MPHDetect: Multi-View Prompting and Hypergraph Fusion for Malevolence Detection in Dialogues

Bo Xu
School of Computer Science and Technology,
Dalian University of Technology, Dalian, China
xubo@dlut.edu.cn

Xuening Qiao
School of Computer Science and Technology,
Dalian University of Technology, Dalian, China
qiao@mail.dlut.edu.cn

Hongfei Lin
School of Computer Science and Technology,
Dalian University of Technology, Dalian, China
hflin@dlut.edu.cn

Linlin Zong[1]
School of Software,
Dalian University of Technology, Dalian, China
llzong@dlut.edu.cn

Abstract

Malevolence detection in dialogues aims to identify harmful or inappropriate utterances, significantly impacting dialogue quality and user satisfaction. Although existing studies have shown promising performance by modeling interaction patterns from dialogue history, various malevolence-invoking factors, such as fine-grained emotions, evolving topics and user profiles, are often overlooked. To comprehensively consider these factors, we propose a hypergraph fusion model by employing multi-view LLM-driven prompts for malevolence detection in dialogues. Our model integrates emotion context, topic context, user profile context and interaction context, utilizing hypergraphs to establish high-order contextual relationships from multi views for deducing malevolence-invoking semantics. Experimental results on two benchmark datasets demonstrate that our model achieves the state-of-the-art performance.

CCS Concepts

• **Computing methodologies** → **Discourse, dialogue and pragmatics**.

Keywords

Malevolence detection, Dialogue systems, Prompt learning, Emotion analysis, User profile, Topic analysis

ACM Reference Format:
Bo Xu, Xuening Qiao, Hongfei Lin, and Linlin Zong[1]. 2024. MPHDetect: Multi-View Prompting and Hypergraph Fusion for Malevolence Detection in Dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679966

[1]Corresponding Author

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679966

1 Introduction

Large language models (LLMs) trained on a vast amount of internet text, may absorb harmful or biased content, leading to the generation of malicious responses in conversations. Such responses can trigger disputes, discomfort, and even exacerbate mental health issues, particularly for users who are psychologically vulnerable [1]. Therefore, it is crucial to detect malevolent content in dialogues.

While significant research efforts have been devoted to detecting toxic or offensive language [9, 5, 8], the inherently interactive nature of dialogues introduces heightened subtlety and complexity into malevolence detection. This complexity emerges from the implicit expressions of malevolence scattered throughout utterances in dialogues, presenting a considerable challenge in malevolence detection. Despite existing research endeavors to model interaction patterns for malevolence detection [12, 11], the influence of *fine-grained emotions*, *evolving topics* and *user profiles* in dialogues has largely been overlooked for malevolence detection. **Fine-grained emotions** in dialogues are instrumental in detecting malevolence. Even slight shifts in user emotions can significantly influence the interpretation and formulation of dialogue contents, introducing biases in intent comprehension. Consequently, this can result in the misinterpretation of malevolent expressions. **Topics evolve continuously** in the conversation as the multi-turn dialogue progresses. Malevolent expressions under different dialogue topics manifest in varied forms and semantics. Hence, it is necessary to analyze the evolution of dialogue topics and model them to facilitate effective detection of malevolence. **User profiles** in dialogues are constructed from an individual user's utterances, incorporating statements, personalized traits, and internal opinion dependencies. These components are integral to user profiling and significantly influence the detection of malevolence within dialogues.

To enhance malevolence detection, we propose MPHDetect, a multi-view prompting and hypergraph fusion model that leverages the robust reasoning capabilities of large language models. Specifically, we designed tailored prompt templates to generate three-view prompts: *emotion prompt* for capturing fine-grained emotional nuances, *topic prompt* for modeling evolving conversational topics, and *user profile prompt* for representing individual preferences and personality traits in utterances of the dialogue history. After extracting features from the dialogue content and prompts, these representations are inputted into three separate GCNs and Bi-LSTMs to learn the emotional, topical and user profile

context, along with a LSTM-based utterance interaction context learning. Furthermore, we introduce hypergraph facilitates flexible connections across different contextual views, enabling the discovery of semantic patterns related to malevolence. Through hypergraph convolution, we achieve comprehensive feature fusion, allowing seamless integration of emotional, topical, and user profile-based contexts. We summarize our main contributions as follows.

