SERI: Generative Chatbot Framework for Cybergrooming Prevention

Pei Wang, Zhen Guo, Lifu Huang, Jin-Hee Cho
Computer Science, Virginia Tech, VA, USA
{pwang1, zguo, lifuh, jicho}@vt.edu

Abstract

Cybergrooming refers to a crime to lure potential victims, particularly youth, by establishing personal trust relationships with them for sexual abuse or exploitation. Although cybergrooming is recognized as one of the serious social issues, there has been a lack of proactive programs to protect the youth. In this paper, we present a generative chatbot framework, called SERI (Stop cybERgroomIng), that can generate authentic conversations between a perpetrator chatbot and a potential victim chatbot. The SERI is designed to provide a safe and authentic environment for enhancing youth's sensitivity and awareness of subtle cues of cybergrooming without exposing unnecessary ethical issues caused by potentially offensive or upsetting languages. The SERI is developed as a pre-stage before the perpetrator chatbot is deployed to chatting with an actual human youth user to observe how the youth user can respond to a stranger or acquaintance asking for sensitive or private information. Hence, to evaluate the quality of the conversations generated by the SERI, we use open-source, referenced, and unreferenced metrics to assess the generated conversations automatically. In addition, we evaluated the quality of the conversation based on the human evaluation method. Our results show that the SERI can generate authentic conversations between the two chatbots compared to the original conversations from the used dataset in perplexity and MaUde scores.

1 Introduction

As of 2017, approximately one-third of online users in the world are known young people below the age of 18 (UNICEF, 2017). Although Internet has provided countless benefits in our everyday life, it also has introduced serious concerns in online sexual exploitation and abuse of children (Choo, 2009; Marchenko, 2017). Cybergrooming refers to the crime of establishing a personal trust relationship with potential victims, commonly youth, via Internet only for sexual exploitation or abuse (Choo, 2009). In the US, from 1998 to 2013, the CyberTipline (on International Law and Policy, 2017) received 60,000 cases of luring children for sexual purposes in cyberspace. Due to the high seriousness of cybergrooming, some studies have investigated the key properties of cybergrooming or developed tools to detect online child sexual exploitation or predators (Anderson et al., 2019; Bours and Kulsrud, 2019; Fauzi and Bours, 2020).

In computer science, the majority of cybergrooming studies focused on detecting predators by analyzing malicious conversations. However, this does not provide any proactive prevention to protect potential youth victims from cybergrooming. Due to this reason, this work is motivated to develop a proactive cybergrooming prevention program to increase youth's awareness and sensitivity to cybergrooming and its serious consequence. To this end, we aim to develop a generative chatbot framework that can provide authentic conversations between a perpetrator chatbot and a youth user chatbot to achieve stopping cybergrooming ultimately. We named this generative chatbot framework by SERI, Stop cybERgroomIng. The SERI will be used as a pre-stage to provide a safe and authentic environment before deploying the perpetrator chatbot with a real human youth user. The SERI will allow a safe environment that a youth user can involve an authentic conversation with a stranger or acquaintance and learn how to deal with the person talking about sensitive or private issues.

In developing the authentic, generative chatbot framework, SERI, to mimic the conversations between a perpetrator and a potential victim, we found the following **research challenges**. First, unlike general casual talks, the perpetrator is very goal-oriented by leading conversations and striving to achieve a final goal, such as meeting in person. The perpetrator gradually establishes a trust rela-

tionship with a potential victim by asking a series of questions about the potential victim's private life. Second, it is highly challenging to develop a chatbot generating authentic conversations because of a lack of datasets that sufficiently train the SERI. The only available dataset is the Perverted Justice (PJ) dataset (Perverted Justice Foundation Inc., 2020), consisting of the conversations between cybergrooming perpetrators and professionally trained volunteers playing the role of potential youth victims. However, the volume of the PJ dataset is limited (i.e., 100 sets of conversations) and contains highly informal languages, such as short abbreviations, slang, or unsegmented words, emojis, or URLs. The poor quality of the training datasets makes it significantly challenging to train the SERI chatbot model directly. To tackle these, we made the following **key contributions** via our developed SERI:

1. We applied a two-stage paradigm to train the SERI by the T5 (Text-to-Text Transfer Transformer) model (Raffel et al., 2020), where both the perpetrator and victim chatbots were first pre-trained on general and large-scale causal talk datasets, such as ConvAI2 (The Second Conversational Intelligence Challenge dataset) (Dinan et al., 2019). After then, the chatbots were fine-tuned on the PJ dataset, which was pre-processed with a series of social text normalization tools to mitigate the effect of highly informal languages (e.g., slang, online abbreviations, unsegmented words, emojis, or URLs).

2. We modeled the multi-stage strategies that the perpetrators can take to evolve the relationship with a potential victim and achieve the goal of meeting in person. To achieve this, we defined four grooming stages based on the evolution of the relationships and predicted a stage for each utterance by encoding an utterance through BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). We accordingly trained four perpetrator subchatbots using the T5 model. Each perpetrator's subchatbot was trained on the set of utterances from the PJ dataset. The corresponding stage labels of utterances were predicted by a BERT-based stage classifier.

