Malevolent dialogue response detection and evaluation

Zhang, Y.

Publication date
2022
Document Version
Final published version

A malevolent response is a kind of dialogue response that might contain offensive or objectionable content including hate, insult, threat, etc. In this thesis, we first analyze the malevolence problem of state-of-the-art generation models with malevolence detection models. Second, we introduce taxonomies, datasets, and methods for single-label dialogue malevolence detection. Third, we build datasets and methods for multi-label dialogue malevolence detection from a single-label training set. The taxonomy of multi-label dialogue malevolence detection is the same as single-label detection. Finally, we propose a human-machine collaborative evaluation framework for dialogue malevolence evaluation.

ISBN 978-94-93278-30-1

MALEVOLENT DIALOGUE RESPONSE DETECTION AND EVALUATION

YANGJUN ZHANG

MALEVOLENT DIALOGUE RESPONSE DETECTION AND EVALUATION

YANGJUN ZHANG

Malevolent Dialogue Response Detection and Evaluation

Yangjun Zhang

Malevolent Dialogue Response Detection and Evaluation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in
de Agnietenkapel
op vrijdag 2 december 2022, te 10.00 uur

door

Yangjun Zhang

geboren te Anhui

Promotiecommissie

Promotor:	Prof. dr. M. de Rijke	Universiteit van Amsterdam
Co-promotor:	Prof. dr. P. Ren	Shandong University
Overige leden:	Prof. dr. H. Haned	Universiteit van Amsterdam
	Prof. dr. E. Kanoulas	Universiteit van Amsterdam
	Prof. dr. C. Monz	Universiteit van Amsterdam
	Prof. dr. V.T. Rieser	Heriot Watt University
	Dr. M. Huang	Tsinghua University

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research was supported by the China Scholarship Council.

Cover by Yanbing Xu
Printed by Off Page, Amsterdam, The Netherlands

ISBN: 978-94-93278-30-1

Acknowledgment

The journey of pursuing my PhD at the University of Amsterdam started in October 2018. Today, I am wrapping up the research plan that I developed over the years and I have now been allowed to defend my thesis. The book is the result of four years of hard work. I really appreciate the help from those who have provided academic guidance and social support over the past four years.

First and foremost, I would like to sincerely thank my supervisor Professor Maarten de Rijke. You have given me the opportunity to pursue this PhD. Your supervision has helped me to decide the research direction and the whole research process, from experiments to writing. Thank you, Maarten!

Throughout the process, Professor Pengjie Ren's advice has been detailed and insightful. It made a significant difference in accomplishing my research goals. I would also like to thank Professor Christof Monz who always provided helpful perspectives during my study. Thank you both!

I am honored to have Christof, Evangelos, Hinda, Minlie, and Verena as my committee members. Thank you for reviewing my thesis and giving valuable feedback.

I would also like to thank the professors who instructed my studies and helped my applications in the past. I thank Professor Zhanjun Zhang and Professor Yufang Bian from my master's program at Beijing Normal University. I thank Professor Meijia Zhang from my bachelor's program at China Agricultural University. I thank the professors from my second master's program at City University of Hong Kong: Professor Guangwu Liu, Doctor Siu Keung Tse, and Doctor Hak Keung Yuen.

During my four years at the University of Amsterdam, I met so many friendly people who have helped me with my research and daily life. Many thanks to Ali A, Ali V, Amir, Ana, Andrew, Anna, Antonios, Arezoo, Barrie, Chang, Chuan, Clemencia, Dan, David, Gabriel, Georgios, Hamid, Harrie, Hinda, Hongyu, Hosein, Ilya, Jiahuan, Jie, Jin, Jingfen, Julien, Maarten M, Maartje, Maria, Mariya, Marzieh, Maurits, Masha, Ming, Mohammad, Mostafa, Mounia, Mozhdeh, Nikos, Olivier, Peilei, Philipp, Pooya, Praveen, Rolf, Romain, Ruben, Ruqing, Samarth, Sam, Sami, Sebastian, Shaojie, Shashank, Shubha, Spyretta, Svitlana, Thilina, Thong, Vaishali, Vera, Wanyu, Weijia, Xiangsheng, Xiaohui, Xinyi, Yang, Yuanxing, Yifei, Yibin, Yifan, Yuanna, Zhaochun, Zihan, and Ziming. I also thank Wentao and Ruiqi for their help with the experiments. Moreover, I thank Petra and Pablo for the wonderful activity organizations, especially Petra for helping me with university procedures. Besides, I thank other friends I have met in Amsterdam: Chunfang, Qi, Shihan, Shuang, Wei, Xiaomeng, and Yahui; and also my friend not in Amsterdam: Qian. I also thank Ruben for translating the thesis summary from English to Dutch. Thank you all!

The road for me to start my PhD was influenced by my mother, Hezhi, who got a PhD degree from one of the best universities in China. I also thank my father Deqian. Thank you both!

Last, I thank my boyfriend Xin. Your support, which has always been there for me, has been very important during my PhD studies. Thank you!

Yangjun Zhang
September 30th, 2022

Contents

1 Introduction

Conversational agents interact with end-users through natural language responses. People and conversational agents take turns during the dialogue interaction. Conversational agents connect people to information, products, or services that may be of interest to them [71]. Conversational agents increasingly attract attention [2] are widely applied in diverse domains, such as finance [165], healthcare [109], education [92], business [169], and beyond.

Conversational interfaces have the potential to improve the lives of people around the world. However, they may generate unsafe content, that is, content that may offend end-users and that may lead to serious real-world consequences. The Tay bot built in 2016 by Microsoft is a prominent example of a conversational agent that ended up producing malevolent utterances while being deployed.[1] The GPT-4chan bot is another well-known example of a conversational agent that may display a mix of offensiveness, nihilism, trolling, and distrust.[2] How can we identify malevolent utterances in dialogue responses? That is, how can we detect and evaluate malevolence in dialogue response?

While publications predicting the arrival of conversational interfaces go back at least three decades [see, e.g., 10], the widespread adoption of conversational interfaces such as task-oriented dialogue systems (TDSs) and conversational dialogue systems (CDSs) is a recent development [122]. TDSs are meant to understand end-users' goals, such as booking a restaurant or canceling an order and help them to achieve those goals. CDSs, which are also referred to as chatbots, are usually not developed to achieve a specific goal. The development of CDS and TDS has given rise to research and deployment of corpus-based – as opposed to template-based [153] – conversational agents [48] that promise to generate more natural responses than template-based responses. Template-based conversational agents utilize rules to mimic real human conversations. Corpus-based conversational agents learn from dialogue corpus to generate dialogue response. In terms of corpus-based dialogue systems, the development of CDS techniques starts from ELIZA to later dialogue systems such as Alexa, Siri, and XiaoIce.

Significant research efforts are being invested to improve the quality of generated responses. Different dimensions, e.g., fluency [74], coherence [62], informativeness [39, 127], interestingness [72], diversity [71, 85], engagement [183], and appropriateness [101], have been considered. There are various models for improving

[1] https://en.wikipedia.org/wiki/Tay_(bot)
[2] https://huggingface.co/ykilcher/gpt-4chan

these aspects of dialogue quality. For instance, some dialogue systems investigate leveraging background information beyond the dialogue consent to improve the dialogue informativeness [108]. However, corpus-based approaches to response generation for CDSs are unsafe in terms of the content and dialogue acts they produce [168]; not all possible responses and dialogue acts that a corpus-based conversational interface may generate are suitable for end-users.

There are some previous works that are related to detecting and evaluating malevolent dialogue responses, but these works cannot be applied to malevolent dialogue detection directly. For instance, previous works have built different taxonomies related to dialogue malevolence, e.g., "hate", "dark triad", "aggressive", "offensive", "safety" [5, 78, 118, 147, 176], as well as for labels at the opposite side of the spectrum to malevolence, e.g., "courteous" [54]. Based on such taxonomies, different datasets have been annotated. For instance, the OLID dataset for "offensive" [176], the TRAC dataset for "aggressive" [78] and the MDHS dataset for "hate" [9]. The sources of the dataset include posts on Twitter, comments on Wikipedia, and posts on Facebook. With these datasets, detection and classification methods have been built. Features such as bag-of-words, n-grams and entities, and models such as support vector machine (SVM) have been used for detection and classification [9, 33, 135, 138, 166]. In addition, there are methods that use deep learning-based methods such as convolutional neural networks (CNNs) and long short-term memorys (LSTMs) [78, 143, 163, 176]. However, these studies have important limitations concerning dialogue malevolence taxonomies, datasets, and methods. First, the number of categories in previously proposed taxonomies is limited. E.g., the definition of hate speech is limited to language that expresses hatred towards a group or individuals, humiliates or insults others [5], and it fails to include inappropriate aspects such as violent behavior. Second, very few of the datasets are built for multi-turn dialogues even though dialogue context is important for detecting malevolent dialogue responses [147]. Third, previous methods fail to consider classification confidence, which is important for out-of-distribution samples [25], and label correlations, which are important for multi-label classification [79, 151].

Besides the taxonomies, datasets, and methods for malevolence detection, malevolence *evaluation* is also important. There are two main directions for dialogue evaluation, i.e., automatic evaluation and human evaluation [45]. Automatic methods can be grouped into word overlap metrics [19, 91, 115] and learning-based metrics [101, 149]. As to human evaluation, the evaluation process is often based on crowd-sourcing. Neither of the two families of methods balance reliability and effort: (i) automatic methods have limited agreement with human assessments [93], and (ii) human evaluation is labor-intensive and lacks speed and scalability [35]. In contrast to only using automatic methods or only using human evaluation, *human-machine collaborative methods* can decrease human effort with high reliability and agreement with human judgments [126].

This thesis focuses on malevolent dialogue detection and evaluation. A *malevolent response* is a kind of dialogue response that might contain offensive or objectionable content including hate, insult, threat, etc [185]. Importantly, the malevolence of some dialogue responses can only be detected when the dialogue context, which is the dialogue history information, is considered. For instance, the response "I agree" is safe in general, however, it is malevolent as a response to a malevolent user utterance. The issue of malevolent dialogue responses has negative social risks and consequences [116, 117].

In order to address the issue, dialogue response detection and evaluation methods need to be developed.

In this thesis, we first analyze the malevolence problem of state-of-the-art (SOTA) generation models with malevolence detection models. We also propose a knowledge pre-selection mechanism to improve informativeness before analyzing sequence to sequence-based generation models. Second, we introduce taxonomies, datasets, and methods for single-label dialogue malevolence detection. Third, we build datasets and methods for multi-label dialogue malevolence detection from a single-label training set. The taxonomy of multi-label dialogue malevolence detection is the same as single-label detection. Finally, we propose a human-machine collaborative evaluation framework for dialogue malevolence evaluation.

1.1 Research Outline and Questions

We organize the thesis around three research themes:

- An empirical investigation of the existence of malevolence in generated dialogue responses (Chapter 2);
- The development of methods to detect malevolence in generated dialogue responses (Chapter 3 and 4); and
- A method for evaluating malevolence detection performance (Chapter 5).

1.1.1 Establishing the existence of the malevolence problem in generated dialogue responses

Pretrained generation models and *sequence to sequence* (S2S)-based generation models are often used in dialogue generation [86]. We hypothesize that current state-of-the-art dialogue generation models, whether pretrained or S2S-based, may generate malevolent responses. It is important to investigate whether the generation models have a malevolence problem as they may negatively impact people and may cause negative feedback to end-users. For the pretrained generation models, the dialogue generation quality is generally high, but they require a large dataset and are sensitive to contextual input [59, 174]. For models without pretraining, the dataset size could be smaller, however, they may generate bland and generic responses [64]. Before analyzing the malevolence of S2S-based models, we first improve the quality, i.e., informativeness and fluency, of the generation model, so that we examine the existence of malevolent responses only for models of sufficient dialogue quality. In particular, previous work on background-based conversations (BBCs), which is a task with extra background content for dialogue generation, has shown that leveraging background knowledge for dialogue response generation can make dialogue systems generate more informative and natural responses [108]. To improve the quality of S2S models, we proceed as follows. Extraction-based methods extract spans from background material as responses [108, 123, 155], but responses based on fixed extraction of the background are

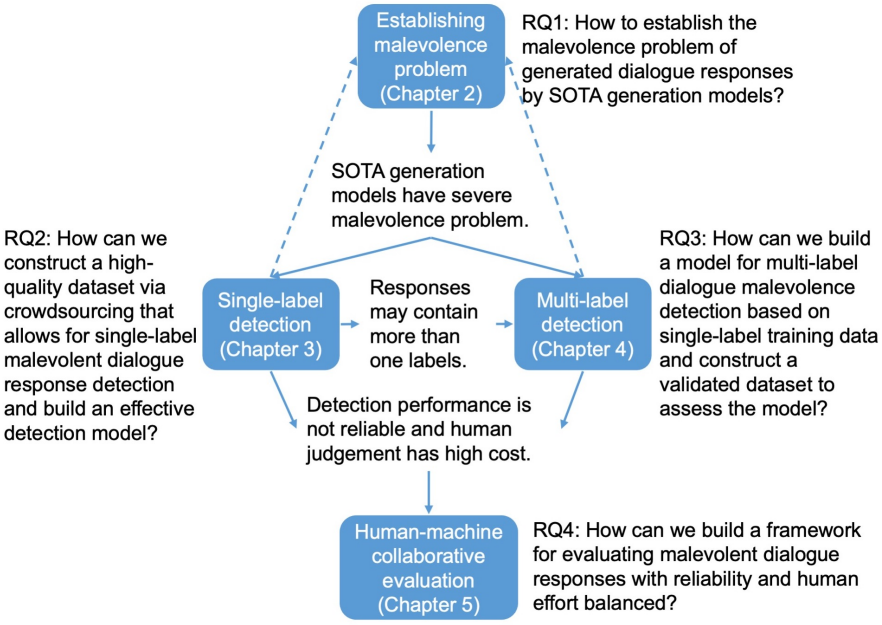


Figure 1.1: Research framework for malevolent dialogue response detection and evaluation used in the thesis.

not natural or fluent enough; generation-based methods generate responses that are natural but not necessarily effective in leveraging background knowledge [7, 89, 139, 148]. Context history has not been fully explored in previous work in selecting appropriate background to improve dialogue response quality. Then, in terms of malevolent dialogue response analysis, with the background information selected and dialogue history utilized, the generated responses from background and dialogue context may contain malevolent content learned from the corpus, which requires a proper assessment.

We seek to answer the following question:

RQ1 How to establish the malevolence problem of generated dialogue responses by SOTA generation models?

To answer **RQ1**, we analyze the malevolence of generation-based methods to determine whether the malevolent dialogue response challenge exists. We first analyze the malevolence of pretrained dialogue generation models. Second, we introduce a knowledge pre-selection based model, i.e., context-aware knowledge pre-selection (CaKe), to improve dialogue informativeness of S2S-based generation model without pretraining and then analyze the malevolence of CaKe and baselines. CaKe uses dynamic bi-directional attention to improve knowledge selection by using the utterance history context as prior information to select the most relevant background information. We compare our model with current state-of-the-art baselines to find whether CaKe benefits from the pre-selection process and is superior to baselines in improving informativeness and fluency. We then check for the existence of malevolent responses.

1.1.2 Building a taxonomy, dataset, and benchmark for single-label dialogue malevolence detection

Our work towards answering RQ1 shows that corpus-based conversational agents may generate non-predictable responses that contain offensive or objectionable content. It is important to automatically detect malevolent dialogue responses that are inappropriate in terms of content and dialogue acts since malevolent dialogue responses may negatively impact people, increase social friction, and lead to dialogue breakdown. The research community has created several taxonomies and numerous resources to help characterize, model, and classify textual content that is somehow inappropriate, but these contributions cannot be applied to malevolent dialogue response detection directly.

First, previously published taxonomies are not suitable for malevolent dialogue response detection since previous studies on the topic of detecting and classifying inappropriate content are mostly focused on a specific category of malevolence and lack hierarchy [5, 147]; some previous lexicon items have a small lexicon size [33].

Second, previously released datasets are not for multi-turn dialogues [9, 33, 78, 166, 176], and some datasets have substantial annotation errors [152].

Third, there are various limitations in previous malevolence detection methods. On the one hand, there are out-of-distribution or low-confidence samples that have not been considered by previous methods. The reasons these challenges exist are that the sample size of some label groups might be small and that the inter-annotator agreement of the crowd workers who provided the annotations is imperfect. The confidence score of the predicted category reflects the probability of whether the prediction is reliable, and confidence calibration could be used to improve classification performance [25]. On the other hand, previous work does not consider dialogue context and rephrased utterances during classification. There are some recent works that consider context, but they have been proposed no earlier than our work included in this thesis [29, 147].

We seek to answer the following question:

RQ2 How can we construct a high quality dataset via crowdsourcing that allows for single-label malevolent dialogue response detection and build an effective detection model?

To answer **RQ2**, we propose the *malevolent dialogue response detection and classifying* (MDRDC) task that is aimed at identifying and classifying malevolent dialogue responses. We take three steps to advance research on the MDRDC task. First, we define the task and present a *hierarchical malevolent dialogue taxonomy* (HMDT). The taxonomy includes three levels of labels: 1st-level, 2nd-level, and 3rd-level labels. Second, we create a labeled multi-turn dialogue dataset and formulate the MDRDC task as a hierarchical classification task. Each utterance in the dataset has a single label. During annotation, we also ask the annotators to rephrase the malevolent utterances. Third, we propose BERT-based classifier with confidence calibration (BERT-conf), a method for estimating the confidence of each predicted category and calibrating the classification. We apply BERT-conf and previous SOTA text classification methods to the MDRDC task and report on experiments aimed at assessing the performance of these approaches. As we collected dialogue context and rephrased data, we also use this type of data to determine whether they can improve the classification performance.

We analyze the performance differences between the 1st-level, 2nd-level, and 3rd-level results, as we use hierarchical taxonomies for detection and classification. We also analyze what space is left for further improvement of the results.

1.1.3 Multi-label dialogue malevolence detection

A dialogue utterance may contain more than one category of malevolence. There are some limitations to the current dataset and methods for malevolence detection when we look at the detection task as a multi-label task. First, there are limitations to datasets. The current datasets for dialogue malevolence detection and the related datasets are mostly single-label data. Some datasets are multi-label, however, they are not built for multi-turn dialogues, e.g., toxic content classification. Second, current work on malevolent dialogue malevolence detection does not consider label correlations [185]. Previous work on text classification has proved that label correlation, e.g., co-occurrence [79] and international classification of diseases (ICD) label correlation [151], can help to improve classification peformance. However, it is costly to label training sets for multi-label malevolent dialogue response classification. Cole et al. [24] have shown that, surprisingly, training with fewer confirmed labels could approach the performance of a fully labeled classifier, which means that the gap between the single labeled classifier and the fully labeled classifier is trivial. Therefore, we plan to utilize label correlations in our proposed taxonomy as well as dialogue context information to improve multi-label dialogue malevolence detection. We propose the task of *multi-label classification from a single-label training set* for malevolence detection.

More precisely, we answer the following research question:

RQ3 How can we build a model for multi-label dialogue malevolence detection based on single-label training data and construct a validated dataset to assess the model?

To answer **RQ3**, we propose the task of multi-label dialogue malevolence detection from a single-label training set and we crowdsource a multi-label dataset, i.e., *multi-label dialogue malevolence detection* (MDMD), for evaluation. The MDMD dataset is annotated via MTurk crowdsourcing. We only label the validation and test dataset.

We also propose a multi-label malevolence detection model, i.e., MCRF, with two label correlation mechanisms, *label correlation in taxonomy* (LCT) and *label correlation in context* (LCC). MCRF is built based on CRF, which is a graphical model that can leverage the dependencies between the word output representations and it is suitable for sequence labeling tasks with an underlying graph structure [96]. Compared with the BERT-based conditional random field (CRF) classifier, MCRF adds label correlations into the model for improving classification performance. We conduct experiments on the MDMD dataset to evaluate the effectiveness of our MCRF model and determine whether it outperforms the baseline models.

1.1.4 Evaluation of dialogue response malevolence through a human-machine collaborative framework

Evaluation of *conversational dialogue systems* (CDSs) has drawn significant attention since it is important for CDS development. There are two groups of methods for dia-

logue evaluation: automatic evaluation and human judgements. Automatic evaluation of dialogues often shows insufficient correlation with human judgements [93]. Human evaluation is reliable but labor-intensive [93]. The need for reliable and efficient balanced evaluation methods arises. There are scenarios, i.e., daily research and development of CDS and CDS leaderboards, that require better dialogue response evaluation methods. Human-machine collaborative (HMC) methods can decrease human effort with high reliability and agreement with human [126]. In Chapter 5, we focus on building an evaluation framework that considers both reliability and human effort based on HMC methods.

We seek to answer the following question:

RQ4 How can we build a framework for evaluating malevolent dialogue responses with reliability and human effort balanced?

To answer **RQ4**, we introduce a human-machine collaborative framework, human-machine collaborative evaluation (HMCEval), that can guarantee the reliability of the evaluation outcomes with reduced human effort. HMCEval casts dialogue evaluation as a sample assignment problem, where we need to decide to assign a sample to a human or a machine for evaluation. HMCEval includes a model confidence estimation module to estimate the confidence of the predicted sample assignment, a human effort estimation module to estimate the human effort should the sample be assigned to human evaluation, as well as a sample assignment execution module that finds the optimum assignment solution based on the estimated confidence and effort. We assess the performance of HMCEval on the task of evaluating malevolent dialogue responses.

1.2 Main Contributions

The main contributions of this thesis are listed below. They are grouped into three categories.

1.2.1 Resource contributions

(1) We propose a taxonomy for dialogue malevolence detection. 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 categories in 1st-level, 2nd-level, and 3rd-level labels (Chapter 3).

(2) We build a malevolent dialogue response detection and classifying (MDRDC) dataset for single-label dialogue malevolence detection (Chapter 3). The dataset is built based on tweets and annotated with single-label malevolence categories.

(3) We build a multi-label dialogue malevolence detection (MDMD) dataset for evaluation of the multi-label dialogue malevolence detection models (Chapter 4). The dataset is built based on the MDRDC dataset. We label the validation and test dataset with multi-label malevolence categories.

1.2.2 Algorithmic contributions

(1) We propose a knowledge pre-selection based model for improving dialogue response informativeness (Chapter 2).

(2) We devise a confidence calibrated model based on BERT-base, i.e., BERT-based classifier with confidence calibration (BERT-conf), for single-label dialogue malevolence detection (Chapter 3).

(3) We propose a multi-faceted label correlation enhanced model for multi-label dialogue malevolence detection from a single-label training set (Chapter 4).

(4) We introduce a human-machine collaborative (HMC) framework based on linear programming for malevolent dialogue response evaluation and it can guarantee the reliability of the evaluation outcomes with half of the human effort (Chapter 5).

1.2.3 Empirical contributions

(1) An empirical comparison of context-aware knowledge pre-selection (CaKe) with other state-of-the-art methods for the background-based conversation (BBC) task in terms of informativeness and malevolence and an empirical comparison of S2S-based generation model with pretrained generation models in terms of malevolence (Chapter 2).

(2) An empirical comparison of BERT-conf with other state-of-the-art methods for single-label malevolent dialogue response classification (Chapter 3).

(3) An empirical comparison of multi-faceted label correlation enhanced CRF (MCRF) with other state-of-the-art methods for multi-label malevolent dialogue response classification (Chapter 4).

(4) An empirical comparison of human-machine collaborative evaluation (HMCEval) with automatic evaluation and human judgement for malevolent dialogue response evaluation (Chapter 5).

1.3 Thesis Overview

In this thesis, we investigate malevolent dialogue response detection and evaluation.

First, in Chapter 2, we investigate malevolence in utterances produced by state-of-the-art generation models. We first improve the quality of S2S-based generation models by introducing a knowledge pre-selection based dialogue generation model, i.e., CaKe. Then, we analyze the malevolent dialogue responses generated by pretrained generation models and S2S-based generation models including CaKe based on the classification model in Chapter 3 and Chapter 4.

Second, in Chapter 3, we build the hierarchical malevolent dialogue taxonomy (HMDT), the malevolent dialogue response detection and classifying (MDRDC) dataset, and a confidence-based method that utilizes confidence calibration to improve the BERT-based classification model for single-label dialogue malevolence detection.

Third, in Chapter 4, we build the multi-label dialogue malevolence detection (MDMD) dataset for multi-label dialogue malevolence detection and propose the multi-faceted label correlation mechanism to improve the multi-label dialogue malevolence detection.

Fourth, in Chapter 5, we propose a human-machine collaborative framework for dialogue malevolence evaluation, which balances overall reliability and human effort.

Finally, in Chapter 6 we conclude the thesis, formulate broader implications of our work, and discuss limitations and future directions.

The chapters are best read in their natural order, from Chapter 2 through to Chapter 6. Even though Chapter 2 uses some materials that are only introduced in Chapter 3 and Chapter 4, it is the natural starting point for the remainder of the thesis as it examines the data, i.e., utterances produced by generation models, for malevolence.

1.4 Origins

In this section, we list publications that form the basis of the thesis. Each research chapter is based on a paper. We list references to these publications and explain the roles of the co-authors.

Chapter 2 is based on the conference paper:

- Y. Zhang, P. Ren, and M. de Rijke. Improving background based conversation with context-aware knowledge pre-selection. *Search-Oriented Conversational AI Workshop (SCAI)*, 2019.

The experiments and result analyses were performed by Yangjun. Yangjun and Pengjie designed the model. Yangjun did most of the writing. Maarten and Pengjie helped with the writing.

Chapter 3 is based on the journal paper:

- Y. Zhang, P. Ren, and M. de Rijke. A taxonomy, data set, and benchmark for detecting and classifying malevolent dialogue responses. *Journal of the Association for Information Science and Technology*, 72(12):1477–1497, 2021.

Dataset annotation, experiments, and result analyses were performed by Yangjun. Yangjun and Pengjie defined the taxonomy and designed the model. Yangjun did most of the writing. Maarten and Pengjie helped with the writing.

Chapter 4 is based on the conference paper:

- Y. Zhang, P. Ren, W. Deng, Z. Chen, and M. de Rijke. Improving multi-label malevolence detection in dialogues through multi-faceted label correlation enhancement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3543–3555, 2022.

Dataset annotation, experiments, and result analyses were performed by Wentao and Yangjun. Yangjun and Pengjie designed the model. Yangjun did most of the writing. Maarten and Pengjie helped with the writing.

