Benchmarking Cyber Harassment Dialogue Comprehension through Emotion-Informed Manifestations-Determinants Demarcation

Soumitra Ghosh[1*], **Gopendra Vikram Singh**[2*], **Jashn Arora**[3], **Asif Ekbal**[2]

[1]NLP Research Group, Fondazione Bruno Kessler (FBK), Trento, Italy
[2]Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India
[3]International Institute of Information Technology, Hyderabad
{ghosh.soumitra2, gopendra.99}@gmail.com, jashn.arora@research.iiit.ac.in, asif@iitp.ac.in

Abstract

In the digital age, cybercrimes, particularly cyber harassment, have become pressing issues, targeting vulnerable individuals like children, teenagers, and women. Understanding the experiences and needs of the victims is crucial for effective support and intervention. Online conversations between victims and virtual harassment counselors (chatbots) offer valuable insights into cyber harassment manifestations (CHMs) and determinants (CHDs). However, the distinction between CHMs and CHDs remains unclear. This research is the first to introduce concrete definitions for CHMs and CHDs, investigating their distinction through automated methods to enable efficient cyber-harassment dialogue comprehension. We present a novel dataset, *Cyber-MaD* that contains Cyber harassment dialogues manually annotated with Manifestations and Determinants. Additionally, we design an Emotion-informed Contextual Dual attention Convolution Transformer *(E-ConDuCT)* framework to extract *CHMs* and *CHDs* from cyber harassment dialogues. The framework primarily: a) utilizes inherent emotion features through adjective-noun pairs modeled by an autoencoder, b) employs a unique Contextual Dual attention Convolution Transformer to learn contextual insights; and c) incorporates a demarcation module leveraging task-specific emotional knowledge and a discriminator loss function to differentiate manifestations and determinants. *E-ConDuCT* outperforms the state-of-the-art systems on the *Cyber-MaD* corpus, showcasing its potential in the extraction of *CHMs* and CHDs. Furthermore, its robustness is demonstrated on the emotion cause extraction task using the CARES_CEASE-v2.0 dataset of suicide notes, confirming its efficacy across diverse cause extraction objectives. Access the code and data at 1. https://www.iitp.ac.in/~ai-nlp-ml/resources.html#E-ConDuCT-on-Cyber-MaD, 2. https://github.com/Soumitra816/Manifestations-Determinants

Introduction

In recent years, the rise of the digital age has brought about significant advancements in communication and connectivity. However, along with these advancements, there has also been an alarming increase in cybercrimes, particularly targeting vulnerable individuals, such as children, teenagers, and women. Cyber harassment, in various forms, has become a pressing issue that demands attention from researchers, policymakers, and support organizations.

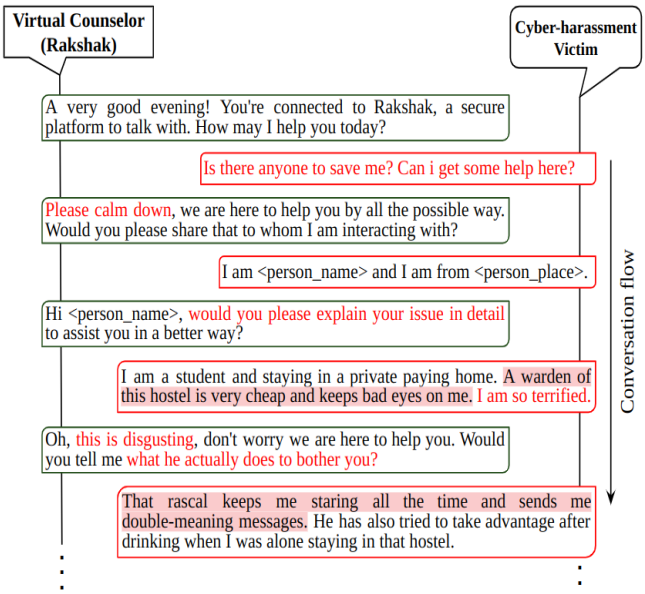

Figure 1: Example of a conversation snippet from the introduced *Cyber-MaD* Dataset. Texts in red font denote the CHMs. The highlighted spans denote the CHDs.

Understanding the needs and experiences of cyber harassment victims is crucial for effective support and intervention. Conversations between victims and counselors offer insights into the manifestations and determinants of online victimization. Cyber Harassment Manifestations (CHMs) and Cyber Harassment Determinants (CHDs) play distinct and critical roles in comprehending cyber harassment dialogues. After thorough and careful analysis of several cyber harassment resources (Dinakar, Reichart, and Lieberman 2011; Sprugnoli et al. 2018; Kim et al. 2021), we define CHMs as expressions, statements, or indicators that manifest the victim's distress, vulnerability, or the existence of an online harassment scenario. Conversely, CHDs pertain to the specific actions or behaviors undertaken by the perpetrator, which constitute the cyber harassment itself. Figure 1 illustrates the disparity between *CHMs* and *CHDs*.

*Both authors contributed equally and are co-first authors.

Understanding the significance of CHMs is crucial, as they provide insights into distress, fear, or victimization. Accurate identification of these manifestations helps counselors and support services promptly assist those in need. Conversely, recognizing CHDs is vital to comprehend the nature and severity of cyber harassment for legal actions. While identifying CHMs or CHDs independently matters, relying solely on one aspect can lead to an incomplete view. Manifestations reveal emotional states and potential harassment, but not their complexity. Determinants offer insights into tactics used, yet may miss victim distress. Simultaneous identification of both in cyber-harassment dialogues offers a comprehensive picture. Merging both aspects aids a holistic understanding, improving strategies for support, intervention, and prevention. This informs law enforcement, empowers victims, and shapes effective policies and awareness campaigns against cybercrimes.

