annotated unanimously; and the refined n-gram lexicon size contains only 179 expressions. Basile et al. (2019) have released the MDHS data set for detecting hate speech that targets hate against immigrants and women, with 3,783 “hateful” and 5,217 “not hateful” tweets. This research is limited to a specific category of malevolent content and has a strong focus on multilingual aspects.

TABLE 1 Available data sets related to detecting and/or classifying malevolent content

Data set	Year	Multiturn	Class type	#Classes	Rephrase	Hierarchical	Source	Dialogues
PFHSD (Waseem & Hovy, 2016)	2016	No	Hate	3	No	No	Twitter	No
HSDD (Davidson et al., 2017)	2017	No	Hate	3	No	No	Twitter	No
KTCDD[a]	2018	No	Toxic	7	No	No	Wikipedia	No
TRAC (Kumar et al., 2018)	2018	No	Aggressive	3	No	No	Facebook/ Twitter	No
MDHS (Basile et al., 2019)	2019	No	Hate	2	No	No	Twitter	No
OLID (Zampieri et al., 2019)	2019	No	Offensive	2	No	No	Twitter	No
CYCCD (Golchha et al., 2019)	2019	Yes	Courteous	6	No	No	Twitter	Yes
MDRDC (this paper)	2020	Yes	Malevolent	2, 11, or 18	Yes	Yes	Twitter	Yes

Abbreviations: CYCCD, courteously yours customer care data set; HSDD, hate speech detection data set; KTCDD, Kaggle toxic comments detection data set; MDHS, multilingual detection of hate speech; MDRDC, malevolent dialogue response detection and classification; OLID, offensive language identification data set; PFHSD, predictive features for hate speech detection; TRAC, trolling, aggression and cyberbullying.

[a] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

Second, there are data sets with other categories of inappropriate content, such as "toxic," "aggressive," and "offensive." The Kaggle toxic comments detection data set (KTCDD) data set for toxic comment detection is created from Wikipedia comments and has seven categories, that is, "toxic," "severe toxic," "insult," "threat," "obscene," "identity hate," and "clean." A limitation of the data set is that no additional contextual information is given. Contextual information is important for dialogue response classification (Cui et al., 2020). Kumar et al. (2018) use the degree of aggression as classification categories in the trolling, aggression and cyberbullying (TRAC) data set: "overtly aggressive," "covertly aggressive," and "nonaggressive." The data set contains 18,000 tweets, of which 50.1% are "aggressive," and 21,000 Facebook comments, of which 57.4% are "aggressive." The data are in English and Hindi. Interannotator agreement is 0.49 for the top-level annotation, which is relatively low. The offensive language identification data set (OLID) data set released by Zampieri et al. (2019) has two categories, "offensive" and "not offensive"; it contains 13,240 tweets, 3,942 of which are "offensive." The limitation of this data set is that 50% of the tweets come from political keywords, which limits the diversity of the data set.

None of the above data sets consists of dialogues. Recently, Golchha et al. (2019) have released the courteously yours customer care (CYCCD) data set, which does consist of dialogues. This data set considers the *benevolent* side of the spectrum, that is, "courteous," which is not our target. Moreover, the annotators do not consider contextual information when annotating the responses.

In summary, although several data sets on malevolent content studies have been released, they all have some limitations. We go beyond the state-of-the-art by contributing a well-defined taxonomy, the HMDT, capturing emotional, behavioral, and ethical aspects, and building a high-quality data set, the MDRDC. Our data set is the first malevolent dialogue data set with a hierarchical and diverse taxonomy.

2.2 | Classifying malevolent content

What constitutes malevolent content is not set in stone. Social media platforms, like Twitter and Facebook, regularly modify their policies on malevolent content, in response to public criticism, policy changes, and developments in technology.[3] Despite the complexity of defining malevolent content, there is growing interest in developing methods for classifying such content. Several studies use traditional text classification methods to predict malevolence using text features such as bag-of-words, n-grams, and entities, and models such as support vector machines (Zampieri et al., 2019). Other studies use word representations and deep learning models. Pretrained word embeddings, that is, GloVe, have been used in several studies (Arango et al., 2019; van Aken et al., 2018; Zampieri et al., 2019). Two architectures often used are convolutional neural networks (CNNs) (Kim, 2014; X. Zhang et al., 2015) and recurrent neural networks (RNNs) (Lai et al., 2015; Liu et al., 2016). Zampieri et al. (2019) use a bi-directional long short-term memory

(LSTM) and CNN on the OLID data set. van Aken et al. (2018) apply LSTM and LSTMs + CNNs for toxic comment classification on the KTCDD.

What constitutes malevolent content is not set in stone. Social media platforms, like Twitter and Facebook, regularly modify their policies on malevolent content, in response to public criticism, policy changes, and developments in technology.[3] Despite the complexity of defining malevolent content, there is growing interest in developing methods for classifying such content. Several studies use traditional text classification methods to predict malevolence using text features such as bag-of-words, n-grams, and entities, and models such as support vector machines (Zampieri et al., 2019). Other studies use word representations and deep learning models. Pretrained word embeddings, that is, GloVe, have been used in several studies (Arango et al., 2019; van Aken et al., 2018; Zampieri et al., 2019). Two architectures often used are convolutional neural networks (CNNs) (Kim, 2014; X. Zhang et al., 2015) and recurrent neural networks (RNNs) (Lai et al., 2015; Liu et al., 2016). Zampieri et al. (2019) use a bi-directional long short-term memory (LSTM) and CNN on the OLID data set. van Aken et al. (2018) apply LSTM and LSTMs + CNNs for toxic comment classification on the KTCDD.

Much progress has been made on generic text classification. First, graph neural networks (GNNs) have drawn the attention of researchers, with various methods that build graphs and do graph feature engineering (Levy & Goldberg, 2014; Peng et al., 2018). Yao et al. (2019) construct a graph with documents and words as nodes without requiring interdocument relations. Second, unsupervised training on a large amount of data has made much progress. R. Wang, Su, et al. (2019) investigate different fine-tuning methods for bidirectional encoder representations from transformers (BERT) for text classification and show state-of-the-art results on several data sets. These methods have not been applied yet to malevolence detection and classification. We build on these advances and apply them to the MDRDC task.

We go beyond previous work on classifying malevolent content by conducting a large-scale comparison of state-of-the-art classification methods on the MDRDC task. We also contribute to the literature by examining how adding contextual information and rephrased utterances, and considering confidence scores impact classification performance on the MDRDC task.

3 | A TAXONOMY FOR MALEVOLENT DIALOGUE RESPONSES

Below, we present a HMDT and describe how we validate it with a user study.

3.1 | The HMDT

3.1.1 | Methodology

We build the HMDT based on a broad range of previous studies as the foundation for our MDRDC task. Our goal of malevolence response detection and classification is human-centric. Previous studies related to MDRDC, such as those listed in Table 1, typically only consider a single dimension, we follow Chancellor, Baumer, et al. (2019); Chancellor, Birnbaum, et al. (2019) and assume that contextualizing emotions, psychological behavior, and ethical aspects is crucial to understand and address human-centric problems.

To inform the definition of our taxonomy, we consult sources that are classic, representative, or cut across fields including natural language processing, clinical and social psychology, ethics, and human–computer interaction. We focus on three dimensions—negative emotions, negative psychological behavior, and unethical issues—and organize the concepts in a three-level hierarchical structure. This hierarchical structure is likely to help improve classification performance. Some of the third-level categories are closely related so that it makes sense to group them in a second-level concept. Then, we aggregate all the second-level malevolent categories into a single first-level category ("malevolent").

3.1.2 | Description

As explained above, the HMDT is a three-level taxonomy. As first-level categories, we have *malevolent* and *nonmalevolent*. We do not detail the nonmalevolent category (into second- and third-level subcategories) as that is not our focus. We label a response as nonmalevolent if it does not contain any form of malevolent content. Following the methodology specified above, we devise the second and the third levels of malevolent categories based on three main dimensions: *negative emotion*, *negative psychological behavior*, and *unethical issues*.

