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

Soumitra Ghosh¹*, Gopendra Vikram Singh²*, Jashn Arora³, Asif Ekbal²

¹NLP Research Group, Fondazione Bruno Kessler (FBK), Trento, Italy
²Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India
³International 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-nlpml/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.

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.

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 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 demonstrating 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; Mahar 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 (Wie- gand, 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 annotators1 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.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Dialogues</td>
<td>996</td>
</tr>
<tr>
<td>Total Utterances</td>
<td>27666</td>
</tr>
<tr>
<td>Average Utterance Length</td>
<td>18.3</td>
</tr>
<tr>
<td>Utterances with one manifestation span</td>
<td>7427</td>
</tr>
<tr>
<td>Utterances with two manifestation span</td>
<td>1220</td>
</tr>
<tr>
<td>Utterances with one determinant span</td>
<td>3260</td>
</tr>
<tr>
<td>Utterances with two determinant span</td>
<td>410</td>
</tr>
<tr>
<td>Manifestation Length (words)</td>
<td>9</td>
</tr>
<tr>
<td>Determinant Length (words)</td>
<td>14.9</td>
</tr>
</tbody>
</table>

Table 1: General Dataset Statistics

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

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.
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 $D = \{u_1, \ldots , u_p\}$, where each utterance $u_i$ is represented as a sequence of words $w = \{\text{word}_{i,1}, \ldots , \text{term}_{i,j}, \ldots, \text{term}_{i,q}\}$, 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 distinctly 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 $k_r$ represent tokens in Query ($Q \in \mathbb{R}^{n \times r}$) and Key ($K \in \mathbb{R}^{m \times r}$), respectively, with dimensions $n$ and $m$ for query and key vectors. The revised scores for column and row vectors are given by:

\[
Sco_r = \sum_{i=1}^{n} \sum_{j=1}^{m} q_i k_r^T (\sum_{i=1}^{n} q_i) (\sum_{j=1}^{m} k_r^T), \quad r \in 1,..,n
\]
The model aims to minimize the difference between the emotion-induced manifestation (Ekman 1992). Utilizing BERT features eliminates the need for further human annotation. A single emotion representation, manifestation-specific emotion embedding and determinant representation, yields emotion-aware manifestation features \( \phi(x) \), and doing the same with \( \chi \) and the original determinant representation generates emotion-aware determinant features \( \phi(x) \). 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:

\[
\mathcal{L}_{ad} = \mathcal{E}_y (\log D(\theta(h_{text})) - \log D(\theta(h_{ANP}) \right)
\]

where \( \mathcal{E}_y \) denotes the two tasks, and the weights \( \omega \) denotes the two tasks, and the weights \( \omega \) denote the two tasks.

**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:

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

**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:

\[
L_D = -\frac{1}{N} \sum_{i=1}^{N} \left[ \log(D(L_{Mar.i})) + \log(1 - D(L_{Cau.i})) \right] \tag{6}
\]

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:

\[
L = \sum_{\omega} W_{\omega} L_{\omega} \tag{7}
\]

where, \( \omega \) denotes the two tasks, and the weights \( W_{\omega} \) are updated using back-propagation for each specific task loss.
We train the model through a variety of baselines. Bold values represent the maximum scores. * values are directly fetched from (Ghosh et al. 2022).

### Results and Analysis

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

$$
\mathcal{L} = \mathcal{L}_{ad} + \mathcal{L}_{zl} + \sum_\omega W_\omega L_\omega \quad (8)
$$

### 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-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 to implement our Dual Transformer model with an amplification value ($\alpha$) 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 $D$ 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.

### 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 cyberharassment 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.

---

1https://pytorch.org/
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.

Comparison with ChatGPT 3.5. We evaluate the performance of ChatGPT⁴ 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 evident 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.

Table 4: Sample predictions from the various systems.

<table>
<thead>
<tr>
<th>Model</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Annotator</td>
<td>Yes, but he forced me to not block/report him. I am very scared.</td>
</tr>
<tr>
<td>CMSEKI</td>
<td>CHM: I am very scared</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>CHM: he forced me to not block/report him.</td>
</tr>
<tr>
<td>E-ConDuCT (ours)</td>
<td>CHM: I am very scared</td>
</tr>
<tr>
<td></td>
<td>CHD: he forced me to not block him</td>
</tr>
<tr>
<td>Example 1:</td>
<td>My uncle is trying to harass me. I don’t know where he got my sensitive documents and he is asking me to sleep with him for at least one night.</td>
</tr>
<tr>
<td>Human Annotator</td>
<td>CHM: My uncle is trying to harass me.</td>
</tr>
<tr>
<td>CMSEKI</td>
<td>CHM: got my sensitive documents</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>CHM: sleep with him for at least one night, I don’t know where he got my sensitive documents.</td>
</tr>
<tr>
<td>E-ConDuCT (ours)</td>
<td>CHM: uncle is trying to harass me.</td>
</tr>
</tbody>
</table>

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

In this study, we utilized the openly accessible Mental Health and Legal Counseling Dataset (MHLCD) (Mishra, Priyankar, 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