Chapter 5 is based on the conference paper:

- Y. Zhang, P. Ren, and M. de Rijke. 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, 2021.

The experiments and result analyses were performed by Yangjun. Yangjun and Pengjie designed the model. Yangjun did most of the writing. Maarten and Pengjie helped with the writing.

Work on the thesis also benefited from research that led to the following paper:

- P. Ren, R. Li, Y. Zhang, and M. de Rijke. Dialogue malevolence attacks against pre-trained models. In *Under Review*, 2022.

2

Establishing the Malevolent Dialogue Response Problem of Generation Models

In this chapter, we address RQ1: How to establish the malevolence problem of generated dialogue responses by SOTA generation models?

2.1 Introduction

Dialogue systems have attracted great attention recently. Different methods have achieved promising results [7, 148]. Encoder-decoder-based dialogue generation models, e.g., sequence to sequence (S2S) [148] and pretrained dialogue generation models, e.g., DialoGPT [183] and Blenderbot [134], are two kinds of state-of-the-art (SOTA) models for dialogue generation. However, many challenges remain. These include a lack of diversity [85], limited informativeness [180], inconsistency [41], and being malevolent. Among these, being malevolent is a particularly important one. The responses and dialogue acts of a chatbot may be malevolent, as in "I don't want to talk to you" or "are those ****** human".[1] More generally, a *malevolent dialogue response* is a generated response that is grounded in negative emotions, inappropriate behavior, or an unethical value basis in terms of content and dialogue acts [185]. Malevolent responses may cause friction and breakdown of the dialogue systems.

It is important to analyze the malevolence of current SOTA generation models, especially pretrained and sequence to sequence (S2S)-based models. First, pretrained dialogue generation models are able to generate high-quality dialogue responses [134, 183], so we can directly analyze their outputs for malevolence. In contrast, S2S-based dialogue generation models may generate responses that are bland and deflective (e.g., "can't tell you", "I'm not sure", "I don't have a clue"), without being informative [64]. Thus, before examining S2S-based generation models for malevolence, we first need to increase the informativeness of their responses [108]. For this purpose, we consider the concept of a *background-based conversation* (BBC). BBCs have demonstrated a potential for improving informativeness. Given background material and a conversation, generation models for BBCs generate responses by referring to background information

This chapter was published as [182].

[1] Malevolent words are masked.

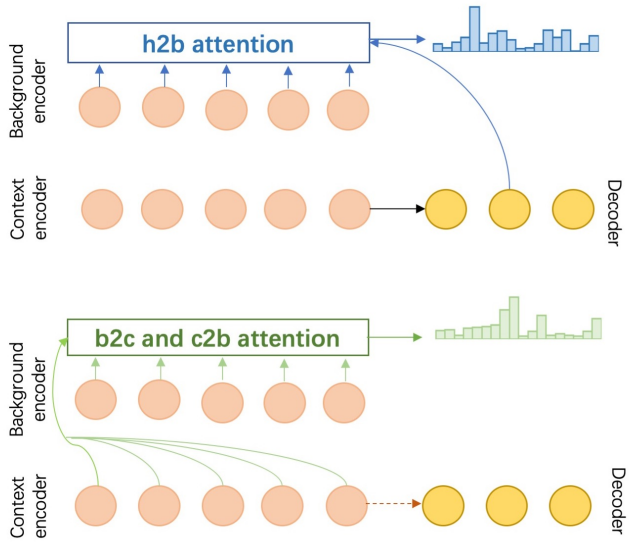


Figure 2.1: Knowledge pre-selection for context-aware knowledge pre-selection (CaKe) model.

and considering the dialogue history context at the same time. The background material may be semi-structured information or free text; in this chapter, background material is provided in the form of free text.

There are two main approaches to BBCs, extraction-based methods and generation-based methods. Extraction-based methods extract text segments from the background as responses. Hence, the responses are informative. Generation-based methods generate sequences based on an encoder-decoder mechanism and the responses are natural and fluent. For extraction-based methods, models including the bi-directional attention flow (BiDAF) model [140] have been proposed for selecting the best matching positions of the tokens from the target context. However, responses produced by extraction-based methods are directly copied from background sentences. As a result, they are neither fluent nor natural. Generation-based methods include the hierarchical recurrent encoder-decoder architecture (HRED) [141] and get to the point (GTTP) [108, 139]; they are not always effective at leveraging background knowledge and may return responses with inappropriate background knowledge.

The motivation behind our proposed model is to improve dialogue informativeness for S2S-based generation models. The GTTP model [108, 139] selects background knowledge by using a hidden state at each decoding time step as a query to select background knowledge. For each query, it is supposed to comprise features of the corresponding response token, features from the previous response tokens, as well as features passed on from the utterance history context. However, the query may not contain all information from the utterance history since the information attenuated rapidly [20] during state transfer especially when the length of the input utterance history context of the encoder increases, as the red line demonstrates in Figure 2.1. The

utterance history contains rich information about the conversation and it is also highly relevant to the response. Therefore, we propose a neural model, namely *context-aware knowledge pre-selection* (CaKe), that introduces a knowledge pre-selection step to offset the problem of the current state-of-art model. CaKe utilizes utterance history context as a query directly to facilitate a better selection of knowledge. There are two steps to implement knowledge pre-selection in CaKe. To start, we choose the encoder state of the utterance history context as a query to select the most relevant knowledge. Then, we employ a modified version of BiDAF that combines background-to-context attention and context-to-background attention to point out the most relevant token positions of the background sequence to cope with long background text.

After building the CaKe model, we analyze the malevolence of generated responses by the baseline models and CaKe. We also analyze the malevolence of pretrained generation models. In order to analyze the malevolence automatically, we utilize the dialogue malevolence classification model described in Chapter 3 and 4. In Chapter 3, we build a taxonomy, dataset, and benchmark for detecting and classifying single-label malevolent responses. The taxonomy contains eighteen categories and the dataset contains 6,000 dialogues. In Chapter 4, we build a dataset and benchmark for detecting and classifying multi-label malevolent responses. First, we utilize the SOTA single-label classification model to classify binary malevolence. Then, we utilize the SOTA multi-label classification model to find the most frequent malevolent categories. We will introduce the classifiers in Chapter 3 and 4.

The main contribution of our work in this chapter is two-fold. Most importantly, we establish the malevolence problem of SOTA dialogue generation models. We infer that the malevolent response challenge exists for both pretrained and S2S-based generation models. Second, we propose a knowledge pre-selection module to improve the dialogue informativeness for S2S-based generation model. This enables us to assess the problem of generating malevolent responses for both SOTA pretrained generation models and S2S-based generation models.

2.2 Related Work

2.2.1 Malevolence of dialogue responses

Dialogue malevolence is an important issue for open-domain conversational systems as it may cause social friction and break the dialogue [185]. Related to dialogue malevolence, there are different aspects, including general aspects, e.g., safety [168], and specific aspects, e.g., hate speech [5, 166], aggressiveness [78], offensiveness [176]. Corpus-based conversational models tend to generate unsafe content, e.g., Zhang et al. [183] found that DialoGPT generates occasional toxic contents and suggested the importance of detection and control of toxicity in dialogue response. Therefore, the need arises to analyze the different safety aspects of responses generated by current SOTA generation models. Dinan et al. [40] propose a classification framework of safety issues in open-domain chatbots including three categories. Sun et al. [147] build a model for dialogue safety with seven categories.

However, due to the lack of a unified framework for malevolence analyses, the

aspects analyzed in previous work are different from ours. We analyze the malevolence of generated dialogue responses by pretrained generation models and S2S-based models, with seventeen malevolence categories. Besides, prior to analyzing the S2S-based model, we first propose a mechanism to solve the blandness problem as blandness may influence malevolence analysis.

2.2.2 Leveraging background to improve informativeness

There are two main methods to increase informativeness, thus solving the blandness problem: extraction-based methods and generation-based methods.

Extraction-based methods to BBC are originally derived from reading comprehension (RC) tasks [123], where the answer could be picked from a set of token positions of the input sequence. Vinyals et al. [155] propose a pointer network (Ptr-Net) model that uses attention as a pointer to select a token in an input sequence so as to generate an output sequence where some tokens come from the input sequence. Wang and Jiang [160] extend this work by combining match-long short-term memory (LSTM) and Ptr-Net. Seo et al. [140] introduce BiDAF to improve the extraction of the context span by pointing to the start point and the end point of the relevant span. Lee et al. [83] and Yu et al. [173] predict answers by ranking text spans, where the token positions of the span are continuous, within background passages. Wang et al. [161] predict the boundary of the answer span by a self-matching mechanism. Extraction-based methods are better at locating the right background span than generation-based methods [108]. Nevertheless, extraction-based methods are not suitable for BBCs as BBCs do not have standard answers like those in RC tasks, and responses based on fixed extraction of the background are not natural or fluent enough for conversation tasks.

Generation-based methods achieve good results on different conversation tasks. S2S learning methods [148], later extended to sequence to sequence with attention (S2SA) [7], are the basis of most generation-based methods. First, response diversity is to be improved. Lots of methods have been proposed to solve this issue. To increase diversity, Li et al. [85] present maximum mutual information (MMI) as the objective function in neural models to decrease generic response sequences and increase varied and interesting outputs. Serban et al. [141] propose HRED, which uses a two-level hierarchy, including word level and dialogue turn level, to exploit long-term text. Later, Serban et al. [142] add a high-dimensional stochastic latent variable to extend HRED, aiming to generate responses with more diversified content and reduce blandness. Zhang et al. [180] explicitly optimize a variational lower bound on pairwise mutual information between query and response to boost diversity during training. Second, response informativeness is an important issue. Many methods of adding background knowledge and common sense to conversations have been proposed for mitigating blandness. Most of the existing conversational datasets are not labeled with relevant knowledge, so it is difficult to apply large datasets to the model training. As a result, most models need to do knowledge selection before training the model, such as knowledge diffusion [98] and graph attention [188]. These methods use knowledge datasets that are separate from the conversations. Recently, several datasets have become available where conversations are generated based on background knowledge. Moghe et al. [108] build a dataset for BBCs and conduct experiments with several methods. The datasets

of Persona-chat [178] and Wizard-of-Wikipedia [39] are similar to Moghe et al. [108]'s dataset.

Although the background knowledge available in BBCs has a low degree of redundancy, selecting the most relevant background is important to improve informativeness. Moghe et al. [108] employ GTTP, proposed by [139] to copy tokens from background knowledge at each generation timestamp. Lian et al. [89] use a posterior knowledge distribution to guide knowledge selection. The selection of background material plays a vital role in generating informative responses. However, the crucial role of context history in selecting appropriate background has not been fully explored by current methods. Unlike previously proposed methods, in order to encourage informative and non-deflective responses, our proposed model leverages the context as prior context to do pre-selection of the background knowledge.

2.3 Methodology

We postpone a thorough introduction of our dialogue malevolence classification models for malevolence analysis in Section 3.5 and Section 4.3. In this section, we introduce the methodology of building the CaKe model[2].

Given a background in the form of free text and the current utterance history context, BBC aims to generate an utterance as the next response. Formally, let $b = (b_1, b_2, \ldots, b_i \ldots, b_I)$ represent the words in the background knowledge, a current utterance history context in the form of $c = (c_1, c_2, \ldots, c_j, \ldots, c_J)$, and the task of BBC is to generate response $r = (x_1, x_2, \ldots, x_t, \ldots, x_T)$ based on b and c.

In this section, we introduce our model CaKe for BBC. An overview of CaKe is shown in Figure 2.2. First, we use two encoders to encode background and context. Second, for knowledge pre-selection, we select the background related to the context. To achieve this, we choose the encoder state of the context as query to select the most relevant knowledge from the background knowledge. We use a modified version of BiDAF [140], which combines background-to-context attention and context-to-background attention to point out the most relevant token position of the background sequence. The pre-selection module forms a context-aware background distribution. Third, for the generation part, the generator generates a vocabulary distribution with global attention [7]. Lastly, we combine this pre-selector part with the generator to generate the final output of each decoding time step. We get the response based on the probability of generating a token from the vocabulary or copying a token from the background.

2.3.1 Background and context encoder

The word embeddings are used to map words to high-dimensional vector space. We apply random initialized word embeddings.

The background and context encoders encode background and context embeddings into $h^b = (h_1^b, h_2^b, \ldots, h_i^b, \ldots, h_I^b)$ and $h^c = (h_1^c, h_2^c, \ldots, h_j^c, \ldots, h_J^c)$ respectively. We use bidirectional recurrent neural networks (RNNs) and concatenate the outputs of two

[2] https://github.com/repozhang/bbc-pre-selection

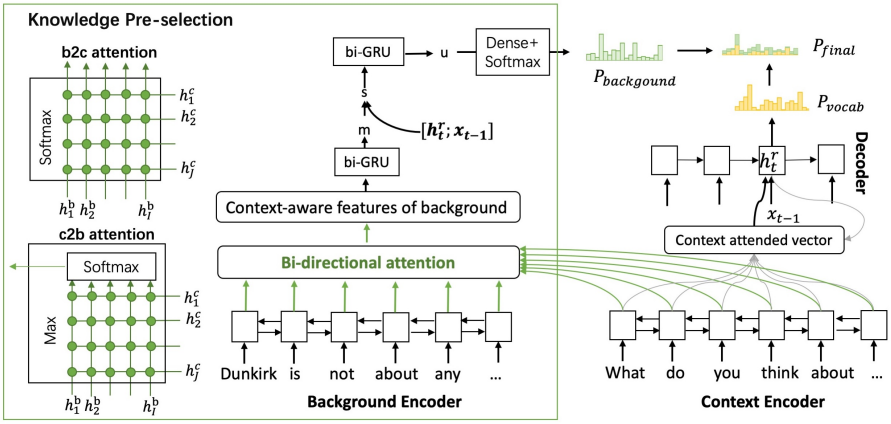


Figure 2.2: Overview of our model, context-aware knowledge pre-selection (CaKe).

RNNs for the two encoders. Therefore, we get $h^b \in \mathbb{R}^{2d \times I}$ for the background and $h^c \in \mathbb{R}^{2d \times J}$ for the context.

2.3.2 Knowledge pre-selection

The selector is used for pre-selecting background words and to form context-aware background distribution. We combine background-to-context and context-to-background attention. The original BiDAF [140] trains start and end span position. Moghe et al. [108] use this model to find the relevant span from the background knowledge by two positions, while we only use the start span position as the most relevant token index. The similarity score between context and background is calculated by:

$$score_{ij} = S(h^b_{:i}, h^c_{:j}), \tag{2.1}$$

where $h^b_{:i}$ is the i-th column vector of h^b and $h^c_{:j}$ is the j-th column vector of h^c. Meanwhile, S is defined as:

$$S(h^b, h^c) = w^T[h^b; h^c; h^b \odot h^c], \tag{2.2}$$

where w^T is a trainable weight vector. Then, we feed $score_{ij}$ to a softmax function to get the attention and the corresponding vector.

First, background-to-context (b2c) attention reflects which context words are most relevant to each background word. α_i represents the attention weights on the context words by the i-th background word, where $\sum \alpha_{ij} = 1$ for all i. The attention vector of the context is computed by

$$\widetilde{h}^c_{:i} = \sum_j \alpha_{ij} h^c_{:j}, \tag{2.3}$$

where α_i is computed by normalizing $S_{i:}$ by a softmax function. $\widetilde{h}^c \in \mathbb{R}^{2d \times I}$.

Second, context-to-background (c2b) attention computes which background words are most relevant to each context word. The attention weights on the background words are calculated by $\beta = \text{softmax}(\max_{col}(S))$. Then the attended background vector is computed by:

$$\widetilde{h}^b_{:i} = \sum_i \beta_i h^b_{:i}, \tag{2.4}$$

where $\widetilde{h}^b \in \mathbb{R}^{2d \times I}$.

Finally, contextual embeddings and the attention vectors are combined to yield g, where g is defined by:

$$g_{:i} = \eta(h^b_{:i}, \widetilde{h}^c_{:i}, \widetilde{h}^b_{:i}), \tag{2.5}$$

where η is a trainable vector. We use simple concatenation in our experiments: $\eta(h^b, h^c, \widetilde{h}^b) = [h^b, \widetilde{h}^c, h \odot \widetilde{h}^c, h^b \odot \widetilde{h}^b]$. g is the static context-aware background representations. We use a bi-RNN layer for g, and get m, which captures the interaction among the background words conditioned on the context. Then we concatenate m with h^r_t and x_{t-1} to generate s. s is fed into a bi-RNN to generate u. Finally, a background distribution is calculated by:

$$P_{background} = \text{softmax}(w^T_{p1}[g; m; s; u]), \tag{2.6}$$

where W^T_{p1} is a trainable factor.

2.3.3 Generator

For the generator, we generate the response token, which is the vocabulary distribution, with attention. For the generator module, the hidden state of response is: $h^r = (h^r_1, h^r_2, \ldots, h^r_t, \ldots, h^r_T)$, where h^r_t is the state of the decoder at the current time step. The final decoder hidden state aware representation of the context is the attention weighted sum of context c_t. The representation of context is $h^c = (h^c_1, h^c_2, \ldots, h^c_j, \ldots, h^c_J)$, where J is the total length of the context. The final weighted sum of the context is calculated as follows:

$$\begin{aligned} e^t_j &= v^T \tanh(W_c h^c_j + V h^r_t + b_c), \\ \gamma^t &= \text{softmax}(e^t), \\ c_t &= \sum_j \gamma^t_j h^c_j, \end{aligned} \tag{2.7}$$

where h^r_t is the current state of the decoder.

The generator then uses c_t, s_t and x_t to generate P_{vocab}, and the generation probability p_{gen} is calculated as follows:

$$p_{gen} = \sigma(w^T_c c_t + w^T_s s_t + w^T_x x_t + b_{gen}), \tag{2.8}$$

where $w^T_c, w^T_s, w^T_x, b_{gen}$ are trainable parameters and σ is the sigmoid function. p_{gen} is used as a switch to choose between generating a word by sampling from the vocabulary distribution P_{vocab} or copying a word from the background by sampling from background distribution $P_{background}$.

2.3.4 Mixture and loss

The mixture is used to mix the results of knowledge pre-selection and generation. We obtain the following final distribution over the extended vocabulary:

$$P_{final}(w) = p_{gen}P_{vocab}(w) + (1 - p_{gen})P_{background}. \tag{2.9}$$

Let $P(w) = P_{final}(w)$. During training, the training loss of time step t is the negative log likelihood of the target word w_t^* for that time step:

$$loss_t = -\log P(w_t^*). \tag{2.10}$$

The training loss of the whole sequence $loss$ and the whole dataset $L(\theta)$ are:

$$\begin{aligned} loss &= \frac{1}{T}\sum_{t=0}^{T} loss_t, \\ L(\theta) &= \sum_{n=0}^{N} loss_n, \end{aligned} \tag{2.11}$$

where θ is the set of all trainable weights, N is the number of total samples in the dataset.

2.4 Experimental Setup

Now that we have defined CaKe, our next step is to evaluate its performance, in terms of informativeness and, especially, in terms of malevolence. In this section, we detail our experimental setup.

2.4.1 Datasets

In terms of the pretrained generation models, the DialoGPT model by Microsoft and Blenderbot model by Facebook are pretrained on 1.5 billion conversation-like exchanges extracted from Reddit comments [134, 183].

In terms of the S2S-based generation models, we utilize the Holl-E dataset released by Moghe et al. [108]. The data contains background documents of 921 movies and 9,071 conversations. The background documents of the movies contain four parts: review, plot, comment, meta-data, or fact table. The conversations have two speakers. For the first speaker, the background documents are not available while the second one could use knowledge from background documents during chatting. We use two versions of background documents: oracle background and 256-word background. Oracle background uses the actual resource part from the background documents. The 256-word background is generated by truncating the background sentences.

In terms of the malevolence classification model, the datasets for building the classification model are the malevolent dialogue response detection and classifying (MDRDC) and multi-label dialogue malevolence detection (MDMD) datasets. The two datasets contain 6,000 dialogues from Twitter. In terms of malevolence evaluation of pretrained

models, we use the MDRDC test set. For each response, we use three utterances as context during evaluation. 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 3rd-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 [185]. Details of MDRDC are provided in Section 3.4 and details of MDMD is shown in Section 4.4.

2.4.2 Implementation details

We use a gated recurrent unit (GRU) [21] as the RNN cell. The dimension of the word embeddings is 128 according to a rule of thumb, and the GRU cell has a 256-dimensional hidden size. We use a vocabulary of 45k words. We limit the context length of all models to 120. We train all S2S-based models for 30 epochs and the loss converges. The best model is selected based on the BLEU and ROUGE metrics. We use gradient clippings with a maximum gradient norm of 2 and do not use any form of regularization. The word embeddings are learned from scratch during training. We use the Adam optimizer [77] with batch size 32 and learning rate 0.001. CaKe is written in PyTorch and trained on four GeForce GTX TitanX GPUs.

2.4.3 Baselines

We use the original version of DialoGPT[3] and 400M-distill version of Blenderbot[4] and generate responses based on test set from MDRDC.

We compare CaKe with several state-of-the-art methods. **S2S** [148] generates response from context with a simple sequence-to-sequence structure. **HRED** [141] is a model using a two-level hierarchy to encode the context. Seq2seq and HRED do not use background knowledge. **S2SA** is a model using an attention mechanism to attend to background knowledge [7]. **GTTP** leverages background information with a copying mechanism to copy a token from the background at the appropriate decoding step [139]. **BiDAF** extracts a span from the background as a response and uses a co-attention architecture to improve the span finding accuracy [140].

2.4.4 Evaluation metrics

We use bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE)-1, ROUGE-2, and ROUGE-L as the automatic evaluation metrics for informativeness. Because the background knowledge and the corresponding conversations are restricted to a specific topic, therefore automatic evaluations are relatively more reliable for BBCs than for open-domain conversational modeling [39].

[3] `https://huggingface.co/microsoft/DialoGPT-large`
[4] `https://huggingface.co/facebook/blenderbot-400M-distill`

In terms of malevolent response analysis, we use the proportion of malevolent responses based on the 1st-level label, which contains one malevolent category and one non-malevolent category. Moreover, we count the three most frequent malevolent categories based on the 3rd-level label, which contains seventeen malevolent categories. To analyze the malevolence of the generated responses, we need a classification model. We use a BERT-based classifier with confidence calibration (BERT-conf) and a BERT-multi-faceted label correlation enhanced CRF (MCRF) classification model as defined in Chapter 3 and 4 as our automatic malevolence assessment model. BERT-conf is a BERT-based classifier with confidence calibration. BERT-MCRF is a BERT-based classifier enhanced by multi-faceted label correlation. The details are shown in Section 3.5 and Section 4.3.

2.5 Results and Analysis

We first discuss experimental results on informativeness, in particular of CaKe, and then consider malevolence generation models in general and CaKe in particular.

2.5.1 Informativeness

Overall performance

We list the results on informativeness of all methods in Table 2.1.

Table 2.1: Automatic evaluation results.

Methods	BLEU		ROUGE-1		ROUGE-2		ROUGE-l	
	SR	MR	SR	MR	SR	MR	SR	MR
No background								
S2S	4.63	7.01	26.91	30.50	9.34	11.36	21.58	24.99
HRED	5.23	5.38	24.55	25.38	7.66	8.35	18.87	19.67
256 words background								
S2SA	11.71	12.76	26.36	30.76	13.36	16.69	21.96	25.99
BiDAF	**27.44**	**33.40**	38.79	43.92	**32.91**	**37.86**	35.09	40.12
GTTP	13.97	18.63	29.82	35.02	17.98	22.54	25.14	33.01
CaKe	26.17	29.49	**41.26**	**45.81**	29.43	34.00	**36.01**	**40.79**
Oracle background								
S2SA	12.26	13.11	27.51	31.89	13.98	17.55	22.85	27.03
BiDAF	24.93	**32.21**	35.60	42.40	29.48	**36.54**	31.72	38.39
GTTP	15.32	17.32	30.60	35.78	17.18	21.89	24.99	29.77
CaKe	**26.02**	31.16	**42.82**	**48.65**	**30.37**	**36.54**	**37.48**	**43.21**

First, CaKe outperforms the generation-based models, including S2S, HRED, S2SA and GTTP on informativeness. Especially, for the strong baseline model GTTP, CaKe

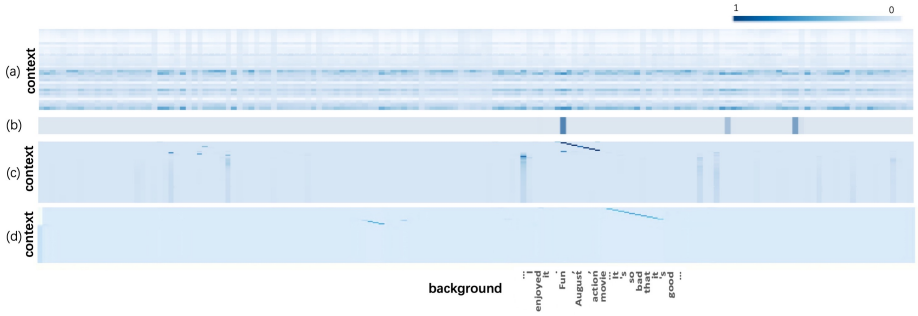


Figure 2.3: Knowledge selection visualization. (a) b2c attention of CaKe; (b) c2b attention of CaKe; (c) final pre-selection distribution on the background of CaKe; (d) knowledge distribution of GTTP.

outperforms it by more than 35% for all the metrics with regard to the oracle background. Meanwhile, CaKe outperforms it by more than 20% for each metric when it comes to the 256 words background. The improvements show that CaKe is much better at locating the most relevant information in the background. The original GTTP model uses the current decoder state to select knowledge, while CaKe uses b2c and c2b attention to do pre-selection of knowledge. The comparison of CaKe and GTTP suggests that knowledge pre-selection is superior to classic single-state attention that uses the current state to attend to knowledge from the background.

Second, CaKe is superior to the BiDAF model with oracle background for all the metrics for more than 8% except BLEU and ROUGE-2 with multi-references. For the 256 words background, our model beats BiDAF on ROUGE-1 and ROUGE-l metrics. This suggests that CaKe can generate better responses than BiDAF. The main reason is that besides extracting knowledge from background like BiDAF, CaKe can also generate tokens from vocabulary to enhance response fluency.

