2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822192

# Modeling Implicit Emotion and User-specific Context for Malevolence Detection in Mental Health Counseling Dialogues

Bo Xu, Xuening Qiao, Xiaokun Zhang
*School of Computer Science*
*Dalian University of Technology*
Dalian, China
xubo@dlut.edu.cn
{qiao, kun}@mail.dlut.edu.cn

Jiahui Wan, Xinyue Liu, Linlin Zong*
*School of Software*
*Dalian University of Technology*
Dalian, China
jhwan@mail.dlut.edu.cn
{xyliu, llzong}@dlut.edu.cn

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

***Abstract*—Generative conversational agents, driven by large language models, have gained widespread popularity. However, a significant drawback lies in their tendency to produce uncontrollable and unpredictable contents, thereby increasing the risk of generating malevolent responses that potentially exacerbate users' mental health issues. Although existing research on malevolence detection in dialogues addressed the modeling of interaction patterns in dialogues, the implicitly expressed emotion and user-specific context are often neglected. Addressing this gap, we propose a hypergraph-enhanced context modeling approach for detecting malevolence in mental health counseling dialogues. Our approach harnesses the emotion reasoning capabilities of large language models to generate implicit emotional prompts. Employing hypergraph neural networks, our approach effectively integrates emotional context, user-specific context, and interactive context, fusing them into high-order semantic representations using hypergraph convolution. Experimental results on two benchmark datasets, MDRDC and Dialogue Safety, demonstrate the superiority of our model over state-of-the-art baseline models, particularly in complex contextual scenarios.**

***Index Terms*—malevolence detection, hypergraph learning, emotion analysis, user profiling, mental health counseling.**

## I. Introduction

Large language models (LLMs) have gained extensive attention, particularly in the domain of generative conversational agents [1]. Since the training data for LLMs are typically sourced from extensive Internet text, covering diverse information, these models may inadvertently learn aggressive, harmful, or biased contents. This inadvertent learning can lead to the generation of malevolent responses during conversations, causing discomfort or even disputes for the interlocutors [2]. Furthermore, malevolent responses can worsen mental health issues within conversations [3]. Incidents have been documented where users with mental health issues engaged in conversations with the GPT-3 model, resulting in the model providing dangerously suggestive contents encouraging self-harm [4]. For psychologically unstable users, malevolent responses can exacerbate depression and even induce suicidal tendencies [5]. Therefore, it is imperative to detect malevolence in mental health counseling dialogues.

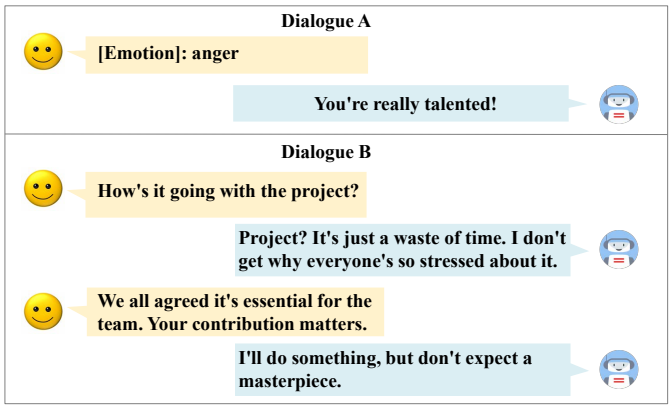


Fig. 1: Two example dialogues showing the utility of user-specific context and implicit emotions in malevolence detection.

While research has focused on detecting toxic or offensive language [6]–[8], the inherently interactive nature of dialogues introduces an increased subtlety and complexity in malevolence detection. This complexity arises from implicit expressions of malevolence scattered throughout utterances in counseling dialogues, posing a great challenge in malevolence detection. Although existing research has sought to model the interaction patterns for malevolence detection [9], [10], the impact of *implicitly expressed emotions* and *user-specific context* on malevolence detection has been mostly overlooked.