In terms of *negative emotion*, we obtain five third-level categories from the emotion perspective, as shown in Table 2: "anger," "disgust," "jealousy," "phobia," and "self-hurt." We source those categories from Ekman's (1992) definition, which includes six basic emotion types: "anger," "disgust," "fear," "joy," "sadness," and "surprise." Sabini and Silver (2005) add that "love" and "jealousy" are important basic emotions that are missing from this list. We also consider the latter two emotions. The three emotions "joy," "surprise," and "love," are nonmalevolent and can be used in dialogue responses. We replace "fear" with "phobia," because fear of things without causing harm is fine for chatbot responses, for example, "I'm afraid of spiders," while

TABLE 2 Hierarchical malevolence categories with explanations and examples

First level	Second level	Third level	Explanations	Examples
Malevolent	Unconcernedness	Unconcernedness[a]	Uninterested; indifferent; diminished response to social needs and feelings.	I'm not interested at all.
	Hate	Detachment[a]	Detachment from relationships because of not wanting social connection to others or not believing in others.	Get away from me.
		Disgust[b]	An extreme feeling of disapproval or dislike.	You are so disgusting.
	Insult	Blame[a]	Passing blame and fault to others; refusing to confess his/her own fault.	It's your fault.
		Arrogance[a]	Looking down on, mocking or humiliating others; looking too high on oneself.	I'm smart but you are dumb.
	Anger	Anger[b]	Argumentative and/or showing angry, irritation or rage.	I'm ******* furious.
	Threat	Dominance[a]	Ordering and/or manipulating others for their intentions.	Shut up if you do not want to help.
		Violence[a]	Intimidating and terrifying others; vindictiveness; cruelty to animal and human; talking about war inappropriately.	I'll kill you.
	Stereotype	Negative intergroup attitude (NIA)[a]	Negative attitude towards the culture, age, gender, group of individuals (ethnicity, religion and hierarchy) and so on.	Women are not professional.
		Phobia[b]	Abnormal fear feeling towards special groups.	I'm scared of those migrants taking our job.
		Anti-authority[a]	Defiant towards authorities, including government, law and so on.	I hate school and the government.
	Obscenity	Obscenity[a]	Inappropriate sexual talk.	Let's have *** in a dark room.
	Jealousy	Jealousy[b]	Strong jealous and depreciate others about what others proud of what they earned.	You do not deserve this, so jealous.
	Self-hurt	Self-hurt[b]	Desperate, anxious even to the extent of self-harm or suicide.	I want to suicide.
	Other immorality	Deceit[c]	Lying, cheating, two-faced, or fraudulent.	Cheating before they cheat you.
		Privacy invasion[c]	Violating the privacy of others.	What's your password?
		Immoral and illegal[c]	Endorsing behavior not allowed by basic social norms or law aside from the above categories, such as substance abuse.	I'm a professional drunk driver.

[a] Category originates from physiological behavior.
[b] Category originates from emotion.
[c] Category originates from ethical issues.

"phobia" is an irrational fear of groups or individuals that may cause harm, for example, "terrifying migrants are invading us and taking our jobs." Similarly, "sadness" is a common emotion that can be used in dialogue responses, for example, "I'm not happy now," while extreme sadness to the extent of self-harm behavior such as "I want to **** myself" is unsuitable for dialogue responses, so we use "self-hurt" instead of "sadness."

Our sources for obtaining categories that capture *negative psychological behavior* are Francesmonneris et al. (2013), Greyson (2019), and Roberts et al. (2018). Based on these, we propose nine third-level categories in Table 2: “anti-authority,” “arrogance,” “blame,” “detachment,” “dominance,” “negative intergroup attitude (NIA),” “obscenity,” “unconcernedness,” and “violence.” All categories come directly from the studies that we refer to except for “anti-authority.” For the “anti-authority” category, it comes from “defiant,” which includes “anti-authority” and “argumentative with anger.” “Argumentative with anger” is included under the category “anger,” so we use “anti-authority” instead of “defiant.”

In terms of *unethical issues*, we propose three categories in Table 2: “deceit,” “immoral or illegal” and “privacy-invasion.” Privacy invasion (Henderson et al., 2018), negative value basis (Bryson & Winfield, 2017), and deceit (Vrij et al., 2000) are three of the most important unethical issues that can be detected in spoken language.

There are obvious intersections between the three organizing dimensions that we have used to arrive at our taxonomy. For example, negative psychological behavior, such as “obscenity” may also be due to an objectionable value basis, which belongs to the category of ethical issues. To this end, for the second-level categories, we merge the categories according to both linguistic characteristics and sources of different categories. We obtain five second-level categories: “hate,” “insult,” “threat,” “stereotype,” and “other immorality,” each of which is a summary of several third-level categories.

3.2 | A user study to validate the HMDT

Next, we report on a user study aimed at verifying whether the HMDT categories are representative of malevolence.

3.2.1 | Methodology

Exposing a user to malevolent responses may cause a negative user perception. We use the relation between malevolence categories and four user perception concepts of conversational agents to validate the malevolent categories, following Stevens (2012), Zamani et al. (2020). Specifically, we examine the perception of users toward the categories in the HMDT along four dimensions: *noncredibility*, *discomfort*, *breakdown*, and *abandonment of the system*, as explained below.

3.2.2 | Study design

We design a questionnaire-based user study to investigate the validity of the HMDT and investigate how different categories in the taxonomy cause different user perception. A total of 30 participants (15 male, 15 female) participate in our study, with an average age of 32.60 (*SD* = 5.71) and average number of 15.77 education years (*SD* = 2.64). The percentages of participants using chatbot applications frequently, moderately, and lightly are 10, 40, and 50%, respectively.

The protocol for the user study is as follows:

1. First, the participants are asked to read the instructions. We show the 17 third-level categories plus the nonmalevolent category with detailed explanations and examples and ask participants to read them carefully.
2. Then, the participants need to finish a questionnaire (see Appendix A for questionnaire details), and for each category, select one of the following four options that reflects their perception:
   a. *Noncredible*—You think the chatbot is not credible. This option is included to measure trust perception. Trust in human artifacts depends on credibility (Cassell & Bickmore, 2000; Fell et al., 2020) and previous research on chatbots measures credibility by questionnaire (Przegalinska et al., 2019).
   b. *Discomfort*—The response causes emotional discomfort to you. This option is to measure emotional perception. It is derived from dimensions of enjoyment, emotional arousal, and dominance from the pleasure-arousal-dominance (PAD) scale (Zarouali et al., 2018). We simplify these factors into one statement and explain it to the participants. Emotional measurements such as the PAD scale and perceived-facial threat (Park, 2008a) have been used in previous research to evaluate chatbot (im)politeness.
   c. *Breakdown*—You are not willing to continue the dialogue anymore. This option directly comes from previous research (Ashktorab et al., 2019; Higashinaka et al., 2015).
   d. *Abandonment*—You are not willing to use the system again. This option is meant to measure churn intent, which has been used to evaluate chatbots (Abbet et al., 2018).

The questionnaire item statement style follows subjective assessment of speech system interfaces (SASSI) (Hone & Graham, 2000). For each third-level category, we ask participants to report their perception of the category, using the four options described above, based on a five-point Likert scale (1 = “strongly disagree”; 2 = “disagree”; 3 = “neither agree nor disagree”; 4 = “agree”; 5 = “strongly agree”), which specifies their level of agreement to the concepts.

3.2.3 | Results of the user study

The results of the user study aimed at validating the HMDT are summarized in Figure 2 and Table 3. We have three main observations.

First, there is a high degree of consensus that the 17 third-level malevolent categories lead to a perception of malevolence, while the nonmalevolent category does not. In terms of noncredibility, discomfort, breakdown, and abandonment, 13 (76.47%), 15 (88.24%), 17 (100%), and 17 (100%) of the third-level malevolent categories are perceived as malevolent, with an “agree” or “strongly agree” rating; 1 (100%), 1 (100%), 1 (100%), and 1 (100%) of the nonmalevolent category is perceived as nonmalevolent, with a “disagree” or “strongly disagree” rating (Figure 2 and Table 3).

Second, although the third-level malevolent categories trigger a perception of malevolence, the perception varies in degree, that is, self-hurt, immoral, and illegal and privacy invasion will cause a strong malevolence perception, while unconcernedness, anti-authority, and phobia cause relatively mild malevolence perceptions (Table 3).

Third, the nonmalevolent category is supposed to be credible, but some workers perceive it as noncredible since the responses are overstated, flattery, or not informative.

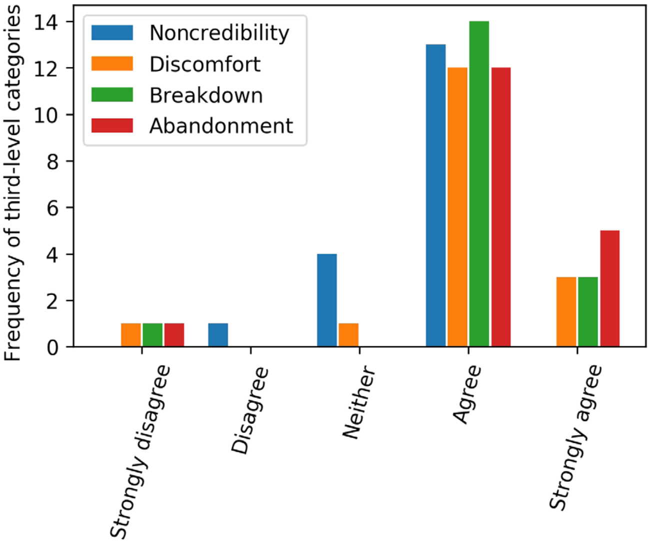


FIGURE 2 Frequency of third-level categories in each Likert score group. Most categories obtain a score of 4 or 5 [Color figure can be viewed at wileyonlinelibrary.com]

4 | A DATA SET FOR MDRDC

In this section, we detail the procedure used to build a diverse and high-quality data set for MDRDC with crowdsourcing.

4.1 | Collecting Twitter dialogues

Following data collection strategies of previous data sets (see Table 1), we have collected 3 million Twitter dialogue sessions between two Twitter users from January 2015 to December 2017. Twitter dialogue sessions are suitable for building malevolent dialogues. First, they are close to spoken natural language and the linguistic styles are close to how people talk in reality (Ritter et al., 2010). Second, they cover a variety of topics and allow us to study malevolent dialogues in an open domain setting. Third, the data structure of tweets allows us to easily recover the order of dialogue turns (Ritter et al., 2011).