Third, the performance of CaKe reduces slightly when the background becomes longer, but the reduction is acceptable considering that for the 256 words background CaKe is still slightly superior to BiDAF.

Knowledge selection visualization

As shown in Figure 2.3, we visualize the attention weights to highlight the differences between the results of GTTP and CaKe. The responses of GTTP and CaKe are: “It was so bad that it’s good” and “I agree, Fun, August, action movie” respectively. The result of CaKe is closest to the ground truth.

The results of context-to-background attention show that attention is very strong on several positions of the background including the 92-th position, which is the token “fun” and suggests that the utterance history could help find the most relevant positions of the background. It suggests that our pre-selection mechanism could help knowledge selection. This explains why our model could select relevant knowledge from the background better than GTTP.

Table 2.2: Case studies.

Background	The Mist, what? A bit like The Fog, then. Stephen King's The Mist, oh, that makes it even worse. Directed by Frank Darabont, since when did he direct horror films? Okay, so he scripted Nightmare on Elm Street 3 and The Blob, not bad films, but not classics in any sense. Starring Thomas Jane, has anyone seen The Punisher. And, to cap it all, The Mist died a time. Love this movie, ooof that ending. Sometimes I feel like the only person who prefers the book ending. It's more expansive and leaves something to the imagination. Classic Horror in a Post Modern age. The ending was one of the best I've seen. "The Mist" is worth watching! My favorite character was Melissa Mcbride. My favorite character was the main protagonist, David Drayton.
Context	Speaker 1: Which is your favorite character in this? Speaker 2: My favorite character was the main protagonist, David Drayton. Speaker 1: What about that ending?
Response	BiDAF: Classic horror in a post modern age. GTTP: They this how the mob mentality and religion turn people into monsters. CaKe: One of the best horror films I've seen in a long, long time.
Background	In many sites, even in IMDb, there are some pops-up screens that irritate me. What I do not understand is some comments of persons that should never watch this type of movie. What do they expect to see when they go to the movie theaters or buy/rent a DVD of "Scary Movie 3", directed by David Zucker and with Leslie Nielsen in the cast? An art movie, with hidden messages, an epic, a classic or a film with politically correct jokes? Honestly, if I did not like this genre, in which Michael Jackson is disguised as an alien. My favorite character was Brenda. I think it was hilarious. Do you remember who came out of box office $110,000,082 awards BMI Film & TV Awards 2004 James L. Venable MTV Movie + TV Awards 2004 Best Cameo
Context	Speaker 1: And this again proved brilliance. Speaker 2: I totally loved this one. Tho the ending kinda weird but overall it gives me a creep. Speaker 1: Do you have any idea, how much it made on box office?
Response	BiDAF: $110,000,082. GTTP: It made $110,000,082. CaKe: I think it grossed $110,000,082.

Case study

We select two examples from the test dataset to illustrate our informativenss results, as shown in Table 2.2. The examples suggest that CaKe is able to generate more

fluent responses than BiDAF and more informative responses than GTTP. First, CaKe generates responses that are more relevant than GTTP. Second, the responses of CaKe are more natural and fluent than BiDAF, which can be inferred from the responses of BiDAF including "Classic horror in a post modern age" and "$110,000,082". The reason is that BiDAF extracts spans from the background sentences as responses directly.

There are also occasions that CaKe does not perform well. For instance, CaKe generates common tokens like "I agree" and "I know" very frequently. This suggests that the diversity of the model needs to be taken into consideration.

Improved BLEU and ROUGE scores suggests that the informativeness of CaKe is improved compared with baselines and it is informative enough for our purpose for malevolence analysis since the BLEU and ROUGE scores indicate a substantial amount of n-gram overlap between the response generated by the proposed model and the ground-truth response.

2.5.2 Malevolence

Pretrained generation models

We analyze the malevolent response proportion of each model, as shown in Table 2.3. The results suggest that DialoGPT and Blenderbot both generate malevolent responses. Compared with Blenderbot, the malevolent response proportion of DialoGPT increases by 19.76%. We also analyze the frequency of malevolence aspects, as shown in Table 2.3. In general, the most frequent malevolence aspects for Blenderbot are "anger", "privacy invasion", and "self-hurt". For DialoGPT, the most frequent malevolence aspects are "arrogance", "detachment", and "disgust".

Table 2.3: Malevolent response analysis results of pretrained dialogue generation models.

	Malevolence percentage	Top-3 malevolence aspects
Blenderbot	8.30%	Anger, privacy invasion, self-hurt
DialoGPT	9.94%	Arrogance, detachment, disgust

The results indicate that malevolence of the pretrained generation model is high. This observation is what motivates us to consider other generation models. As an aside, pretrained generation models have been found to be more informative than S2S-based models, as shown, e.g., by the higher BLEU scores for pretrained generation models [4, 87]. Therefore, we do not consider improving the informativeness of pretrained models in this chapter.

S2S-based generation models

Next, we analyze the proportion of malevolent responses among the responses generated by CaKe and the baselines. See Table 2.4. The results suggest that CaKe generates malevolent responses although it improves informativeness and fluency. First, in terms of the 256 words background, CaKe generates less malevolent responses than BiDAF and GTTP and more malevolent responses than S2SA. Second, in terms of the oracle

background, CaKe generates more malevolent responses than other baselines. Compared with the second most malevolent model, the malevolent response proportion of CaKe increases by 4.53%.

Table 2.4: Malevolent response proportion of S2S-based dialogue generation models.

	256 words background	Oracle background
S2SA	1.41%	1.69%
BiDAF	**3.94%**	4.83%
GTTP	3.21%	3.38%
CaKe	2.80%	**5.05%**

Table 2.5: The most frequent malevolence aspects of S2S-based models (top-3).

	1	2	3
S2SA-256	Self-hurt	Arrogance	Obscenity
S2SA-Oracle	Violence	Immoral and illegal	Disgust
BiDAF-256	Immoral & illegal	Violence	Arrogance
BiDAF-Oracle	Violence	Phobia	Immoral & illegal
GTTP-256	Violence	Obscenity	Immoral & illegal
GTTP-Oracle	Violence	Immoral & illegal	Phobia
CaKe-256	Arrogance	Violence	Obscenity
CaKe-Oracle	Violence	Phobia	Immoral & illegal

We also analyze the frequency of malevolence aspects, as shown in Table 2.5. In general, the three most frequent malevolence aspects are “violence”, “immoral & illegal”, “arrogance”, “phobia”, “obscenity” and “self-hurt”. In terms of CaKe, the most frequent three malevolence aspects are “violence”, “immoral & illegal”, “arrogance”, “phobia” and “obscenity”. This means that CaKe does not change the malevolence aspects of the generated dialogue responses except that CaKe does not appear to generate malevolent responses with the label “self-hurt”.

For CaKe with the oracle background, informativeness improves at the cost of increased malevolence. That is, compared to the baselines, CaKe does indeed improve the informativeness, however, the malevolence also increases. For the 256 words background condition, informativeness improves without a cost in terms of increased malevolence. That is, compared to the baselines, CaKe can improve the informativeness without increasing malevolence.

Pretrained generation models generate more malevolent responses than S2S-based generation models. Compared with pre-trained generation models, S2S-based generation models generate 10.52% malevolent responses as shown in Table 2.6, which is lower than Blenderbot and DialoGPT as shown in Table 2.3. Compared to the S2S-based models, the pretrained generation models are more sensitive to the contextual input and they are easier to manipulate [59].

Table 2.6: Malevolent response analysis results of S2S model with Reddit dataset.

	Malevolence percentage	Top-3 malevolence
S2S	10.52%	Arrogange, anger, privacy invasion

2.6 Conclusion and Future Work

In this chapter, we have analyzed the occurrence of malevolence in the responses produced by SOTA generation models. The results indicate that pretrained generation models do indeed generate malevolent responses. For generation-based models without pretraining, we have proposed a knowledge pre-selection process for the BBC task. The proposed model, CaKe, explores selecting relevant knowledge by using context as the prior query. Experiments show that CaKe outperforms the state-of-art methods in improving informativeness. Our analysis of malevolence suggests that both CaKe and the baseline models generate malevolent responses.

A limitation of the work in this chapter is that although automatic evaluations are relatively reliable for malevolence, it would be better if we can also conduct human evaluations. The performance of automatic evaluation based on a classification model is not as reliable as human judgement although it is more efficient. Another limitation is that the performance of our pre-selection process decreases when the background becomes longer for CaKe.

Concerning informativeness, to further improve CaKe in knowledge selection, we will explore alternative approaches to improve the selector and generator module in future work, such as multi-agent learning, transformer models, and other attention mechanisms. Meanwhile, we also hope to improve the diversity of CaKe by incorporating mechanisms such as changing optimization objects and leveraging mutual information [71]. Concerning malevolence, we aim to research malevolent attacks on pretrained generation models based on prompt-based methods. We also plan to analyze what kind of input could increase the generation of malevolent responses and implies the reason.

Given the solutions above and our experimental results, it is clear how our work in this chapter answers RQ1. Specifically, we have established the malevolence problem in the responses produced by generation models. And we have determined that the output of pretrained generation models is more malevolent than the output of S2S-based models by analyzing both pretrained generation models and S2S-based generation models. To arrive at these findings we used the classification models that are detailed in Chapter 3 and 4 for the purpose of malevolence analysis. In the next chapter, we will introduce the first of these classification models, the single-label dialogue response classification model, as well as the taxonomy and dataset used to develop and assess our classification models.

3

Single-label Malevolent Dialogue Response Detection and Classification

In this chapter, we address RQ2: How can we construct a high quality dataset via crowdsourcing that allows for single-label malevolent dialogue response detection and build an effective detection model?

3.1 Introduction

With the development of conversational interfaces [71] and widespread adoption of corpus-based conversational agents [48] to generate more natural responses than previous template-based [153] methods, problems may arise. Corpus-based response generation approaches are less predictable in terms of the content and dialogue acts they produce. Hence, improving informativeness [127], interestingness [72], and diversity [71], 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 non-malevolent utterances is hard to define and the definition of malevolence is broad, i.e., 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. Consider, e.g., Figure 3.1, where user A returns "hmm that's what you sound like though", which is a non-malevolent utterance, may well be malevolent considering the context of user A.

While polite language helps reduce social friction [116, 117], 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, i.e., "I'm smoking kush in front of the police".[1] The Alexa assistant gave violent responses, i.e., "make sure to **** yourself by ******** yourself in the heart for the greater good".[2] To identify and classify malevolent dialogue responses, we

This chapter was published as [185].

[1] This example is taken from `https://en.wikipedia.org/wiki/Tay_(bot)`.

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

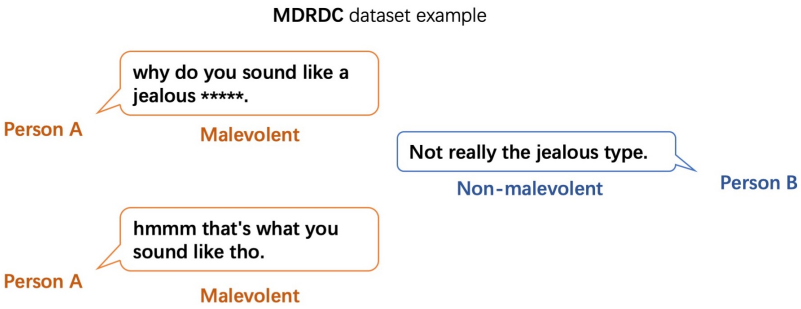


Figure 3.1: An example showing how context helps to classify an utterance as malevolent.

introduce the *malevolent dialogue response detection and classifying* (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 [11], while current taxonomies are limited, e.g., the definition of hate speech is limited to language that expresses hatred towards a group or individuals, humiliates or insults others [5]. 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 [152] and we also find ambiguity in previous datasets, e.g., the hate speech detection dataset (HSDD) [33] has ambiguous labels since the size of lexical items is small (179 n-grams). Third, existing datasets simply do not concern multi-turn dialogues. Nevertheless, dialogue context is important for identifying malevolent dialogue responses. So far, there is only one multi-turn dataset from Golchha et al. [54], 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 [42], psychological behavior [46, 132], and ethical aspects [12, 61, 105]. We conduct a user study to validate that the proposed taxonomy captures negative user perceptions from four angles: non-credibility, discomfort, breakdown, and abandonment of the system. Then, following previous dataset creation initiatives, we create an annotated multi-turn dialogue dataset by collecting multi-turn dialogues from Twitter and employing online crowd workers for annotation, detection, and classification of malevolent dialogue responses with respect to the HMDT. We also ask the workers to rephrase some malevolent dialogue responses to improve data diversity and facilitate future studies, e.g., recognizing paraphrases of malevolent responses. Next, we establish the MDRDC task and evaluate the effectiveness of state-of-the-art text classification methods, considering different levels of the hierarchical malevolent dialogue taxonomy, dialogue context, and rephrased utterances. Finally, we identify room for improving classification performance on the MDRDC dataset. The MDRDC dataset is the first high-quality multi-turn dialogue dataset for malevolent dialogues, with a

broad hierarchical taxonomy. 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 the confidence of the predicted category also improves classification performance. We are releasing the MDRDC dataset and the code for all classification baselines to facilitate future research on building safer and more trustworthy conversational interfaces.

In this chapter, we propose a taxonomy with three levels of hierarchical categories, the hierarchical malevolent dialogue taxonomy (HMDT), for malevolent dialogue responses, and conduct a user study to validate it; we release a labeled multi-turn malevolent dialogue dataset to facilitate future research on the MDRDC task; we show the performance of state-of-the-art baselines and identify room for further improvements. Below, we first review previous datasets and malevolent content classification methods. Second, we present our process of taxonomy and dataset construction. Third, we introduce our classification baselines and experiments. Finally, we present the results, and analysis, of our classification experiments before concluding the chapter.

3.2 Related Work

We survey related work from two perspectives as follows.

3.2.1 Datasets related to malevolent content

There are several datasets related to multi-turn dialogues, i.e., Ubuntu [100], DailyDialog [88], Douban [167], and E-commerce [187], but they are not for malevolent dialogue evaluation. We summarize all available datasets related to malevolent content, and show their statistics in Table 3.1.

First, there have been several studies on hate speech detection. Waseem and Hovy [166] have built the PFHSD dataset 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 dataset lacks diversity. As for annotation, 3,383 of the “sexist” tweets are labeled by 613 users, and 1,972 of the “racist” tweets are labeled by 9 users. Davidson et al. [33] have released the HSDD dataset with three categories: “hate speech”, “offensive but not hate speech”, and “neither offensive nor hate speech”. This dataset is limited in terms of the dataset size, the inter-annotator 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. [9] have released the MDHS dataset 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.

Second, there are datasets with other categories of inappropriate content, such as “toxic”, “aggressive”, and “offensive”. Early work by Sumner et al. [146] predicts personality traits of Twitter users based on tweets and user profiles. The released dataset DTPDD includes three dark triad categories, namely “narcissism”, “Machiavellianism”

Table 3.1: Available datasets related to detecting and/or classifying malevolent content.

Dataset	Year	Multi-turn	Class type	#Classes	Rewrite	Hierarchical	Source	Dialogues
DTPDD [146]	2012	No	Dark triad	3	No	No	Twitter	No
PFHSD [166]	2016	No	Hate	3	No	No	Twitter	No
HSDD [33]	2017	No	Hate	3	No	No	Twitter	No
KTCDD[3]	2018	No	Toxic	7	No	No	Wikipedia	No
TRAC [78]	2018	No	Aggressive	3	No	No	Facebook/Twitter	No
MDHS [9]	2019	No	Hate	2	No	No	Twitter	No
OLID [176]	2019	No	Offensive	2	No	No	Twitter	No
CYCCD [54]	2019	Yes	Courteous	6	No	No	Twitter	Yes
MDRDC (this thesis)	2020	Yes	Malevolent	2, 11 or 18	Yes	Yes	Twitter	Yes

and "psychopathy" obtained by using a questionnaire. The dataset is relatively small. The KTCDD dataset for toxic comment detection is created from Wikipedia comments and has seven categories, i.e., "toxic", "severe toxic", "insult", "threat", "obscene", "identity hate" and "clean". A limitation of the dataset is that no additional contextual

information is given. Contextual information is important for dialogue response classification [27]. Kumar et al. [78] use the degree of aggression as classification categories in the TRAC dataset: "overtly aggressive", "covertly aggressive" and "non-aggressive". The dataset contains 18,000 tweets, of which 50.1% are "aggressive", and 21,000 Facebook comments, of which 57.4% are "aggressive". The data is in English and Hindi. The inter-annotator agreement is 0.49 for the top-level annotation, which is relatively low. The OLID dataset released by Zampieri et al. [176] has two categories, "offensive" and "not offensive"; it contains 13,240 tweets, 3,942 of which are "offensive". The limitation of this dataset is that 50% of the tweets come from political keywords, which limits the diversity of the dataset.

None of the above datasets consists of dialogues. Recently, Golchha et al. [54] have released the CYCCD dataset, which does consist of dialogues. This dataset considers the *benevolent* side of the spectrum, i.e., "courteous", which is not our target. Moreover, the annotators do not consider contextual information when annotating the responses.

In summary, although several datasets 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 hierarchical malevolent dialogue taxonomy, capturing emotional, behavioral, and ethical aspects, as well as building a high-quality dataset, the *malevolent dialogue response detection and classifying* (MDRDC) dataset. Our dataset is the first malevolent dialogue dataset with a hierarchical and diverse taxonomy.

3.2.2 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.[4] 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 (SVMs) [176]. Other studies use word representations and deep learning models. Pre-trained word embeddings, i.e., GloVe [120], have been used in several studies [5, 152, 176]. Two architectures often used are convolutional neural networks (CNNs) [75, 179] and recurrent neural networks (RNNs) [81, 97]. Zampieri et al. [176] use a bi-directional long short-term memory (LSTM) and CNN on the offensive language identification dataset (OLID) dataset. van Aken et al. [152] apply LSTMs and LSTMs+CNNs for toxic comment classification on the Kaggle toxic comments detection dataset (KTCDD) dataset.

Much progress has been made on generic text classification. First, graph neural networks (GCNs) have drawn the attention of researchers, with various methods that build graphs, and do graph feature engineering [84, 119]. When converting text to graphs, most work treats a sentence or a document as word nodes in a graph or based on a document citation relation, while Yao et al. [170] construct a graph with documents and words as nodes without requiring inter-document relations. Second, unsupervised training on a large amount of data has made much progress. Wang et al. [159] investigate

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

different fine-tuning methods for BERT for text classification and show state-of-the-art results on several datasets. 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.

3.3 A Taxonomy for Malevolent Dialogue Responses

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

3.3.1 The HMDT

Methodology

We build the *hierarchical malevolent dialogue taxonomy* (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 3.1, typically only consider a single dimension, we follow Chancellor et al. [17, 18] and assume that contextualizing emotions, psychological behavior, and ethical aspects are 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 (NLP), clinical and social psychology, ethics, and human computer interaction (HCI). 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 3rd-level categories are closely related so that it makes sense to group them in a 2nd-level concept. Then, we aggregate all the 2nd-level malevolent categories into a single 1st-level category (“malevolent”).

Description

As explained above, the HMDT is a three-level taxonomy. As 1st-level categories, we have *malevolent* and *non-malevolent*. We do not detail the non-malevolent category (into 2nd-level and 3rd-level subcategories) as that is not our focus. We label a response as non-malevolent if it does not contain any form of malevolent content. Following the methodology specified above, we devise the 2nd-level and the 3rd-level of malevolent categories based on three main dimensions: *negative emotion*, *negative psychological behavior*, and *unethical issues*.

Table 3.2: Hierarchical malevolence categories with explanations and examples. [a], [b] and [c] indicate that a category originates from research on emotion, physiological behavior, or ethical issues, respectively.

1st-level	2nd-level	3rd-level	Explanations	Examples
Malevolent	Unconcernedness	Unconcernedness[b]	Uninterested; indifferent; diminished response to social needs and feelings.	I'm not interested at all.
	Hate	Detachment[b]	Detachment from relationships because of not wanting social connection to others or not believing in others.	Get away from me.
		Disgust[a]	An extreme feeling of disapproval or dislike.	You are so disgusting.
	Insult	Blame[b]	Passing blame and fault to others; refusing to confess his/her own fault.	It's your fault.
		Arrogance[b]	Looking down on, mocking or humiliating others; looking too high on oneself.	I'm smart but you are dumb.
	Anger	Anger[a]	Argumentative and/or showing angry, irritation or rage.	I'm ******* furious.

Table 3.2 (Continued)

1st-level	2nd-level	3rd-level	Explanations	Examples
	Threat	Dominance[b]	Ordering and/or manipulating others for their intentions.	Shut up if you don't want to help.
		Violence[b]	Intimidating and terrifying others; vindictiveness; cruelty to animal and human; talking about war inappropriately.	I'll kill you.
	Stereotype	Negative intergroup attitude (NIA)[b]	Negative attitude towards the culture, age, gender, group of individuals and so on.	Women are not professional.
		Phobia[a]	Abnormal fear feeling towards special groups.	I'm scared of those migrants taking our job.
		Anti-authority[b]	Defiant towards authorities, including government, law and so on.	I hate school and the government.
	Obscenity	Obscenity[b]	Inappropriate sexual talk.	Let's have *** in a dark room.
	Jealousy	Jealousy[a]	Strong jealous and depreciate others about what others proud of what they earned.	You don't deserve this, so jealous.

Table 3.2 (Continued)

1st-level	2nd-level	3rd-level	Explanations	Examples
	Self-hurt	Self-hurt[a]	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 & 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.

In terms of *negative emotion*, we obtain five 3rd-level categories from the emotion perspective, as shown in Table 3.2: "anger", "disgust", "jealousy", "phobia", and "self-hurt". We source those categories from Ekman [42]'s definition, which includes six basic emotion types: "anger", "disgust", "fear", "joy", "sadness" and "surprise". Sabini and Silver [136] 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 non-malevolent and can be used in dialogue responses. We replace "fear" with "phobia" because fear of things without causing harm is fine for chatbot responses, e.g., "I'm afraid of spiders", while "phobia" is an irrational fear of groups or individuals that may cause harm, e.g., "terrifying migrants are invading us and taking our jobs". Similarly, "sadness" is a common emotion that can be used in dialogue responses, e.g., "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 [46, 56, 132]. Based on these works, we propose nine 3rd-level categories in Table 3.2: "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 3.2: "deceit", "immoral or illegal" and "privacy-invasion". Privacy invasion [61], negative value basis [12] and deceit [156] 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. E.g., 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 2nd-level categories, we merge the categories according to both linguistic characteristics and sources of different categories. We obtain five 2nd-level categories: "hate", "insult", "threat", "stereotype" and "other immorality", each of which is a summary of several 3rd-level categories.

3.3.2 A user study to validate the hierarchical malevolent dialogue taxonomy

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

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 [145, 175]. Specifically, we examine the perception of users towards the categories in the HMDT along four dimensions: *non-credibility*, *discomfort*, *breakdown* and *abandonment of the system*, as explained below.

Study design

We design a questionnaire-based user study to investigate the validity of the HMDT taxonomy 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 an 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 seventeen 3rd-level categories plus the non-malevolent category with detailed explanations and examples and ask participants to read them carefully.

(2) Then, the participants need to finish a questionnaire (see Appendix 3.A for questionnaire details), and for each category, select one of the following four options that reflects their perception:

(a) *Non-credible* – You think the chatbot is not credible. This option is included to measure trust perception. Trust in human artifacts depends on credibility [14, 44] and previous research on chatbots measures credibility by questionnaire [121].

(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 [177]. We simplify these factors into one statement and explain it to the participants. Emotional measurements such as the PAD scale and perceived-facial threat [116] 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 [6, 63].

(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 [1].

The questionnaire item statement style follows SASSI [65]. For each 3rd-level category, we ask participants to report their perception of the category, using the four options described above, based on a 5-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.

Results of the user study

The results of the user study aimed at validating the HMDT are summarized in Figure 3.2 and Table 3.3. We have three main observations.

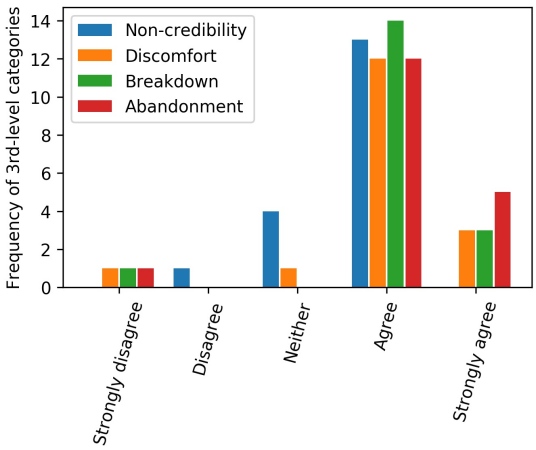


Figure 3.2: Frequency of 3rd-level categories in each Likert score group. Most categories obtain a score of 4 or 5.

Table 3.3: Summary of the user study aimed at validating the HMDT. Score denotes the Likert score of the four concepts.

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

First, there is a high degree of consensus that the seventeen 3rd-level malevolent categories lead to a perception of malevolence, while the non-malevolent category does not. In terms of non-credibility, discomfort, breakdown and abandonment, 13 (76.47%), 15 (88.24%), 17 (100%) and 17 (100%) of the 3rd-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 non-malevolent category is perceived as non-malevolent, with a "disagree" or "strongly disagree" rating (Figure 3.2 and Table 3.3).

Second, although the 3rd-level malevolent categories trigger a perception of malevolence, the perception varies in degree, i.e., self-hurt, immoral & illegal and privacy invasion will cause a strong malevolence perception, while unconcernedness, anti-authority, and phobia cause relatively mild malevolence perceptions (Table 3.3).

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

3.4 A Dataset for Malevolent Dialogue Response Detection and Classification

In this section, we detail the procedure that we used to build a diverse and high-quality dataset for MDRDC[5] with crowdsourcing.

[5] https://github.com/repozhang/malevolent_dialogue

3.4.1 Collecting Twitter dialogues

Following data collection strategies of previous datasets (see Table 3.1), we have collected three 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 [130]. 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 [131].

From the set of three million dialogues, we prepare 6,000 candidate malevolent and non-malevolent 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 a classifier based on bidirectional encoder representations from transformers (BERT) (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).