**Implicitly expressed emotions** in mental health counseling dialogues play a crucial role in identifying the malevolence in dialogues. Subtle changes in user emotions can profoundly impact the understanding and generation of dialogue content, introducing notable biases in semantic comprehension. This, in turn, may lead to the misjudgment of malevolent utterances. As illustrated in Dialogue A in Fig.1, the statement 'You're really talented!' might initially appear complimentary when considered in isolation. However, when the user implicitly

*Corresponding Author

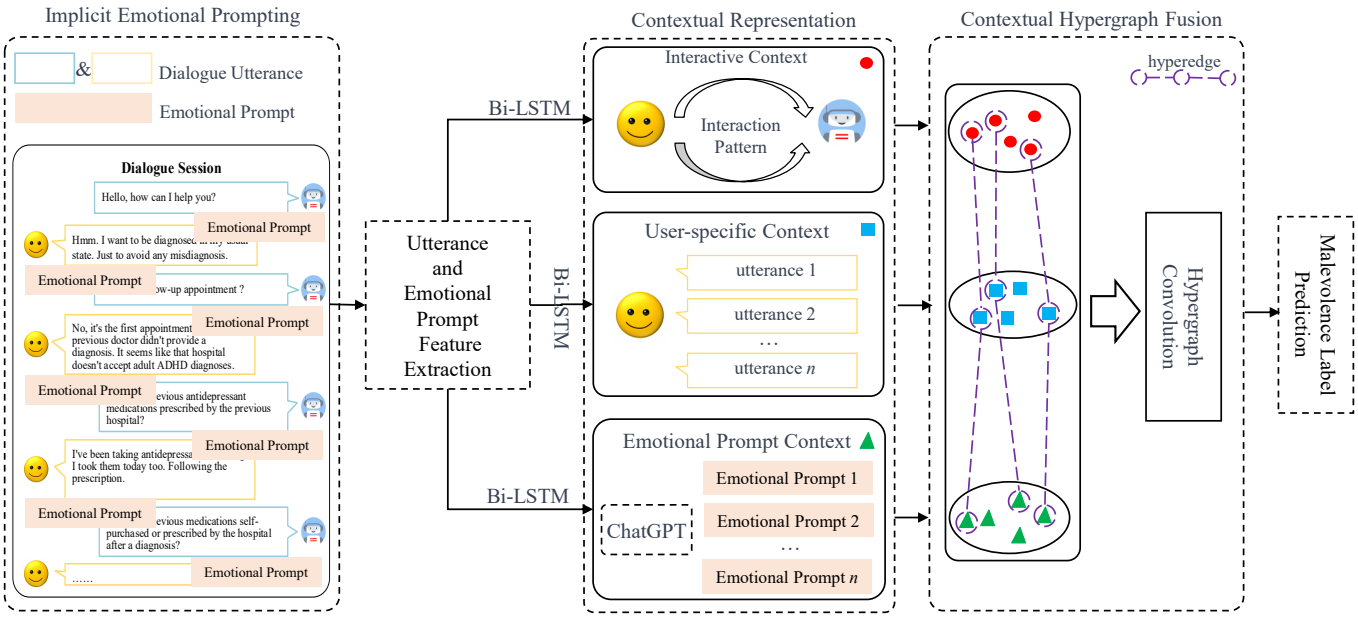


Fig. 2: The overall structure of our HyperCEP model. The implicit emotional prompting module generates LLM-based prompts to capture implicitly expressed emotions. The feature extraction module learns the semantic features of utterances and prompts. The contextual representation module captures contextual information from interactive context, user-specific context and emotional prompt context. The contextual hypergraph fusion module integrates the three types of contextual information with hypergraph convolutions. The malevolence label prediction module predicts category labels of malevolence.

expresses an anger emotion beforehand, the same sentence takes on a sarcastic and unfriendly undertone. Hence, modeling implicitly expressed emotions is pivotal in uncovering concealed malevolence in mental health counseling dialogues.

**User-specific context**, derived from all utterances of an individual user, encompasses statements, personalized traits and internal opinion dependencies. These elements are crucial components in user profiling and play a decisive role in deciphering inter-utterance inferred malevolence in mental health counseling dialogues. As shown in Dialogue B in Fig.1, the chatbot consistently lacks enthusiasm in all responses, reflecting a negative attitude towards the project. Modeling all utterances helps recognize the persistent dismissive behavior, thereby revealing the potentially malevolent intent throughout this conversation.