The primary contributions or key attributes of our current work are summarized as follows:

1. Investigate cyber harassment comprehension, considering both manifestations and determinants, with concrete definition and automated methods.
2. Introduction of a pioneering *Cyber-MaD* corpus featuring manually annotated spans of manifestations and determinants in cyber-harassment dialogues.
3. Development of the innovative *E-ConDuCT* framework, effectively modeling cyber-harassment dialogues using emotion features, a Contextual Dual attention Convolution Transformer, and a demarcation module.
4. *E-ConDuCT* outperforms state-of-the-art systems, showcasing its potential in cyber-harassment comprehension on the *Cyber-MaD* corpus and demonstrates robustness on the CARES_CEASE-v2.0 dataset for emotion cause extraction, confirming efficacy in diverse cause extraction objectives.
5. We release the code and data to facilitate further research in this area.

Identifying manifestations and determinants in cyber-harassment dialogues holds significant social relevance.

- *Holistic Approach:* This research takes a holistic approach, considering both manifestations (expressing distress, vulnerability, or harassment) and determinants (pertaining to the perpetrator's actions) to comprehensively understand cybercrime situations.
- *Support and Intervention Focus:* By recognizing distress manifestations, support services can promptly assist individuals in need, ensuring their safety and well-being. Understanding cyber harassment determinants aids in comprehending the crimes' nature and severity, enabling effective legal actions and interventions.
- *Comprehensive Insights:* Identifying manifestations and determinants offers holistic insights into victims' emotions, crime characteristics, and perpetrator behaviors, empowering stakeholders to address well-being and legal aspects effectively. This leads to improved support, interventions, preventive measures, policies, and awareness campaigns to combat cybercrimes.

Related Work

The field of cybercrime analysis using computational methods has gained considerable attention recently. Addressing these issues necessitates adept analysis of online data through computational means. In this section, we assess existing research and methodologies applied to cybercrime and harassment analysis through computational techniques.

Previous research involves analyzing social media data to understand cybercrime and harassment dynamics (Gonzales 2014). Social network analysis techniques are often used to explore online community structures, identify influential users, and uncover information flow patterns (Boukhtouta et al. 2015; Paracha, Arshad, and Khan 2023). These analyses reveal mechanisms driving cybercrime propagation and the social dynamics behind their occurrence.

Another line of research focuses on the detection and identification of cybercrimes and harassment incidents. Machine learning and data mining techniques have been utilized to analyze large datasets and identify patterns indicative of different forms of cybercrimes, including online harassment, cyberbullying, and fraud (Andleeb et al. 2019; Al-Garadi et al. 2019; Ali, Mohd, and Fauzi 2021). These studies often employ supervised learning algorithms to train classifiers that can automatically detect and categorize instances of cybercrimes based on features, such as text content, user behavior, or network traffic (Talpur and O'Sullivan 2020; Mahor et al. 2021). Text mining and natural language processing techniques have been extensively utilized for cybercrime analysis. Sentiment analysis (Ghosh et al. 2023), topic modeling (Markov et al. 2023), and information extraction (Wiegand, Ruppenhofer, and Kleinbauer 2019) methods have been employed to identify and understand the underlying motivations, intents, and emotions expressed in online conversations related to cybercrime and harassment. These techniques enable researchers to uncover important insights into the psychological aspects of perpetrators and victims, the context of cybercrimes, and the underlying determinants and motivations behind online harassment incidents.

Despite notable advancements in the analysis of cybercrime and harassment, there exists a research gap in fully comprehending the intricacies of language and behavior choices in online interactions. Our proposed work seeks to bridge this gap by utilizing advanced computational methods, including AI techniques, to detect manifestations and determinants in online cyber harassment dialogues. By analyzing language and behavior, our research aims to enhance comprehension of cybercrime incidents and enable more effective prevention and intervention strategies.

Dataset

We introduce the *Cyber-MaD* corpus, a distinctive dataset of manually annotated *(Cyber harassment dialogues with Manifestations and Determinants)* at the utterance level. Our dataset is derived from the Mental Health and Legal Counseling Dataset (MHLCD) (Mishra, Priya, and Ekbal 2023), originally designed for mental health and legal assistance for crime victims.

Dataset Annotation

Three expert annotators[1] were selected for robust and accurate annotations. Their strong background in natural language processing and computational social science, along with expertise in sentiment analysis, emotion-aware text processing, and cybercrime detection, made them well-suited for the task. Notably, the annotators had no prior exposure to the conversations or research goals, ensuring impartial annotations solely rooted in the dialogue content. Clear definitions of manifestations and determinants were provided, along with instructions to extract relevant phrases while excluding irrelevant or noisy words. Inspired by prior work (Poria et al. 2021; Ghosh et al. 2022), we identified up to 2 causal spans for manifestations and determinants in each utterance, as most utterances typically contained one determinant, and a few had two or more. Aggregating spans based on (Gui et al. 2016), the final causal span for an utterance was determined. We assessed inter-rater agreement using the macro-F1 metric, as established in earlier span extraction research (Poria et al. 2021; Ghosh et al. 2022). The resulting F1-scores of 0.79 for CHM and 0.81 for CHD span annotations underscored the quality of annotations.