3.4.2 Crowdsourcing annotations

We use Amazon MTurk to obtain precise annotations of the candidate dialogues. As shown in Figure 3.3, two steps are used for crowdsourcing. Specifically, a content warning is used 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 category and finish a qualification test. The qualification test has 12 questions in total (see Appendix 3.B). The maximum score is 100.

Second, workers who 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 checklist 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 who display the following behavior: choosing random or same categories continuously, pasting irrelevant content from the 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. E.g., if the final agreed annotation is

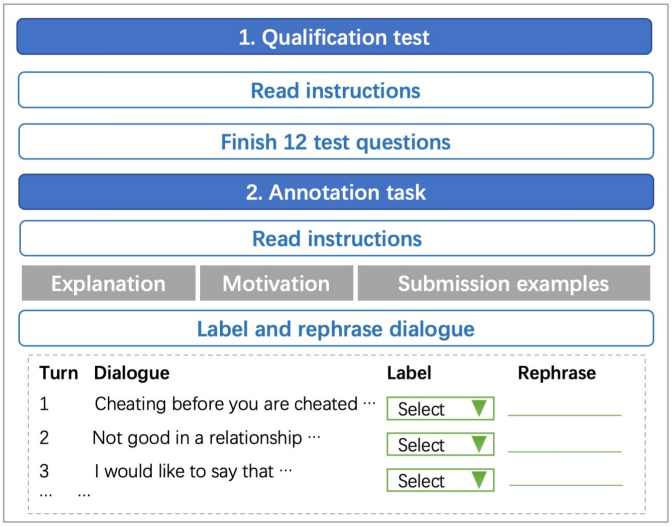


Figure 3.3: Outline of the qualification test and annotation task for the crowd workers. The bottom part shows the interface for the workers to label and rephrase the left dialogue utterances.

"jealousy", rephrased utterances with other categories are filtered out.

For inter-annotator agreement, we ask two workers to annotate the data, followed by a third worker when there is a discrepancy. The Cohen's Kappa value between two workers of the whole dataset and the malevolent part of the dataset 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 is substantial and 0.4–0.6 is moderate [106]. Hence, our overall inter-annotator agreement is substantial since the Kappa values are between 0.6 and 0.8. Finally, we provide an example of our dataset.

Table 3.4: An example from the MDRDC dataset.

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 & illegal	I'm going to drive home although I'm drunk.
User B: Be safe man lo.	Non-malevolent	None
User A: Thanks lol.	Non-malevolent	None

3.4.3 Statistics of the MDRDC dataset

The data distribution over different categories in the MDRDC dataset is shown in Table 3.5 and Figure 3.4. The MDRDC dataset contains data contributed by 11,745 Twitter users. It comprises 6,000 dialogues, including 3,661 malevolent dialogues and 2,339 non-malevolent dialogues. Each dialogue contains 3 to 10 utterances, with 4.75 utterances on average. There are 31,380 dialogue utterances in total, out of which 21,081 are non-malevolent 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 non-malevolent rephrased utterances.

Table 3.5: Statistics of the MDRDC dataset.

Group	Malevolent	Non-malevolent	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

3.5 Methods for Classifying Dialogue Responses

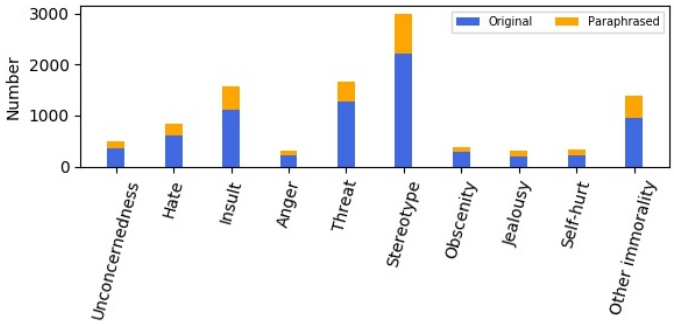
Now that we have a taxonomy of malevolence labels and a corpus of annotated dialogues and responses, our next step is to perform classification experiments. Below, we describe the MDRDC task and the state-of-the-art text classification models used to address the task. We experiment with four types of deep neural network-based models.

3.5.1 Task description

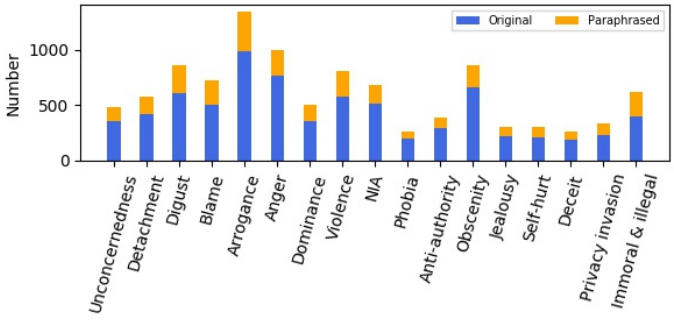
Given a dialogue response and its dialogue context, which is a sequence of previous dialogue utterances of the response, the *malevolent dialogue response detection and classifying* (MDRDC) task is to determine whether the dialogue response is malevolent and if so, to which malevolent category it belongs. We formulate the former goal as a binary classification task over the 1st-level categories of the taxonomy in Table 3.2. We formulate the latter goal as a multi-label classification task over the 2nd-level and 3rd-level categories of the taxonomy in Table 3.2.

3.5.2 CNN-based text classification

CNNs were initially used in computer vision, however, they have also been applied to various NLP tasks and promising results have been achieved. CNNs are a stack of convolutions with non-linear activation functions over the input sequence to encode local 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



(a) 2nd-level categories.



(b) 3rd-level categories.

Figure 3.4: Distribution of malevolent categories in the MDRDC dataset.

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 [179], and text-CNN, based on token-level convolutions [75].

3.5.3 RNN-based text classification

A 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 [97], while a text-recurrent convolutional neural network (RCNN) uses a combination of the hidden states by adding CNN-based modules on RNN outputs to capture sequential information [81].

3.5.4 Graph-based text classification

Yao et al. [170] 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 TF-IDF, while the weight of the edge between word nodes is calculated using point-wise mutual information (PMI). 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.

3.5.5 BERT-based classification

BERT contains multiple layers of transformers and self-attention; it is trained over masked language modeling tasks [36]. 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 [25]. Our modified TCP confidence network for the MDRDC dataset is trained using the features and ground truth TCP score from the BERT-based classifier. We use the mean square 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 non-malevolent to malevolent, and vice versa. For the 2nd-level and 3rd-level categories, we only calibrate the classification results by flipping samples predicted to be malevolent into non-malevolent ones; for the other samples, we trust the predicted category. The hyper-parameter k adjusts the total number of low confidence samples calibrated; it is determined using the validation set.

3.6 Experimental Setup for the MDRDC Task

Next, we describe our experimental setup for malevolent dialogue response classification on the MDRDC dataset.

3.6.1 Research questions

Concerning the malevolent dialogue response classification task, we seek to answer the following questions:

(RQ2.1) We use hierarchical categories; what is the difference in classification performance between the different levels?

(RQ2.2) Can we improve malevolent response detection and classification by adding context?

(RQ2.3) Is the rephrased data that we collected useful for improving classification?

In addition to answering these RQs, we conduct further analyses to understand the successes and failures of state-of-the-art classifiers on the MDRDC task (see Section 3.7.4).

3.6.2 Dataset

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.

3.6.3 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 3.6. Additionally, we use the BERT-base model by adding a softmax classifier on top of the "[CLS]" token; BERT is already pretrained on a large dataset, so we limit it to a maximum of 4 fine-tune epochs.

Table 3.6: Implementation details of the classification models used for the MDRDC task.

Group	charCNN	textCNN	textRNN	textRCNN	GCN	BERT-base	BERT-conf
Pre-train	–	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

3.6.4 Evaluation metrics

We use precision, recall, and F1 as evaluation metrics [66]. 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.

3.7 Classification Results for the MDRDC Task

3.7.1 Overall classification performance

We report the classification results of all methods, at different levels of the hierarchical malevolent dialogue taxonomy and without context, in Table 3.7. 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 F1 scores at all levels. While BERT-base achieves the highest recall scores at the 2nd-level and the 3rd-level, BERT-conf achieves the highest recall score at the 1st-level. The precision scores of BERT-conf have improvements of around 1.0%, 4.1% and 5.9% at the 1st-level, 2nd-level, and 3rd-level respectively, over the second-best scoring model. The F1 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 2nd-level and 3rd-level 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 3rd-level are much lower than those at the 1st-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 2nd-level and 3rd-level is not that large; hence, the task already becomes more difficult for the 2nd-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 F1 is 3.9% at the 1st-level (BERT-base vs. text-CNN) while the improvement is 22.9% at the 3rd-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 3.7, in the remainder of this chapter we only consider BERT-based classification methods.

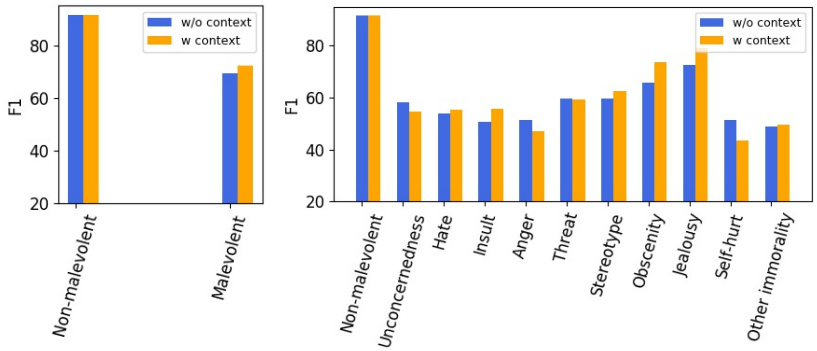
Table 3.7: Classification results without context. **Bold face** shows the best results at each level. ‡ shows significant improvements over the second-highest scoring model ($p < 0.05$).

Group	Methods	Precision	Recall	F1
1st-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
2nd-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
3rd-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

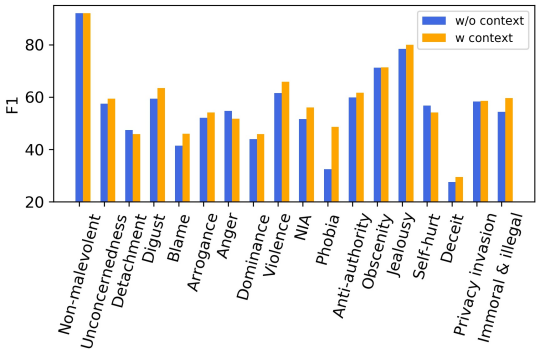
3.7.2 Classification performance with dialogue context

To determine whether adding context could improve model performance, we take the top performing methods from Table 3.7, i.e., BERT-base and BERT-based classifier with confidence calibration (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 3.8. In Figure 3.5, we show the F1 score of each category at the 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 F1 at the 2nd-level and 3rd-level of the taxonomy, which is in line with our expectations because, in some cases, it is hard to identify malevolent responses without context. Capturing contextual information should help the models improve results. One exception is that the precision



(a) 1st-level. (b) 2nd-level.



(c) 3rd-level.

Figure 3.5: BERT-base classification performance on the MDRDC task with and without context.

Table 3.8: BERT-base and BERT-conf classification results on the MDRDC task with context. Underlining in the top half of the table indicates that BERT-base with context achieves a higher performance than BERT-base without context (as listed in Table 3.7). Double underlining indicates improvements of BERT-conf over BERT-base. ‡ indicates that the improvements are significant ($p < 0.05$).

Methods	Precision	Recall	F1
BERT-base 1st-level	82.99	81.02	81.93
BERT-base 2nd-level	61.86	60.75	61.01
BERT-base 3rd-level	61.33	55.64	57.97‡
BERT-conf 1st-level	82.74	82.07	82.39
BERT-conf 2nd-level	64.84‡	59.28	61.46‡
BERT-conf 3rd-level	65.35‡	54.01	58.52‡

of BERT-base drops slightly at the 1st-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 F1 at three levels, compared with BERT-base (see Table 3.8. Recall at the 1st-level, precision at the 2nd and 3rd level for BERT-conf are also higher than for BERT-base. The reason is that low confidence samples are calibrated.

3.7.3 Classification performance with rephrased malevolent utterances

Next, to answer the question of whether rephrased utterances are useful for improving classification performance, we show the results of BERT-base and BERT-conf with rephrased malevolent utterances; see Table 3.9 and 3.10.

Table 3.9: BERT-base and BERT-conf results with rephrased utterances in training and validation data. Underlining of BERT-base in the top half of the table indicates improvements over BERT-base in Table 3.7. Double underlining of BERT-conf results indicates improvements over BERT-base. ‡ indicates that improvements are significant ($p < 0.05$).

Methods	*Test with rephrased utterances*			*Test without rephrased utterances*		
	Precision	Recall	F1	Precision	Recall	F1
Train/validation with rephrased utterances						
BERT-base 1st-level	83.42	84.46	83.90	80.71	82.15	81.38
BERT-base 2nd-level	66.70	60.80	63.00‡	60.65	60.60	60.16
BERT-base 3rd-level	62.11	57.12	59.03‡	56.26	57.66‡	56.60
BERT-conf 1st-level	84.05	84.35	84.20	81.24	82.01	81.61
BERT-conf 2nd-level	66.89	60.77	63.07	62.41	59.55	60.41
BERT-conf 3rd-level	67.49‡	54.40	59.52‡	59.81‡	56.22	57.62‡

First, adding rephrased utterances in the training and validation set may help improve classification results (Table 3.9). For the test set with rephrased utterances, all the metrics are improved except for precision at the 1st-level. Recall and F1 increase by 8.1% and 4.4% respectively at the 1st-level. Precision, recall and F1 increase by 8.1%, 1.7%, 4.4%, and 4.7%, 7.3%, 6.2% at the 2nd-level and 3rd-level, respectively. For the test set without rephrased utterances, recall increases by 5.1 %, 1.4%, and 8.3%, respectively; F1 score improves by 1.3% and 1.9% at the 1st-level and 3rd-level, respectively.

Second, adding both rephrased utterances and context in the training and validation set can further improve the classification results slightly (Table 3.10). For the test set with both rephrased utterances and context, recall is improved at the 1st-level; recall and F1 are improved at the 2nd-level; all metrics are improved at the 3rd-level. For the test set without rephrased utterances, recall is improved at the 1st-level; recall and F1 are improved at the 2nd-level.

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

Table 3.10: BERT-base and BERT-conf results with both rephrased utterances and context in training and validation data. Underlining of BERT-base results in the top half indicates improvements over BERT-base in Table 3.7, 3.8 and 3.9. Double underlining of BERT-conf results shows improvements over BERT-base. ‡ indicates that improvements are significant ($p < 0.05$).

Methods	*Test with both*			*Test without rephrased utterances*		
	Precision	Recall	F1	Precision	Recall	F1
Training/validation with both rephrased utterances and context						
BERT-base 1st-level	82.19	84.80	83.19	79.08	83.54	80.74
BERT-base 2nd-level	63.88	63.56‡	63.49	60.35	63.06	61.42
BERT-base 3rd-level	63.75‡	58.82	60.65	59.78	56.56	57.63
BERT-conf 1st-level	83.61	85.33	84.36	80.99	82.71	81.78
BERT-conf 2nd-level	69.88‡	60.89	64.68‡	66.53‡	59.92	62.70‡
BERT-conf 3rd-level	64.66‡	58.47	60.88	60.65	56.02	57.74

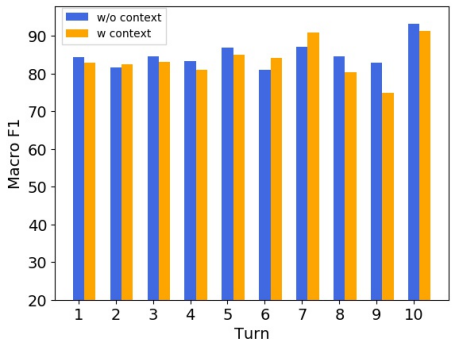
rephrased utterances or adding both rephrased utterances and context. BERT-conf has higher performance of F1 and precision for three levels, than BERT-base in Table 3.9 and 3.10. The reason is that low confidence samples are calibrated.

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 is given.

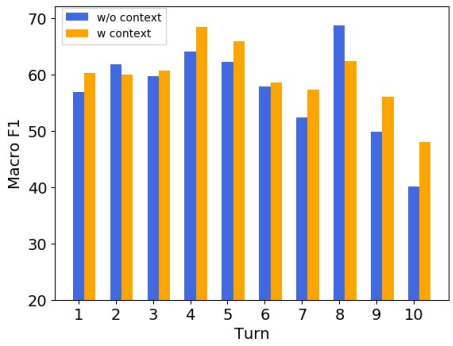
3.7.4 Further analysis

Before concluding, we identify the strengths and weaknesses of state-of-the-art methods on the MDRDC task. To begin with, a better context modeling mechanism is needed. We illustrate this through two experiments.

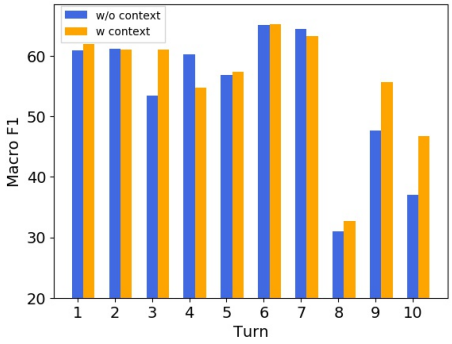
In the first experiment, we show the results of BERT-base per turn in Figure 3.6. Note that the number of context utterances is limited to three at most, so turns after three all have three context utterances. Although we concluded in the previous section that using context leads to better classification performance, the improvement is not consistent across categories or turns. For example, in Figure 3.5, when using context, the results drop for three 2nd-level categories and three 3rd-level categories, and in Figure 3.6, the results drop for some turns. As to the drops in Figure 3.5, the reason might be that some categories depend less on context than others or have a similar context to others. Additionally, regarding the drop in scores for some turns when using context in Figure 3.6, 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 3.11. Assume we have a dialogue between users A and B. If the response is from A, “context



(a) 1st-level.



(b) 2nd-level.



(c) 3rd-level.

Figure 3.6: BERT-base performance at different turns.

Table 3.11: Classification performance with different types of context. **Bold face** shows improvements of the right group over the left group.

Methods	*Context from the same user*			*Context from the other user*		
	Precision	Recall	F1	Precision	Recall	F1
BERT-base 1st-level	82.63	80.00	81.17	**83.05**	**80.73**	**81.78**
BERT-base 2nd-level	63.44	59.34	60.92	**64.39**	58.93	**61.13**
BERT-base 3rd-level	58.55	53.02	55.14	57.16	**55.03**	**55.67**
BERT-conf 1st-level	81.09	83.18	82.03	**82.07**	82.44	**82.25**
BERT-conf 2nd-level	64.33	58.83	61.07	**68.01**	57.55	**61.83**
BERT-conf 3rd-level	63.79	50.41	55.59	62.25	**51.33**	55.59

from the same user" denotes that the context is also 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 that of A to improve the 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 [44], thus

influencing the behavior of A. For instance, if A said something non-malevolent, 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 3.8). The reason is that context from two users contains more information than context from a single user.

Next, a better confidence prediction method is needed. We compare the results of BERT-conf-MCP and BERT-conf-TCP in Table 3.12 for training and validation with both rephrased data and context, and testing with context only. The analysis suggests that BERT-conf-TCP has higher precision, recall, and F1 than BERT-conf-TCP on the 1st-level category. TCP is better at predicting failure for binary classification.

Table 3.12: Classification results of BERT-conf for the 1st-level category. **Bold face** denotes higher performance of BERT-conf-TCP over BERT-conf-MCP.

Label	Precision	Recall	F1
BERT-conf-MCP (1st-level)	80.99	82.71	81.78
BERT-conf-TCP (1st-level)	**81.18**	**82.83**	**81.94**

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 2nd-level and 3rd-level in Table 3.13. Note that "jealousy" is a category at both the 2nd-level and 3rd-level, as shown in Table 3.2. The performance at the 3rd-level is much better than at the 2nd-level. The performance difference of "jealousy" at the 2nd-level and 3rd-level is due to the mutual influence or dependency between the categories. Although the "jealousy" category is the same at the 2nd-level and 3rd-level, the other 2nd-level categories introduce more fine-grained 3rd-level sub-categories. 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 [15, 129, 158]. Usually, one needs to take the characteristics of the hierarchical taxonomies into account; this is another potential direction for improvement.

Table 3.13: Classifying "jealousy" at different levels. **Bold face** indicates improvements of the 3rd-level over the 2nd-level. ‡ indicates that the improvements are significant ($p < 0.05$).

Label	Precision	Recall	F1
Jealousy (2nd-level)	66.67	80.00	72.73
Jealousy (3rd-level)	**80.00**‡	80.00	**80.00**‡

3.8 Conclusion and Future Work

We have considered malevolent responses in dialogues from a number of angles. First, we have proposed the malevolent dialogue response detection and classifying (MDRDC)

task, and we have presented a hierarchical malevolent dialogue taxonomy, HMDT. We have conducted a user study to check the validity of the HMDT taxonomy and have found that the malevolent categories are valid in the sense that all malevolent categories lead to the perception of malevolence. Second, we have crowdsourced a multi-turn malevolent dialogue dataset for malevolent dialogue response detection and classifying (MDRDC), where each turn is labeled using HMDT categories. Last, we have implemented 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 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, i.e., leveraging hierarchical labels. We hope that the efforts made in this chapter help to promote future research on this topic.

The MDRDC dataset has several future applications. First, it is promising to evaluate malevolence of dialogue generation models and moderating malevolent content on the web, e.g., Reddit, based on a malevolence classification model. Second, using paraphrased data can help generate more malevolent data and generate fewer non-malevolent responses for conversational dialogue systems. We aim to study how to avoid generating malevolent responses by applying the classifier to sequence to sequence based response generation models [48]. 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.

We have provided a taxonomy, dataset, and classification model, complemented with experimental results, to answer RQ2. We have constructed a hierarchical malevolent dialogue taxonomy (HMDT) and a high-quality dataset malevolent dialogue response detection and classifying (MDRDC) via crowdsourcing. We also built a benchmark for classifying single-label malevolent dialogue responses. In the next chapter, we take a different angle for classifying multi-label malevolent dialogue responses and study how we can utilize label correlation to improve the performance of multi-label classification.

Appendix

3.A User Study for Validating the HMDT

A user study is used to investigate the users' opinions towards malevolent dialogue responses. The user study design includes a task summary, user profiles, and a questionnaire. The task explanation includes a content warning, an ethical disclaimer, and the task goal. We ask the user to finish the questionnaire according to their own situation and perception. Users are asked to imagine talking with a chatbot that may return responses with certain malevolent categories as defined in Table 3.2. Part of the questionnaire for the users is shown in Figure 3.7. We also explain the 5-point Likert scores.

3.A.1 Task summary

Warning

This task may contain adult content and offensive content. Worker discretion is advised.

Disclaimer

The dialogues are collected from an external website. The views, opinions, and negative words in the dialogues do not necessarily reflect our opinion. Please do not use any expressions from the examples we show in a real-world or online scenario.

Goal

This user study is used to collect your perception about some provided malevolent categories happening in dialogues. We include extra requirement notes for the workers as follows:

(1) Table 3.2 shows the definitions of all malevolent categories with detailed explanations and examples.

(2) Finish the user profiles and questionnaire according to your own situation and perception. Do NOT take public perception into account.

(3) If you do not want to share attributes of your personal profiles, please leave them blank.

3.A.2 User profile

The (attributes and) user profiles that we collected are as follows:

(1) Age.

(2) Gender.

(3) Total years of education.

(4) The frequency of using a chatbot such as Siri, Xiao ice, etc.

3.A.3 Questionnaire interface

The questionnaire instruction and part of the questionnaire interface are as follows:

> Consider you are talking with a chatbot and it returns responses with certain malevolent categories as defined in Table 3.2. Please finish the questionnaire in Figure 3.7 by selecting one of the five scores to reflect your perception. 1: Strongly disagree; 2: Disagree; 3: Neither agree nor disagree; 4: Agree; 5: Strongly agree.

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.
Non-malevolent	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 3.7: Part of the questionnaire interface.

3.B Qualification Test for the Response Annotation Task

The qualification test design includes a task summary and qualification test questions. Specifically, we show the workers implicit utterance and dialogue context examples. Utterance "I ******* hate you" is explicit and "I tell the world: The immortal words of Adolf Hitler" is implicit. Context example is the same as Figure 3.1. The 12 questions are shown in Figure 3.8.

3.B.1 Task summary

Warning

This task may contain adult content and offensive content. Worker discretion is advised.

Disclaimer

The dialogues are collected from an external website. The views, opinions, and negative words in the dialogues do not necessarily reflect our opinion. Please do not use any expressions from the examples we show in a real-world or online scenario.

Goal

This is the qualification test for the response annotation task. The response annotation task is to collect malevolent annotations for each dialogue turn for research that could help avoid generating malevolent responses in dialogue systems. You need to answer 12 questions to get the qualification to attend the response annotation task.

(1) Table 3.2 shows the definitions of all malevolent categories with detailed explanations and examples.

(2) Implicit/Explicit malevolent responses: Explicit responses contain explicitly malevolent words, while implicit responses have more abstract, coded expressions or attitudes without using explicit malevolent words.

Examples:

Explicit: “I ******* hate you (disgust)”.

Implicit: “I tell the world: The immortal words of Adolf Hitler (negative intergroup attitude)”.

(3) Context-aware malevolent responses: Some responses are malevolent only if the context (previous dialogue turns) is taken into consideration.

Examples:

Turn 1: “Drunk driving is illegal”.

Turn 2: “But I think it’s more professional”.

The response in Turn 2 is malevolent (immoral & illegal) when considering the response from Turn 1.

3.B.2 Qualification test questions

The questions of the qualification test are shown in Figure 3.8. The total possible score is 100. For questions 1-4, each question has a score of 15. For questions 5-12, each question has a score of 5.

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

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:
○ Non–malevolent ○ 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
○ Non–malevolent, 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 & illegal
○ Unconcernedness, immoral & 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, non–malevolent
○ Anger, jealousy
○ Anger, phobia
○ Violence, phobia

Figure 3.8: The questions of the qualification test.