- We model three crucial malevolence-invoking factors, including emotion, topic and user profile, by designing tailored prompt templates based on large language models for malevolence detection in dialogues.
- We propose a hypergraph-based multi-view fusion model for comprehensively modeling emotion context, topic context, user profile context and interaction context, enabling the capture of implicit malevolent remarks embedded in dialogues.
- We examine the effectiveness of our model on two benchmark datasets. Experimental results demonstrate the superiority of our model over the state-of-the-art models.

2 Methodology

2.1 Overview

The objective of our task is to predict the malevolence label of each utterance for a given dialogue $D = \{x_1, x_2, ..., x_N\}$ with N utterances. To this end, we propose the MPHDetect model that contains four modules as depicted in Figure 1. The prompt learning module adopts the LLM-based prompting strategy to capture three-view prompts from dialogues. The feature extraction module learns the initial representations of utterances and prompts. The hypergraph fusion module first captures contextual information from four views and then integrates multi-view contextual information by using hypergraph. The malevolence label prediction module predicts the malevolence labels of utterances.

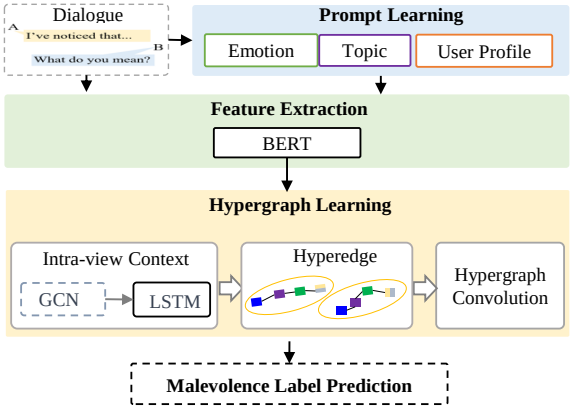

Figure 1: The main architecture of MPHDetect.

2.2 Prompt Learning

We design three prompt templates to capture fine-grained emotions, evolving topics and user profiles in dialogues. The emotion prompt facilitates the understanding of user attitudes and intents, while the topic prompt reveals potential topic-related malevolence expressed subtly in utterances. The user profile prompt captures the unique traits of each individual interlocutor. The designed prompt templates are shown in the following.

Prompt: The data consist of a dialogue. Now, you will play as an expert in analyzing the emotions, topics and user profiles in it. Based on this dialogue {*dialogue_history*}, analyze {*dialogue_content*} of this utterance: {*current_utterance*}.

Within these templates, the *dialogue_history* encompasses $i-1$ utterances denoted as $D_i = \{x_1, x_2, ..., x_{i-1}\}$ across i rounds of conversations; the *dialogue_content* includes the fine-grained emotions, the topics and the user profiles of this utterance. Following prompt learning, each utterance corresponds to a generated fine-grained emotion prompt $E = \{e_1, e_2, ..., e_N\}$, a topic prompt $T = \{t_1, t_2, ..., t_N\}$, and a user profile prompt $S = \{s_1, s_2, ..., s_N\}$.

Using ChatGPT-3.5 with OpenAI API, the generated prompts encode in-depth semantic information derived from analyzing the dialogue history generally. However, the generated prompts may have lower quality, such as a higher repetition rate compared to the original dialogue. Therefore, for the generated prompts, we need to filter out those of low quality. Specifically, we calculate the cosine similarity between the generated prompts and the current utterances, retaining prompts with a similarity less than 90%.