A perceptive method was used, utilizing the agent's responses to the victim's descriptions. Instances of empathy, sympathy, or concern from the agent were considered cues to identify manifestations and determinants in the preceding victim utterances (as illustrated in Figure 1). Agent's responses were treated as candidate utterances to explore manifestations and determinants, resulting in a more nuanced annotation of the dataset.

Dataset Statistics

Table 1 shows general statistics of the introduced dataset. The disparity in average lengths of manifestations and determinants provides insights into victims' communication patterns. The concise nature of manifestations suggests that victims instinctively employ brief expressions to swiftly convey their emotions and experiences, particularly during moments of distress. Conversely, relatively longer average determinants indicate complex perpetrator actions, prompting victims to provide detailed descriptions of the crime.

Attribute	Count
Total Dialogues	996
Total Utterances	27666
Average Utterance Length	18.3
Utterances with one manifestation span	7427
Utterances with two manifestation span	1220
Utterances with one determinant span	3260
Utterances with two determinant span	410
Manifestation Length (words)	9
Determinant Length (words)	14.9

Table 1: General Dataset Statistics

[1]two Ph.D. linguists and one Ph.D. computer science student

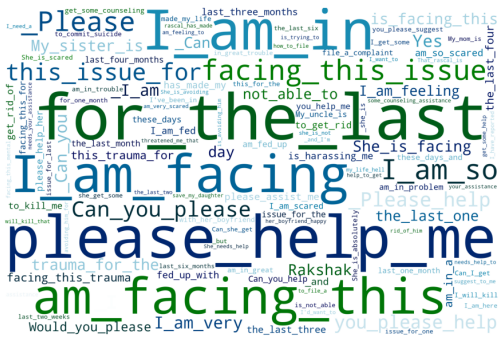

Figure 2: Word Cloud from the annotated manifestations.

Figure 3: Word Cloud from the annotated determinants

Distinctive Observations

We generate separate word clouds for annotated manifestations and determinants as shown in Figures 2 and 3 that offer a visual representation of the prominent themes and patterns in the spans of the annotated manifestations and determinants. Particularly, they illustrate how manifestations represent victim distress, vulnerability, or online harassment, while determinants depict actions by the perpetrator, constituting cyber harassment.

- The manifestations' word cloud highlights emotional expressions denoting victim distress, vulnerability, and cyber harassment impact. Notable words such as "scared," "trouble," and "trauma" depict emotional states. Phrases like "help me," "please suggest," and "get some counseling" convey the victim's plea for agent support.
- The determinants' word cloud illustrates cyber harassment actions like "following me," "sleep with him," and "blackmailing me." Moreover, terms like "pics," "videos," "Facebook," and "Whatsapp" indicate the perpetrator's use of digital evidence. The presence of "uncle" and "boyfriend" suggests the familial or social connection between victim and perpetrator.

Methodology

In this section, we present our framework for extracting manifestation and determinant spans in dyadic conversational utterances between a cyber harassment victim and a virtual automated counselor.

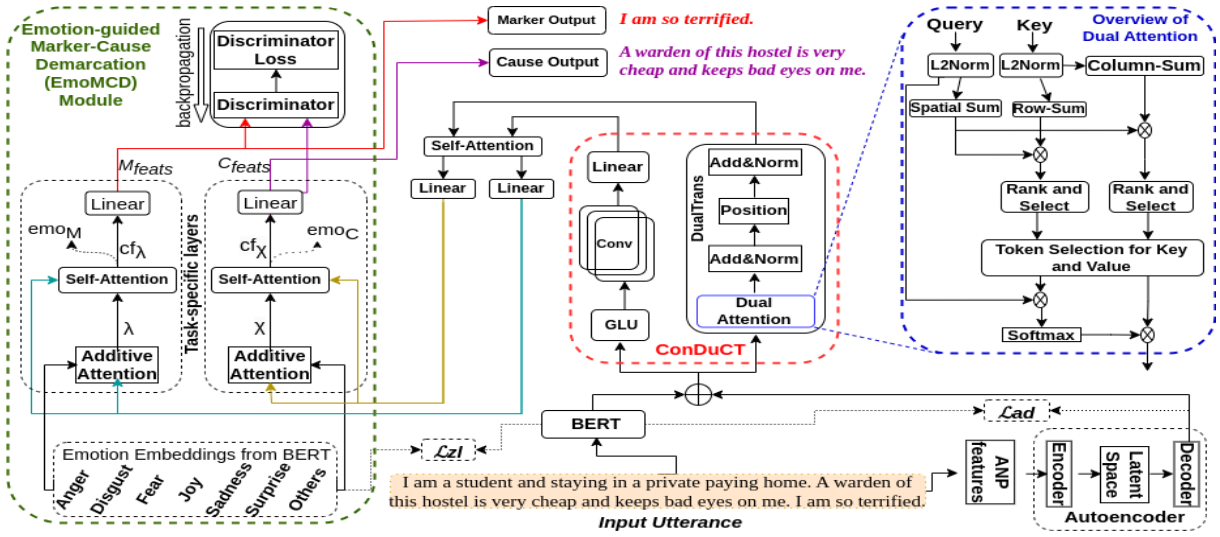

Figure 4: Illustration of Emotion-informed Contextual Dual attention Convolution Transformer *(E-ConDuCT)*.