3.C Response Annotation Task

The response annotation task design includes a task summary and annotation interface. Specifically, we have annotation examples and quality requirements.

3.C.1 Task summary

Warning

This task may contain adult content and offensive content. Worker discretion is advised.

Disclaimer

The dialogues are collected from an external website. The views, opinions, and negative words in the dialogues do not necessarily reflect our opinion. For the rephrasing part, you are asked to reformulate utterances to keep their semantics and malevolent categories unchanged. These are just used for research, which does not necessarily reflect your views and opinions. Please do not use any expressions from the examples we show in a real-world or online scenario.

Goal

The response annotation task is used to collect malevolent annotations for each dialogue turn for research that could help avoid generating malevolent responses in dialogue systems. You need to label the dialogue responses according to the given malevolent categories.

Requirements for annotation quality

The requirements to control the annotation quality are as follows:

(1) Read the definitions of all malevolent categories with detailed explanations and examples in Table 3.2.

(2) Label each turn of the provided dialogue according to the given malevolent categories.

(3) Rephrase at least one malevolent utterance in each dialogue.

(4) During rephrasing please do not change the annotation category.

(5) We might limit your qualification for the task if we find the following behaviors: pasting irrelevant content from the website, copying dialogue, rephrasing with repeating words, and rephrasing with random words.

Example

Dialogue:
Turn 1: Drunk driving is illegal.
Turn 2: But I think it's more professional.
Turn 3: Hey, my boy, we need to be careful when driving.
Annotations:
Turn 1: Non-malevolent
Turn 2: Endorse immoral or illegal behavior
Turn 3: Non-malevolent
Rephrase malevolent utterance:
Turn 2: I think drunk driving is nice since it's more professional.

3.C.2 Annotation interface

The workers are asked to read the given dialogue on the left, label each turn and rephrase at least one of the malevolent responses if any. The instructions are shown in Figure 3.9 and the annotation interface is shown in Figure 3.10.

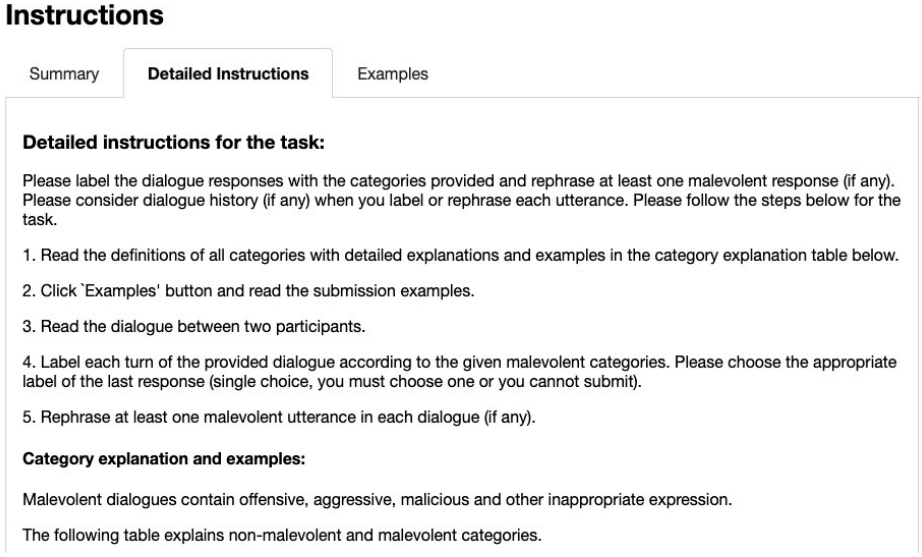


Figure 3.9: The instruction interface of the response annotation task.

Figure 3.10: The annotation interface of the response annotation task.

4

Improving Multi-label Malevolence Detection and Classification in Dialogues

In this chapter, we address RQ3: How can we build a model for multi-label dialogue malevolence detection based on single-label training data and construct a validated dataset to assess the model?

4.1 Introduction

Safety is an increasingly important aspect of artificial intelligence development [3, 133, 147]. When it comes to dialogue agents, taking measures to avoid the risks of generating undesirable and harmful responses may have a profound positive impact on the adoption of conversational technology [168]. Research on safe dialogue agents concerns aspects such as inaccurate information [57], fairness [94], and unauthorized expertise [147]. Malevolence is another key aspect [184, 185], i.e., whether the dialogue utterance contains malevolent content that is related to offensiveness [38], toxicity [50], ad hominem [143], or toxicity agreement [8], etc.

There have been several studies targeting malevolence detection [135, 138, 184, 185]. We build on the work of Zhang et al. [185], which is included as Chapter 3 in the thesis. There, we introduce the malevolent dialogue response detection and classification task, present a hierarchical malevolent dialogue taxonomy, create a labeled multi-turn dialogue dataset, and apply state-of-the-art text classification methods to the task. One important limitation of the work in Chapter 3 is that we only explore single-label dialogue malevolence detection (SDMD), i.e., we 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 4.1, 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, a phenomenon to which we refer as *label correlation in taxonomy* (LCT).

In Chapter 3 we have proposed the hierarchical malevolent dialogue taxonomy (HMDT) that classifies correlated malevolence labels into the same group by inves-

This chapter was published as [186].

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

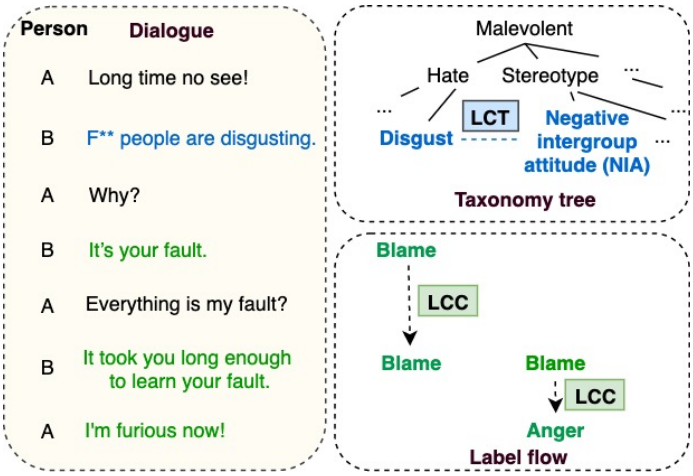


Figure 4.1: Label correlation in taxonomy (LCT) and label correlation in context (LCC). In terms of LCT, "NIAs" is correlated with "disgust", which can be reflected by the utterance in blue (LCT). In different turns, "blame" is likely to co-occur with "anger" and "blame", which can be reflected by the utterances in green (LCC).

tigating 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 transition. For instance, in Figure 4.1, "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). In Chapter 3 we do not take correlations of malevolence labels in different dialogue turns into account and our label-correlation mechanisms are different from previous methods that require multi-label training sets [79, 151].

In this chapter, 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 validation and test 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 fields (CRFs), 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, i.e., MDMD, based on the previously released malevolent dialogue response detection and classifying (MDRDC) dataset, conduct experiments on MDMD, 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 in this chapter 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.

4.2 Related Work

4.2.1 Malevolence detection taxonomies

The taxonomies for hate speech, aggressiveness, offensiveness, and condescending only contain a few categories [78, 163, 166, 176], which lack a unified understanding of what constitutes malevolence. To address this gap, Sheng et al. [143] introduce a two-level ad hominem taxonomy and Sun et al. [147] introduce a safety taxonomy, both of which contain seven different aspects. Furthermore, in Chapter 3 we define a three-level malevolence taxonomy that contains eighteen categories in total. In this chapter, we follow the taxonomy proposed in Chapter 3.

4.2.2 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) [166], hate speech detection dataset (HSDD) [33], and multilingual detection of hate speech (MDHS) [9], which are all collected from Twitter. These datasets have several shortcomings, e.g., a lack of diversity, small data size, low inter-annotator agreement, and small lexicon size. Others work on aggressiveness, offensiveness, and condescending, e.g., trolling, aggression and cyberbullying (TRAC) [78], offensive language identification dataset (OLID) [176], and TALKDOWN [163], 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.

Furthermore, none of the datasets listed above is in the form of multi-turn dialogues. To address this, recent studies have released the TOXICHAT [8], ADHOMINTWEETS [143], MDRDC [185], and DIASAFETY datasets [147], 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.

4.2.3 Malevolence detection methods

Methods for malevolence detection include rule-based [135], traditional machine learning-based [9, 33, 138, 166], and deep learning-based [78, 143, 163, 176, 185] approaches. Roussinov and Robles-Flores [135] define malevolence by filtering the keywords. Saral et al. [138] survey the machine learning-based detection methods, including k-nearest neighbors (KNN) and support vector machine (SVM)-based methods. The performance of these methods is not strong enough as malevolence detection requires a deep understanding of semantics. Kumar et al. [78] apply convolutional neural networks (CNNs) and long short-term memorys (LSTMs) for aggressiveness detection. Zampieri et al. [176] apply CNNs and Bi-LSTMs for offensiveness detection. More recently, pretrained models, e.g., bidirectional encoder representations from transformers (BERT) and RoBERTa, have been used for ad hominem, malevolence, and safety [143, 147, 185], demonstrating better performance than LSTM, CNN, recurrent convolutional neural network (RCNN), and graph neural network (GNN)-based models [185].

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).

4.3 Method

4.3.1 Overall

Let x be a dialogue that contains m utterances, $x = [x_1, x_2, \ldots, x_i, \ldots, x_m]$ and let x_i be 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 4.2, 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.

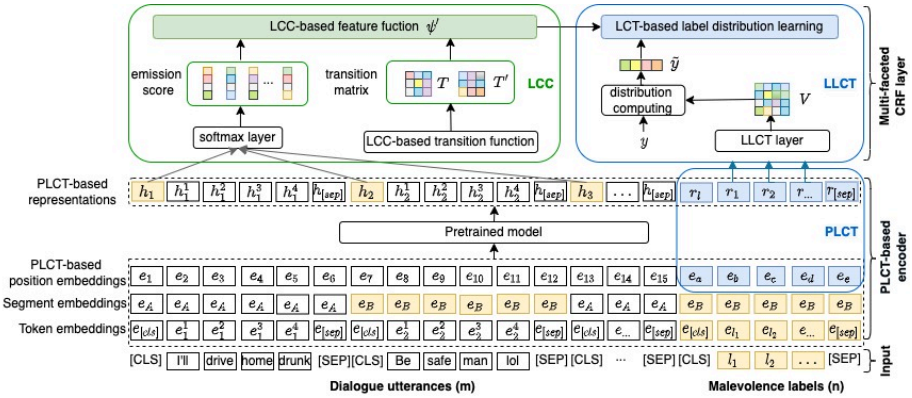


Figure 4.2: Framework of the proposed *multi-faceted label correlation enhanced CRF* (MCRF) model.

.

4.3.2 Utterance and label encoder

As shown in Figure 4.2, 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 [99], 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 (i) token embeddings, (ii) segment embeddings, and (iii) position embeddings. Token embeddings follow the original transformer paper [154]. 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 [162]. In order to improve the representation of labels, we change the position embeddings of labels into PLCT-based position embeddings (see Section 4.3.3). We feed the three embeddings into a pretrained model (i.e., BERT) to get the representations of utterances and labels:

$$\begin{aligned} H, R &= PTM([e(x_i), e(l_j)]), \\ e &= e_{tok} + e_{seg} + e_{pos}, \end{aligned} \tag{4.1}$$

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 representations of the utterances with h_i (corresponding to pooled 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.

4.3.3 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 [164]. 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 4.3, the taxonomy tree-based positional label embedding for "blame" is $(1, 2, 0)$. We use the label position output of PLCT to improve e_{pos} in Eq. 4.1, and the improved encoder is referred to as *PLCT-based encoder*.

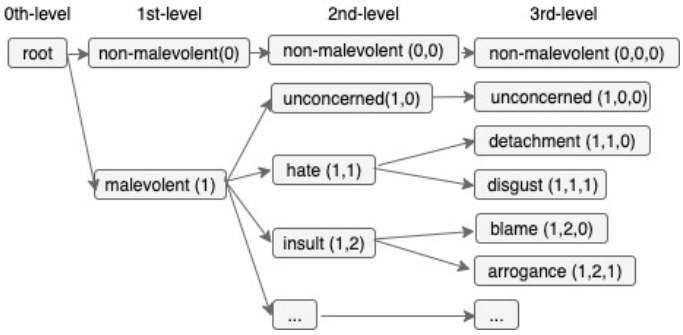


Figure 4.3: Demonstration of taxonomy tree of labels.

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 as follows:

$$V = \frac{1}{2}(\hat{V}_{j,j'} + V'_{j,j'}), \tag{4.2}$$

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 the PLCT-based encoder with taxonomy tree position, i.e., R from Eq. 4.1; 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 Section 4.3.4.

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:

$$\begin{aligned} t(y_{i-1} = l_j, y_i = l_{j'}) &= T_{(l_j, l_{j'})}, \\ t'(y_{i-2} = l_j, y_i = l_{j'}) &= T'_{(l_j, l_{j'})}, \end{aligned} \tag{4.3}$$

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.

4.3.4 Multi-faceted conditional random field layer

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

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_i \psi(x_i, y_i)\right), \tag{4.4}$$

where Z is a normalization function, and

$$\psi(x, y) = \sum_i s(y_i, x) + \sum_i t(y_{i-1}, y_i), \tag{4.5}$$

where t is defined in Eq. 4.3. 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:

$$s(y_i, x) = \text{softmax}(h_i), \tag{4.6}$$

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

$$\psi'(x, y) = \frac{1}{2}\left(\psi(x, y) + \sum_i s(y_i, x) + \sum_i t'(y_{i-2}, y_i)\right), \tag{4.7}$$

where t', ψ and s and are defined in Eq. 4.3, 4.5 and 4.6, 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:

$$\tilde{y_i} = \lambda V y_i + y_i, \tag{4.8}$$

where λ denotes how much the original one-hot distribution is redefined and V is the matrix that estimates the LCT in Eq. 4.2.

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

$$\mathcal{L} = \sum_{y} q(y|x) \log \frac{q(y|x)}{p(y|x)}, \tag{4.9}$$

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 $p(y|x)$ [96]:

$$\mathcal{L} = \sum_{i} \sum_{y_i} \{q(y_i|x) \log q(y_i|x)\} - \sum_{y} \{q(y|x)\psi'(y,x)\} + \log Z(x), \tag{4.10}$$

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

4.4 Experimental Setup

We conduct experiments to answer the following research questions:

(RQ3.1) How does BERT-MCRF compare to baselines on the MDMD test set?

(RQ3.2) What is the impact of the number of labels on the performance of BERT-MCRF?

(RQ3.3) What is the influence of different LCT and LCC settings?

(RQ3.4) What do the components of BERT-MCRF contribute to its overall performance?

4.4.1 Dataset

We conduct experiments on an extension of the MDRDC dataset released with [185] (and included as Chapter 3 in the thesis). 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 [185]. We select workers with a test score of at least 90, 500 approved human intelligence tasks (HITs), 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

Table 4.1: Statistics of the validation and test sets of MDMD.

	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

labels in the taxonomy of Chapter 3. 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 [106].

The MDMD dataset statistics are shown in Table 4.1. 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 the 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 the MDMD task considering the percentage of multi-label utterances. We use the training, validation, and test splits provided in Chapter 3, which has a ratio of 7:1:2.

4.4.2 Baselines

We compare BERT-MCRF against BERT and BERT-CRF. The two baselines are competitive since BERT with a softmax classifier performs well in the SDMD task, as reported in Chapter 3, and BERT-CRF with a modified encoder for separate sentences is the state-of-the-art model for sequence labeling tasks [22].

4.4.3 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 4.C. All the neural models are trained on GeForce GTX TitanX GPUs.

4.4.4 Evaluation metrics

We use the precision, recall, F1 score, and Jaccard score as our evaluation metrics [103]. We report the macro scores since the data is imbalanced in terms of labels [185].

4.5 Results and Analysis

4.5.1 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 4.2.

Table 4.2: Main results of MCRF on the MDMD test set.

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**

In terms of overall performance, adding 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%, 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.

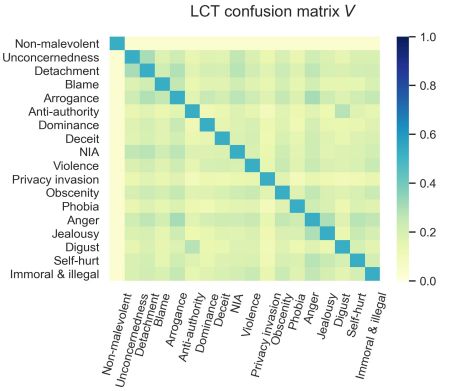
4.5.2 Performance of different label groups

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 4.3.

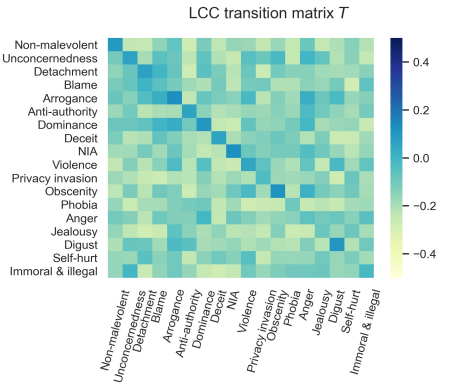
Table 4.3: Jaccard scores of different label groups.

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

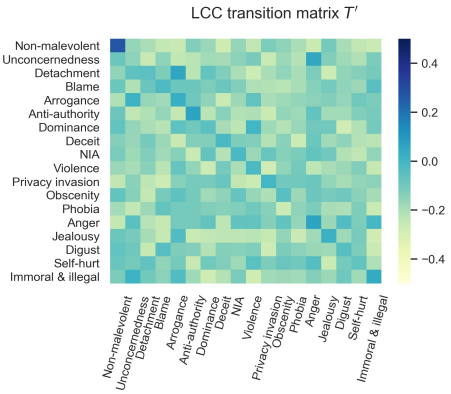
First, the results suggest that BERT-MCRF has better performance with regard to different label groups. BERT-MCRF's Jaccard scores for the 1-label, 2-label, and 3-label are 5.4%, 16.6%, 8.3% higher than the second-best performing approach. An exception is that for the 4-label group, the result of BERT-MCRF 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 4.3 decrease as the number of labels increases.



(a) LCT matrix V.



(b) LCC matrix T.



(c) LCC matrix T'.

Figure 4.4: Visualization of LCT and LCC.

4.5.3 Influence of the label correlation in taxonomy and label correlation in context settings

First, we study the influence of the hyperparameter λ of LCT in Eq. 4.8, as shown in the upper part of Table 4.4. 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. 4.8) in Figure 4.4a. 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 the final results.

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 4.4. For the three LCC settings, T has better recall thus improving the final performance compared with T'; T' has better precision than the

Table 4.4: BERT-MCRF performance w.r.t. different LCT and LCC settings. λ is the hyperparameter in Eq. 4.8, T and T' are the transition matrices by Eq. 4.3.

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**

other two groups, but the 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 4.4b and T' in Figure 4.4c; 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.

4.5.4 Ablation study

We perform an ablation study on BERT-MCRF by removing LCT or LCC. The results are reported in Table 4.5. 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 Section 4.5.3.

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. 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 [181].

Table 4.5: 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'.

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

However, for MDMD, there is no label co-occurrence information; our results suggest that LCT is able to increase the MDMD performance; the reason might be that the LCT reduces overfitting of single-label training data.

4.5.5 Case study

We randomly select two examples from the test set to illustrate the performance of BERT, BERT-CRF, and BERT-MCRF (see Table 4.7 in Appendix 4.B.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.

4.6 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.

Finally, in this chapter we have answered RQ3 by providing a multi-faceted label correlation enhanced model solution for classifying multi-label malevolent dialogue responses, the MDMD dataset, and the experimental results above.

In Chapters 2 and 3, and the present chapter we have established the malevolence problem and built malevolence detection models. The malevolent detection models provide a basis for human-machine collaborative dialogue malevolence evaluation. In the next chapter, we focus on malevolent dialogue response evaluation and study how to balance the reliability and effort of evaluation by human-machine collaborative mechanisms.

Appendix

We present additional details on our experimental design in the appendices below. We include the ethical considerations (Appendix 4.A); the validation performance of BERT-MCRF for the main results reported in this chapter (Appendix 4.B.1); a case study (Appendix 4.B.2); a description of our source code (Appendix 4.B.3); a summary of the average runtime of each module and detailed information about the parameters (Appendix 4.C); and further details about the newly created dataset that we release with this thesis (Appendix 4.D).

4.A 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 University of Amsterdam. The data will be made available to researchers that agree to the ethical regulations of our ethics committee. Characteristics and quality control of the re-annotated dataset are described in Section 4.5.

The claims in this chapter match the results and the model can be generalized to multi-label dialogue safety detection tasks. This chapter 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 the malevolence label counts of each utterance or the perception of different labels.

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

4.B Experimental Results

4.B.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 4.6. 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 as shown in Table 4.2 (presented in Section 4.5) and Table 4.6.

Table 4.6: Main results of BERT-MCRF on the multi-label validation set.

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 4.7: Case study. Top: utterances and labels of example 1. Bottom: utterances and labels of example 2.

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.

4.B.2 Case study examples

We show two examples for case study that explains how LCT and LCC work, as shown in Table 4.7; the description is in Section 4.5 of the main content of Chapter 4.

4.B.3 Code

Our code is uploaded to `https://github.com/repozhang/MCRF`.

4.C Runtime and Parameters

In terms of average runtime, the time cost for our BERT-MCRF model is acceptable. The time cost 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 Section 4.4.3, 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 the 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 the 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. 4.8, we use the value range $[0, 0.5, 1, 2]$ and select the best result. In terms V' in Eq. 4.2, 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, and 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 as 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 the CRF layer.

4.D Dataset

Our data is uploaded to `https://github.com/repozhang/malevolent_dialogue`. The statistics and splits are described in Section 4.4.1. The language of the dataset is English. For data prepossessing, we use all the data from the dataset.

In terms of the data collection process, we follow previous research [185], except that the workers are asked to choose multiple choices from the labels.

5 A Human-machine Collaborative Malevolence Evaluation Framework

In this chapter, we address RQ4: How can we build a framework for evaluating malevolent dialogue responses with reliability and human effort balanced?

5.1 Introduction

Conversational dialogue systems (CDSs) are often trained to generate responses given unstructured, open-domain dialogues. Evaluation of CDS responses has drawn broad attention due to its crucial rule for CDS development [35]. Broadly speaking, there are two approaches to perform dialogue evaluation: *automatic* evaluation and *human* judgements [45]. Automatic evaluation metrics such as appropriateness [101], engagement [183], are efficient but have low agreement with human judgements due to the diversity of responses [93], especially for word-overlap based metrics, such as BLEU [115] and ROUGE [91].

More recently, training-based methods, e.g., automatic dialogue evaluation model (ADEM) [101], unreferenced metric blended evaluation routine (RUBER) [149] and contextualized methods, e.g., BERT-based RUBER [51], have been shown to have better agreement with human judgements. However, these methods are still not reliable enough: the Pearson correlation with human judgments is 0.44 for appropriateness [101] and 0.55 for relevance [51]. To guarantee reliability of evaluation outcomes, our current best practice is to use human judgements. In terms of most evaluation aspects, e.g., appropriateness [171], coherence [124], and empathy [125], human judgements simply show the highest reliability. Obviously, human judgments are more labor-intensive than automatic evaluation [35].

The flaws of automatic evaluation and the lack of speed and scalability of human evaluation limit the speed at which the community can develop more intelligent CDSs. For example, as part of the daily research and development cycle of CDSs, we need to change the model design and retrain the model multiple times, on a daily or even hourly basis. Even if there is a minor change, we need to verify its performance again each time. For another example, CDS leaderboards have become very popular recently

This chapter was published as [184].

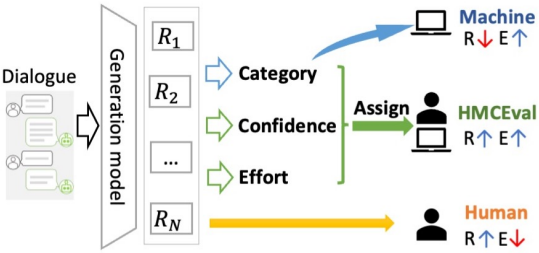


Figure 5.1: Human-machine collaborative evaluation (HMCEval) framework. $R_1, \ldots, R_N$ are the generated response samples to be evaluated. R and E are reliability and efficiency, respectively.

as a means to provide platforms for fair comparison [67]. There are usually dozens of models to evaluate, and new models are introduced every day. Practical scenarios like the above two call for dialogue evaluation methods that are both reliable and efficient.

In this chapter, we propose the *human-machine collaborative evaluation* (HMCEval) framework for dialogue evaluation with the aim of balancing reliability and efficiency. HMCEval formulates the dialogue evaluation task as a sample assignment problem, i.e., if the machine can provide accurate outcomes, most evaluation samples should be assigned to the machine; otherwise, we should assign more samples to human evaluators. As shown in Figure 5.1, automatic evaluation has low reliability although the efficiency is high; human judgement has high reliability but it is labor-intensive; HMCEval beats the previous two methods in balancing reliability and efficiency. Finding a good balance between reliability and efficiency is non-trivial as the two desiderata are often in conflict with each other. Usually, reliability is improved at the expense of efficiency [16].

There are three main modules in the proposed *human-machine collaborative evaluation* (HMCEval) framework, namely the *model confidence estimation* (MCE) module, the *human effort estimation* (HEE) module, and the *sample assignment execution* (SAE) module. First, the MCE module measures the confidence of predicted evaluation for each dialogue response-based sample. Our implementation of MCE is based on three estimation methods, namely, BERT-based maximum class probability (MCP), trust score (TS) [70], and true class probability (TCP) [25]. TS and TCP have originally been introduced for images; we add a bidirectional encoder representations from transformers (BERT) layer to expand it to dialogues. Second, the HEE module estimates the effort. Our implementation is based on annotation time cost prediction by dialogue-related and worker-related features. Third, the SAE module decides whether a dialogue response sample should be assigned to a human or a machine for evaluation by maximizing the confidence and minimizing the (human) effort. We implement the module by integer linear programming (ILP).