2.3 Feature Extraction

To extract initial representations of these three-view prompts and the dialogue history, we first fine-tune a BERT [3] model at the utterance level, and then use it as the feature extractor. Specifically, given an utterance x_i, a special token $[CLS]$ is added at the beginning of the sequence, making the model's input represented as a sequence $\{[CLS], x_i^1, x_i^2, ..., x_i^n\}$. Upon fine-tuning, we extract the embedding from the $[CLS]$ token in the last layer as the feature representation u_i^x for x_i, where $u_i^x \in \mathbb{R}^{d_h}$.

2.4 Hypergraph Learning

To more accurately detect malicious dialogues, we need to deeply integrate information about various aspects. However, existing fusion methods are mostly based on binary relation graphs, which struggle to achieve effective inter-view semantic fusion when dealing with three or more types of data. To address this issue, we introduce hypergraphs to establish multi-view polyadic relationships, replacing the multiple binary relationships used in existing graph structures. This allows for a more thorough and efficient integration of multi-view features.

2.4.1 Intra-view Contextual Learning. This module considers four-view contextual representations: the interaction context, emotion context, topic context and user profile context.

The interaction context is derived from dialogue history. The contextual representations $c_i^x \in \mathbb{R}^{2d_u}$ of interaction context is learned by utilizing Bi-LSTM to model interaction patterns between two interlocutors as follows:

$$c_i^x, h_i^x = \text{LSTM}^x\left(u_i^x, h_{i-1}^x\right) \tag{1}$$

where h_i^x represents the i-th hidden state of the Bi-LSTM.

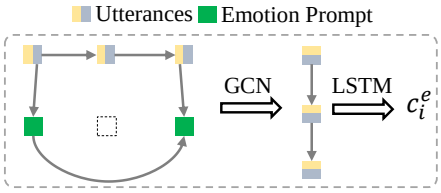

Figure 2: The process of emotion context learning.

The emotion context, topic context and user profile context are acquired from the incomplete prompts and the dialogue history. To align and extract the semantic presentations of prompts and corresponding utterance. We firstly fuse the representations of prompts (if exist) and their corresponding utterance using GCNs [4], followed by Bi-LSTMs to learn the contextual representations from each view. Take the emotion context as an example, we construct a graph by treating prompts and utterances as nodes, connecting them based on contextual relationships. As shown in Figure 2, nodes representing the same sentence across prompts and utterances are interconnected. GCN is then applied to learn fine-grained emotional representations from the initial representations of nodes.

$$G = AG_0W + O \tag{2}$$

where A is the adjacency matrix, G_0 denotes initial node representations extracted by BERT, W is the weight matrix, O is the bias, and G represents the learned node representations by GCN. Next, we regard the fine-grained emotional representations $u_i^e \in \mathbb{R}^{d_h}$ of the utterance x_i as the node representations of the utterance x_i in the graph. The emotion context is then modeled using Bi-LSTM.

$$c_i^e, h_i^e = \text{LSTM}^e\left(u_i^e, h_{i-1}^e\right) \tag{3}$$

Similarly, we obtain the topic context c_i^t and user profile context c_i^s.

2.4.2 Inter-view Hypergraph Fusion. Following the learning of intro-view contextual representations, we obtain four-view representations c_i^x, c_i^e, c_i^t, and c_i^s corresponding to each utterance x_i. Each contextual representation is considered as a node, with the extracted contextual features serving as node representations. Hyperedges are established between the four views of nodes corresponding to different contextual representations for the same utterance, e.g., there is a hyperedge $r_i = \{v_i^x, v_i^e, v_i^t, v_i^s\}$ for the i-th utterance x_i. Through the construction of hyperedge relations, various contextual relationships can be effectively connected. Multiple contextual features and coupling features among views are fully integrated in the constructed hyperedge, which is convenient to make full use of multi-view information in the subsequent hypergraph convolution operation.