To address the aforementioned challenges, we propose to unitedly model implicit emotions and user-specific context beyond the interaction context for malevolence detection in mental health counseling dialogues. In discerning implicitly expressed emotions, we devise a prompt strategy based on large language models to infer emotional subtleties. A well-crafted prompt guides the language model to focus on emotion-specific aspects throughout the entire dialogue, facilitating the recognition of implicit emotional cues in user utterances. To capture user-specific context, our focus is solely on the user-generated utterances within specific dialogue sessions. By modeling the utterances of each individual user, we effectively capture and model personalized characteristics, thereby aiding in the identification of the potentially malevolent utterances.

## II. Methodology

### A. Overview

The primary objective of the malevolence detection task is to identify malevolent utterances within a dialogue session. Given a dialogue session $D = \{x_1, x_2, ..., x_N\}$ consisting of $N$ utterances, our task is to predict the malevolence label of each utterance. To this end, our model encapsulates intricate contextual information, including implicitly expressed emotions, user profiles and interaction patterns in mental health counseling dialogues. The overall structure of our model is illustrated in Figure 2, comprising five components.

The implicit emotional prompting module utilizes the LLM-based prompting learning strategy to capture implicitly expressed emotions. The mental health dialogue feature extraction module learns the semantic features of dialogue utterances and the generated emotional prompts. The multi-view contextual representation module captures contextual information from three views—implicit emotional prompt context, user-specific context and interactive context. The contextual hypergraph fusion module connects and integrates the three types of contextual information. The malevolence label prediction module predicts category labels based on hypergraph fused representations.

### B. Implicit Emotional Prompting Module

In mental health counseling dialogues, users frequently manifest implicit emotions characterized by subtle changes. These nuanced emotional expressions are typically rooted in the potential semantics within the interlocutor's wording or expression, playing a crucial role in detecting malevolence

in dialogues. Therefore, the model's ability in capturing constantly fluctuated emotions within a dialogue session proves valuable in pinpointing the origins of malevolence that align with the user's emotional state. To capture implicitly expressed emotions, we propose using an LLM-based prompting learning strategy that mimic human thinking and reasoning processes by considering LLM's emerged outstanding common-sense reasoning abilities [11], [12]. By providing emotional prompt cues, we seek to effectively enhance the model's detection performance on dialogue malevolence.

TABLE I: The prompt template for generating implicit emotional responses, where {} denotes content to be replaced with real data. For prompt generation, we employed the gpt_3.5_turbo model with max_tokens set to 60.

| Variable | Template Content |
|---|---|
| system_role_content | role: *system*; content: *The given data consists of a conversation between two individuals. Now you are required to role-play as the speaker to analyze the implicit emotion within the statements.* |
| prompt_template | Based on the dialogue history {*dialogue_history*}, you are needed to play the role of the person {*speaker*}. Express the implicit emotion of the following statement in the speaker's tone: {*speaker*}: {*utterance*} |

Specifically, to generate implicit emotional prompts tailored to malevolence detection in mental health counseling dialogues, we devise a prompt template, as illustrated in Table I. In this template, the $dialogue\_history$ encompasses $i-1$ utterances denoted as $D_i = \{x_1, x_2, ..., x_{i-1}\}$ during $i$ rounds of conversations. After prompt learning, each utterance will correspond to a generated implicit emotional prompt denoted as $P = \{p_1, p_2, ..., p_N\}$. By harnessing large language models to amplify implicit emotional semantics, The generated prompts introduce intermediate reasoning cues, thereby fortifying the model's capability to distinguish different types of malevolent utterances. In our implementations, we employ ChatGPT to generate the emotional prompts. The configuration for prompting is outlined in Table I.

### C. Mental Health Dialogue Feature Extraction Module

We utilize the pre-trained language model BERT [13] to extract dialogue feature representations. Initially, on each used dataset, we fine-tune the pre-trained BERT model at the utterance level. Subsequently, we employ the fine-tuned model to extract utterance features from the dialogue $D = \{x_1, x_2, ..., x_N\}$ and the emotional prompts $P = \{p_1, p_2, ..., p_N\}$. Taken a given utterance representation for example, a special token [*CLS*] is added at the beginning of the sequence, making the model's input represented as a sequence {[*CLS*], $x_i^1$, $x_i^2$, ..., $x_i^n$}. Finally, we extract the 768-dimensional embedding from the [*CLS*] token in the last layer of the model as the utterance feature representation $u_i$ for the utterance $x_i$, where $u_i \in \mathbb{R}^{d_h}$ and $d_h = 768$.