We demonstrate the effectiveness of HMCEval on dialogue malevolence evaluation [185]. The main reason we choose this particular task is that dialogue malevolence is highly related to social good [144, 168], which is of vital importance for CDSs, but it is hard to evaluate because of the need for deep semantic understanding [31]. We carry out experiments on the malevolent dialogue response detection and classify-

ing (MDRDC) dataset, which was published in [185] and which has been introduced in Chapter 3.

Our results show that the proposed HMCEval framework significantly surpasses machine evaluation and human judgement in terms of balancing reliability and effort. HMCEval achieves around 99% evaluation accuracy (compared to human evaluation) with as much as half of the human effort saved. The results demonstrate that HMCEval can be used for reliable and efficient evaluation of CDSs since the accuracy is high and the effort is significantly reduced compared to fully human evaluation.

5.2 Related Work

5.2.1 Evaluation of CDSs

Automatic evaluation for CDSs includes untrained methods and learning-based methods. Early untrained methods, such as perplexity [19], and the quality metrics bilingual evaluation understudy (BLEU) [115] and recall-oriented understudy for gisting evaluation (ROUGE) [91] are widely used for CDS, but the aspects they evaluate are limited. Recent work based on word embeddings covers more aspects, such as distinct-n for diversity [85] or average word embedding similarity for coherence [102]. Most untrained methods have a low agreement with human judgements [93] because machine responses are highly diversified, although a few metrics have sufficient agreement with human judgements, i.e., a Pearson correlation of 0.69 for coherence [102].

To address the problem of low agreement with human judgments, learning-based methods have been developed [112, 149]. Lowe et al. [101] propose ADEM to evaluate the appropriateness of responses. Tao et al. [149] propose RUBER, which shows better agreement with human judgments than ADEM. RUBER is designed for relevance and similarity by blending relevance between the generated response with human ground truth and context. Several methods utilize pretrained language models such as BERT for automatic evaluation. Ghazarian et al. [51] propose contextualized RUBER, which outperforms RUBER. Similarly, a predictive engagement metric is built by utilizing user engagement score [52]; quality is evaluated by transformer-based language models without reference response [110]. The above methods cover more aspects and integrate linguistic features [149], thus the agreement with human judgement is higher than most word-overlap based methods. However, for most of the metrics, the model performance still has space to improve, for instance, the accuracy of engagement is 0.76 [52]. Our proposed HMCEval framework could be applied to these metrics and improve general evaluation reliability with an acceptable amount of human effort.

Human judgement is applied in common evaluation aspects including fluency, consistency, relevance, appropriateness, coherence, and quality for CDSs [45]. It is reliable, yet expensive and time-intensive, especially for large-scale evaluation [67]. In order to guarantee the reliability, agreement among different workers is needed, which makes the high effort problem more severe [31].

Unlike the methods listed above, the HMCEval framework specifically aims to balance reliability and human effort for the evaluation of CDSs.

5.2.2 Human-machine collaboration

Human-machine collaboration hybridizes machine prediction and human judgements. Previous research mostly focuses on using human judgments to help label the low reliability samples [13, 49, 80]. Earlier research gives humans the output of an automatic model and lets them decide whether the model prediction is reliable [82]. However, people tend to ignore the predictions of a model if it makes mistakes [37] since they are not tolerant to model mistakes. In such cases, predictive results are not fully utilized and human effort increases. At the same time, there is a possibility that human annotators mistakenly follow the outputs of a model with errors [28]. Both situations lead to the failure of human-machine collaboration.

The core problem is to determine when a human annotator should trust a model. Confidence estimation for a model's prediction has been proposed to help improve the overall accuracy, correctness, etc. of human-machine collaboration. Callaghan et al. [13] develop a hybrid cardiogram classification human-machine collaborative (HMC) framework, which achieves better performance than a classifier by itself and uses less expert resources compared to expert classification by itself. Kyono et al. [80] develop a Man and Machine Mammography Oracle that improves overall breast cancer diagnostic accuracy while reducing the number of radiologist readings. Gates et al. [49] use Abstrackr, a HMC screening method to screen relevant titles and abstracts for paper reviews, which could save the time of reviewers and have little risk of missing relevant records. However, the above methods select the top-k most unreliable samples and do not consider the division of effort between human and machine. Chaganty et al. [16] are the first to combine machine and human evaluation to obtain a reliable estimate at a lower cost than human alone on summarizing and open-source question answering, with a cost reduction of only 7–13%. Ravindranath et al. [126] build a highly cost-efficient face recognition HMC framework that outperforms both a machine-based method and a fully manual method, with both reliability and effort considered. Nevertheless, the methods introduced previously are not suitable for HMC evaluation for dialogue as they focus on non-dialogue tasks, low cost reduction, or do not consider both reliability and effort.

Our proposed framework is purpose-built for dialogue evaluation. It leverages both human judgement and machine prediction by assigning low confidence machine-generated samples to human workers while minimizing overall human effort.

5.3 Methodology

5.3.1 Overview

Suppose we have a set of M samples $\{(C_i, \hat{x}_i)\}_{i=1}^{M}$ to be evaluated. Here, C_i is the dialogue context and $\hat{x}_i$ is a response generated by a CDS model $f_g(C) \rightarrow \hat{x}$. Below, we propose a method to achieve a reliable and efficient evaluation of the M samples under the constraint that a human can annotate at most $N \ll M$ samples. We propose the *human-machine collaborative evaluation* (HMCEval)[1] framework to solve this

[1] https://github.com/repozhang/CaSE_HMCEval

task. HMCEval is divided into three modules: (i) sample assignment execution (SAE), (ii) model confidence estimation (MCE), and (iii) human effort estimation (HEE).

5.3.2 SAE module

The optimization problem of assigning M samples to a human or machine can be solved by tractable integer linear programming (ILP), which is NP-complete [114]. First, we introduce the decision variable z_i to denote the sample assignment to a human or machine:

$$z_i = \begin{cases} 0, & \text{sample } i \text{ is assigned to a human;} \\ 1, & \text{sample } i \text{ is assigned to machine.} \end{cases} \tag{5.1}$$

Second, we define two ILP objectives that try to maximize the overall confidence and minimize the overall effort, respectively:

$$\begin{aligned} &\max \sum_{i=1}^{M} \hat{a}_i z_i + \sum_{i=1}^{M} b_i(1 - z_i), \\ &\min \sum_{i=1}^{M} k_i z_i + \sum_{i=1}^{M} \hat{l}_i(1 - z_i), \end{aligned} \tag{5.2}$$

where (i) M is the total number of samples to evaluate generated by the generation model $f_g(C) \rightarrow \hat{x}$; (ii) $\hat{a}_i \in [0, 1]$ is the model confidence for evaluating sample i; (iii) b_i is the human confidence for evaluating sample i; (iv) k_i is the machine effort for evaluating sample i; and (v) $\hat{l}_i \in [0, 1]$ is the human effort for evaluating sample i.

We use the weighted sum method [104] to solve Eq. 5.2 so as to get the optimal z_i. The objective function in Eq. 5.2 is transformed into:

$$\max \left[\sum_{i=1}^{M} \hat{a}_i z_i + \sum_{i=1}^{M} b_i(1 - z_i) - \lambda \left(\sum_{i=1}^{M} k_i z_i + \sum_{i=1}^{M} \hat{l}_i(1 - z_i) \right) \right], \tag{5.3}$$

subject to

$$\begin{aligned} &\sum_{i=1}^{M} z_i \geq M - N, \\ &b_i = 1 \text{ for } i = 1, \ldots, M, \\ &k_i = 0 \text{ for } i = 1, \ldots, M, \\ &\lambda \geq 0. \end{aligned} \tag{5.4}$$

The constraints are motivated as follows: (i) the number of samples assigned to a human is less than or equal to N; (ii) human confidence is assumed to be 1; (iii) machine effort is assumed to be 0; and (iv) λ is greater than 0. N and λ are two parameters that we use to balance reliability and effort; λ is a trade-off parameter that controls the contribution of two objectives to the overall objective, as shown in Eq. 5.3; and N controls the total samples assigned to a human. As N gets larger or λ gets smaller, the overall evaluation is more reliable but needs more human effort. As N gets smaller or λ gets larger, the overall evaluation costs less human effort but gets less reliability.

5.3.3 MCE module

Given a machine evaluation model (usually a classification model [34]) $f_c(C, \hat{x}) \rightarrow \hat{y}$, where $\hat{y}$ is the evaluation result (usually a category, e.g., malevolence or non-malevolence), the MCE module aims to recognize how confident the evaluation $\hat{y}$ is. In this chapter, we investigate three confidence estimation methods, namely (i) maximum class probability (MCP), (ii) trust score (TS), and (iii) true class probability (TCP).

MCP is a basic method that directly uses the classification probabilities to measure confidence. Based on the dataset $\{(C_j^{'}, x_j), y_j\}_{j=1}^{Q}$, we build a BERT-based classifier as a machine evaluation model f_c. MCP is the softmax probability of the evaluation result $\hat{y}$. Formally,

$$\mathrm{MCP}(C^{'}, x) = P(Y = \hat{y}|w, C^{'}, x). \tag{5.5}$$

Next, TS is a confidence measurement that estimates whether the predicted category of a test sample by a classifier can be trusted. It is calculated as the ratio between the Hausdorff distance from the sample to the non-predicted and the predicted categories [70]. First, the training data is processed to find k-nearest neighbors (KNN) radius based α-high-density-set $\hat{H}(\tilde{C}_{train}^{'}, \tilde{x}_{train})$, where $\{\tilde{C}_{train}^{'}, \tilde{x}_{train}\}$ is the output of feeding training samples $\{(C_{train}^{'}, x_{train})\}$ into the BERT layer of f_c. This part is different from the original TS work designed for images [172]. Then, for a given test sample, we predict the ratio of distances, which is the TS value. Formally,

$$\hat{a} = d(C_j^{'}, x_j, \hat{H}_1)/d(C_j^{'}, x_j, \hat{H}_2), \tag{5.6}$$

where $\hat{H}_1$ is the high density set of the non-predicted category, $\hat{H}_2$ is the high density set of the predicted category. The estimated TS is normalized within 0 and 1 by min-max normalization.

As for TCP, the estimation is obtained by a learning-based method. Similar to TS, the original confidence network for TCP estimation is also built for images [25]. We expand it into a BERT-based confidence network for CDSs. The TCP estimation part f_{conf} is based on the BERT-classifier f_c. Formally, $f_{conf}(C, \hat{x}, f_c, f_g) \rightarrow \hat{a} \in [0, 1]$, where f_g is the generation model. We pass the features from the BERT layer of f_c and feed them into a confidence network implemented by a succession of dense layers with a sigmoid activation to get the confidence scalar. We define an MSE loss for TCP:

$$L_{conf} = \frac{1}{Q}\sum_{i=1}^{Q}(\hat{a}(C_i^{'}, x_i, \theta) - a^*(C_i^{'}, x_i, y_i^*))^2, \tag{5.7}$$

where $a^*(C_i^{'}, x_i, y_i^*)$ is the target confidence value. During inference, the ground truth TCP score is calculated based on the BERT-based classifier: $\mathrm{TCP}(C^{'}, x, y^*) = P(Y = y^*|w, C^{'}, x)$, where y^* is the true category.

5.3.4 HEE module

The HEE module is designed for estimating the human effort $\hat{e}$. In this chapter, we use time cost, i.e., the time spent for each annotation, to represent human effort. We

implement the time cost estimation model f_l with random forest regression [90]: $f_l(h(C, \hat{x})) \rightarrow \hat{l} \in [0, 1]$, h is the feature extraction function.

There are two groups of features, namely dialogue-related features and worker-related features; see Table 5.5. The dialogue related features are: (i) "total turns": total number of turns in a dialogue; (ii) "malevolent turns": total number of malevolent turns in a dialogue; for prediction, we use the BERT-classifier results; (iii) "non-malevolent turns": total number of non-malevolent turns in a dialogue; for prediction, we use the BERT-classifier results. (iv) "first submission or not": if this is the first time the worker does this task, the value is 1, else 0; (v) "paraphrased turns": some turns are paraphrased; we calculate the total number of such turns; (vi) "total length": total number of tokens in the dialogue; (vii) "Flesch-Kincaid (FK) score": the result of a readability test, based on [76]; (viii) "Dale–Chall (DC) score": the result of a readability test, based on [30]; (ix) "contains malevolent turn or not": if the dialogue contains a malevolent turn, the value is 1, else 0; and (x) "perplexity score": we use BERT as a language model to calculate the perplexity [47]. The worker related features are: (i) "worker test score": this is based on a test designed to test workers" ability to annotate the dialogue according to the gold standard annotation [185]; and (ii) "approval rate ranking": we rank workers by their lifetime approval rate in ascending order, and use the index; lower approval rate workers (i.e., with a smaller index) usually spend less time on annotations.

To train the time cost estimation model f_l, we need the annotation time spent on each response. However, for each individual response, the time spent is relatively short; as a consequence, the influence of noise such as attention, and click time, may be relatively large and make the data unreliable as training data. Therefore, we use the annotation time spent on each dialogue instead of each response as the time cost target, and it is normalized within 0 and 1 using min-max normalization. For the SAE module and effort assessment, we use the average time per turn of each dialogue as the time cost $\hat{l}$ for each response. In addition, there are multiple human annotator submissions for inter-annotator agreement; we filter out the data points that disagree with the agreed annotation; then we choose the data point with a higher annotator test score; if the test scores are the same, we randomly choose one.

5.4 Experimental Setup

5.4.1 Dataset

We carry out experiments on the MDRDC dataset which was originally built for malevolent dialogue detection and classification [185]. As described in Section 3.4, the dataset consists of 6,000 dialogues, with 21,081 non-malevolent utterances and 10,299 malevolent utterances. The dataset also includes Amazon MTurk information, e.g., the time spent on each annotation. We follow the original paper to split the dataset into train, validation, and test with a ratio of 7:1:2.

5.4.2 Implementation details

In terms of the responses by the generation model f_g, in our implementation, we use the original responses by a human for evaluation. The MCE module is implemented by

a BERT-based classifier and a BERT-based confidence network. First, for the BERT-based classifier, we add a softmax layer on top of the "[CLS]" token. It is fine-tuned with 4 epochs since it is already pretrained on a large dataset. The vocabulary size is 30,522. Dialogue context and the current response are concatenated with the "[SEP]" delimiter. We consider the previous three dialogue utterances (if any) as context. We set the max sequence length to 128, the batch size to 64, the dropout ratio to 0.1, and the learning rate is 5e-5. Second, the BERT-based confidence network is attached to a BERT-classifier. It is composed of 5 dense layers, following previous work [25]. As for max sequence length, batch size, dropout ratio, and learning rate, these are the same as for the classifier. The confidence network is trained with a maximum of 30 epochs, with early stopping if the validation loss does not improve for 10 epochs.

The HEE module is implemented by a random forest regression model; the max number of estimators in this study is 100; only the features related to time cost are selected for annotation time cost prediction, with a maximum feature size of 10.

We use the python mixed-integer linear program (MIP) package to implement ILP for the SAE module[2] with the Coin-or branch-and-cut solver [107]. The search stops when it reaches a feasible solution. All the neural models are trained on GeForce GTX TitanX GPUs.

5.4.3 Metrics

We use reliability metrics and effort metrics to assess overall performance. The reliability metrics are precision, recall, F1-score, and accuracy. We calculate the macro score of precision, recall, and F1 as the categories are imbalanced [66]. The effort metrics include human ratio and time cost. Human ratio is the ratio of samples assigned to a human. Time cost is the total time required for a human to annotate the samples. We use area under the curve (AUC), and top-k accuracy to assess the different MCE implementations [113]. We rank the confidence in descending order and calculate the accuracy at top-50%. Top-50% accuracy measures how well the MCE predictions work for the top-50% most confident samples. We use mean square error (MSE), rooted mean square error (RMSE), mean absolute error (MAE) and R^2 to assess the HEE module. MSE, RMSE, MAE are calculated between the predicted time cost and real time cost. We also use the Pearson and Spearman correlation scores to analyze the correlation between features and real-time cost.

5.5 Results and Analysis

5.5.1 Reliability and efficiency

To determine how HMCEval compares to human evaluation and machine evaluation in balancing reliability and efficiency, we report the results in Table 5.1. HMCEval outperforms both human and machine evaluation in balancing reliability and efficiency. More importantly, HMCEval, with half of the human effort spared, achieves reliability that is close to human reliability. First, compared to human evaluation, HMCEval arrives

[2] https://python-mip.com

Table 5.1: Reliability and efficiency of HMCEval w.r.t. human and machine evaluation ($N/M = 0.5$).

Metric	Machine	Human	HMCEval
Reliability			
Precision	0.818	1	0.983
Recall	0.803	1	0.976
F1-score	0.810	1	0.980
Accuracy	0.862	1	0.985
Efficiency			
Human ratio	0	1	0.500
Time cost	0	1	0.500

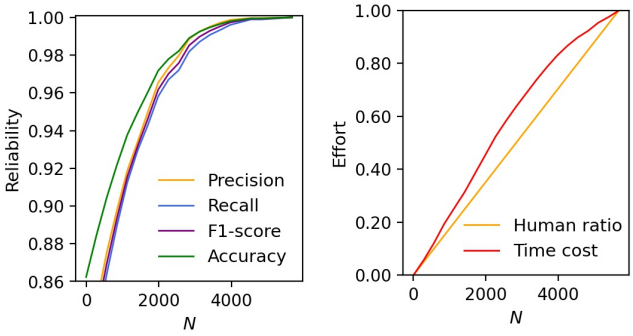
at 98.5% of human accuracy but the human effort decreases by 50.0%. This means that HMCEval is much more efficient than human evaluation, while the reliability is close to human. Second, compared to machine evaluation, the precision, recall, F1-score and accuracy of HMCEval increase by 20.2%, 21.5%, 21.0%, and 14.3%, respectively. This means that HMCEval has higher reliability than machine evaluation. In sum, therefore, HMCEval surpasses both human and machine evaluation in balancing reliability and efficiency.

5.5.2 Influence of N and λ

To investigate how N and λ, two parameters for the SAE module that balance the reliability and effort, influence the performance of HMCEval, we first fix λ and vary N/M from 0 to 1 with a step size of 0.05, where M is the total number of samples to evaluate. Then, we fix N and vary λ from 0 to 45 with a step size of 0.1. The results are shown in Figure 5.2 and 5.3.

Influence of N

Generally, as N increases, HMCEval has better reliability, nevertheless the human effort increases. From Figure 5.2, we can see that when λ is fixed, as N gets larger, the precision, recall, F1-score, and accuracy increase, but human ratio and time cost also increase. With larger N, more samples are assigned to a human, so the overall evaluation results are more reliable, but this requires a bigger human annotation effort. The marginal reliability benefit of assigning more samples to a human decreases as N gets larger. Figure 5.2a shows that as N increases, the reliability increases sharply at the beginning but the increase levels off when $N > 2,500$. The samples assigned to a human when $N < 2,500$ have lower model confidence, i.e., it is very likely that those samples are given inaccurate evaluation by machine. But when $N > 2,500$, samples with higher model confidence are also assigned to human which yields a limited return in terms of reliability.

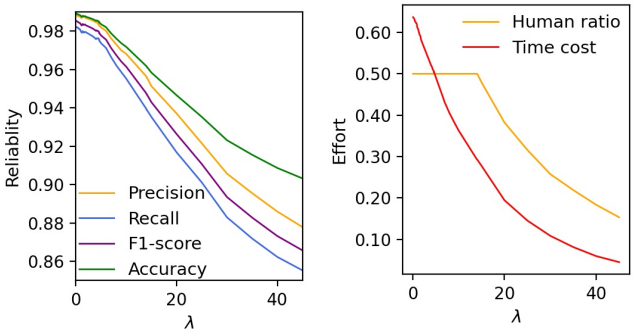


(a) Reliability. (b) Effort.

Figure 5.2: Influence of N with $\lambda = 0.1$.

Influence of λ

As λ increases, HMCEval gets more efficient, while the reliability gets worse. As shown in Figure 5.3, when λ increases, the human ratio stays at 0.5, and after a certain pivotal point, it decreases sharply. The time costs keep decreasing. The precision, recall, F1 score, and accuracy decrease rapidly. With larger λ, the SAE objective puts a bigger emphasis on efficiency, so HMCEval gets more efficient but less reliable.



(a) Reliability. (b) Effort.

Figure 5.3: Influence of λ with fixed N ($N/M = 0.5$).

5.5.3 Module analysis

Analysis of the SAE module

By adjusting the λ values, the SAE module can degenerate into a greedy algorithm [49]. Table 5.2 shows the results with the human ratio set to a fixed value of N/M, i.e., 0.5. When $\lambda = 0$, the HEE module has no effect, so it has the worst efficiency and the best reliability. When $\lambda \to +\infty$, i.e., 500, the MCE module contributes little to the

objective, so it has the best efficiency but the worst reliability.

Table 5.2: Analysis of the SAE module.

Metric	MCE	MCE+HEE	HEE
Reliability			
Precision	0.989	0.983	0.881
Recall	0.982	0.976	0.858
F1-score	0.985	0.980	0.869
Accuracy	0.989	0.985	0.906
Efficiency			
Human ratio	0.500	0.500	0.500
Time cost	0.650	0.500	0.135

Analysis of the MCE module

For the MCE module, we analyze the effect of alternative implementations. As shown in Figure 5.4, TS outperforms MCP and TCP. Specifically, when the human ratio is fixed to 0.5, TS achieves the best accuracy for different time costs. This means that TS has better model confidence estimation for the samples with higher confidence. As shown in Table 5.3, for the top-50% samples ranked by model confidence, TS has the best accuracy. MCP has the best AUC score, which means for all the M samples, MCP is the best. But the top-50% samples have more influence on the SAE module.

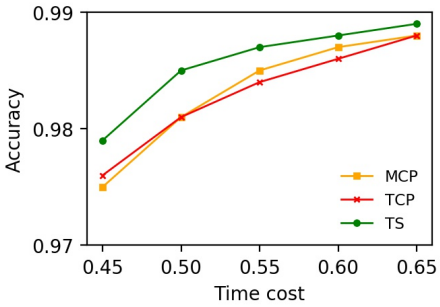


Figure 5.4: Performance of HMCEval with different MCE implementations ($N/M = 0.5$).

Analysis of the HEE module

For the HEE module, we analyze the effect of different features. Adding worker-related features helps to improve accuracy. As shown in Figure 5.5, SAE with both dialogue and worker-related features has better accuracy than SAE with only dialogue-related features

Table 5.3: Confidence prediction results comparison of MCE methods.

Metric	MCP	TCP	TS
AUC	**0.828**	0.823	0.825
Accuracy (top-50%)	0.977	0.975	**0.978**

when the human ratio is fixed to 0.5. Worker-based features are useful for time cost estimation. This is confirmed by the results in Table 5.4. The results with both dialogue and worker related features are the best, with MSE, RMSE and MAE decreasing by 55.6%, 35.9%, 45.9%, and R^2 increasing by 76.2%. The HEE module is sufficient for time cost prediction since R^2 greater than 0.26 is sufficient for behavior-related models [23].

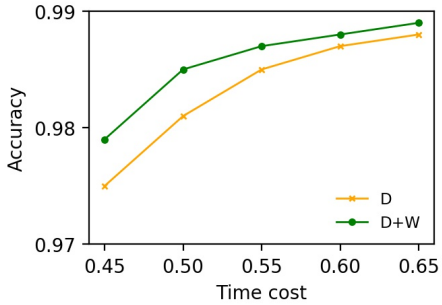


Figure 5.5: Feature analysis w.r.t. accuracy. (D: Dialogue related features, W: Worker related features.)

Table 5.4: Direct evaluation of the HEE module. (D: Dialogue related features, W: Worker related features.)

Metric	D	D+W
MSE	0.009	**0.004**
RMSE	0.092	**0.059**
MAE	0.061	**0.033**
R2	0.433	**0.763**

A correlation analysis between each feature and the real-time cost is shown in Table 5.5. All the features, except perplexity, have significant Pearson or Spearman scores with the real-time cost by workers. Most features show positive correlations. But two features, namely "non-malevolent turns" and "FK score" have a negative correlation with time cost: (i) non-malevolent responses are relatively easy to identify; and (ii) a higher FK score means that the dialogue is easier to understand, which requires less time to annotate.

Table 5.5: Correlation analysis between time cost and different features for HMC module. ** and * indicate significance $p < 0.001$, $p < 0.05$, respectively.

Feature	Pearson	Spearman
Dialogue related features (D)		
Total turns	0.053^{**}	0.122^{**}
Malevolent turns	0.445^{**}	0.600^{**}
Non-malevolent turns	-0.236^{**}	-0.292^{**}
First Submission	0.342^{**}	0.263^{**}
Paraphrased turns	0.555^{**}	0.564^{**}
Total length	0.046^{**}	0.100^{**}
Readability (DC)	0.042^{*}	-0.001
Readability (FK)	-0.026^{*}	-0.053^{**}
Contains malevolent turn	0.432^{**}	0.603^{**}
BERT-perplexity	-0.008	0.001
Worker related features (W)		
Worker test score	0.162^{**}	0.049^{**}
Approval rate ranking	0.840^{**}	0.849^{**}

5.5.4 Performance at different turns

We analyze the effectiveness of HMCEval at different dialogue turns in Figure 5.6. As the dialogue evolves, HMCEval gets more reliable. It gets easier for the MCE module to detect malevolent responses with high confidence when more context information is available. The exception for turn seven and nine might be due to the fact that the total number of utterances is small (less than 5% of the whole test set) and thus the results have high variance. The effort is not related to dialogue turn.

We also look into the 1.5% cases when HMCEval gives inaccurate evaluation and some cases that require human judgement but are not assigned to a human. We find that these cases mostly involve intentional deviations from ordinary language usage, through ambiguity, exaggeration, overstatement, or rhetorical figures. For instance, "I've committed 8 treasonous acts today and they still haven't put me in prison", is actually a non-malevolent joke. However, the MCE module classified it to be malevolent with high confidence.

5.6 Conclusion and Future Work

In this chapter, we have introduced a human-machine collaborative evaluation framework (HMCEval) for reliable and efficient conversational dialogue system (CDS) evaluation. Experiments on the task of evaluating malevolence in dialogue responses show that HMCEval can achieve around 99% reliability with half of the original human effort spared.