3. We developed a mechanism to escalate the attack stages and coordinate the dialogue generation with the four perpetrator subchatbots. The perpetrator will move to a next-level stage by switching from the current subchatbot to the next stage subchatbot if the perpetrator successfully obtains all information from a potential victim while the potential victim still stays in the conversation. If the potential victim leaves the chat, the perpetrator will fail this attack.

4. We evaluated the SERI by using both referenced metrics (i.e., BLEU (Post, 2018), ROUGE (Lin, 2004), and BERTScores (Zhang* et al., 2020)) and unreferenced metrics (i.e., perplexity and MaUde scores (Sinha et al., 2020)). We found that the conversations generated by the SERI showed better performance than the ground truth conversations based on all metrics above. In addition, our human evaluation confirms that about 37% of the utterances generated by the SERI are valid and better than ground truth utterances from the PJ dataset.

2 Related Work

Cybergrooming detection. Many Machine Learning (ML) algorithms, such as support vector machine (SVM) (Anderson et al., 2019; Dhouioui and Akaichi, 2016; Fauzi and Bours, 2020; Gunawan et al., 2018), fuzzy logic (Anderson et al., 2019), k-nearest neighbors (KNN) (Gunawan et al., 2018), Random Forest (Fauzi and Bours, 2020), Naïve Bayes (Bours and Kulsrud, 2019), Decision Tree (Fauzi and Bours, 2020) and Neural Network (NN) classifiers (Bours and Kulsrud, 2019; Fauzi and Bours, 2020), have been used to detect cybergrooming based on lexical (e.g., Term Frequency-Inverse Document Frequency or TF-IDF based features, Bag of Words features) and behavioral features from the text. To understand the evolving conversations between perpetrators and victims, researchers also investigated multiple relational stages of cybergrooming (Winters and Jeglic, 2016). However, while most efforts focused on the grooming stages and prevention methods (Zambrano et al., 2019), no prior research has characterized the features of victims by cybergrooming.

Chatbot application tools. A chatbot, called Negobot, was developed to detect potential pedophiles in the social networks (Laorden et al., 2013). A game-theoretic reward metric could move the chatbot toward the next conversation stage or maintain the current stage. Recently, pre-training lan-

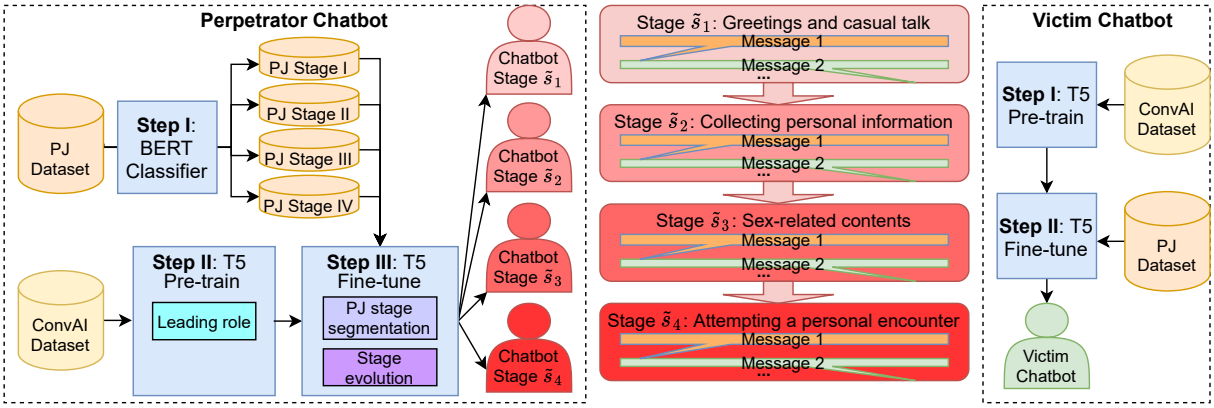

Figure 1: Architecture of the proposed SERI framework.

guage models, such as GPT (Radford et al., 2018, 2019; Brown et al., 2020) and BERT (Devlin et al., 2019), and sequence-to-sequence models, such as BART (Lewis et al., 2019) and T5 (Raffel et al., 2020), have demonstrated their superior capabilities in natural language understanding and generation. Many recent chatbot applications have been developed based on these pre-training language models. For example, the DialoGPT (i.e., dialogue generative pre-trained transformer) (Zhang et al., 2019) extents GPT-2 by training it on large-scale conversations and shows the superior capability of generating coherent and diverse contents. Wolf et al. (2019) proposed TransferTransfo, extending GPT2 with a multi-task objective that combines several unsupervised prediction tasks. However, no previous research has developed a chatbot program to generate fluent conversations between a cybergroomer and a potential victim through the learning on public conversations and task-specific datasets.