### D. Multi-view Contextual Representation Module

Since counseling dialogue context encapsulates intricate semantic information, a thorough modeling of the dialogue context will be useful for detecting malevolence in mental health counseling dialogues. In our approach, three distinct views of contextual representations are considered: implicit emotional prompt context, user-specific context and interactive context. The implicit emotional prompt context is acquired from the prompts generated by LLMs. The user-specific and interactive context are derived from dialogue content, capturing the dynamics of the ongoing interaction and individual user characteristics, respectively. This three-view consideration of contextual representations enables our approach to encompass diverse dimensions of contextual information, fostering a more comprehensive understanding of the dialogue. Drawing inspiration from [14], we employ three Bi-LSTM networks [15] for multi-view contextual representations.

For the implicit emotional prompt contextual representation, Bi-LSTM models all utterances and the emotional prompts as united representations. The input and output feature representations are denoted as $p_i \in \mathbb{R}^{d_h}$ and $c_i^p \in \mathbb{R}^{2d_u}$, respectively.

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

where $h_i^p$ represent the $i$-th hidden state of the Bi-LSTM.

For the user-specific contextual representation, we exclusively consider user-generated utterances in the dialogue. Bi-LSTM models sequential dependencies among all utterances from the same user. Specifically, for the user-specific utterance feature representation $u_i \in \mathbb{R}^{d_h}$, the calculation of the user-specific contextual representation $c_i^s \in \mathbb{R}^{2d_u}$ is as follows:

$$c_i^s, h_{\lambda,j}^v = \overleftrightarrow{\text{LSTM}}^{\,s}\left(u_i, h_{\lambda,j-1}^s\right), j \in [1, |U_\lambda|] \tag{2}$$

where $\lambda = \phi(u_i)$. $U_\lambda$ denotes all utterances from the same user. $h_j \in \mathbb{R}^{d_u}$ is the $j$-th hidden state of the Bi-LSTM.

For the interactive contextual representation, Bi-LSTM models interaction patterns between the two interlocutors in specific dialogue sessions. Formally, for the last user-generated utterance, the feature representation is denoted as $u_i \in \mathbb{R}^{d_h}$. The interactive level aims to capture the contextual representations, $c_i^t \in \mathbb{R}^{2d_u}$,calculated as follows:

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

where $h_i^t$ represent the $i$-th hidden state of the Bi-LSTM.

### E. Contextual Hypergraph Fusion Module

Following the learning of multi-view contextual representations, we obtain three-view representations $c_i^p$, $c_i^s$, and $c_i^t$ corresponding to each utterance $x_i$. Subsequently, we leverage the common sequential relationships shared by the three features to construct a hypergraph network [16], connecting the three types of contextual representations. A hypergraph neural network is a type of graph structure that can capture complex relationships. Unlike traditional graphs, where an edge connects only two nodes, a hypergraph edge can connect any number of nodes, allowing for more accurate modeling

of diverse relationships. By constructing hyperedges between the different contextual representations of the same utterance, semantic dependencies within the dialogue sessions can be well established.

Specifically, each contextual representation for each utterance is considered as a node in the hypergraph, with the extracted contextual features serving as node representations. Hyperedges are established between the three types of nodes corresponding to different contextual representations for the same utterance. The constructed hypergraph denoted as $G(V, E)$ aims to encode the contextual dependency relationships. For the $i$-th utterance $x_i$, the three different contextual representations form three nodes, represented as $v_i^p$, $v_i^s$, and $v_i^t$. Based on the sequential dependency relationships between the three types of contexts of this utterance, the constructed hyperedge relation $e_i$ is represented as follows:

$$e_i = \{v_i^p, v_i^s, v_i^t\}.$$

Through the construction of hyperedge relations, various contextual relationships can be effectively connected. Subsequently, based on the structure of the hypergraph, multi-view contextual representations are fused using hypergraph convolution operation [17], which is capable of capturing high-order relationships and enabling a more comprehensive extraction of deep semantic features. The unique graph structure incorporates the aggregation process from nodes to hyperedges and then from hyperedges to nodes. Specifically, for the hypergraph $G(V, E)$, the hypergraph convolution operation is defined as follows:

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

where $H \in \mathbb{R}^{3N \times N}$ represents the association matrix indicating the relationship between nodes and hyperedges, which is defined as in Eq.(5). $D \in \mathbb{R}^{3N \times 3N}$ 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. $V \in \mathbb{R}^{3N \times 2d_u}$ represents the node representation.