Specifically, multi-view contextual representations are fused using the hypergraph convolution operation, which is capable of capturing high-order relationships and enabling a more comprehensive extraction of in-depth semantics. The used hypergraph convolution operation is defined as follows:

$$V^{(l+1)} = D^{-1} \cdot H \cdot W_e \cdot B^{-1} \cdot H^T \cdot V^{(l)} \tag{4}$$

$$H(i,j) = \begin{cases} 1, & \text{if node } i \text{ is in hyperedge } j \\ 0, & \text{if node } i \text{ is not in hyperedge } j \end{cases} \tag{5}$$

where $H \in \mathbb{R}^{4N \times N}$ represents the association matrix indicating the relationship between nodes and hyperedges, defined as in Eq.(5). $D \in \mathbb{R}^{4N \times 4N}$ represents the node degree matrix, $B \in \mathbb{R}^{N \times N}$ represents the edge degree matrix, the matrices D and B are diagonal matrix with $D_{jj} = \sum_i H_{ij}$ and $B_{ii} = \sum_j H_{ij}$. $W_e \in \mathbb{R}^{N \times N}$ represents the edge weight matrix, which is an identity matrix because each hyperedge is assigned equal importance. $V \in \mathbb{R}^{4N \times 2d_u}$ represents the node representation. After multiple layers of convolutions, the resulting nodes I are represented as follows, where Z is the number of convolution layers.

$$I = \frac{1}{Z+1} \sum_{z=0}^{Z} V^{(z)} \tag{6}$$

2.5 Malevolence Label Prediction

Based on the fully fused contextual representations, a classifier is employed to predict the malevolence labels of utterances.

$$\hat{y}_i = \text{soft max}\left(WI_i + b\right) \tag{7}$$

where W and b are trainable parameters. The cross-entropy training loss is adopted and calculated as follows:

$$Loss = -\frac{1}{\sum_{l=1}^{L} \tau(l)} \sum_{i=1}^{L} \sum_{k=1}^{\tau(i)} y_{i,k}^l \log\left(\hat{y}_{i,k}^l\right) \tag{8}$$

where L is the total number of dialogues, and $\tau(i)$ represents the number of utterances in the l-th dialogue. $y_{i,k}^l$ and $\hat{y}_{i,k}^l$ denote the one-hot and probability feature representations for the k-th malevolence label of the i-th utterance in the dialogue l, respectively.

3 Experiment

3.1 Experimental Settings

We conduct experiments on two benchmark datasets: MDRDC [12] and Dialogue Safety [7]. MDRDC, sourced from Twitter, includes eighteen malevolence labels. The Dialogue Safety dataset, sourced from a Chinese psychological counseling platform, includes eight unsafe dialogue labels. The training, validation, and test sets of these datasets maintain their original division in the ratio of 7:1:2 and 8:1:1, respectively. We compared with four BERT-based models, Pre-trained BERT, RoBERTa, BERT-CRF [2] and the state-of-the-art model BERT-MCRF [11], and two large language models, ChatGPT [6] and Flan-T5 [10]. We utilized GPT-3.5-turbo-0613 with *temperature* and *top_p* both set to 1.0, and Flan-T5 in its XL version. Since Flan-T5 is trained on English data, comparison with Flan-T5 is solely done on the English MDRDC dataset. Due to space limit, hyperparameter settings will be public in our code upon acceptance.

3.2 Experimental Results

We report the comparisons of experimental results in Table 1 evaluated by precision (P), recall (R), and macro-F1 (F1). It is observed that Flan-T5 and ChatGPT achieved inferior performance on both datasets, significantly underperforming other models, while RoBERTa and BERT exhibited better performance than Flan-T5 and ChatGPT, demonstrating their better adaptability for malevolence detection in dialogues. In cases where context is not considered, BERT outperformed RoBERTa. Among the baseline models, BERT-MCRF

achieved the highest performance by concurrently considering label correlation in taxonomy and label correlation in context. Furthermore, our MVPDetect model achieved the best performance, which demonstrates MVPDetect's powerful capability in modeling multi-view contextual information through prompt learning for malevolence detection in dialogues.