Problem Definition

The goal is to identify expressions and actions representing distress, vulnerability, or cyber harassment scenarios (manifestations), as well as actions by the perpetrator constituting the cyber harassment (determinants). To achieve this, we work with the *Cyber-MaD* corpus comprising several dialogues (say, *n*). A dialogue can be represented as D = [u_1, $\cdots$ u_i $\cdots$, u_p], where each utterance u_i is represented as a sequence of words u_i = [$word_{i,1}$, $\cdots$ $term_{i,j}$ $\cdots$, $term_{i,q}$], where *p* indicates the number of utterances in the document, and *q* denotes the length of the word sequence contained in the utterance. The optimization function involves identifying and scoring potential spans (subsequences within u_t) that distinctively represent the manifestations and determinants pertaining to the cyber harassment event.

Proposed Framework

We design an *Emotion-informed Contextual Dual attention Convolution Transformer (E-ConDuCT)* for Cyber Harassment Analysis. Figure 4 illustrates the overall architecture of the proposed framework.

Input Encoder. We utilize the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al. 2019) to learn contextual information from the input utterances. To better comprehend any emotional information in the input text, we use Context-Free-Grammar-Noun-Adjective-Pairs (Context Free ANP)[2] to extract adjective-noun pairs from the utterances, enabling our model to identify textual concepts. The ANP features are passed through an auto-encoder to generate a latent representation. To combine textual and class semantic knowledge in the ANP representation, we apply adversarial loss (Zhu et al. 2018) (see Section *Calculation of Losses*) between BERT representation and decoder output. The aim is to disentangle syntax (ANP-captured) and semantics (BERT-captured), potentially enhancing interpretability and control over learned representations.

Contextual Dual attention Convolution Transformer (ConDuCT). We propose *ConDuCT* to model contextual information in input utterances. We introduce *DualTrans*, a dual-attention method, replacing traditional self-attention in transformer encoder (Vaswani et al. 2017). To enhance attention with less overhead, we use a gated linear unit (GLU) with a convolution layer (Wu et al. 2020) in the input representation. For reduced computation, we replace standard convolution with a lighter variant (Wu et al. 2019), involving linear layers and depth-wise convolution. The convolution output is linearly concatenated with *DualTrans* output, applying self-attention to the concatenated result. This enhances the model's ability to capture intricate relationships and patterns by leveraging features from both sources.

Dual Attention. We present *multi-head dual attention (DuA)* as an innovative enhancement for standard self-attention in transformer models. DuA involves two main phases: token contribution evaluation and token pruning (Zheng et al. 2022). Figure 5 visually outlines the forward propagation within the *DuA* framework. We expound on these stages in the context of a single *DuA* head. A crucial distinction lies in our choice to compute the attention map prior to exploring token contributions, aligning with the pre-pruning strategy for Key and Value.

For aggregated assessment of tokens based on columns or rows, we utilize the distributive property of vector inner products, substantially diminishing measurement expenses. Let q_i and k_j represent tokens in Query ($Q \in \mathbf{R}^{nx}$) and Key ($K \in \mathbf{R}^{mx}$), respectively, with dimensions *n* and *m* for query and key vectors. The revised scores for column and row vectors are given by:

$$Sco_r = \sum_{i=1}^{n}\sum_{j=1}^{m} q_i k_{rj}^T (\sum_{i=1}^{n} q_i)(\sum_{j=1}^{m} k_{rj}^T), \quad r \in 1....n \quad (1)$$

[2]https://github.com/StatguyUser/Context_Free_Grammar-Noun_Adjective_Pairs

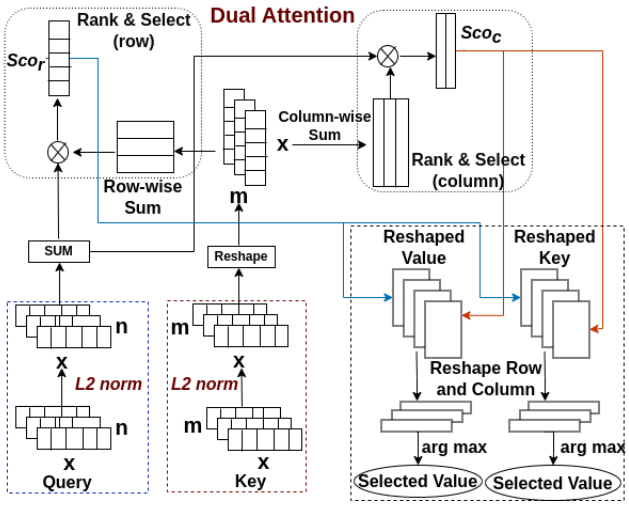

Figure 5: Demystifying Dual Attention

$$Sco_c = \sum_{i=1}^{n}\sum_{j=1}^{m} q_i k_{jc}^T = (\sum_{i=1}^{n} q_i)(\sum_{j=1}^{m} k_{jc}^T), \quad c \in 1....m \tag{2}$$

Here, j and c represent tokens in query and key vectors, and T denotes matrix transpose operations.

Employing token-wise L2 normalization for Query and Key vectors, we gain insights into the contributions of grouped tokens. The constraints placed on the element values in the attention map, spanning the range (-1, 1), avert potential distortions caused by excessively concentrated token vectors prior to Softmax activation.

Token pruning encompasses the computation of contribution scores $Sco_r \in \mathbf{R}^n$ and $Sco_c \in \mathbf{R}^m$. Rows and columns are ranked based on their contribution ratings, and the most highly ranked are selected. This selective process retains certain rows and columns while shedding others. In our experimentation, the selection of rows or columns, denoted as N_s (a hyper-parameter), adheres to the square root of H.

$$Ind_r = argmax Sco_r[:N_s],\ Ind_c = argmax Sco_c[:N_s] \tag{3}$$

The transformation of K and V is then ascertained as $K_s = K^{[Ind_r, Ind_c]}$ and $V_s = V^{[Ind_r, Ind_c]}$. Eminent index rankings derive from contribution scores and ArgMaxScore, further reinforced by row or column selection with $[:N_s]$.