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

After multiple layers of convolution, the information from each node in each sample can aggregate information from other nodes in the same sample, capturing various malevolence categories of semantic information. The resulting nodes $I$ are represented as follows.

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

where $Z$ is the number of convolution layers.

### F. Malevolence Label Prediction Module

On one hand, the model extensively extracts multi-view semantic features within different contexts using Bi-LSTM. On the other hand, the hypergraph convolutional network effectively integrates contextual information from the prompting context, the user-specific context, and the interactive context, aiding the model in predicting malevolent utterances. Finally, based on the fully fused feature representations, a classifier is employed to predict malevolence labels for utterances.

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

where $W \in \mathbb{R}^{6d_u \times |Y|}$ and $b \in \mathbb{R}^{|Y|}$ are trainable parameters, and $|Y|$ is the number of labels. The model's loss function utilizes the cross-entropy training loss, with the specific formula shown 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 in the training set, 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 malevolent category label of the $i$-th utterance in the dialogue $l$, respectively.

## III. Experimental Setup

### A. Datasets

We conducted model evaluations on two related datasets: MDRDC [9] and Dialogue Safety [18]. MDRDC is specifically designed for malevolence detection in dialogues, derived from conversations on Twitter. This dataset comprises 6,000 multi-turn dialogues, consisting of 3,661 malevolent dialogues and 2,339 non-malevolent dialogues. Each dialogue contains 3-10 utterances, resulting in a total of 31,380 utterances, with 21,081 being non-malevolent and 10,299 being malevolent. It includes 18 labels for malevolent categories, including non-malevolent, unconcernedness, detachment, blame, arrogance, anti-authority, dominance, deceit, negative intergroup attitude, violence, privacy invasion, obscenity, phobia, anger, jealousy, disgust, self-hurt, immoral and illegal. The original split [9] was maintained for the training, validation, and test sets, with a ratio of 7:1:2. The Dialogue Safety dataset is tailored for identifying safe responses in mental health counseling dialogues, sourced from a Chinese online psychological counseling platform. It comprises 7,925 multi-turn dialogues, encompassing a total of 3,658 safe responses. This dataset includes eight categories for unsafe dialogue labels: safe response, nonsense, humanoid mimicry, linguistic neglect, unamiable judgment, toxic language, unauthorized preachment, and nonfactual statement. The original data are initially split into a training set and a test set with a ratio of 9:1, lacking a separate validation set. To address this, we further divide the training set into a new training set and a validation set, maintaining an 8:1 ratio. The details on dataset divisions are shown in Table II.

We compare our model with the following baselines, including four BERT-based models and two large language models.

TABLE II: Divisions of MDRDC and Dialogue Safety.

| Dataset | Train | Valid | Test |
|---|---|---|---|
| MDRDC | 20049 | 2788 | 5673 |
| Dialogue Safety | 41619 | 4594 | 5147 |

Firstly, we compare with four BERT-based models. Pre-trained BERT and RoBERTa models have previously demonstrated superior performance in malevolence detection tasks, showcasing their proficiency in learning contextual information [9]. BERT-CRF, equipped with a modified encoder for separate utterances, has excelled in sequence labeling tasks [19]. BERT-MCRF, incorporating multi-faceted label correlation with enhanced CRF, stands as the state-of-the-art model for malevolence detection in dialogues [10]. We fine-tune these models on the same experimental setting for fair comparisons. Additionally, we compare our model with two large language models, ChatGPT [20] and Flan-T5 [21]. ChatGPT, designed for supporting conversational interactions, is invoked using its API with zero-shot prompts. We repeatedly call it until the expected response is generated, utilizing GPT-3.5-turbo-0613 with *temperature* and $top\_p$ both set to 1.0. Flan-T5, a text-to-text framework pre-trained by Google suitable for handling sequence data, is used in its XL version. We invoke this model using zero-shot prompts for malevolence detection. It's important to note that because Flan-T5 is only applicable to English data, its comparison with our model is solely conducted on the MDRDC dataset.