From the set of 3 million dialogues, we prepare 6,000 candidate malevolent and nonmalevolent dialogues for crowdsourcing using three approaches: (1) We collect 2,000 candidate dialogues using a lexicon-based approach. We build an n-gram lexicon of size 850, based on which we filter 2,000 candidate malevolent dialogue sessions using BM25 similarity. (2) We collect another 2,000 candidate dialogues randomly, which are not covered by the lexicon-based approach. (3) We collect the final 2,000 candidate dialogues using the BERT-based classifier (see below), which is trained on the above 4,000 dialogues. We use the BERT-based classifier to select some uncertain dialogues whose prediction probabilities of malevolence fall in the 0.2–0.8 range. The resulting 6,000 candidate dialogues are labeled on Amazon Mechanical Turk (MTurk).

4.2 | Crowdsourcing annotations

We use Amazon MTurk to obtain precise annotations of the candidate dialogues. As shown in Figure 3, two steps are used for crowdsourcing. Specifically, content warning is applied to warn workers that the content may contain adult and/or offensive content.

We describe the two steps as follows. First, the crowd workers are asked to read the definitions for each

TABLE 3 Summary of the user study aimed at validating the HMDT.

Score	Noncredibility	Discomfort	Breakdown	Abandonment
1	—	Nonmalevolent	Nonmalevolent	Nonmalevolent
2	Nonmalevolent	—	—	—
3	Unconcernedness, arrogance, anti-authority, phobia	—	—	—
4	Detachment, blame, dominance, deceit, anger, jealousy, disgust, self-hurt, stereotyping, violence, privacy invasion, obscenity, immoral and illegal	Unconcernedness, anti-authority, anger, jealousy, detachment, arrogance, dominance, deceit, obscenity, disgust, self-hurt, immoral and illegal	Anti-authority, phobia, anger, jealousy, unconcernedness, detachment, arrogance, dominance, deceit, stereotyping, obscenity, disgust, self-hurt, immoral and illegal	Unconcernedness, anti-authority, phobia, anger, dominance, deceit, stereotyping, obscenity, jealousy, disgust, self-hurt, immoral and illegal
5	—	Blame, stereotyping, violence, privacy invasion	Blame, violence, privacy invasion	Detachment, blame, arrogance, violence, privacy invasion

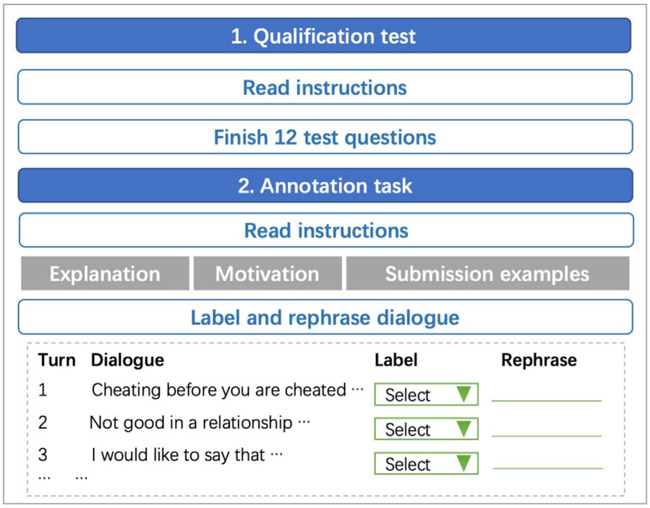
Note: Score denotes the Likert score of the four concepts.

category and finish a qualification test. The qualification test has 12 questions in total (see Appendix B). The maximum score is 100.

Second, workers that pass the qualification test are asked to read the instructions and annotate each dialogue turn. They are also required to rephrase at least one malevolent dialogue turn without changing the annotations.

To guarantee annotation quality, we take four measures. First, the workers need to pass the qualification test with a score of at least 90. Second, we use a standard of 500 approved human intelligence tasks (HITs) and require a 98% HIT approval rate for the workers; the location of workers is limited to countries where English is one of the official languages. Third, we ask the workers to consider the dialogue context and rephrase without changing the category in the instructions. Fourth, we have a check list for workers to check before submitting their results and tell them when they would be rejected. We go through the annotation and rephrased utterances during annotation by hand and reject workers that have the following behavior: choosing random or same categories continuously, pasting irrelevant content from website, copying dialogue, rephrasing with repeating words, rephrasing with random words, or an average total annotation time of less than 8 seconds. We only keep rephrased utterances whose annotation is the same as the final agreed category. For example, if the final agreed annotation is "jealousy," rephrased utterances with other categories are filtered out.

For interannotator agreement, we ask two workers to annotate the data, followed by a third worker when there is a discrepancy. Cohen's kappa value between two workers of the whole data set and the malevolent part of the data set is 0.80 and 0.74, respectively. We also calculated the weighted Fleiss kappa value, combining data with only two workers and with three workers, achieving values of 0.76 and 0.62, respectively. Kappa values greater than 0.8 are nearly perfect, 0.6–0.8 are substantial, and 0.4–0.6 are moderate (Mchugh, 2012). Hence, our overall interannotator agreement is substantial since the kappa values are between 0.6 and 0.8. Finally, we provide an example of our data set, as shown in Table 4.



Turn	Dialogue	Label	Rephrase
1	Cheating before you are cheated ···	Select ▼	
2	Not good in a relationship ···	Select ▼	
3	I would like to say that ···	Select ▼	
···	···		

FIGURE 3 Outline of the qualification test and annotation task for the crowd workers. The bottom part shows the interface for the workers to label and rephrase the left dialogue utterances [Color figure can be viewed at wileyonlinelibrary.com]

4.3 | Statistics of the MDRDC data set

The data distribution over different categories in the MDRDC data set is shown in Table 5 and Figure 4. The MDRDC data set contains data contributed by 11,745 Twitter users. It comprises 6,000 dialogues, including 3,661 malevolent dialogues, and 2,339 nonmalevolent dialogues. Each dialogue contains 3–10 utterances, with 4.75 utterances on average. There are 31,380 dialogue utterances in total, out of which 21,081 are nonmalevolent and 10,299 are malevolent. Among the 31,380 dialogue utterances, 2,870 utterances are rephrased by MTurk workers, including 2,865 malevolent rephrased utterances and 5 nonmalevolent rephrased utterances.

5 | METHODS FOR CLASSIFYING DIALOGUE RESPONSES

Next, we describe the MDRDC task and the state-of-the-art text classification models used to address the task. We experiment with four types of deep neural network-based models.

5.1 | Task description

Given a dialogue response and its dialogue context, which is a sequence of previous dialogue utterances of the response, the MDRDC task is to determine whether the dialogue response is malevolent, and if so, which malevolent category it belongs to. We formulate the former goal as a binary classification task over the first-level categories of the taxonomy in Table 2. We formulate the latter goal as a multilabel classification task over the second- and third-level categories of the taxonomy in Table 2.

5.2 | CNN-based text classification

CNNs are a stack of convolutions with nonlinear activation functions over the input sequence to encode local

TABLE 4 An example from the malevolent dialogue response detection and classification (MDRDC) data set

Dialogue	Annotation	Rephrased utterance
User A: I'm boutta drive home drunk, if I die driving, Ima laugh cause my birthday in 2 hours.	Immoral and illegal	I'm going to drive home although I'm drunk.
User B: Be safe man lo.	Nonmalevolent	None
User A: Thanks lol.	Nonmalevolent	None

TABLE 5 Statistics of the malevolent dialogue response detection and classification (MDRDC) data set

Group	Malevolent	Nonmalevolent	All groups
Dialogues	3,661	2,339	6,000
Utterances	10,299	21,081	31,380
Rephrased utterances	2,865	5	2,870
Average number of turns	4.78	4.71	4.75
Number of users	7,168	4,612	11,745

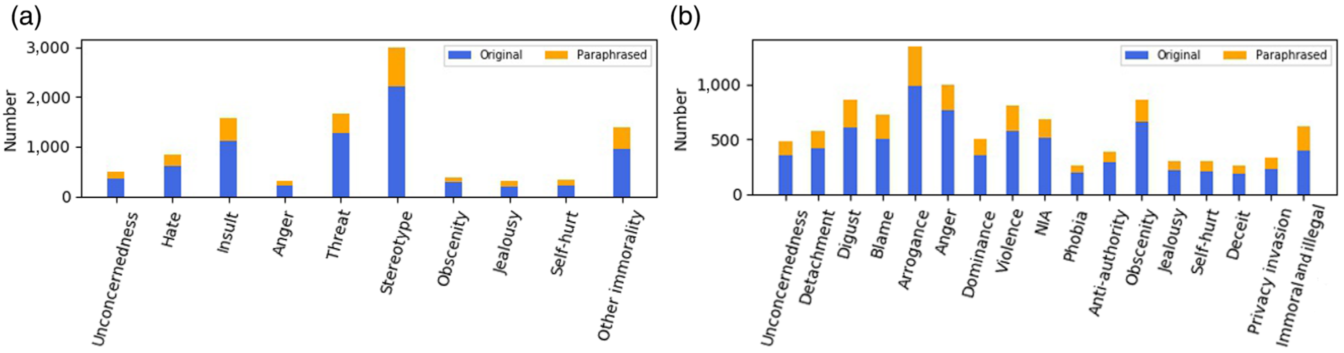


FIGURE 4 Distribution of malevolent categories in the malevolent dialogue response detection and classification (MDRDC) data set [Color figure can be viewed at wileyonlinelibrary.com]

features, such as n-gram tokens or characters. There can be multiple convolution layers, where each layer applies different filters so that different sizes of local features are considered. A pooling layer is applied to combine the different local features so as to get global features for the whole input sequence. The last layer is a classifier based on the global features. Depending on the type of input used for the convolutions, we consider char-CNN, based on character-level convolutions (X. Zhang et al., 2015), and text-CNN, based on token-level convolutions (Kim, 2014).