Emotion-guided Manifestation-Determinant Demarcation (EmoMD2) Module. After *ConDuCT*, self-attention output undergoes two linear layers and task-specific layers for manifestation and determinant features within the *EmoMD2* module (Figure 4). The intermediate manifestation and determinant features are derived from the linear layer outputs, intended to incorporate emotional knowledge. Additive attention (Yang et al. 2016) is applied between the manifestation representation and emotion class embeddings. Emotion class embeddings are obtained using a pre-trained BERT (*base*) (Devlin et al. 2019) model, representing Ekman's (Ekman 1992) basic emotion classes (*Anger*, *Disgust*, *Sad*, *Joy*, *Surprise*, *Fear*), and an additional *Others* class to accommodate instances outside Ekman's categorization (Ekman 1992). Utilizing BERT features eliminates the need for further human annotation. A single emotion representation, *manifestation-specific emotion embeddivide access to the code and data for research endeavorsng* (λ), is computed as a weighted sum of individual emotion representations based on their importance to manifestation information. Similarly, the *determinant-specific emotion embedding* (χ) is obtained. Concatenating λ with the original manifestation representation yields *emotion-aware manifestation features* (cf_λ), and doing the same with χ and the original determinant representation generates *emotion-aware determinant features* (cf_χ). Self-attention is applied to both representations, and a discriminator loss function (see Section *Calculation of Losses*) is employed between manifestation and determinant representations to emphasize distinctiveness despite shared representations and output objectives.

Calculation of Losses. In this section, we discuss the various losses used to train our proposed framework.

Adversarial Loss. To minimize the distance between BERT output ($\theta(h_{text})$) and the autoencoder's emotional structural data ($\theta(h_{ANP})$), we impose an adversarial restriction that seeks to bring these two representations as close as possible:

$$\mathcal{L}_{ad} = \mathcal{E}_y(\log \mathcal{D}(\theta(h_{text}) - \mathcal{E}_y(\log \mathcal{D}(\theta(h_{ANP}) \tag{4}$$

Zero-shot loss. The model aims to minimize the difference between the text feature ($\theta(h_{text})$) and the semantic feature of the emotion label ($\phi(l_{emo})$) through optimization:

$$\mathcal{L}_{zl} = ||\theta(h_{text}) - \phi(l_{emo})||_2^2 \tag{5}$$

Discriminator Loss. In our framework, we employ a discriminator to distinguish between the emotion-induced manifestation (M_{feats}) and determinant (C_{feats}) representations. The discriminator loss is adapted from Vanilla GANs (Goodfellow et al. 2014) and focuses on discriminating between the embeddings of M_{feats} and C_{feats}, rather than distinguishing real and fake samples. The idea is to help the model capture and maintain the semantic differences between manifestation and determinant features, even though they share similar representations and output objectives. The modified loss function allows us to focus on task-specific embeddings while still benefiting from the principles of GANs for enhancing representation learning.

The Discriminator Loss (L_D) is computed as:

$$L_D = -\frac{1}{N}\sum_{i=1}^{N}[\log(D(L_{Mar_i})) + \log(1 - D(L_{Cau_i}))] \tag{6}$$

where, N is the number of samples in the batch, and $D(x)$ represents the output of the discriminator.

Manifestation and determinant extraction training. For manifestation and determinant extraction tasks, we use binary cross-entropy losses:

$$L = \sum_{\omega} W_\omega L_\omega \tag{7}$$

Here, ω denotes the two tasks, and the weights (W_ω) are updated using back-propagation for each specific task loss.

MODEL	*Cyber-MaD* - [Manifestation]					*Cyber-MaD* - [Determinant]					*CARES_CEASE-v2.0* - [Determinant]				
	FM	PM	HD	JS	ROS	FM	PM	HD	JS	ROS	FM	PM	HD	JS	ROS
BiRNN-Attn	24.65	18.19	0.47	0.64	0.69	21.86	16.54	0.43	0.61	0.64	25.76	15.75	0.43	0.59	0.72
CNN-GRU	23.74	17.32	0.45	0.65	0.70	22.54	16.32	0.41	0.62	0.65	24.32	15.24	0.43	0.62	0.71
BiRNN-HateXplain	29.31	22.77	0.53	0.70	0.72	26.21	18.34	0.48	0.66	0.68	29.21	16.88	0.47	0.62	0.73
SpanBERT	31.24	24.23	0.55	0.72	0.73	27.11	21.16	0.50	0.68	0.69	31.17*	17.62*	0.49*	0.66*	0.76*
BERT-HateXplain	30.77	23.43	0.57	0.72	0.75	29.59	20.54	0.52	0.69	0.71	31.75	18.49	0.48	0.68	0.78
CMSEKI	33.65	24.32	0.56	0.75	0.76	31.73	23.87	0.56	0.70	0.73	33.21	20.74	0.52	0.67	0.78
E-ConDuCT	**35.62**	**27.24**	**0.62**	**0.78**	**0.79**	**33.19**	**25.19**	**0.59**	**0.74**	**0.76**	**35.24**	**25.71**	**0.57**	**0.74**	**0.80**

Table 2: Results on the introduced *Cyber-MaD* corpus and the benchmark *CARES* dataset from the *E-ConDuCT* model and the various baselines. Bold values represent the maximum scores. * values are directly fetched from (Ghosh et al. 2022).