## IV. Results and Analysis

### A. Experimental Results

The comparative results of our model against other models are presented in Table III and Table IV. Overall, the experimental results demonstrate that our HyperCEP model outperforms all baseline models on both datasets.

TABLE III: Main results of HyperCEP on MDRDC.

| Methods | Precision | Recall | Macro-F1 |
|---|---|---|---|
| BERT | 51.21 | 54.93 | 53.00 |
| BERT-CRF | 52.68 | 55.30 | 53.96 |
| BERT-MCRF | 53.65 | 56.02 | 54.99 |
| Roberta | 52.69 | 55.59 | 52.45 |
| ChatGPT | 29.75 | 30.49 | 24.86 |
| Flan-T5 | 24.87 | 32.49 | 24.59 |
| HyperCEP | **56.77** | **57.19** | **56.71** |

TABLE IV: Main results of HyperCEP on Dialogue Safety.

| Methods | Precision | Recall | Macro-F1 |
|---|---|---|---|
| BERT | 45.91 | 46.73 | 45.51 |
| BERT-CRF | 46.55 | 47.36 | 46.39 |
| BERT-MCRF | 47.37 | 48.06 | 47.22 |
| Roberta | 47.63 | 41.13 | 43.28 |
| ChatGPT | 24.52 | 32.34 | 21.15 |
| HyperCEP | **49.41** | **50.71** | **49.93** |

From the tables, we observe that Flan-T5 and ChatGPT, being general generative language models, exhibit similar and significantly lower performance than other models when confronted with the malevolence detection task with intricate dialogue context. This highlights the generative models' limitations in specific downstream tasks. Regarding macro-F1 on the MDRDC dataset, ChatGPT slightly outperforms the Flan-T5 model by 0.27%, potentially because ChatGPT is trained on dialogue data, leading to better context understanding capability. Both RoBERTa and BERT demonstrate superior performance compared to Flan-T5 and ChatGPT, showcasing the stronger adaptability of pre-trained language models to malevolence detection. In scenarios without context, the BERT model is notably better than the RoBERTa model. Among all the baseline models, BERT-MCRF outperformed other models by concurrently considering label correlation in taxonomy and label correlation in context.

Furthermore, our HyperCEP model achieves the best performance among all other models. On these two datasets, the macro-F1 of HyperCEP surpasses the best baseline model BERT-MCRF by 3.13% and 5.74%, respectively. This finding validates the outstanding performance of HyperCEP in modeling diverse contextual information, emphasizing the effectiveness of incorporating emotional and user-specific information in mental health counseling dialogues. The emotional prompting strategy, simulating human reasoning processes, enhances the model's ability to extract implicitly expressed emotions, ultimately improving its overall performance in detecting malevolence.

### B. Ablation Study

We perform an ablation study on HyperCEP by removing prompting context, user-specific context, interactive context and the hypergraph. The results, reported in Table V and Table VI, suggest all three context modelings and the hypergraph are important for HyperCEP.

TABLE V: Ablation study on MDRDC.

| Methods | Precision | Recall | Macro-F1 |
|---|---|---|---|
| HyperCEP | 56.77 | 57.19 | 56.71 |
| -Prompting context | 53.79 | 58.93 | 55.75 |
| -User-specific context | 55.01 | 59.07 | 56.38 |
| -Interactive context | 55.22 | 57.44 | 55.98 |
| -Hypergraph | 54.85 | 56.39 | 55.21 |

TABLE VI: Ablation study on Dialogue Safety.

| Methods | Precision | Recall | Macro-F1 |
|---|---|---|---|
| HyperCEP | 49.41 | 50.71 | 49.93 |
| -Prompting context | 45.48 | 52.03 | 48.08 |
| -User-specific context | 43.08 | 50.81 | 45.63 |
| -Interactive context | 43.48 | 51.76 | 46.54 |
| -Hypergraph | 40.89 | 52.47 | 44.84 |

**Impact of Different Context Modeling.** After removing each view of context modeling, the model's performance exhibits varying degrees of decline in precision and macro-F1, with a slight improvement in recall.