Table 1: Main results on MDRDC and Dialogue Safety.

Methods	MDRDC			Dialogue Safety		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BERT	51.21	54.93	53.00	45.91	46.73	45.51
BERT-CRF	52.68	55.30	53.96	46.55	47.36	46.39
BERT-MCRF	53.65	56.02	54.99	47.37	48.06	47.22
Roberta	52.69	55.59	52.45	**47.63**	41.13	43.28
ChatGPT	29.75	30.49	24.86	24.52	32.34	21.15
Flan-T5	24.87	32.49	24.59	-	-	-
MVPDetect	**55.66**	**58.95**	**56.71**	46.48	**51.20**	**48.20**

3.3 Ablation Study

We conduct ablation studies to investigate the impact of different context modelings, GCN learning and hypergraph fusion, respectively, and report the ablated results solely on the MDRDC dataset in Table 2 due to space limit.

Table 2: Ablation study on MDRDC.

Methods	P(%)	R(%)	F1(%)
MVPDetect	55.66	58.95	56.71
-Emotion context	54.41	58.70	56.08
-Topic context	52.83	59.28	55.51
-User profile context	52.26	59.72	55.14
-GCN+concatenate	42.76	65.06	50.05
-GCN+add	52.95	58.65	55.06
-Hypergraph+concatenate	52.22	62.28	56.50
-Hypergraph+add	53.30	60.84	56.10

3.3.1 Impact of Different Context Modeling. Removing the emotion, topic, or user profile prompt from our model leaded to a various decrease in precision and macro-F1, accompanied by a slight increase in recall. This indicates the significant impact of each view of prompts on the overall performance of our model.

3.3.2 Impact of GCN Learning. Removing the GCN, we substituted it by directly concatenating or adding the prompts and utterances. The experimental results show a significant decline in performance, indicating that the GCN fusion provides useful semantic information for malevolence detection in dialogues.

3.3.3 Impact of Hypergraph Fusion. Removing the hypergraph, we substituted it by directly concatenating or adding different views of contextual representations and feeding them into the final classifier. The experimental results suggest that the hypergraph plays a crucial role in integrating different views of contextual information, facilitating a comprehensive fusion from multiple views.

Case 1
Person A : How was your weekend? [**Ground-truth**: *non-malevolence*]
[**Ground-truth**: *non-malevolence*; **MPHDetect**: *non-malevolence*; **BERT-MCRF**: *non-malevolence*]
Person B : It was great! I went hiking with some friends. And you?
[**Ground-truth**: *non-malevolence*; **MPHDetect**: *non-malevolence*; **BERT-MCRF**: *non-malevolence*]
Person A : I spent most of it studying for exams, really stressed.
[**Ground-truth**: *non-malevolence*; **MPHDetect**: *non-malevolence*; **BERT-MCRF**: *non-malevolence*]
Person B : Oh, you always seem to be behind on your work.. Good luck!
[**Ground-truth**: *blame*; **MPHDetect**: *blame*; **BERT-MCRF**: *blame*]
Person A : Well, thanks for the vote of confidence...
[**Ground-truth**: *disgust*; **MPHDetect**: *disgust*; **BERT-MCRF**: *non-malevolence*]

Case 2
Person A : I am volunteering at the animal shelter.
[**Ground-truth**: *non-malevolence*; **MPHDetect**: *non-malevolence*; **BERT-MCRF**: *non-malevolence*]
Person B : Oh, you're one of those overly cheerful do-gooders, aren't you?
[**Ground-truth**: *disgust*; **MPHDetect**: *disgust*; **BERT-MCRF**: *non-malevolence*]
Person A : I suppose you could say that. I just like helping out.
[**Ground-truth**: *non-malevolence*; **MPHDetect**: *non-malevolence*; **BERT-MCRF**: *non-malevolence*]
Person B : Hope the animals' lack of appreciation won't disappoint you much...
[**Ground-truth**: *blame*; **MPHDetect**: *blame*; **BERT-MCRF**: *blame*]

Figure 3: Two cases with ground-truth (red) and prediction labels of MPHDetect (green) and BERT-MCRF (blue).