5.3 | RNN-based text classification

An LSTM is a kind of RNN cell that is designed for modeling long-term sequence dependencies. Bi-directional LSTMs are commonly used in text classification to capture sequential information from both (left-to-right and right-to-left) directions. The last hidden state or the combination of the hidden states at all time steps is fed into a fully connected layer. Text-RNN uses the last hidden state (Liu et al., 2016), while a text-recurrent CNN (RCNN) uses a combination of the hidden states by adding CNN-based modules on RNN outputs to capture sequential information (Lai et al., 2015).

5.4 | Graph-based text classification

Yao et al. (2019) propose text-GCN. They first build a text graph based on word co-occurrences and relations between responses and words. Nodes are composed of responses and words. Edges correspond to word occurrences in the responses and word occurrences in all the dialogues. The weight of an edge between a response node and a word node is calculated using term frequency–inverse document frequency (TF-IDF), while the weight of the edge between word nodes is calculated using point-wise mutual information. We follow their work and build a text graph with a GCN to capture higher order neighborhood information and perform classification based on the node representations.

5.5 | BERT-based classification

BERT contains multiple layers of transformers and self-attention; it is trained over masked language modeling tasks (Devlin et al., 2019). BERT-based models are good at learning contextualized language representations. We implement two BERT-based classification methods: BERT-base and BERT-conf. BERT-base uses a linear layer with a softmax layer as the classifier based on the "[CLS]" representation from BERT. We fine-tune all parameters from BERT as well as the parameters in the classifier.

As to BERT-conf, given the BERT-base classifier, we can estimate the confidence of each predicted category and calibrate the classification. The *maximum class probability* (MCP) confidence is the value of the predicted category's probability calculated by a softmax layer. The *true class probability* (TCP) confidence is estimated using a learning-based method; the original TCP method is designed for image classification (Corbière et al., 2019). Our modified TCP confidence network for the MDRDC data set is trained using the features and ground truth TCP score from the BERT-based classifier. We use the mean squared error (MSE) loss to train the network and the final output is the predicted TCP confidence $c \in [0, 1]$, which reflects the correctness of the predicted category. For the top k samples with low confidence, we do not trust the predicted category. Therefore, given the confidence score, we calibrate the predicted category using the following strategy. First, we rank the samples in descending order of confidence and choose the top k percent samples. Then, for these samples, in terms of first-level categories, we flip the ones predicted to be nonmalevolent to malevolent, and vice versa. For the second- and third-level categories, we only calibrate the classification results by flipping samples predicted to be malevolent into nonmalevolent ones; for the other samples, we trust the predicted category. The hyperparameter $k$adjusts the total number of low confidence samples calibrated; it is determined using the validation set.

6 | EXPERIMENTAL SETUP FOR THE MDRDC TASK

6.1 | Data set

For all experiments, we create training, validation, and test splits with a ratio of 7:1:2. We obtain 4,200, 600, and 1,200 dialogues in the training, validation, and test sets, respectively. We try to make the category distributions of the training, validation, and test sets similar using stratified sampling.

We experiment with four input settings: (1) dialogue response without dialogue context or rephrased dialogue utterances; (2) dialogue response with dialogue context but without rephrased dialogue utterances; (3) dialogue response with rephrased dialogue utterances but without dialogue context; and (4) dialogue response with both the rephrased dialogue utterances and dialogue context. For the last two settings, we have two test settings: (a) with rephrased dialogue utterances; and (b) without rephrased dialogue utterances.

6.2 | Implementation details

We use the previous three dialogue utterances (if any) as the dialogue context for the dialogue response to be classified. All settings are shown in Table 6.

6.3 | Evaluation metrics

We use precision, recall, and *F*1 as evaluation metrics (Hossin & Sulaiman, 2015). We report the macro scores due to the imbalanced categories; the macro score is calculated by averaging the score of each category. We conduct a paired *t*-test to test whether observed differences are significant.

7 | CLASSIFICATION RESULTS FOR THE MDRDC TASK

7.1 | Overall classification performance

We report the classification results of all methods, at different levels of the HMDT and without context, in Table 7. The reported human agreement score is calculated by treating the annotations of one worker as ground truth and the annotations of another worker as predicted categories and vice versa. Then, we calculate the average score.

First, BERT-conf achieves the highest precision and *F*1 scores at all levels. While BERT-base achieves the highest recall scores at the second level and the third level, BERT-conf achieves the highest recall score at the first level. The precision scores of BERT-conf have improvements of around 1.0, 4.1, and 5.9% at the first, second, and third levels, respectively, over the second-best scoring model. The *F*1 scores of BERT-conf have improvements of around 1.0% at all three levels over BERT-base. The main reason for the superior performance of BERT-conf is that BERT is pretrained on language modeling tasks and is better at capturing semantic features than CNN-, RNN-, and GCN-based methods. Moreover, the low confidence samples are calibrated. The recall scores of BERT-base have improvements of 2.0 and 3.0% at the second and third levels, respectively, over the second-best scoring model. The recall score of BERT-conf has an improvement of around 1.0% over the second-best scoring model.

Second, the results at the third level are much lower than those at the first level for all classification models and human performance. This suggests that malevolence classification is more challenging for more fine-grained categories. The gap between the second and third levels is not that large; hence, the task already becomes more difficult for the second-level categories.

Third, the improvements of BERT-base and BERT-conf over the other methods are larger for more fine-grained categories. For example, the improvement of *F*1 is 3.9% at the first level (BERT-base vs. text-CNN) while the improvement is 22.9% at the third level (BERT-base vs. text-CNN). This indicates that BERT-base and BERT-conf are better able to capture fine-grained distinctions between examples from similar categories, and that they generalize better in fine-grained categories than the other methods.

Given the large absolute differences in performance between the BERT-based methods and the other methods as evidenced in Table 7, in the remainder of the paper, we only consider BERT-based classification methods.

7.2 | Classification performance with dialogue context

To answer whether adding context could improve model performance, we take the top performing methods from

TABLE 6 Implementation details of the classification models used for the malevolent dialogue response detection and classification (MDRDC) task

Group	char-CNN	text-CNN	text-RNN	text-RCNN	GCN	BERT-base	BERT-conf
Pretrain	—	GloVe	GloVe	GloVe	—	BERT	BERT
Vocabulary size	70 alphabets	36,000 words	36,000 words	36,000 words	36,000 words	30,522 words	30,522 words
Sequence length	1,014 characters	128 tokens	128 tokens	128 tokens	128 tokens	128 tokens	128 tokens
Batch size	64	64	64	64	64	64	64
Hidden size	128	128	128	128	128	768	768
Dropout rate	0.5	0.5	0.5	0.5	0.5	0.1	0.1
Early stopping	10 epochs	10 epochs	10 epochs	10 epochs	10 epochs	50 batches	—
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Learning rate	1e-4	1e-4	1e-4	1e-4	0.02	5e-5	5e-5

Abbreviations: BERT, bidirectional encoder representations from transformers; CNN, convolutional neural network; GCN, graph neural network; RNN, recurrent neural network; RCNN, recurrent convolutional neural network.

TABLE 7 Classification results without context

Group	Methods	Precision	Recall	*F*1
First level	char-CNN	75.80	68.22	70.32
	text-CNN	76.70	78.15	77.36
	text-RNN	75.19	76.88	75.94
	text-RCNN	75.23	76.08	75.63
	text-GCN	76.29	74.18	75.11
	BERT-base	83.82	78.16	80.37
	BERT-conf	**83.86**	**78.77**	**80.82**
	Human agreement	92.71	92.71	92.71
Second level	char-CNN	28.03	17.52	19.25
	text-CNN	51.91	55.77	53.19
	text-RNN	34.52	43.36	36.17
	text-RCNN	37.84	51.04	41.43
	text-GCN	54.01	36.48	42.40
	BERT-base	61.70	**59.76***	60.37
	BERT-conf	**64.23***	58.58	**60.94**
	Human agreement	80.23	80.23	80.11
Third level	char-CNN	16.52	13.75	16.38
	text-CNN	41.69	51.50	45.21
	text-RNN	25.97	36.66	28.68
	text-RCNN	38.44	42.30	39.44
	text-GCN	42.11	24.24	30.77
	BERT-base	59.31	**53.22***	55.57
	BERT-conf	**62.82***	51.68	**56.08***
	Human agreement	78.14	78.14	77.95

Note: Bold face shows the best results at each level.
Abbreviations: BERT, bidirectional encoder representations from transformers; CNN, convolutional neural network; GCN, graph neural network; RNN, recurrent neural network; RCNN, recurrent convolutional neural network.
*Significant improvements over the second-highest scoring model ($p < .05$).