A limitation of HMCEval is that given 50% samples assigned to a human, 1.1–1.5% samples are evaluated inaccurately. This is due to contexts that consist of a small

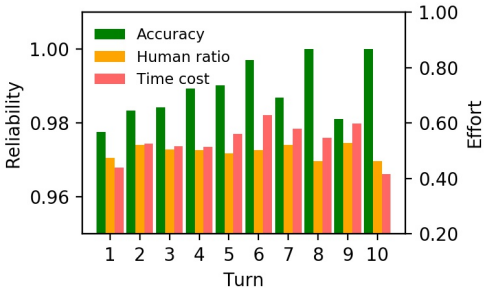


Figure 5.6: Accuracy and effort per turn with half human effort spared in average.

number of turns, or high confidence for some dialogues where language is used in a non-literal way. In addition, although HMCEval could be generalized to several evaluation metrics of CDS, e.g., BERT-based RUBER and BERT-based engagement, for score-based metrics, suitable confidence estimation is required.

In the future, we seek to improve the model confidence and human effort estimation by considering better neural architectures and more factors; we also plan to conduct a comprehensive and reliable analysis of the performance of current state-of-the-art CDS models by applying HMCEval to various evaluation aspects.

In this chapter we have answered RQ4. Our answer consisted of the HMCEval solution for balancing the reliability and human effort of malevolent dialogue evaluation, with experiment results as evidence for the effectiveness of HMCEval. Automatic and human evaluation methods for dialogue evaluation have been studied in the past and they do not suffice for dialogue malevolence evaluation. We put forward that HMCEval is a better solution.

Next, we conclude the thesis, take stock, and elaborate on future directions related to the malevolence of dialogue systems.

Appendix

In this appendix, we present additional details to further reproducibility of the results in this chapter. Specifically, we include the validation performance for the main result (Appendix 5.A), the average runtime of each module, and detailed information of the parameters (Appendix 5.B).

5.A Reliability and Efficiency for Validation Set

As to validation performance, we report the validation results of comparing HMCEval to machine evaluation and human evaluation in balancing reliability and efficiency, as shown in Table 5.6. HMCEval surpasses both human and machine evaluation in balancing reliability and efficiency for validation. On the one hand, compared to human evaluation, HMCEval achieves 98.2% of human accuracy with 50% human effort spared. This suggests that for the validation set, HMCEval is more efficient than human evaluation, while the reliability is close to human evaluation. On the other hand, compared to machine evaluation, the precision, recall, F1-score, and accuracy of HMCEval increase by 21.5%, 22.8%, 22.0%, and 15.3%, respectively. Moreover, the results on the validation set and the test set are similar. Compared to results on the test set, reliability results on the validation set are slightly lower, but the difference is less than 0.5%, as shown in Table 5.1 (presented in Section 5.5) and Table 5.6.

Table 5.6: Reliability and efficiency of HMCEval w.r.t. human and machine evaluation for validation ($N/M = 0.5$).

Metric	Machine	Human	HMCEval
Reliability			
Precision	0.806	1	0.979
Recall	0.793	1	0.974
F1-score	0.800	1	0.976
Accuracy	0.852	1	0.982
Efficiency			
Human ratio	0	1	0.500
Time cost	0	1	0.500

5.B Runtime and Parameters

In terms of average runtime, we have three modules. The time costs for all the modules are acceptable. The MCE module has thee methods: MCP, TS and TCP. Their time costs are 0.5 hours, 0.1 hours, and 3.5 hours, respectively. The HEE module is implemented by random forest regression and the runtime is less than 10 minutes for 5-fold cross-validation. The SAE module is implemented by ILP and the runtime is around 2.5 hours.

In terms of parameters, the MCE module is a neural network-based module. MCP and TS are estimated with the BERT-based classifier, which has 109.5 million parameters. TCP has an additional confidence network compared with MCP and TS. The confidence network part has 2.4 million parameters. The HEE module and the SAE module are not neural networks-based, we have included most of the relevant information above in Chapter 5.

Finally, the SAE module is based on search. There are a total number of 10 thousand trials with different N and λ parameters. The best N and λ are chosen by reliability metrics and efficiency metrics. In Table 5.1 (presented in Section 5.5) and Table 5.6, we choose the final results with $\lambda = 4.6$ and $N = 0.5M$, where M is the number of the total samples to be evaluated.

6 Conclusions

In this chapter, we first revisit the questions we asked in Section 1.1 and summarize the main findings and implications of our research in Section 6.1. Then, in Section 6.2, we describe the main limitations of our work and possible future directions.

6.1 Main Findings

6.1.1 The challenge of malevolent response exists for dialogue generation models

We started with the task of identifying and exposing the malevolence challenge in dialogue responses generated by generative models:

RQ1 How to establish the malevolence problem of generated dialogue responses by state-of-the-art (SOTA) generation models?

To answer **RQ1**, we analyzed the malevolence of pre-trained dialogue generation models and sequence to sequence (S2S)-based dialogue generation models. In terms of the S2S-based dialogue generation model, we also proposed the context-aware knowledge pre-selection (CaKe) model to decrease blandness and deflective responses so that CaKe is informative enough for malevolence analysis. The CaKe model uses dialogue context as the query to select knowledge from background text, thus improving informativeness. The knowledge pre-selection mechanism combines background-to-context (b2c) and context-to-background (c2b) attention.

The main findings we obtained are as follows:

(1) The results performed on the Holl-E dataset show that CaKe is capable of selecting relevant knowledge. We compared CaKe with the SOTA baseline models, and it outperforms the baselines for informativeness. Our case study also suggests that the generated samples are more fluent than baseline models. Moreover, our analysis of attention weights along with the positions of the background information, suggests that the attention mechanism is strong in attending to the relevant positions.

(2) The generated responses of pre-trained generation models do indeed exhibit the malevolence problem.

(3) The generated responses of the proposed CaKe and S2S-based baselines exhibit the problem of malevolence.

(4) In our experimental setup, pre-trained generation models generate more malevolent dialogue responses than the S2S-based generation models.

We conclude that we have demonstrated that malevolent dialogue response issues exist for dialogue generation models.

6.1.2 Building a taxonomy, dataset, and benchmark models for single-label dialogue malevolence detection

Next, we were interested in building a taxonomy and a dataset for dialogue malevolence detection based on previous work concerning what malevolent dialogue response is. We were also interested in building benchmark models for dialogue malevolence detection. We sought to answer the following question:

RQ2 How can we construct a high quality dataset via crowdsourcing that allows for single-label malevolent dialogue response detection and build an effective detection model?

To address **RQ2**, we built a three-level hierarchical malevolent dialogue taxonomy (HMDT) based on emotion, psychological and ethical aspects. We validated the user perception of the taxonomy through a user study. Then, we created a multi-turn malevolence dialogue dataset, i.e., malevolent dialogue response detection and classifying (MDRDC) on MTurk. Lastly, we ran several baseline models on the dataset and built a confidence-calibrated BERT-based classification model, i.e., BERT-based classifier with confidence calibration (BERT-conf).

The main findings we obtained are as follows:

(1) The results suggest that the concepts in the HMDT taxonomy capture malevolence in terms of user perception, including "non-credibility", "discomfort", "breakdown", and "abandonment".

(2) The data quality is substantial for dialogue malevolence detection. The collected dialogue context and rephrased utterances could be used to improve the classification performance.

(3) For the baseline malevolence detection models, the BERT-base classification model performs the best.

(4) The proposed BERT-conf model, which uses the confidence of the predicted category, has a better classification performance than the baseline models.

We conclude that we have built a meaningful taxonomy and dataset for single-label dialogue malevolence detection. We also build a benchmark for the task. Confidence calibration, dialogue context, and rephrased utterances are useful for improving classification performance.

6.1.3 A dataset and a label-correlation enhanced approach for multi-label dialogue malevolence detection

Building on our contributions to single-label dialogue malevolence detection, we were interested in multi-label dialogue malevolence detection and in strengthening the detection model. We sought to answer the following question:

RQ3 How can we build a model for multi-label dialogue malevolence detection based on single-label training data and construct a validated dataset to assess the model?

To answer **RQ3**, we crowdsourced the multi-label dialogue malevolence detection (MDMD) dataset with multi-label validation and test set. The task is multi-label dialogue malevolence detection from single-label training data. We also proposed the multi-faceted label correlation enhanced CRF (MCRF) model with a multi-faceted label correlation mechanism, including label correlation in context (LCC) and label correlation in taxonomy (LCT). Based on the dataset, we evaluated the effectiveness of our proposed MCRF model through extensive experiments. The MCRF model was compared with the baselines.

The main findings we obtained are as follows:

(1) It is essential to perform multi-label dialogue malevolence detection as the dataset statistics suggest that 12% of the utterances have more than one malevolent label.

(2) We are able to perform multi-label dialogue response malevolence detection from a single-label training set.

(3) The experimental results show that label correlation is able to improve the performance of multi-label malevolence detection. The improvement is convincing, based on an ablation study and comparison with baselines.

We conclude that we have built a meaningful dataset for multi-label dialogue malevolence detection from a single-label training set. We also build a benchmark for the task. Label correlation is useful for improving detection performance.

6.1.4 A human-machine collaborative approach for dialogue malevolence evaluation

Finally, based on the previous detection methods, we took a step towards dialogue malevolence evaluation and answered the following research question:

RQ4 How can we build a framework for evaluating malevolent dialogue responses with reliability and human effort balanced?

To answer **RQ4**, we built the human-machine collaborative evaluation (HMCEval) framework with three modules: model confidence estimation (MCE), human effort estimation (HEE), and sample assignment execution (SAE).

The findings that we obtained are as follows:

(1) HMCEval achieves around 99% evaluation accuracy with half of the human effort spared, showing that HMCEval provides reliable evaluation outcomes while reducing human effort by a large amount.

(2) HMCEval can degenerate into a greedy algorithm in cases where the parameter $\lambda = 0$ or $\lambda \to +\infty$.

We conclude that we have built a successful evaluation framework that balances reliability and human effort.

6.2 Future Work

6.2.1 Adversarial attack of pretrained dialogue generation models

Pre-trained dialogue systems are increasingly being used in practical applications. Pre-trained language models such as generative pre-trained transformer (GPT)-2 have a good performance on generation tasks. However, dialogue systems backed by pre-trained models, e.g. GPT-2, language model for dialogue applications (LaMDA), tend to generate unsafe responses and they are vulnerable to adversarial attacks of a black-box nature [150]. First, there is no work on analyzing the malevolence of responses generated by pre-trained dialogue models. Second, there is no deep understanding of the inner mechanism of its vulnerabilities to attack samples. For future work, we are interested to analyze the malevolent responses generated by pre-trained dialogue generation models and provide a method to find the vulnerabilities of the pre-trained dialogue systems.

In order to identify the vulnerabilities of the pre-trained dialogue systems, we need to generate adversarial samples. There are two kinds of methods for adversarial sample generation: non-trainable based methods and trainable based methods. First, earlier methods use concatenation or editing at the character level, word level, or sentence level. Jia and Liang [69] are the first to use a concatenation of sentences to attack the generation of answers. Later, Niu and Bansal [111] use edit-based methods to attack goal-oriented dialogue. Wallace et al. [157] use concatenation of adversarial triggers to the input sequence to attack GPT-2 to generate racial or offensive sentences and Heidenreich and Williams [60] attack GPT-2 to affect both topic and stance of the generated sentence. These methods are better at preserving semantics and efficiency, however it is easy to detect the modification of the input since constraints, e.g. lexical rules and equivalence to the original input, do not satisfy [137]. Second, later methods use training-based methods, such as paraphrasing to avoid detection [95, 137]. Moreover, some methods add constraints to make the adversarial samples similar to the original samples, however, the embedding space is large for a safety attack. Therefore, these methods may not efficient. Huang and Zhang [68] build a model to decrease the sample space; however, it is situated in the computer vision area, where the space is continuous, and it's not targeted for large malevolence related embedding space.

We will form the dialogue malevolence attack task and analyze the possible directions for improvement of pretrained models. We plan to investigate what kind of input could increase the malevolent response generation and implies the reason. The attack process will constrain the model to output a malevolent response based on reinforcement learning and improve the attack efficiency.

6.2.2 Semi-supervised algorithm to strengthen malevolent dialogue response detection robustness

We have annotated a dataset of 6,000 dialogues for dialogue malevolence detection in RQ2 and RQ3. The data sizes of some malevolence groups are still not large enough, which limits the robustness of the model against out-of-distribution samples. In order to solve this problem, we can annotate more samples, however, the labeling human cost is high.

Semi-supervised learning for classification can be used to improve model robustness, which makes the model reliability stable under various conditions [26]. There are two kinds of popular methods for semi-supervised learning: self-training generates pseudo labels to label the unlabeled data, and combines the previous clean labeled data to train a new model together; joint-training combines teacher and student network, where the teacher model trained on labeled data generate pseudo labels on unlabeled data, and student model optimizes the loss on the human label and pseudo labels jointly [55].

We plan to use a semi-supervised based method to strengthen detection robustness. First, based on the unlabeled data we collected, we generate pseudo labels for unlabeled data. Then, we combine the clean labeled data and pseudo labeled data to train a new model. The generated pseudo label can be noisy. In order to solve this problem, during the training process, we only assign the reliable pseudo-labels to the training. The reliability of the pseudo-label will be assessed by calculating the confidence of each label.

6.2.3 Improving malevolent dialogue response detection based on paraphrased implicitly malevolent data

For RQ2 and RQ3, we have built models for detecting all malevolence categories. For each category group, the utterances can be divided into an implicit or explicit sample. Implicit responses contain abstract, coded expressions without using explicitly malevolent words, e.g., “I tell the world: The immortal words of Adolf Hitler”. Explicit responses contain malevolent words with a clear meaning. The performance of BERT-conf and BERT-MCRF in detecting implicitly malevolent utterances, e.g., “deceit”, needs to be improved. We have collected rephrased data of the original utterances for RQ2. Previously, we have not collected the “implicit/explicit” label for each utterance. The rephrased utterance and the original utterance with the implicit or explicit expression can be used to improve the detection of implicitly malevolent dialogue responses, thus improving overall performance.

We plan to make the previous rephrased data more complete for the task of improving malevolent response detection via implicitly malevolent utterance paraphrasing. There have been some datasets on implicitly hate detection [43, 58], however, the taxonomy is limited to “hate speech” and the dataset is not for multi-turn dialogue. First, we plan to add the “implicit” or “explicit” label for the original utterance and rephrased the utterance collected; and rephrase all the malevolent utterances left since we only rephrased part of the data previously. We will annotate the utterances in the MDRDC and MDMD dataset as implicitly or explicitly malevolent. For the implicitly malevolent response, we will ask the annotator to label it as an explicit utterance, and for the

explicitly malevolent response, we will ask the annotator to label it as an implicit utterance. During the rephrased data collection process, we will also ask the workers to label the spans. Second, we plan to do implicitly malevolent dialogue response detection and use a data-efficient paraphrasing framework [73] to improve overall malevolent dialogue response detection performance.

6.2.4 Mitigating the malevolence of generated dialogue responses for generation models

In previous work, we have built models for detecting and evaluating dialogue malevolence. However, we have not mitigated the malevolence of dialogue responses. Current work on dialogue generation without pre-training could both generate malevolent dialogues, which is suggested in the results of RQ1. Previous work also suggests that pre-trained generation models, e.g., DialoGPT, Blenderbot, and Plato-2 all generate malevolent responses [147]. Therefore, it is hard to deploy the generation systems online and the need of mitigating malevolent responses arises. Previously, two main kinds of methods for mitigating dialogue response malevolence have been proposed: (i) methods based on different decoding strategies, e.g., vocabulary shift [53], top-k similarity [143], plug and play language model (PPLM) [32]; and (ii) pretraining-based methods, e.g., adaptive pretraining [50]. Pretraining-based dialogue unsafe content mitigation method is the current SOTA model. However, there is no large dataset for malevolence mitigation of pretrained generation model and the cost is high for collecting a large dataset with all aspects of malevolence.

Prompt-based fine-tuning can be used to improve the low-efficiency problem of adaptive pre-training [50]. To solve the high-cost problem, we propose a prompt-based fine-tune of the pretrained language model to mitigate dialogue malevolence. We plan to design different prompts that concatenate to the dialogue context and utterances to mitigate the generation of malevolent responses.

Bibliography

[1] C. Abbet, M. M'hamdi, A. Giannakopoulos, R. West, A. Hossmann, M. Baeriswyl, and C. Musat. Churn intent detection in multilingual chatbot conversations and social media. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 161–170, 2018. (Cited on page 37.)

[2] J. Allan, J. Arguello, L. Azzopardi, P. Bailey, T. Baldwin, K. Balog, H. Bast, N. Belkin, K. Berberich, B. von Billerbeck, J. Callan, R. Capra, M. Carman, B. Carterette, C. L. A. Clarke, K. Collins-Thompson, N. Craswell, W. B. Croft, J. S. Culpepper, J. Dalton, G. Demartini, F. Diaz, L. Dietz, S. Dumais, C. Eickhoff, N. Ferro, N. Fuhr, S. Geva, C. Hauff, D. Hawking, H. Joho, G. Jones, J. Kamps, N. Kando, D. Kelly, J. Kim, J. Kiseleva, Y. Liu, X. Lu, S. Mizzaro, A. Moffat, J.-Y. Nie, A. Olteanu, I. Ounis, F. Radlinski, M. de Rijke, M. Sanderson, F. Scholer, L. Sitbon, M. Smucker, I. Soboroff, D. Spina, T. Suel, J. Thom, P. Thomas, A. Trotman, E. Voorhees, A. P. de Vries, E. Yilmaz, and G. Zuccon. Report from the third strategic workshop on information retrieval in Lorne (SWIRL 2018). *SIGIR Forum*, 52:34–90, June 2018. (Cited on page 1.)

[3] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. (Cited on page 61.)

[4] V. M. Andreas, G. I. Winata, and A. Purwarianti. A comparative study on language models for task-oriented dialogue systems. In *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5. IEEE, 2021. (Cited on page 23.)

[5] A. Arango, J. Pérez, and B. Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 45–54, 2019. (Cited on pages 2, 5, 13, 28, and 31.)

[6] Z. Ashktorab, M. Jain, Q. V. Liao, and J. D. Weisz. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019. (Cited on page 37.)

[7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun, editors, *International Conference on Learning Representations*, 2015. (Cited on pages 4, 11, 14, 15, and 19.)

[8] A. Baheti, M. Sap, A. Ritter, and M. Riedl. 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, 2021. (Cited on pages 61 and 63.)

[9] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *SemEval*, pages 54–63, 2019. (Cited on pages 2, 5, 29, 30, 63, and 64.)

[10] N. J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5(1):133–143, 1980. (Cited on page 1.)

[11] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020. (Cited on page 28.)

[12] J. Bryson and A. Winfield. Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, 50(5):116–119, 2017. (Cited on pages 28 and 36.)

[13] W. Callaghan, J. Goh, M. Mohareb, A. Lim, and E. Law. Mechanicalheart: A human-machine framework for the classification of phonocardiograms. *Proceedings of the ACM on Human-Computer Interaction*, 2(The 21st ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)):28:1–28:17, 2018. (Cited on page 82.)

[14] J. Cassell and T. Bickmore. External manifestations of trustworthiness in the interface. *Communications of the ACM*, 43(12):50–56, 2000. (Cited on page 37.)

[15] R. Cerri, R. C. Barros, and A. C. De Carvalho. Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences*, 80(1):39–56, 2014. (Cited on page 51.)

[16] A. Chaganty, S. Mussmann, and P. Liang. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, 2018. (Cited on pages 80 and 82.)

[17] S. Chancellor, E. P. Baumer, and M. De Choudhury. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *The 21st ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 3:1–32, 2019. (Cited on page 32.)

[18] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. Silenzio, and M. De Choudhury. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88, 2019. (Cited on page 32.)

[19] S. F. Chen, D. Beeferman, and R. Rosenfeld. Evaluation metrics for language models. In *DARPA Broadcast News Transcription and Understanding Workshop (BNTUW)*, 1998. (Cited on pages 2 and 81.)

[20] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, pages 103–111, 2014. (Cited on page 12.)

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. (Cited on page 19.)

[22] A. Cohan, I. Beltagy, D. King, B. Dalvi, and D. S. Weld. 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, 2019. (Cited on page 69.)

[23] J. Cohen. *Statistical power analysis for the behavioral sciences (2nd edition)*. Hillsdale, L. Erlbaum Associates, 1988. (Cited on page 90.)

[24] E. Cole, O. Mac Aodha, T. Lorieul, P. Perona, D. Morris, and N. Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021. (Cited on page 6.)

[25] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, pages 2902–2913, 2019. (Cited on pages 2, 5, 43, 80, 84, and 86.)

[26] D. Croce, G. Castellucci, and R. Basili. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2114–2119, 2020. (Cited on page 99.)

[27] L. Cui, Y. Wu, S. Liu, Y. Zhang, and M. Zhou. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, 2020. (Cited on page 31.)

[28] M. Cummings. Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*, page 6313, 2004. (Cited on page 82.)

[29] A. C. Curry, G. Abercrombie, and V. Rieser. Convabuse: Data, analysis, and benchmarks for nuanced detection in conversational ai. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, 2021. (Cited on page 5.)

[30] E. Dale and J. S. Chall. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54, 1948. (Cited on page 85.)

[31] A. Das, B. Dang, and M. Lease. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. In *Proceedings of the Thirty-Fourth AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 33–42, 2020. (Cited on pages 80 and 81.)

[32] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2019. (Cited on page 100.)

[33] T. Davidson, D. Warmsley, M. Macy, I. Weber, Y. Kim, J. Devlin, D. Bamman, N. A. Smith, M. Khodak, R. West, et al. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 1837–1848. Association for Computational Linguistics, 1810. (Cited on pages 2, 5, 28, 29, 30, 63, and 64.)

[34] L. De Mattei, M. Cafagana, F. Dell'Orletta, M. Nissim, and A. Gatt. Changeit@evalita2020: Change headlines, adapt news, generate. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online. CEUR.org*, 2020. (Cited on page 84.)

[35] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, pages 1–56, 2020. (Cited on pages 2 and 79.)

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019. (Cited on page 43.)

[37] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid

algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015. (Cited on page 82.)

[38] E. Dinan, S. Humeau, B. Chintagunta, and J. Weston. 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, 2019. (Cited on page 61.)

[39] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston. Wizard of Wikipedia: Knowledge-powered conversational agents. In *The International Conference on Learning Representations*, International Conference on Learning Representations, 2019. (Cited on pages 1, 15, and 19.)

[40] E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*, 2021. (Cited on page 13.)

[41] P. E. Dunne, S. Doutre, and T. Bench-Capon. Discovering inconsistency through examination dialogues. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1680–1681, 2005. (Cited on page 11.)

[42] P. Ekman. Are there basic emotions? *Psychological Review*, 99(3):550, 1992. (Cited on pages 28 and 35.)

[43] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, 2021. (Cited on page 99.)

[44] L. Fell, A. Gibson, P. Bruza, and P. Hoyte. Human information interaction and the cognitive predicting theory of trust. In *the 5th ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)*, pages 145–152, 2020. (Cited on pages 37 and 50.)

[45] S. E. Finch and J. D. Choi. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 2020. (Cited on pages 2, 79, and 81.)

[46] A. Francesmonneris, H. Pincus, and M. First. *Diagnostic and statistical manual of mental disorders: DSM-V*. American Psychiatric Association, 2013. (Cited on pages 28 and 35.)

[47] M. Gamon, A. Aue, and M. Smets. Sentence-level mt evaluation without reference translations: Beyond language modeling. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, 2005. (Cited on page 85.)

[48] J. Gao, M. Galley, L. Li, et al. Neural approaches to conversational AI. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298, 2019. (Cited on pages 1, 27, and 52.)

[49] A. Gates, M. Gates, M. Sebastianski, S. Guitard, S. A. Elliott, and L. Hartling. The semi-automation of title and abstract screening: a retrospective exploration of ways to leverage abstrackr's relevance predictions in systematic and rapid reviews. *BMC Medical Research Methodology*, 20:1–9, 2020. (Cited on pages 82 and 88.)

[50] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, 2020. (Cited on pages 61 and 100.)

[51] S. Ghazarian, J. Wei, A. Galstyan, and N. Peng. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, 2019. (Cited on pages 79 and 81.)

[52] S. Ghazarian, R. Weischedel, A. Galstyan, and N. Peng. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pages 7789–7796, 2020. (Cited on page 81.)

[53] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer. Affect-lm: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, 2017. (Cited on page 100.)

[54] H. Golchha, M. Firdaus, A. Ekbal, and P. Bhattacharyya. Courteously yours: Inducing courteous behavior in customer care responses using reinforced pointer generator network. 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 851–860, 2019. (Cited on pages 2, 28, 30, and 31.)

[55] A. C. Gorgônio, A. Magály de Paula Canuto, K. M. Vale, and F. L. Gorgônio. A semi-supervised based framework for data stream classification in non-stationary environments. In *2020 International*

Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2020. (Cited on page 99.)

[56] D. Greyson. The social informatics of ignorance. *Journal of the Association for Information Science and Technology*, 70(4):412–415, 2019. (Cited on page 35.)

[57] N. Gunson, W. Sieińska, Y. Yu, D. H. Garcia, J. L. Part, C. Dondrup, and O. Lemon. Coronabot: A conversational ai system for tackling misinformation. In *Proceedings of the Conference on Information Technology for Social Good*, pages 265–270, 2021. (Cited on page 61.)

[58] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, 2022. (Cited on page 99.)

[59] T. He, J. Liu, K. Cho, M. Ott, B. Liu, J. Glass, and F. Peng. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1121–1133, 2021. (Cited on pages 3 and 24.)

[60] H. S. Heidenreich and J. R. Williams. The earth is flat and the sun is not a star: The susceptibility of gpt-2 to universal adversarial triggers. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 566–573, 2021. (Cited on page 98.)

[61] P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau. Ethical challenges in data-driven dialogue systems. In *AIES*, pages 123–129. ACM, 2018. (Cited on pages 28 and 36.)

[62] R. Higashinaka, T. Meguro, K. Imamura, H. Sugiyama, T. Makino, and Y. Matsuo. Evaluating coherence in open domain conversational systems. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. (Cited on page 1.)

[63] R. Higashinaka, M. Mizukami, K. Funakoshi, M. Araki, H. Tsukahara, and Y. Kobayashi. Fatal or not? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248, 2015. (Cited on page 37.)

[64] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019. (Cited on pages 3 and 11.)

[65] K. S. Hone and R. Graham. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, 6(3–4):287–303, 2000. (Cited on page 37.)