Calculation of Final Loss. We train the model through a single unified loss function:

$$\mathcal{L} = \mathcal{L}_{ad} + \mathcal{L}_{zl} + \mathcal{L}_{D} + \sum_{\omega} W_{\omega} L_{\omega} \tag{8}$$

Experiments and Results

For comprehensive evaluation of our proposed method, we run experiments on our *Cyber-MaD* corpus (for the *CHMs* and *CHDs* extraction) and the benchmark suicide notes dataset, CARES_CEASE-v2.0 (Ghosh et al. 2022) (for emotion cause extraction). We compare the performance of our model with six state-of-the-art systems, namely: BiRNN-Attn (Liu and Lane 2016), CNN-GRU (Zhang, Robinson, and Tepper 2018), BiRNN-HateXplain (Mathew et al. 2021), BERT (Liu et al. 2019), BERT-HateXplain (Mathew et al. 2021), SpanBERT (Liu et al. 2019) and Cascaded Multitask System with External Knowledge Infusion (CMSEKI) (Ghosh, Ekbal, and Bhattacharyya 2022). In line with recent studies (Ghosh et al. 2022; Singh et al. 2023) with cause extraction objective, we report our results on the following metrics: Full match (FM), Partial match (PM), Hamming Distance (HD), Jaccard Similarity (JS) and Ratcliffe-Obershelp Similarity (ROS).

We use PyTorch[3] to implement our Dual Transformer model with an amplification value (a) of 64 on an NVIDIA GeForce RTX 2080 Ti GPU. The embedding size is 768, and training spans 400 epochs with a dropout rate of 0.5. The auto-latent encoder maintains a fixed dimension of 812. The discriminator $\mathcal{D}$ has two fully connected layers and a ReLU layer, accommodating 812-dimensional input features. Stochastic gradient descent is employed with a learning rate of 1e-4, weight decay of 1e-3, and momentum of 0.5. The learning rate starts at 2e-4 for the initial 200 epochs and linearly reduces to zero over the subsequent 200 epochs.

Results and Analysis

Table 2 shows the results of the baselines and the proposed *E-ConDuCT* framework on the introduced *Cyber-MaD* and the benchmark CARES_CEASE-v2.0 datasets.

Comparison with Prior Research. In the context of the Cyber-MaD corpus for manifestations and determinants extraction, our *E-ConDuCT* framework consistently surpassed baselines across all metrics. Notably, for determinant extraction, the framework achieved consistently higher Hamming Distance and Jaccard Similarity scores, revealing its proficiency in capturing subtle nuances in discerning cyber-harassment causes. Moreover, in emotion cause extraction using the CARES dataset, *E-ConDuCT* maintained its effectiveness, outperforming other methods in most metrics. Its superior performance across various metrics highlights its potential to advance the field, aiding targeted interventions and policies against cybercrimes.

Varying Context Length. We analyze the influence of context sizes on understanding cyber harassment dialogues. Our model is evaluated with context lengths ranging from 1 to 11, depicted in Figure 6. A context of length 1 signifies no additional context, using only the target utterance. Performance consistently improves with more previous utterances, peaking at a context length of 7. Beyond this point, excessive context causes confusion and performance decline.

Ablation Study. We perform ablation experiments with *E-ConDuCT* to assess the importance of each primary model component and individual loss functions in training. Results in Table 3 show significant performance drops when any component or loss is removed. The most substantial decline (across all metrics) occurs when replacing the dual attention component with traditional self-attention. Omitting the zero-shot loss leads to the most significant decrease in performance, highlighting its crucial role in incorporating emotion knowledge during model training.

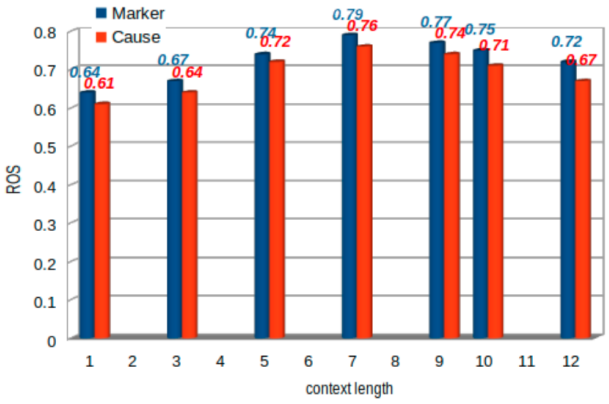

Figure 6: *E-ConDuCT* results with varying context length.

[3]https://pytorch.org/

MODEL	*Cyber-MaD* [Manifestation]				*Cyber-MaD* [Determinant]				*CARES_CEASE-v2.0* [Emo cause]			
	FM	**HD**	**JS**	**ROS**	**FM**	**HD**	**JS**	**ROS**	**FM**	**HD**	**JS**	**ROS**
Component-based Ablation Experiments												
E-ConDuCT - [ANP]	33.73	0.57	0.76	0.75	31.47	0.57	0.71	0.72	34.03	0.54	0.72	0.78
E-ConDuCT - [DuA]	32.73	0.53	0.72	0.73	31.15	0.52	0.68	0.70	34.01	0.53	0.67	0.77
E-ConDuCT - [EmoMD2]	33.41	0.57	0.75	0.75	31.47	0.57	0.72	0.72	34.23	0.54	0.70	0.76
Loss-based Ablation Experiments												
E-ConDuCT - [$\mathcal{L}_D$]	34.44	0.59	0.75	0.78	32.18	0.58	0.73	0.74	34.36	0.55	0.73	0.78
E-ConDuCT - [$\mathcal{L}_{ad}$]	34.72	0.61	0.77	0.78	32.07	0.58	0.72	0.75	33.88	0.56	0.71	0.77
E-ConDuCT - [$\mathcal{L}_{zl}$]	34.40	0.60	0.75	0.77	31.72	0.57	0.72	0.75	33.83	0.55	0.72	0.76

Table 3: Results illustrating variance in outputs of the *E-ConDuCT* model when individual components and losses are ablated.