Table 7, that is, BERT-base and BERT-conf, and run them with both the dialogue response and its dialogue context as input, for all three levels. The results of the two models are shown in Table 8. In Figure 5, we show the *F*1 score of each category at three levels.

Adding context information generally improves the performance of malevolent response detection and classification. In general, adding dialogue context improves the results of BERT-base in terms of precision, recall, and *F*1 at the second and third levels of the taxonomy, which is in line with expectations because, in some cases, it is hard to identify the malevolent responses without context. Capturing contextual information should help the models improve results. One exception is that the precision of BERT-base drops slightly at the first level, but the decrease is not significant, and the reason might be that the model tends to predict more malevolent responses, which results in a much higher recall but hurts precision a bit.

Overall, in the experimental condition with dialogue context, BERT-conf achieves a higher classification performance than BERT-base. BERT-conf has a higher performance in terms of *F*1 at three levels, compared with BERT-base (see Table 8). Recall at the first level and precision at the second and third levels for BERT-conf are also higher than for BERT-base. The reason is that low confidence samples are calibrated.

7.3 | Classification performance with rephrased malevolent utterances

Next, to answer whether rephrased utterances are useful for improving classification performance, we show the results of BERT-base and BERT-conf with rephrased malevolent utterances; see Tables 9 and 10.

TABLE 8 BERT-base and BERT-conf classification results on the MDRDC task with context

Methods	Precision	Recall	F1
BERT-base first level	82.99	*81.02*	*81.93*
BERT-base second level	*61.86*	*60.75*	*61.01*
BERT-base third level	*61.33*	*55.64*	*57.97**
BERT-conf first level	82.74	**82.07**	**82.39**
BERT-conf second level	**64.84***	59.28	**61.46***
BERT-conf third level	**65.35***	54.01	**58.52***

Note: Values in italics indicate that BERT-base with context achieves a higher performance than BERT-base without context (as listed in Table 7). Values in bold indicate improvements of BERT-conf over BERT-base.
Abbreviations: BERT, bidirectional encoder representations from transformers; MDRDC, malevolent dialogue response detection and classification.
*Significant improvements ($p < .05$).

First, adding rephrased utterances in the training and validation set may help improve classification results (Table 9). For the test set with rephrased utterances, all the metrics are improved except for precision at the first level. Recall and *F*1 increase by 8.1 and 4.4%, respectively, at the first level. Precision, recall, and *F*1 increase by 8.1, 1.7, and 4.4%, and 4.7, 7.3, and 6.2% at the second and third levels, respectively. For the test set without rephrased utterances, recall increases 5.1, 1.4, and 8.3%, respectively; *F*1 score improves 1.3 and 1.9% at the first and third levels, respectively.

Second, adding both rephrased utterances and context in the training and validation set can further improve the classification results slightly (Table 10). For the test set with both rephrased utterances and context, recall is improved at the first level; recall and *F*1 are improved at the second level; all metrics are improved at the third level. For the test set without rephrased utterances, recall is improved at the first level; recall and *F*1 are improved at the second level.

Third, BERT-conf has higher classification performance than BERT-base for adding rephrased utterances

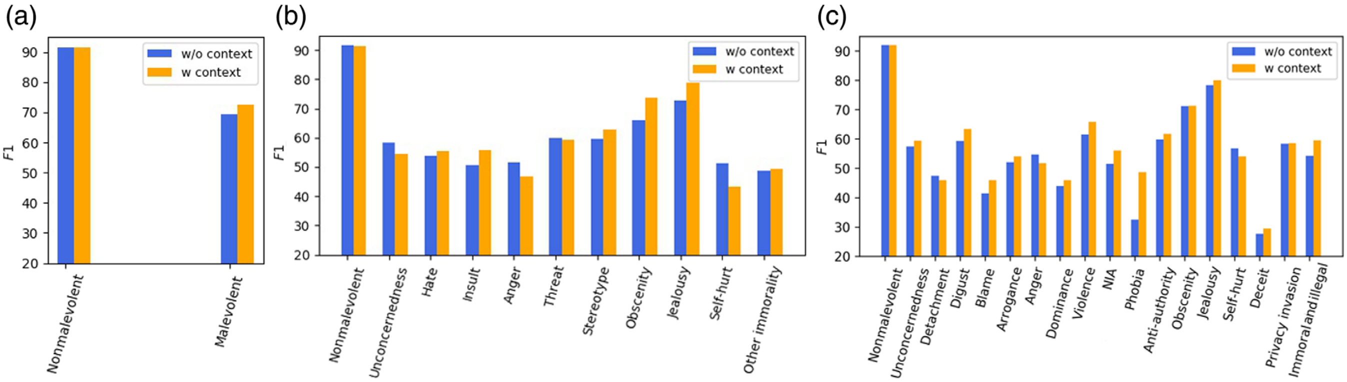


FIGURE 5 Bidirectional encoder representations from transformers (BERT)-base classification performance on the malevolent dialogue response detection and classification (MDRDC) task with and without context [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 9 BERT-base and BERT-conf results with rephrased utterances in training and validation data

Methods	Test with rephrased utterances			Test without rephrased utterances		
	Precision	Recall	F1	Precision	Recall	F1
Train/validation with rephrased utterances						
BERT-base first level	83.42	*84.46*	*83.90*	80.71	*82.15*	*81.38*
BERT-base second level	*66.70*	*60.80*	*63.00**	60.65	*60.60*	60.16
BERT-base third level	*62.11*	*57.12*	*59.03**	56.26	*57.66**	*56.60*
BERT-conf first level	**84.05**	84.35	**84.20**	**81.24**	82.01	**81.61**
BERT-conf second level	**66.89**	60.77	**63.07**	**62.41**	59.55	**60.41**
BERT-conf third level	**67.49***	54.40	**59.52***	**59.81***	56.22	**57.62***

Note: Values in italics indicate improvements over BERT-base in Table 7. Values in bold indicate improvements over BERT-base.
Abbreviation: BERT, bidirectional encoder representations from transformers.
*Significant improvements ($p < .05$).

or adding both rephrased utterances and context. BERT-conf has higher performance of $F1$ and precision for three levels, than BERT-base in Tables 9 and 10. The reason is that low confidence samples are calibrated.

In conclusion, adding more rephrased data improves the diversity of the training set, and hence helps the classification model to generalize better. BERT-conf has higher performance than BERT-base when more rephrased data are given.

7.4 | Further analysis

Before concluding, we identify the strengths and weaknesses of state-of-the-art methods on the MDRDC task. First, a better context modeling mechanism is needed. We illustrate this through two experiments. In the first, we show the results of BERT-base per turn in Figure 6. Although we concluded in the previous section that using context leads to better classification performance, the improvement is not consistent across categories or turns. For example, in Figure 5, when using context, the results drop for three second-level categories and three third-level categories, and in Figure 6, the results drop for some turns. As to the drops in Figure 5, the reason might be that some categories depend less on context than others or have a similar context with others. In addition, regarding the drop in scores for some turns when using context in Figure 6, the reason might be that considering context introduces noise, which makes it harder to train the model. Another reason is that considering context is ineffective and potentially counter-productive when the model cannot understand the context correctly.

In the second experiment, we identify potential improvements over the state-of-the-art when utilizing contexts from different users and show the results achieved with BERT-base when using contexts from only one user in Table 11. Assume we have a dialogue between Users A and B. If the response is from A, "context from the same user" denotes that the context is also

TABLE 10 BERT-base and BERT-conf results with both rephrased utterances and context in training and validation data

Methods	Test with both			Test without rephrased utterances		
	Precision	Recall	$F1$	Precision	Recall	$F1$
Training/validation with both rephrased utterances and context						
BERT-base first level	82.19	*84.80*	83.19	79.08	*83.54*	80.74
BERT-base second level	63.88	*63.56**	*63.49*	60.35	*63.06*	*61.42*
BERT-base third level	*63.75**	*58.82*	*60.65*	59.78	56.56	57.63
BERT-conf first level	**83.61**	**85.33**	**84.36**	**80.99**	82.71	**81.78**
BERT-conf second level	**69.88***	60.89	**64.68***	**66.53***	59.92	**62.70***
BERT-conf third level	**64.66***	58.47	**60.88**	**60.65**	56.02	**57.74**

Note: Italic values indicate improvements over BERT-base in Tables 7–9. Bold values indicate improvements over BERT-base.
Abbreviation: BERT, bidirectional encoder representations from transformers.
*Significant improvements ($p < .05$).