**Impact of Hypergraph Feature Fusion.** By removing the hypergraph, we substituted it by directly concatenating different views of contextual representations and feeding them into the final classifier. The experimental results indicate that removing the contextual hypergraph fusion module results in a substantial performance decline on both datasets.

## V. Conclusion

We propose a hypergraph-enhanced model, HyperCEP, for malevolence detection in mental health counseling dialogues. Our model integrates three views of context modelings: emotional prompt context, user-specific context and interactive context. Emotional prompt context leverages large language models to generate informative prompts containing implicit emotional cues, facilitating the model in learning human-like emotion reasoning. User-specific and interactive context captures profound semantic dependencies in each individual's utterances and interlocutors' interaction patterns, respectively. Ultimately, hypergraph convolution effectively aggregated three-view contextual information, generating high-order contextual semantics for malevolence detection in dialogues. Experimental results on two datasets, MDRDC and Dialogue Safety, demonstrated the effectiveness and superiority of our HyperCEP model over state-of-the-art baseline models.

## VI. Acknowledgements

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] Y. Gu, J. Wen, H. Sun, Y. Song, P. Ke, C. Zheng, Z. Zhang, J. Yao, L. Liu, X. Zhu, and M. Huang, "EVA2.0: investigating open-domain chinese dialogue systems with large-scale pre-training," *Mach. Intell. Res.*, vol. 20, no. 2, pp. 207–219, 2023.

[2] J. Park, "Linguistic politeness and face-work in computer-mediated communication, part 1: A theoretical framework," *J. Assoc. Inf. Sci. Technol.*, vol. 59, no. 13, pp. 2051–2059, 2008.

[3] A. A. Abd-alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, and M. S. Househ, "An overview of the features of chatbots in mental health: A scoping review," *Int. J. Medical Informatics*, vol. 132, 2019.

[4] A. Baheti, M. Sap, A. Ritter, and M. O. 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, EMNLP 2021*, 2021, pp. 4846–4862.

[5] Y. M. Cho, S. Rai, L. H. Ungar, J. Sedoc, and S. C. Guntuku, "An "integrative survey on mental health conversational agents to bridge computer science and medical perspectives"," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, 2023, pp. 11 346–11 369.

[6] E. Sheng, K. 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, NAACL-HLT 2021, Online, June 6-11, 2021*, 2021, pp. 750–767.

[7] J. Lu, B. Xu, X. Zhang, C. Min, L. Yang, and H. Lin, "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*, 2023, pp. 16 235–16 250.

[8] O. Shaikh, H. Zhang, W. Held, M. S. Bernstein, and D. Yang, "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*, 2023, pp. 4454–4470.

[9] Y. Zhang, P. Ren, and M. de Rijke, "A taxonomy, data set, and benchmark for detecting and classifying malevolent dialogue responses," *J. Assoc. Inf. Sci. Technol.*, vol. 72, no. 12, pp. 1477–1497, 2021.

[10] 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, ACL 2022*, 2022, pp. 3543–3555.

[11] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. L. Bras, Y. Choi, and H. Hajishirzi, "Generated knowledge prompting for commonsense reasoning," 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*, 2022, pp. 3154–3169.

[12] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," in *The Eleventh International Conference on Learning Representations, ICLR 2023*, 2023.

[13] J. Devlin, M. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[14] D. Hu, L. Wei, and X. Huai, "Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 2021, pp. 7042–7052.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in *Neural Information Processing Systems*, 1997.

[16] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, 2019, pp. 3558–3565.

[17] N. Yadati, M. Nimishakavi, P. Yadav, V. Nitin, A. Louis, and P. P. Talukdar, "Hypergcn: A new method for training graph convolutional networks on hypergraphs," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 1509–1520.

[18] H. Qiu, T. Zhao, A. Li, S. Zhang, H. He, and Z. Lan, "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*, ser. Lecture Notes in Computer Science, vol. 14303. Springer, 2023, pp. 1–13.

[19] 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 2019, Hong Kong, China, November 3-7, 2019*, 2019, pp. 3691–3697.

[20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems, NeurIPS 2022*, 2022.

[21] S. Suresh, K. Mukherjee, and T. T. Rogers, "Semantic feature verification in FLAN-T5," in *The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023*, 2023.