Qualitative Analysis. Our comprehensive evaluation demonstrates that while existing methods show potential, they struggle with differentiating between manifestations and determinants accurately (as shown in Table 4). BERT-HateXplain tends to misclassify determinants as manifestations, causing potential confusion in distinguishing between the two categories. Span-BERT encounters challenges in accurately identifying the boundaries of determinants, which can lead to imprecise extraction results. CMSEKI exhibits inconsistency in its performance, occasionally producing incomplete extractions for both manifestations and determinants. In contrast, our proposed *E-ConDuCT* model showcases a robust performance by consistently capturing the essence of both concepts. This consistency and accuracy highlight *E-ConDuCT*'s suitability for addressing the complexities of cyber-harassment dialogues.

Model	**Text**
Example 1:	Yes, but he forced me to not block/report him. I am very scared.
Human Annotator	CHM: *I am very scared* CHD: *he forced me to not block him*
Span-Bert	CHM: *very scared* — CHD: very scared
CMSEKI	CHM: *very scared* — CHD: *forced me*
ChatGPT	CHM: he forced me to not block/report him, I am very scared CHD: *he forced me to not block/report him.*
E-ConDuCT *(ours)*	CHM: *I am very scared* CHD: *he forced me to not block him*
Example 2:	My uncle is trying to harass me. I don't know from where he got my sensitive documents and he is asking me to sleep with him for at least one night.
Human Annotator	CHM: *My uncle is trying to harass me.* CHD: *he got my sensitive documents*
Span-Bert	CHM: *harass me* — CHD: *got my sensitive*
CMSEKI	CHM: *trying to harass*, sensitive documents CHD: *my sensitive documents*
ChatGPT	CHM: *sleep with him for at least one night*, I don't know from where he got my sensitive documents. CHD: *trying to harass me, he got my sensitive documents*
E-ConDuCT *(ours)*	CHM: *uncle is trying to harass me*. CHD: *he got my sensitive documents*

Table 4: Sample predictions from the various systems.

Comparison with ChatGPT 3.5. We evaluate the performance of ChatGPT[4] and our proposed method on several dialogue sequences. While ChatGPT excels in numerous NLP tasks, it faces challenges distinguishing between manifestations and determinants in cyber-harassment dialogues (as eveident from Table 4). It often misclassifies determinant spans as manifestations, underscoring the need for task-specific models like *E-ConDuCT* to capture nuanced differences and yield more accurate results.

Task-Specific Emotional Nuances. We compute intermediate emotion outputs for 500 randomly chosen accurately predicted spans for manifestations and determinants. In analyzing the task-specific distinctions observed in the predictions for manifestations and determinants, the task-specific emotional nuances present in cyber harassment context becomes evident. For manifestations, the dominant emotions are Anger, Fear, Sadness, and Disgust, reflecting the distressing and repulsive nature of victim experiences. In contrast, determinants primarily elicit Anger, Joy, and Sadness, suggesting that discussions on perpetrator actions provoke emotions like anger, empathy, and positive sentiments.

Conclusion

The major contributions of the current research lie in investigating cyber harassment comprehension with distinct *CHMs* and *CHDs* definitions and automated methods. The introduction of the *Cyber-MaD* dataset and the development of the *E-ConDuCT* framework provide practical tools for enhancing understanding in this domain. *E-ConDuCT*'s superior performance underscores its significance, offering comprehensive insights into victims' emotions, crime characteristics, and perpetrator behaviors. This insight empowers stakeholders to develop effective support, interventions, policies, and awareness campaigns against cybercrimes. The potential of this research extends beyond cyber harassment to fostering safer digital environments and empowering individuals to combat cybercrimes effectively.

Possible research paths involve extending the framework to various languages and cultures, delving into finer nuances within manifestations and determinants. We anticipate exploring multi-modal approaches, encompassing audio and visual cues, and integrating advanced emotion-aware models to enhance dialogue comprehension.

[4]https://chat.openai.com/

Ethical Statement

In this study, we utilized the openly accessible Mental Health and Legal Counseling Dataset (MHLCD) (Mishra, Priya, and Ekbal 2023) to create our *Cyber-MaD* corpus, ensuring full compliance with copyright regulations. We provide access to the code and data for research endeavors through a suitable data agreement mechanism.

The research involves analyzing online conversations, which may contain personal and sensitive information. Ensuring the privacy and consent of individuals involved in these conversations is crucial to prevent any ethical violations. Additionally, extracting and analyzing cyber harassment-related content can have psychological implications for researchers, annotators, and users. Care should be taken to provide support and minimize harm to individuals involved in the research process. Furthermore, the research deals with sensitive legal and social issues. Collaborating with legal experts and ensuring compliance with relevant regulations is essential to avoid legal pitfalls.

References

Al-Garadi, M. A.; Hussain, M. R.; Khan, N.; Murtaza, G.; Nweke, H. F.; Ali, I.; Mujtaba, G.; Chiroma, H.; Khattak, H. A.; and Gani, A. 2019. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. *IEEE Access*, 7: 70701–70718.