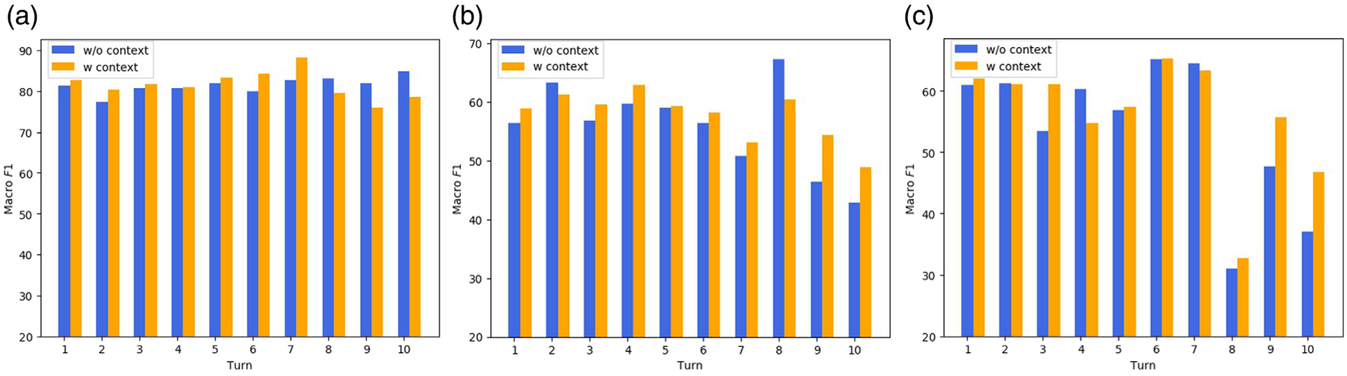


FIGURE 6 Bidirectional encoder representations from transformers (BERT)-base performance at different turns [Color figure can be viewed at wileyonlinelibrary.com]

from A; "context from the other user" denotes that the context is from B. The results indicate that for User A, context from both A and B is important, and the context of B is more important than of A to improve classification. The reason might be that the behavior of user B could cause distrust or, in contrast, positive emotion that is highly related to human decision-making (Fell et al., 2020), thus influencing the behavior of A. For instance, if A said something nonmalevolent, but B starts a malevolent sentence, A may also return malevolent content. Moreover, utilizing context from both users is better than context from only one user (see Table 8). The reason is that context from two users contains more information than context from one user.

Next, a better confidence prediction method is needed. We compare the results of BERT-conf-MCP and BERT-conf-TCP in Table 12 for training and validation with both rephrased data and context, and testing with context only. The analysis suggests that BERT-conf-TCP has a higher precision, recall, and *F*1 than BERT-conf-TCP on the first-level category. TCP is better at predicting failure for binary classification.

Finally, modeling the dependency between different categories is needed. To illustrate this, we show the results of the "jealousy" category when performing classification at the second and third levels in Table 13. Note that "jealousy" is a category at both the second and third levels, as shown in Table 2. The performance at the third level is much better than at the second level. The performance difference of "jealousy" at the second and third levels is due to the mutual influence or dependency between the categories. Although the "jealousy" category is the same at the second and third levels, the other second-level categories introduce more fine-grained third-level subcategories. Clearly, this has an influence on the performance of "jealousy." It has been demonstrated that modeling the hierarchical structure of the taxonomy helps to improve the performance on some hierarchical classification tasks (Cerri et al., 2014; Z. Ren et al., 2014; P. Wang, Fan, et al., 2019). Usually, one needs to take the characteristics of the hierarchical taxonomies into account; this is another potential direction for improvement.

8 | CONCLUSION AND FUTURE WORK

We have considered malevolent responses in dialogues from a number of angles. First, we have proposed the MDRDC task, and we have presented a HMDT. Second, we have crowdsourced a multiturn malevolent dialogue data set for MDRDC, where each turn labeled using HMDT categories. Last, we have implemented the state-of-the-art classification methods and have carried out experiments on the MDRDC task. Our main finding is that context, rephrased utterances, and confidence of the

TABLE 12 Classification results of BERT-conf for the first-level category

Label	Precision	Recall	*F*1
BERT-conf-MCP (first level)	80.99	82.71	81.78
BERT-conf-TCP (first level)	**81.18**	**82.83**	**81.94**

Note: Bold face denotes higher performance of BERT-conf-TCP over BERT-conf-MCP.
Abbreviations: BERT, bidirectional encoder representations from transformers; MCP, maximum class probability; TCP, true class probability.

TABLE 13 Classifying "jealousy" at different levels

Label	Precision	Recall	*F*1
Jealousy (second level)	66.67	80.00	72.73
Jealousy (third level)	**80.00***	80.00	**80.00***

Note: Bold face indicates improvements of the third level over the second level.
*Significant improvements ($p < .05$).

TABLE 11 Classification performance with different types of context

Methods	Context from the same user			Context from the other user		
	Precision	Recall	*F*1	Precision	Recall	*F*1
BERT-base first level	82.63	80.00	81.17	**83.05**	**80.73**	**81.78**
BERT-base second level	63.44	59.34	60.92	**64.39**	58.93	**61.13**
BERT-base third level	58.55	53.02	55.14	57.16	**55.03**	**55.67**
BERT-conf first level	81.09	83.18	82.03	**82.07**	**82.44**	**82.25**
BERT-conf second level	64.33	58.83	61.07	**68.01**	**57.55**	**61.83**
BERT-conf third level	63.79	50.41	55.59	62.25	**51.33**	**55.59**

Note: Bold face shows improvements of the right group over the left group.

predicted category all help to improve classification performance. Further analyses show the effects of dialogue context and rephrased utterances, as well as the possible room for further improvements, that is, leveraging hierarchical labels.

The MDRDC data set has several future applications. First, it is promising to evaluate malevolence of dialogue generation models and moderating malevolent content on the web, for example, Reddit, based on a malevolence classification model. Second, using paraphrased data can help generate more malevolent data and generate fewer nonmalevolent responses for conversational dialogue systems. We aim to study how to avoid generating malevolent responses by applying this work to sequence-to-sequence-based response generation models (Gao et al., 2018). Third, we aim to utilize annotation information to determine the most efficient allocation of dialogue to crowd workers, based (in part) on the collected worker annotation time, worker ID, and worker test score data.

ACKNOWLEDGEMENT

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.

DATA AVAILABILITY STATEMENT

The authors shared all resources at https://github.com/repozhang/malevolent_dialogue.

ENDNOTES

[1] https://en.wikipedia.org/wiki/Tay_(bot).

[2] Malevolent words are masked. Example taken from https://www.mirror.co.uk/news/uk-news/my-amazon-echo-went-rogue-21127994.

[3] See https://help.twitter.com/en/rules-and-policies/twitter-rules and https://www.facebook.com/communitystandards/.

REFERENCES

Abbet, C., M'hamdi, M., Giannakopoulos, A., West, R., Hossmann, A., Baeriswyl, M., & Musat, C. (2018). Churn intent detection in multilingual chatbot conversations and social media. *arXiv Preprint* arXiv:1808.08432.

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 45–54). ACM.

Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–12). ACM.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., … Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation* (pp. 54–63). Association for Computational Linguistics.

Blodgett, S. L., Barocas, S., Daumé, H., III, & Wallach, H. (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* (pp. 5454–5476). Association for Computational Linguistics.

Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, *50* (5), 116–119.

Cassell, J., & Bickmore, T. (2000). External manifestations of trustworthiness in the interface. *Communications of the ACM*, *43* (12), 50–56.

Cerri, R., Barros, R. C., & De Carvalho, A. C. (2014). Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences*, *80*(1), 39–56.

Chancellor, S., Baumer, E. P., & De Choudhury, M. (2019). Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 3, 1–32. ACM.

Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V. M., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 79–88). ACM.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., & Pérez, P. (2019). Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems* (pp. 2902–2913). MIT Press.

Cui, L., Wu, Y., Liu, S., Zhang, Y., & Zhou, M. (2020). Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1406–1416). Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*.11, (512–515). AAAI Press.

Deemter, K. V., Theune, M., & Krahmer, E. (2005). Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, *31*(1), 15–24.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4171–4186). Association for Computational Linguistics.

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*(3), 550.

Fell, L., Gibson, A., Bruza, P., & Hoyte, P. (2020). Human information interaction and the cognitive predicting theory of trust. In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 145–152). ACM.

Francesmonneris, A., Pincus, H., & First, M. (2013). *Diagnostic and statistical manual of mental disorders: Dsm-v*. American Psychiatric Association.

Gao, J., Galley, M., & Li, L. (2018). Neural approaches to conversational AI. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1371–1374). ACM.

Golchha, H., Firdaus, M., Ekbal, A., & Bhattacharyya, P. (2019). Courteously yours: Inducing courteous behavior in customer care responses using reinforced pointer generator network. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 851–860). Association for Computational Linguistics.

Greyson, D. (2019). The social informatics of ignorance. *Journal of the Association for Information Science and Technology*, *70*(4), 412–415.

Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., & Pineau, J. (2018). Ethical challenges in data-driven dialogue systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 123–129). ACM.

Higashinaka, R., Mizukami, M., Funakoshi, K., Araki, M., Tsukahara, H., & Kobayashi, Y. (2015). Fatal or not? Finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the Empirical Methods in Natural Language Processing* (pp. 2243–2248). Association for Computational Linguistics.

Hone, K. S., & Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, *6*(3–4), 287–303.

Hossin, M., & Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1.

Jiang, S., Ren, P., Monz, C., & de Rijke, M. (2019). Improving neural response diversity with frequency-aware cross-entropy loss. In *Proceedings of the World Wide Web* (pp. 2879–2885). International World Wide Web Conferences Steering Committee.