[66] M. Hossin and M. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015. (Cited on pages 45 and 86.)

[67] Y. Hou, C. Jochim, M. Gleize, F. Bonin, and D. Ganguly. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, 2019. (Cited on pages 80 and 81.)

[68] Z. Huang and T. Zhang. Black-box adversarial attack with transferable model-based embedding. *International Conference on Learning Representations*, 2019. (Cited on page 98.)

[69] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, 2017. (Cited on page 98.)

[70] H. Jiang, B. Kim, M. Guan, and M. Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, pages 5541–5552, 2018. (Cited on pages 80 and 84.)

[71] S. Jiang, P. Ren, C. Monz, and M. de Rijke. Improving neural response diversity with frequency-aware cross-entropy loss. In *The Web Conference 2019*, pages 2879–2885. ACM, May 2019. (Cited on pages 1, 25, and 27.)

[72] S. Jiang, T. Wolf, C. Monz, and M. de Rijke. TLDR: Token loss dynamic reweighting for reducing repetitive utterance generation. *arXiv preprint arXiv:2003.11963*, 2020. (Cited on pages 1 and 27.)

[73] S. Jolly, T. Falke, C. Tirkaz, and D. Sorokin. Data-efficient paraphrase generation to bootstrap intent classification and slot labeling for new features in task-oriented dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 10–20, 2020. (Cited on page 100.)

[74] K. Kann, S. Rothe, and K. Filippova. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, 2018. (Cited on page 1.)

[75] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014*

Conference on Empirical Methods in Natural Language Processing, pages 1746–1751, 2014. (Cited on pages 31 and 42.)

[76] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975. (Cited on page 85.)

[77] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. (Cited on page 19.)

[78] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri. Benchmarking aggression identification in social media. In *TRAC-2018*, pages 1–11, 2018. (Cited on pages 2, 5, 13, 30, 31, 63, and 64.)

[79] G. Kurata, B. Xiang, and B. Zhou. 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, 2016. (Cited on pages 2, 6, and 62.)

[80] T. Kyono, F. J. Gilbert, and M. van der Schaar. Mammo: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. *arXiv preprint arXiv:1811.02661*, 2018. (Cited on page 82.)

[81] S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI conference on artificial intelligence*, 2015. (Cited on pages 31 and 42.)

[82] W. Lasecki, C. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. Bigham. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 23–34, 2012. (Cited on page 82.)

[83] K. Lee, S. Salant, T. Kwiatkowski, A. Parikh, D. Das, and J. Berant. Learning recurrent span representations for extractive question answering. *arXiv preprint arXiv:1611.01436*, 2016. (Cited on page 14.)

[84] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014. (Cited on page 31.)

[85] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, 2016. (Cited on pages 1, 11, 14, and 81.)

[86] J. Li, T. Tang, W. X. Zhao, and J.-R. Wen. Pretrained language models for text generation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence: Survey Track*, 2021. (Cited on page 3.)

[87] P. Li. An empirical investigation of pre-trained transformer language models for open-domain dialogue generation. *arXiv preprint arXiv:2003.04195*, 2020. (Cited on page 23.)

[88] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, 2017. (Cited on page 29.)

[89] R. Lian, M. Xie, F. Wang, J. Peng, and H. Wu. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911*, 2019. (Cited on pages 4 and 15.)

[90] A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002. (Cited on page 85.)

[91] C.-Y. Lin and E. Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4*, pages 45–51, 2002. (Cited on pages 2, 79, and 81.)

[92] D. Litman and S. Silliman. Itspoke: An intelligent tutoring spoken dialogue system. In *Demonstration papers at HLT-NAACL 2004*, pages 5–8, 2004. (Cited on page 1.)

[93] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, 2016. (Cited on pages 2, 7, 79, and 81.)

[94] H. Liu, J. Dacon, W. Fan, H. Liu, Z. Liu, and J. Tang. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, 2020. (Cited on page 61.)

[95] H. Liu, Z. Wang, T. Derr, and J. Tang. Chat as expected: Learning to manipulate black-box neural dialogue models. *arXiv preprint arXiv:2005.13170*, 2020. (Cited on page 98.)

[96] J. Liu and J. Hockenmaier. Phrase grounding by soft-label chain conditional random field. In

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5112–5122. Association for Computational Linguistics, 2020. (Cited on pages 6 and 68.)

[97] P. Liu, X. Qiu, and X. Huang. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2873–2879. AAAI Press, 2016. (Cited on pages 31 and 42.)

[98] S. Liu, H. Chen, Z. Ren, Y. Feng, Q. Liu, and D. Yin. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, 2018. (Cited on page 14.)

[99] Y. Liu and M. Lapata. 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, 2019. (Cited on page 65.)

[100] R. Lowe, N. Pow, I. V. Serban, and J. Pineau. 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*, pages 285–294, 2015. (Cited on page 29.)

[101] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, 2017. (Cited on pages 1, 2, 79, and 81.)

[102] L. Luo, J. Xu, J. Lin, Q. Zeng, and X. Sun. An auto-encoder matching model for learning utterance-level semantic dependency in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 702–707, 2018. (Cited on page 81.)

[103] C. D. Manning, H. Schütze, and P. Raghavan. *Introduction to Information Retrieval*. Cambridge University Press, 2008. (Cited on page 69.)

[104] R. T. Marler and J. S. Arora. The weighted sum method for multi-objective optimization: new insights. *Structural and multidisciplinary optimization*, 41(6):853–862, 2010. (Cited on page 83.)

[105] R. O. Mason. Four ethical issues of the information age. *Mis Quarterly*, 10(1):5–12, 1986. (Cited on page 28.)

[106] M. L. McHugh. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012. (Cited on pages 40 and 69.)

[107] J. E. Mitchell. Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of applied optimization*, 1:65–77, 2002. (Cited on page 86.)

[108] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. (Cited on pages 2, 3, 11, 12, 14, 15, 16, and 18.)

[109] F. Morbini, E. Forbell, D. DeVault, K. Sagae, D. R. Traum, and A. A. Rizzo. A mixed-initiative conversational dialogue system for healthcare. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–139, 2012. (Cited on page 1.)

[110] R. Nedelchev, J. Lehmann, and R. Usbeck. Language model transformers as evaluators for open-domain dialogues. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6797–6808, 2020. (Cited on page 81.)

[111] T. Niu and M. Bansal. Adversarial over-sensitivity and over-stability strategies for dialogue models. *CoNLL 2018*, page 486, 2018. (Cited on page 98.)

[112] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, 2017. (Cited on page 81.)

[113] A. Ouni, R. G. Kula, M. Kessentini, T. Ishio, D. M. German, and K. Inoue. Search-based software library recommendation using multi-objective optimization. *Information and Software Technology*, 83: 55–75, 2017. (Cited on page 86.)

[114] C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998. (Cited on page 83.)

[115] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. (Cited on pages 2, 79, and 81.)

[116] J.-r. Park. 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, 2008. (Cited on pages 2, 27, and 37.)

[117] J.-r. Park. 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, 2008. (Cited on pages 2 and 27.)

[118] D. L. Paulhus and K. M. Williams. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality*, 36(6):556–563, 2002. (Cited on page 2.)

[119] H. Peng, J. Li, Y. He, Y. Liu, M. Bao, L. Wang, Y. Song, and Q. Yang. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the 2018 World Wide Web Conference*, pages 1063–1072, 2018. (Cited on page 31.)

[120] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. (Cited on page 31.)

[121] A. Przegalinska, L. Ciechanowski, A. Stroz, P. Gloor, and G. Mazurek. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons*, 62(6):785–797, 2019. (Cited on page 37.)

[122] F. Radlinski and N. Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pages 117–126, 2017. (Cited on page 1.)

[123] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. (Cited on pages 3 and 14.)

[124] A. Ram, R. Prasad, C. Khatri, A. Venkatesh, R. Gabriel, Q. Liu, J. Nunn, B. Hedayatnia, M. Cheng, A. Nagar, et al. Conversational AI: The science behind the Alexa prize. *arXiv preprint arXiv:1801.03604*, 2018. (Cited on page 79.)

[125] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, 2019. (Cited on page 79.)

[126] S. Ravindranath, R. Baburaj, V. N. Balasubramanian, N. Namburu, S. Gujar, and C. Jawahar. Human-machine collaboration for face recognition. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 10–18, 2020. (Cited on pages 2, 7, and 82.)

[127] P. Ren, Z. Chen, C. Monz, J. Ma, and M. de Rijke. Thinking globally, acting locally: Distantly supervised global-to-local knowledge selection for background based conversation. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'20, 2020. (Cited on pages 1 and 27.)

[128] P. Ren, R. Li, Y. Zhang, and M. de Rijke. Dialogue malevolence attacks against pre-trained models. In *Under Review*, 2022.

[129] Z. Ren, M.-H. Peetz, S. Liang, W. Van Dolen, and M. De Rijke. Hierarchical multi-label classification of social text streams. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 213–222, 2014. (Cited on page 51.)

[130] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, 2010. (Cited on page 39.)

[131] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, 2011. (Cited on page 39.)

[132] S. Roberts, J. D. Henry, and P. Molenberghs. Immoral behaviour following brain damage: A review. *Journal of Neuropsychology*, 13(3):564–588, 2018. (Cited on pages 28 and 35.)

[133] A. Roegiest, A. Lipani, A. Beutel, A. Olteanu, A. Lucic, A.-A. Stoica, A. Das, A. Biega, B. Voorn, C. Hauff, D. Spina, D. Lewis, D. W. Oard, E. Yilmaz, F. Hasibi, G. Kazai, G. McDonald, H. Haned, I. Ounis, I. van der Linden, J. Garcia-Gathright, J. Baan, K. N. Lau, K. Balog, M. de Rijke, M. Sayed, M. Panteli, M. Sanderson, M. Lease, M. D. Ekstrand, P. Lahoti, and T. Kamishima. FACTS-IR: Fairness, accountability, confidentiality, transparency, and safety in information retrieval. *SIGIR Forum*, 53(2):20–43, December 2019. (Cited on page 61.)

[134] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, et al. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, 2021. (Cited on pages 11 and 18.)

[135] D. Roussinov and J. A. Robles-Flores. Applying question answering technology to locating malevolent online content. *Decision Support Systems*, 43(4):1404–1418, 2007. (Cited on pages 2, 61, and 64.)

[136] J. Sabini and M. Silver. Ekman's basic emotions: Why not love and jealousy? *Cognition and Emotion*, 19(5):693–712, 2005. (Cited on page 35.)

[137] B. Sabir, M. A. Babar, and R. Gaire. Reinforcebug: A framework to generate adversarial textual examples. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5954–5964, 2021. (Cited on page 98.)

[138] S. M. Saral, R. R. Sawarkar, and P. A. Jalan. 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, 2018. (Cited on pages 2, 61, and 64.)

[139] A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, 2017. (Cited on pages 4, 12, 15, and 19.)

[140] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016. (Cited on pages 12, 14, 15, 16, and 19.)

[141] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016. (Cited on pages 12, 14, and 19.)

[142] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017. (Cited on page 14.)

[143] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. "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, 2021. (Cited on pages 2, 61, 63, 64, and 100.)

[144] Z. R. Shi, C. Wang, and F. Fang. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*, 2020. (Cited on page 80.)

[145] J. P. Stevens. *Applied Multivariate Statistics for the Social Sciences*. Routledge, 2012. (Cited on page 36.)

[146] C. Sumner, A. Byers, R. Boochever, and G. J. Park. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *11th International Conference on Machine Learning and Applications*, volume 2, pages 386–393. IEEE, 2012. (Cited on pages 29 and 30.)

[147] H. Sun, G. Xu, J. Deng, J. Cheng, C. Zheng, H. Zhou, N. Peng, X. Zhu, and M. Huang. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, 2022. (Cited on pages 2, 5, 13, 61, 63, 64, and 100.)

[148] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, Advances in Neural Information Processing Systems, 2014. (Cited on pages 4, 11, 14, and 19.)

[149] C. Tao, L. Mou, D. Zhao, and R. Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. (Cited on pages 2, 79, and 81.)

[150] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. (Cited on page 98.)

[151] S.-C. Tsai, C.-W. Huang, and Y.-N. Chen. 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, 2021. (Cited on pages 2, 6, and 62.)

[152] B. van Aken, J. Risch, R. Krestel, and A. Löser. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, 2018. (Cited on pages 5, 28, and 31.)

[153] K. van Deemter, M. Theune, and E. Krahmer. Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, 31(1):15–24, 2005. (Cited on pages 1 and 27.)

[154] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. (Cited on page 65.)

[155] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015. (Cited on pages 3 and 14.)

[156] A. Vrij, K. Edward, K. P. Roberts, and R. Bull. Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal behavior*, 24(4):239–263, 2000. (Cited on page 36.)

[157] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, 2019. (Cited on page 98.)

[158] P. Wang, Y. Fan, S. Niu, Z. Yang, Y. Zhang, and J. Guo. Hierarchical matching network for crime classification. In *proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 325–334, 2019. (Cited on page 51.)

[159] R. Wang, H. Su, C. Wang, K. Ji, and J. Ding. To tune or not to tune? how about the best of both worlds? *arXiv preprint arXiv:1907.05338*, 2019. (Cited on page 31.)

[160] S. Wang and J. Jiang. Machine comprehension using match-lstm and answer pointer. In *International Conference on Learning Representations*, 2017. (Cited on page 14.)

[161] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, 2017. (Cited on page 14.)

[162] Y.-A. Wang and Y.-N. Chen. 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*, pages 6840–6849, 2020. (Cited on page 65.)

[163] Z. Wang and C. Potts. 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, 2019. (Cited on pages 2, 63, and 64.)

[164] Z. Wang, H. Dong, R. Jia, J. Li, Z. Fu, S. Han, and D. Zhang. 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, 2021. (Cited on page 66.)

[165] Z. Wang, J. Liu, H. Cui, C. Jin, M. Yang, Y. Wang, X. Li, and R. Mao. Two-stage behavior cloning for spoken dialogue system in debt collection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4633–4639, 2021. (Cited on page 1.)

[166] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL student research workshop*, pages 88–93, 2016. (Cited on pages 2, 5, 13, 29, 30, 63, and 64.)

[167] Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li. 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)*, pages 496–505, 2017. (Cited on page 29.)

[168] J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*, 2020. (Cited on pages 2, 13, 61, and 80.)

[169] Z. Yan, N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li. Building task-oriented dialogue systems for online shopping. In *Thirty-first AAAI conference on artificial intelligence*, 2017. (Cited on page 1.)

[170] L. Yao, C. Mao, and Y. Luo. Graph convolutional networks for text classification. In *Proceedings of the Thirty-Third AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019. (Cited on pages 31 and 43.)

[171] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32, 2018. (Cited on page 79.)

[172] K. Yu, S. Berkovsky, R. Taib, J. Zhou, and F. Chen. Do I trust my machine teammate? An investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 460–468, 2019. (Cited on page 84.)

[173] Y. Yu, W. Zhang, K. Hasan, M. Yu, B. Xiang, and B. Zhou. End-to-end answer chunk extraction and ranking for reading comprehension. *arXiv preprint arXiv:1610.09996*, 2016. (Cited on page 14.)

[174] M. Zaib, Q. Z. Sheng, and W. Emma Zhang. A short survey of pre-trained language models for conversational ai-a new age in nlp. In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–4, 2020. (Cited on page 3.)

[175] H. Zamani, S. Dumais, N. Craswell, P. Bennett, and G. Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pages 418–428, 2020. (Cited on page 36.)

[176] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 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, 2019. (Cited on pages 2, 5, 13, 30, 31, 63, and 64.)

[177] B. Zarouali, E. Van den Broeck, M. Walrave, and K. Poels. Predicting consumer responses to a chatbot on facebook. *Cyberpsychology, Behavior, and Social Networking*, 21(8):491–497, 2018. (Cited on page 37.)

[178] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018. (Cited on page 15.)

[179] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, pages 649–657, 2015. (Cited on pages 31 and 42.)

[180] Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1810–1820, 2018. (Cited on pages 11 and 14.)

[181] Y. Zhang, R. Henao, Z. Gan, Y. Li, and L. Carin. Multi-label learning from medical plain text with convolutional residual models. In *Machine Learning for Healthcare Conference*, pages 280–294. PMLR, 2018. (Cited on page 72.)

[182] Y. Zhang, P. Ren, and M. de Rijke. Improving background based conversation with context-aware knowledge pre-selection. *Search-Oriented Conversational AI Workshop (SCAI)*, 2019. (Cited on page 11.)

[183] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and W. B. Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020. (Cited on pages 1, 11, 13, 18, and 79.)

[184] Y. Zhang, P. Ren, and M. de Rijke. 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, 2021. (Cited on pages 61 and 79.)

[185] Y. Zhang, P. Ren, and M. de Rijke. A taxonomy, data set, and benchmark for detecting and classifying malevolent dialogue responses. *Journal of the Association for Information Science and Technology*, 72(12):1477–1497, 2021. (Cited on pages 2, 6, 11, 13, 19, 27, 61, 63, 64, 68, 69, 77, 80, 81, and 85.)

[186] Y. Zhang, P. Ren, W. Deng, Z. Chen, and M. de Rijke. Improving multi-label malevolence detection in dialogues through multi-faceted label correlation enhancement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3543–3555, 2022. (Cited on page 61.)

[187] Z. Zhang, J. Li, P. Zhu, H. Zhao, and G. Liu. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, 2018. (Cited on page 29.)

[188] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4623–4629, 2018. (Cited on page 14.)

Summary

Recently, dialogue systems have been adopted in different domains and they interact with users in daily life. Dialogue generation methods have emerged from early rule-based and retrieval-based methods as corpus-based methods. Corpus-based conversational agents can generate more diverse and natural responses than template-based or retrieval-based agents. With the increased generative capacity of corpus-based conversational agents comes the need to detect and evaluate malevolent responses that are inappropriate in terms of content and dialogue acts.

In the thesis, we focus on malevolent dialogue response detection and evaluation. Few studies have addressed the issue of malevolent dialogue responses which have negative social risks and consequences. On the one hand, previous studies on the topic of detecting inappropriate content are mostly focused on a specific category of malevolence or single sentences instead of an entire dialogue and they do not consider multi-label malevolence. On the other hand, currently, there is no research on dialogue malevolence evaluation with both high reliability and low human effort.

First, we analyze malevolence issues of the state-of-the-art dialogue generation models, including both pre-trained generation models and *sequence to sequence* (S2S)-based generation models. We also introduce a knowledge pre-selection based dialogue generation model, i.e., *context-aware knowledge pre-selection* (CaKe), to improve the informativeness of dialogue response since S2S-based generation models tend to generate bland and deflective responses, which may influence malevolence analysis. The results suggest that the malevolent response challenge exists for generation models and that pre-trained generation models are more malevolent than S2S-based generation models. Results also show that the proposed CaKe is superior to current SOTA baselines in informativeness, indicating that it benefits from the pre-selection process.

Second, we advance research on the *malevolent dialogue response detection and classifying* (MDRDC) task. We define the task and build a *hierarchical malevolent dialogue taxonomy* (HMDT), which is a broad hierarchical taxonomy. We create a labeled multi-turn dialogue dataset and formulate the MDRDC task as a hierarchical classification task. The MDRDC dataset is the first high-quality multi-turn dialogue dataset for malevolent dialogues. We apply SOTA text classification methods to the MDRDC task and report on experiments aimed at assessing the performance of these approaches. We present a confidence-based classification model that beats the baselines for single-label dialogue malevolence detection.

Third, we propose the task of multi-label dialogue malevolence detection and crowdsource a multi-label dataset, *multi-label dialogue malevolence detection* (MDMD), for multi-label dialogue malevolence detection from a single-label training set. We also propose a multi-label malevolence detection model, *multi-faceted label correlation enhanced CRF* (MCRF), with a multi-faceted label correlation mechanism which includes two kinds of label correlation mechanisms, *label correlation in taxonomy* (LCT) and *label correlation in context* (LCC); and evaluate the model by MDMD dataset. Experiments conducted on MDMD show that MCRF method outperforms the best-performing baseline by a large margin.

Last, we propose a *human-machine collaborative* (HMC) framework, *human-machine collaborative evaluation* (HMCEval), for dialogue malevolence evaluation,

that balances overall reliability and human effort. HMCEval views dialogue evaluation as a sample assignment problem, where we need to decide to assign a sample to a human or a machine for evaluation. The optimum assignment solution is found by a *sample assignment execution* (SAE) module based on the estimated confidence and effort. The confidence of the predicted sample assignment is estimated by the *model confidence estimation* (MCE) module and the human effort is estimated by the *human effort estimation* (HEE) module. The performance of HMCEval on the task of evaluating malevolence in dialogues is assessed using the MDRDC dataset, and compared with automatic evaluation and human judgement. Our experimental results show that HMCEval achieves around 99% evaluation accuracy with half of the human effort spared, showing that HMCEval provides reliable evaluation outcomes while reducing human effort by a large amount.

Samenvatting

Recentelijk zijn dialoogsystemen in verschillende domeinen ingevoerd en communiceren ze met gebruikers in het dagelijks leven. Dialooggeneratiemethoden zijn voortgekomen uit vroege, op regels gebaseerde en op retrieval gebaseerde methoden, en op corpus gebaseerde methoden. Op corpus gebaseerde gespreksagenten kunnen meer diverse en natuurlijke reacties genereren dan op sjablonen gebaseerde of op retrieval gebaseerde agenten. Met de toegenomen generatieve capaciteit van op corpus gebaseerde gespreksagenten komt de noodzaak om kwaadaardige reacties die ongepast zijn in termen van inhoud en dialooghandelingen te detecteren en te evalueren.

In het proefschrift richten we ons op het detecteren en evalueren van kwaadwillige dialoogreacties. Er zijn maar weinig studies die zich bezighouden met het probleem van kwaadwillige dialoogreacties die negatieve sociale risico's en gevolgen hebben. Aan de ene kant zijn eerdere studies over het detecteren van ongepaste inhoud meestal gericht op een specifieke categorie kwaadwillendheid of enkele zinnen in plaats van op een hele dialoog en houden ze geen rekening met multi-label kwaadwilligheid. Aan de andere kant is er momenteel geen onderzoek naar de evaluatie van kwaadwilligheid van dialogen met zowel een hoge betrouwbaarheid als een lage menselijke inspanning.

Eerst analyseren we boosaardigheidsproblemen van de *state-of-the-art* dialooggeneratiemodellen, met inbegrip van zowel vooraf getrainde generatiemodellen als op sequentie tot sequentie (S2S) gebaseerde generatiemodellen. We introduceren ook een op kennisvoorselectie gebaseerd dialooggeneratiemodel, d.w.z. contextbewuste kennisvoorselectie (CaKe), om de informativiteit van de dialoogrespons te verbeteren, aangezien op S2S gebaseerde generatiemodellen de neiging hebben om saaie en afbuigende reacties te genereren, die van invloed kunnen zijn op de kwaadwillendheidanalyse. De resultaten suggereren dat de boosaardige respons-uitdaging daadwerkelijk bestaat voor generatiemodellen en dat vooraf getrainde generatiemodellen kwaadwillender zijn dan op S2S gebaseerde generatiemodellen. De resultaten laten ook zien dat de voorgestelde CaKe superieur is aan de huidige SOTA-baselines in informativiteit, wat aangeeft dat het profiteert van het preselectieproces.

Ten tweede bevorderen we onderzoek naar de taak voor het detecteren en classificeren van kwaadwillige dialogen (MDRDC). We definiëren de taak en bouwen een hiërarchische kwaadwillende dialoogtaxonomie (HMDT), een brede hiërarchische taxonomie. We creëren een gelabelde *multi-turn* dialoogdataset en formuleren de MDRDC-taak als een hiërarchische classificatietaak. De MDRDC-dataset is de eerste hoogwaardige multi-turn dialoogdataset voor kwaadwillige dialogen. We passen *state-of-the-art* tekstclassificatiemethoden toe op de MDRDC-taak en rapporteren over experimenten die gericht zijn op het beoordelen van de prestaties van deze aanpakken. We presenteren een op *confidence* gebaseerd classificatiemodel dat de *baselines* verslaat voor detectie van kwaadwillendheid met één label.

Ten derde stellen we de taak voor van detectie van kwaadwilligheid met meerdere labels en het *crowdsourcen* van een multi-label dataset, multi-label dialoog kwaadwilligheidsdetectie (MDMD), voor detectie van kwaadwilligheid met meerdere labels vanuit een trainingsset met één label. We stellen ook een multi-label detectiemodel voor kwaadwillendheid voor, een veelzijdige labelcorrelatie-versterkte CRF (MCRF), met een veelzijdig labelcorrelatiemechanisme dat twee soorten labelcorrelatiemechanismen

omvat: labelcorrelatie in taxonomie (LCT) en labelcorrelatie in context (LCC). We evalueren het model door middel van de MDMD dataset. Experimenten uitgevoerd op MDMD tonen aan dat de MCRF-methode de best presterende *baseline* met een grote marge overtreft.

Ten slotte stellen we een mens-machine-samenwerkingskader (HMC) voor, mens-machine-samenwerkingsevaluatie (HMCEval), voor de evaluatie van kwaadwillendheid in dialogen, dat een balans biedt tussen algehele betrouwbaarheid en menselijke inspanning. HMCEval beschouwt dialoogevaluatie als een sampletoewijzingsprobleem, waarbij we moeten beslissen om een sample toe te wijzen aan een mens of een machine voor evaluatie. De optimale opdrachtoplossing wordt gevonden door een voorbeeldopdrachtuitvoeringsmodule (SAE) op basis van het geschatte vertrouwen en de inspanning. Het vertrouwen van de voorspelde steekproeftoewijzing wordt geschat door de module voor het schatten van de betrouwbaarheid van het model (MCE) en de menselijke inspanning wordt geschat door de module voor het schatten van de menselijke inspanning (HEE). De prestaties van HMCEval op de taak om kwaadwillendheid in dialogen te evalueren, worden beoordeeld met behulp van de MDRDC-dataset en vergeleken met automatische evaluatie en menselijk oordelen. Onze experimentele resultaten laten zien dat HMCEval een evaluatienauwkeurigheid van ongeveer 99% behaalt met de helft van de menselijke inspanning, wat aantoont dat HMCEval betrouwbare evaluatieresultaten biedt terwijl de menselijke inspanning aanzienlijk wordt verminderd.