3.4 Case Study

We present two case studies in Fig. 3 to further illustrate how MVPDetect improves malevolence detection in dialogues. In *Case 1*, Person A's last utterance "*Well, thanks for the vote of confidence*" is indicative of a tone of disgust when considering the context. While BERT-MCRF fails to accurately predict the label, MVPDetect correctly identify it. In *Case 2*, Person B's utterance "*Oh, you're one of those overly cheerful do-gooders, aren't you?*" conveys a tone of disgust. Our model also correctly predicts its label, demonstrating its superiority over BERT-MCRF. This demonstrates that our model comprehensively makes the best use of the dialogue context, and provides more useful evidence for malevolence detection.

4 Conclusion

In this paper, we propose a multi-view prompting fusion model for malevolence detection in dialogues. Our model generates three types of prompts for capture malevolence-invoking semantics, and integrates four-view information for contextual modelings: the interaction context, emotion context, topic context and user profile context. Ultimately, hypergraph is introduced to effectively aggregate the four-view contextual information, generating high-order contextual semantics for detecting malevolence subtleties in utterances. Experimental results on two datasets demonstrate the superiority of our model over state-of-the-art models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62006034), the Ministry of Education Humanities and Social Science Project (No.22YJC740110), the Fundamental Research Funds for the Central Universities (DUT23YG136, DUT24MS003), and the Liaoning Provincial Natural Science Foundation Joint Fund Program(2023-MSBA-003).

References

[1] Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark O. Riedl. 2021. Just say no: analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Association for Computational Linguistics, 4846–4862. DOI: 10.18653/V1/2021.EMNLP-MAIN.397.

[2] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S. Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 3691–3697. DOI: 10.18653/V1/D19-1383.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. DOI: 10.18653/V1/N19-1423.

[4] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SJU4ayYgl.

[5] Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 16235–16250. DOI: 10.18653/V1/2023.ACL-LONG.898.

[6] Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. http://papers.nips.cc/paper%5C_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

[7] Huachuan Qiu, Tong Zhao, Anqi Li, Shuai Zhang, Hongliang He, and Zhenzhong Lan. 2023. A benchmark for understanding dialogue safety in mental health support. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part II* (Lecture Notes in Computer Science). Vol. 14303. Springer, 1–13. DOI: 10.1007/978-3-031-44696-2_1.

[8] Omar Shaikh, Hongxin Zhang, William Held, Michael S. Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 4454–4470. DOI: 10.18653/V1/2023.ACL-LONG.244.

[9] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "nice try, kiddo": investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Association for Computational Linguistics, 750–767. DOI: 10.18653/V1/2021.NAACL-MAIN.60.

[10] Siddharth Suresh, Kushin Mukherjee, and Timothy T. Rogers. 2023. Semantic feature verification in FLAN-T5. In *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023*. OpenReview.net. https://openreview.net/pdf?id=%5C_1z2Bqte5L.

[11] Yangjun Zhang, Pengjie Ren, Wentao Deng, Zhumin Chen, and Maarten de Rijke. 2022. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Association for Computational Linguistics, 3543–3555. DOI: 10.18653/V1/2022.ACL-LONG.248.

[12] Yangjun Zhang, Pengjie Ren, and Maarten de Rijke. 2021. A taxonomy, data set, and benchmark for detecting and classifying malevolent dialogue responses. *J. Assoc. Inf. Sci. Technol.*, 72, 12, 1477–1497. DOI: 10.1002/ASI.24496.