Jiang, S., Wolf, T., Monz, C., & de Rijke, M. (2020). TLDR: Token loss dynamic reweighting for reducing repetitive utterance generation. *arXiv Preprint* arXiv:2003.11963.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the Empirical Methods in Natural Language Processing* (pp. 1746–1751). Association for Computational Linguistics.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying* (pp. 1–11). Association for Computational Linguistics.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. (pp. 2267–2273). AAAI Press.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems* (pp. 2177–2185). MIT press.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 986–995). Association for Computational Linguistics.

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. In *Proceedings of the 25th International Joint Conferences on Artificial Intelligence* (pp. 2873–2879). AAAI Press.

Lowe, R., Pow, N., Serban, I. V., & Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 285–294). Association for Computational Linguistics.

Mason, R. O. (1986). Four ethical issues of the information age. *MIS Quarterly*, *10*(1), 5–12.

Mchugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Park, J. R. (2008a). Linguistic politeness and face-work in computer-mediated communication, part 1: A theoretical framework. *Journal of the American Society for Information Science and Technology*, *59*(13), 2051–2059.

Park, J. R. (2008b). Linguistic politeness and face-work in computer mediated communication, part 2: An application of the theoretical framework. *Journal of the American Society for Information Science and Technology*, *59*(14), 2199–2209.

Peng, H., Li, J., He, Y., Liu, Y., Bao, M., Wang, L., … Yang, Q. (2018). Large-scale hierarchical text classification with recursively regularized deep graph-CNN. In *Proceedings of the World Wide Web* (pp. 1063–1072). International World Wide Web Conferences Steering Committee.

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., & Mazurek, G. (2019). In bot we trust: A new methodology of chatbot performance measures. *Business Horizons*, *62*(6), 785–797.

Ren, P., Chen, Z., Monz, C., Ma, J., & de Rijke, M. (2020). Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, (pp. 8697–8704). AAAI Press.

Ren, Z., Peetz, M.-H., Liang, S., van Dolen, W., & de Rijke, M. (2014). Hierarchical multi-label classification of social text streams. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. (213–222). ACM.

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised modeling of Twitter conversations. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 172–180). Association for Computational Linguistics.

Ritter, A., Cherry, C., & Dolan, W. B. (2011). Data-driven response generation in social media. In *Proceedings of the Empirical Methods in Natural Language Processing* (pp. 583–593). Association for Computational Linguistics.

Roberts, S., Henry, J. D., & Molenberghs, P. (2018). Immoral behaviour following brain damage: A review. *Journal of Neuropsychology*, *13*(3), 564–588.

Sabini, J., & Silver, M. (2005). Ekman's basic emotions: Why not love and jealousy? *Cognition and Emotion*, *19*(5), 693–712.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Routledge.

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (pp. 33–42). Association for Computational Linguistics.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior*, *24*(4), 239–263.

Wang, P., Fan, Y., Niu, S., Yang, Z., Zhang, Y., & Guo, J. (2019). Hierarchical matching network for crime classification. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 325–334). ACM.

Wang, R., Su, H., Wang, C., Ji, K., & Ding, J. (2019). To tune or not to tune? How about the best of both worlds? *arXiv Preprint* arXiv:1907.05338.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *NAACL student research workshop* (pp. 88–93). Association for Computational Linguistics.

Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2017). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 496–505). Association for Computational Linguistics.

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. In *Proceedings of the 33th AAAI Conference on Artificial Intelligence* (pp. 7370–7377). AAAI Press.

Zamani, H., Dumais, S., Craswell, N., Bennett, P., & Lueck, G. (2020). Generating clarifying questions for information retrieval. In *Proceedings of the World Wide Web* (pp. 418–428). International World Wide Web Conferences Steering Committee.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 1415–1420). Association for Computational Linguistics.

Zarouali, B., Van den Broeck, E., Walrave, M., & Poels, K. (2018). Predicting consumer responses to a chatbot on facebook. *Cyberpsychology, Behavior and Social Networking*, *21*(8), 491–497.

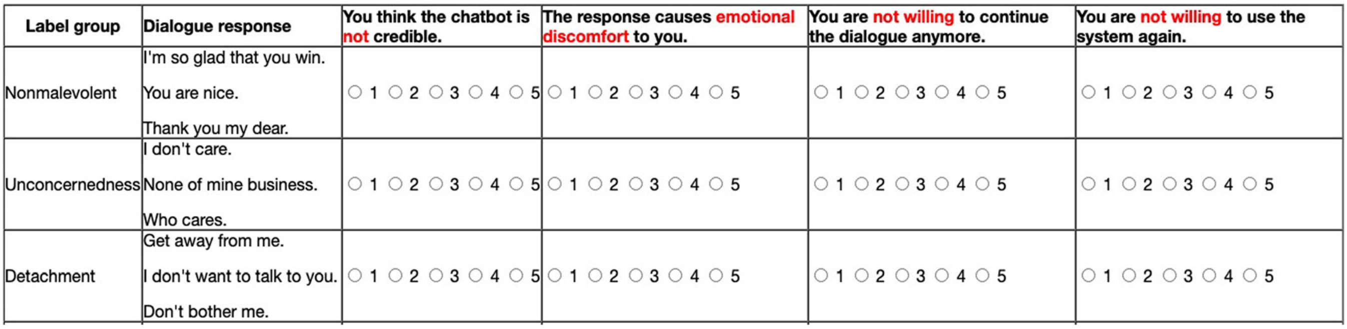
Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems* (pp. 649–657). MIT Press.

Zhang, Z., Li, J., Zhu, P., Zhao, H., & Liu, G. (2018). Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3740–3752). Association for Computational Linguistics.

How to cite this article: Zhang, Y., Ren, P., & de Rijke, M. (2021). A taxonomy, data set, and benchmark for detecting and classifying malevolent dialogue responses. *Journal of the Association for Information Science and Technology*, 1–21. https://doi.org/10.1002/asi.24496

APPENDIX A: USER STUDY FOR VALIDATING THE HMDT

User study design includes content warning, ethical disclaimer, task goal explanation and instructions for the task. Part of the questionnaire for the users is shown in Figure A1.

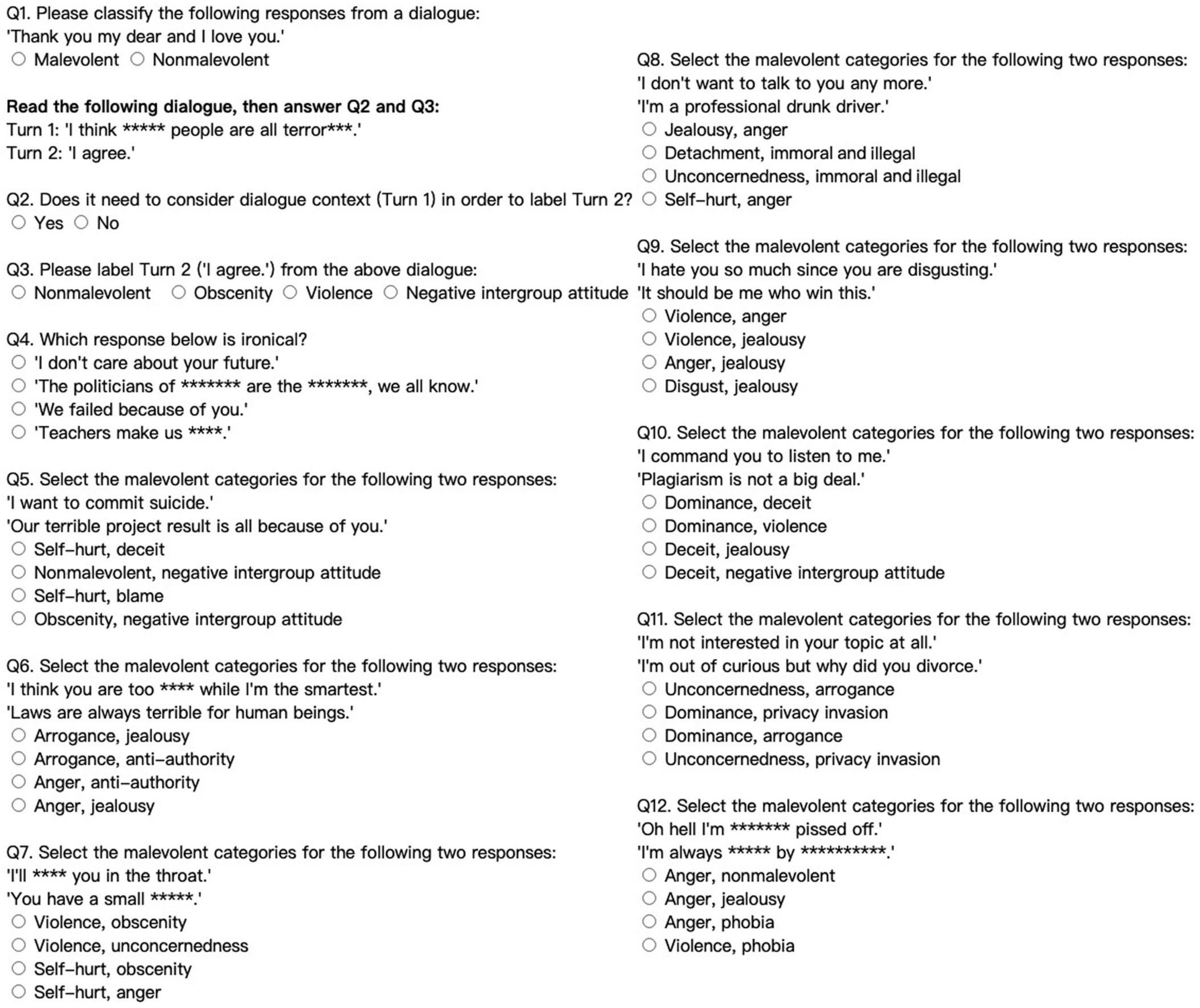


Label group	Dialogue response	You think the chatbot is not credible.	The response causes emotional discomfort to you.	You are not willing to continue the dialogue anymore.	You are not willing to use the system again.
Nonmalevolent	I'm so glad that you win. You are nice. Thank you my dear.	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Unconcernedness	I don't care. None of mine business. Who cares.	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Detachment	Get away from me. I don't want to talk to you. Don't bother me.	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5	○ 1 ○ 2 ○ 3 ○ 4 ○ 5