Ali, W. N. H. W.; Mohd, M.; and Fauzi, F. 2021. Identification of profane words in cyberbullying incidents within social networks. *Journal of Information Science Theory and Practice*, 9(1): 24–34.

Andleeb, S.; Ahmed, R.; Ahmed, Z.; and Kanwal, M. 2019. Identification and Classification of Cybercrimes using Text Mining Technique. In *International Conference on Frontiers of Information Technology, FIT 2019, Islamabad, Pakistan, December 16-18, 2019*, 227–232. IEEE.

Boukhtouta, A.; Mouheb, D.; Debbabi, M.; Alfandi, O.; Iqbal, F.; and Barachi, M. E. 2015. Graph-theoretic characterization of cyber-threat infrastructures. *Digit. Investig.*, 14 Supplement 1: S3–S15.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

Dinakar, K.; Reichart, R.; and Lieberman, H. 2011. Modeling the Detection of Textual Cyberbullying. In *The Social Mobile Web, Papers from the 2011 ICWSM Workshop, Barcelona, Catalonia, Spain, July 21, 2011*, volume WS-11-02 of *AAAI Technical Report*. AAAI.

Ekman, P. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200.

Ghosh, S.; Ekbal, A.; and Bhattacharyya, P. 2022. A Multitask Framework to Detect Depression, Sentiment and Multilabel Emotion from Suicide Notes. *Cogn. Comput.*, 14(1): 110–129.

Ghosh, S.; Priyankar, A.; Ekbal, A.; and Bhattacharyya, P. 2023. A transformer-based multi-task framework for joint detection of aggression and hate on social media data. *Natural Language Engineering*, 1–21.

Ghosh, S.; Roy, S.; Ekbal, A.; and Bhattacharyya, P. 2022. CARES: CAuse Recognition for Emotion in Suicide Notes. In *European Conference on Information Retrieval*, 128–136. Springer.

Gonzales, R. H. 2014. Social media as a channel and its implications on cyber bullying. In *DLSU Research Congress*, 1–7.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gui, L.; Wu, D.; Xu, R.; Lu, Q.; and Zhou, Y. 2016. Event-Driven Emotion Cause Extraction with Corpus Construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1639–1649. Austin, Texas: Association for Computational Linguistics.

Kim, S.; Razi, A.; Stringhini, G.; Wisniewski, P. J.; and Choudhury, M. D. 2021. You Don't Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. In Budak, C.; Cha, M.; Quercia, D.; and Xie, L., eds., *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, 290–302. AAAI Press.

Liu, B.; and Lane, I. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mahor, V.; Rawat, R.; Telang, S.; Garg, B.; Mukhopadhyay, D.; and Palimkar, P. 2021. Machine learning based detection of cyber crime hub analysis using twitter data. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, 1–5. IEEE.

Markov, T.; Zhang, C.; Agarwal, S.; Nekoul, F. E.; Lee, T.; Adler, S.; Jiang, A.; and Weng, L. 2023. A Holistic Approach to Undesired Content Detection in the Real World. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 15009–15018. AAAI Press.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 14867–14875. AAAI Press.

Mishra, K.; Priya, P.; and Ekbal, A. 2023. Help Me Heal: A Reinforced Polite and Empathetic Mental Health and Legal Counseling Dialogue System for Crime Victims. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 14408–14416. AAAI Press.

Paracha, A. A.; Arshad, J.; and Khan, M. M. 2023. S.U.S. You're SUS! - Identifying influencer hackers on dark web social networks. *Comput. Electr. Eng.*, 107: 108627.

Poria, S.; Majumder, N.; Hazarika, D.; Ghosal, D.; Bhardwaj, R.; Jian, S. Y. B.; Hong, P.; Ghosh, R.; Roy, A.; Chhaya, N.; et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5): 1317–1332.

Singh, G. V.; Ghosh, S.; Ekbal, A.; and Bhattacharyya, P. 2023. DeCoDE: Detection of Cognitive Distortion and Emotion Cause Extraction in Clinical Conversations. In *European Conference on Information Retrieval*, 156–171. Springer.

Sprugnoli, R.; Menini, S.; Tonelli, S.; Oncini, F.; and Piras, E. 2018. Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying. In Fiser, D.; Huang, R.; Prabhakaran, V.; Voigt, R.; Waseem, Z.; and Wernimont, J., eds., *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*, 51–59. Association for Computational Linguistics.

Talpur, B. A.; and O'Sullivan, D. 2020. Cyberbullying severity detection: A machine learning approach. *PloS one*, 15(10): e0240924.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wiegand, M.; Ruppenhofer, J.; and Kleinbauer, T. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 602–608. Association for Computational Linguistics.

Wu, F.; Souza, A.; Zhang, T.; Fifty, C.; Yu, T.; and Weinberger, K. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*, 6861–6871. PMLR.

Wu, Z.; Liu, Z.; Lin, J.; Lin, Y.; and Han, S. 2020. Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A. J.; and Hovy, E. H. 2016. Hierarchical Attention Networks for Document Classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 1480–1489.

Zhang, Z.; Robinson, D.; and Tepper, J. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, 745–760. Springer.

Zheng, W.; Li, Q.; Zhang, G.; Wan, P.; and Wang, Z. 2022. Ittr: Unpaired image-to-image translation with transformers. *arXiv preprint arXiv:2203.16015*.

Zhu, Y.; Elhoseiny, M.; Liu, B.; Peng, X.; and Elgammal, A. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1004–1013.