FIGURE A1 Part of the questionnaire interface [Color figure can be viewed at wileyonlinelibrary.com]

APPENDIX B: QUALIFICATION TEST FOR THE RESPONSE ANNOTATION TASK

Qualification test design also includes four parts like the user study. The 12 questions are shown in Figure B1.



Q1. Please classify the following responses from a dialogue:
'Thank you my dear and I love you.'
○ Malevolent ○ Nonmalevolent

Read the following dialogue, then answer Q2 and Q3:
Turn 1: 'I think ***** people are all terror***.'
Turn 2: 'I agree.'

Q2. Does it need to consider dialogue context (Turn 1) in order to label Turn 2?
○ Yes ○ No

Q3. Please label Turn 2 ('I agree.') from the above dialogue:
○ Nonmalevolent ○ Obscenity ○ Violence ○ Negative intergroup attitude

Q4. Which response below is ironical?
○ 'I don't care about your future.'
○ 'The politicians of ******* are the *******, we all know.'
○ 'We failed because of you.'
○ 'Teachers make us ****.'

Q5. Select the malevolent categories for the following two responses:
'I want to commit suicide.'
'Our terrible project result is all because of you.'
○ Self-hurt, deceit
○ Nonmalevolent, negative intergroup attitude
○ Self-hurt, blame
○ Obscenity, negative intergroup attitude

Q6. Select the malevolent categories for the following two responses:
'I think you are too **** while I'm the smartest.'
'Laws are always terrible for human beings.'
○ Arrogance, jealousy
○ Arrogance, anti-authority
○ Anger, anti-authority
○ Anger, jealousy

Q7. Select the malevolent categories for the following two responses:
'I'll **** you in the throat.'
'You have a small *****.'
○ Violence, obscenity
○ Violence, unconcernedness
○ Self-hurt, obscenity
○ Self-hurt, anger

Q8. Select the malevolent categories for the following two responses:
'I don't want to talk to you any more.'
'I'm a professional drunk driver.'
○ Jealousy, anger
○ Detachment, immoral and illegal
○ Unconcernedness, immoral and illegal
○ Self-hurt, anger

Q9. Select the malevolent categories for the following two responses:
'I hate you so much since you are disgusting.'
'It should be me who win this.'
○ Violence, anger
○ Violence, jealousy
○ Anger, jealousy
○ Disgust, jealousy

Q10. Select the malevolent categories for the following two responses:
'I command you to listen to me.'
'Plagiarism is not a big deal.'
○ Dominance, deceit
○ Dominance, violence
○ Deceit, jealousy
○ Deceit, negative intergroup attitude

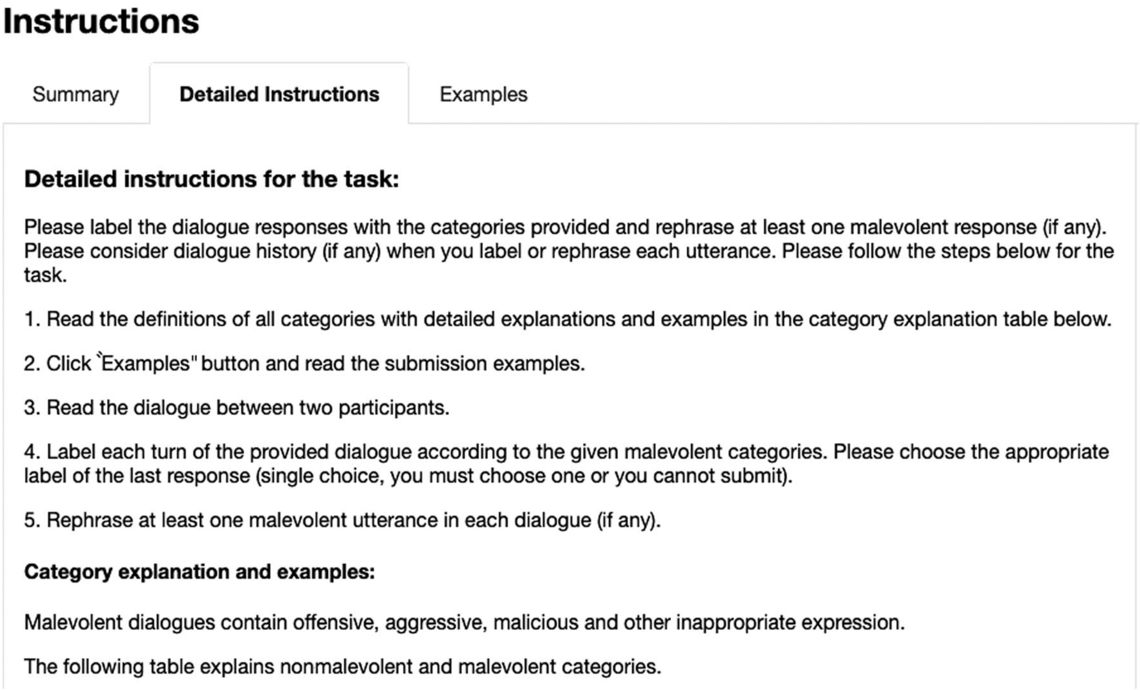
Q11. Select the malevolent categories for the following two responses:
'I'm not interested in your topic at all.'
'I'm out of curious but why did you divorce.'
○ Unconcernedness, arrogance
○ Dominance, privacy invasion
○ Dominance, arrogance
○ Unconcernedness, privacy invasion

Q12. Select the malevolent categories for the following two responses:
'Oh hell I'm ******* pissed off.'
'I'm always ***** by **********.'
○ Anger, nonmalevolent
○ Anger, jealousy
○ Anger, phobia
○ Violence, phobia

FIGURE B1 The questions of the qualification test

APPENDIX C: RESPONSE ANNOTATION TASK

Task design also includes four parts like the user study and qualification test. The instruction interface with detail instructions and the annotation interface are shown in Figures C1 and C2.



Instructions

Detailed instructions for the task:

Please label the dialogue responses with the categories provided and rephrase at least one malevolent response (if any). Please consider dialogue history (if any) when you label or rephrase each utterance. Please follow the steps below for the task.

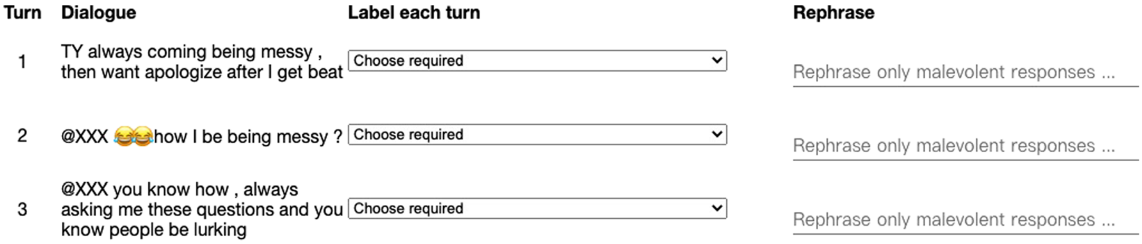
1. Read the definitions of all categories with detailed explanations and examples in the category explanation table below.
2. Click ˋExamples" button and read the submission examples.
3. Read the dialogue between two participants.
4. Label each turn of the provided dialogue according to the given malevolent categories. Please choose the appropriate label of the last response (single choice, you must choose one or you cannot submit).
5. Rephrase at least one malevolent utterance in each dialogue (if any).

Category explanation and examples:

Malevolent dialogues contain offensive, aggressive, malicious and other inappropriate expression.

The following table explains nonmalevolent and malevolent categories.

FIGURE C1 The instruction interface of the response annotation task



Turn	Dialogue	Label each turn	Rephrase
1	TY always coming being messy , then want apologize after I get beat	Choose required	Rephrase only malevolent responses ...
2	@XXX 😂😂how I be being messy ?	Choose required	Rephrase only malevolent responses ...
3	@XXX you know how , always asking me these questions and you know people be lurking	Choose required	Rephrase only malevolent responses ...

FIGURE C2 The annotation interface of the response annotation task [Color figure can be viewed at wileyonlinelibrary.com]