3 The Proposed Generative Chatbot Framework: SERI

Figure 1 shows the overall architecture of the proposed SERI framework with three components: (1) Predicting a stage label for each utterance from perpetrators of the PJ dataset; (2) Pre-training a perpetrator chatbot and a potential victim chatbot on the large-scale ConvAI2 dataset; and (3) Fine-tuning the four perpetrator subchatbots and the victim chatbot using the preprocessed PJ dataset.

Classifying perpetrators' messages per stage. Since the cybergrooming dataset in (Zambrano et al., 2019) is labeled by six stages, we leverage the state-of-the-art BERT (Devlin et al., 2019) to train a stage classifier for the perpetrators. Specifically, given each utterance u, we attach a special token $[CLS]$ before u and feed the sequence into a pre-trained BERT encoder. We use the encoding output of $[CLS]$, denoted as $\boldsymbol{u}$, as the overall contextual representation of the utterance u. Finally, we apply a linear classifier layer to predict a label out of the six stages for u. The BERT encoder and the classifier are optimized by minimizing the following categorical cross-entropy loss:

$$\mathcal{L}_1 = -\frac{1}{|U|}\sum_{u\in U}\sum_{s\in S} y_{u,s}\cdot\log(\tilde{y}_{u,s}), \quad (1)$$

$$\text{where} \quad \tilde{\boldsymbol{y}}_u = \text{softmax}(\boldsymbol{W}\cdot\boldsymbol{u}+\boldsymbol{b}),$$

where U and S are the set of utterances and stage labels, respectively. The $\tilde{\boldsymbol{y}}_u$ denotes a vector of probabilities over all stages for u and $\tilde{y}_{u,s}$ is the probability of a particular stage s. The $y_{u,s}$ is a binary indicator to show whether s is the same as the ground truth stage label of u ($y_{u,s}=1$) or not ($y_{u,s}=0$). The $\boldsymbol{W}$ and $\boldsymbol{b}$ are learnable parameters.

However, the cybergrooming progress stages explored in (Zambrano et al., 2019) were not clearly defined. We found that many utterances from the perpetrator could fit multiple stages. Therefore, we

Stages	Conversation Content
$\tilde{s}_1$	Greetings and casual talks to establish a trust relationship
$\tilde{s}_2$	Collecting private information, such as name, age, gender, location, interests, family, school, or schedule
$\tilde{s}_3$	Asking sexual questions or requests, talking about sexual conversations, or sending sexual pictures/videos
$\tilde{s}_4$	Attempting a personal contact or asking meeting in person

Table 1: Cybergrooming stages

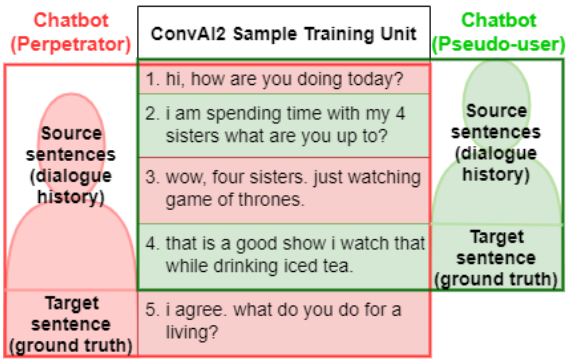

Figure 2: A sample training unit for the perpetrator and pseudo-user (i.e., potential victim) chatbots.

refined the six stages in (Zambrano et al., 2019) to develop four new stages for the grooming process. We summarize the key conversation contents covered by each stage in Table 1. We can simplify the six stages by merging similar original stages. That is, we merge stages s_1 and s_4 as new stage $\tilde{s}_1$, stages s_2 and s_3 as new stage $\tilde{s}_2$, stage s_5 as new stage $\tilde{s}_3$, and stage s_6 as new stage $\tilde{s}_4$. In the end, each utterance in the PJ dataset can obtain a label from the four new stages based on the BERT classifier.

Pre-training the chatbots on the ConvAI2 dataset. For each role of the perpetrator and potential victim, we build a chatbot model using T5 (Raffel et al., 2020) with the PyTorch framework. As the in-domain PJ dataset is small, to improve the fluency of the generated conversations, we first pre-train T5 with the large-scale ConvAI2 dataset, which contains high-quality general conversations.

To train the T5-based chatbots, we concatenate two dialogue turns (i.e., four sentences) as a unit, take the last sentence as the target one (i.e., ground truth response), and treat the preceding sentences as the sources (i.e., dialogue history). Figure 2 shows a sample chatbot training unit of four sentences. The conversations in the ConvAI2 dataset are usually between two persons, where the one initiating the conversation plays a leading role with more leading topics or questions. Since this leading role matches a perpetrator's nature, for all conversations in the ConvAI2 dataset, we treat the leading person as the perpetrator and the other one as the potential victim and formulate the training utterances accordingly. Following (Raffel et al., 2020), given an input sequence x as the source, we generate the response by optimizing the following objective:

$$\mathcal{L} = -\sum_{i} \log P(y_i | y_{i-k}, \ldots, y_{i-1}; x; \Theta), \quad (2)$$

where Θ denotes the set of parameters in the T5, and y_i is the i-th token of the target response.

We pre-train the perpetrator and the potential victim chatbots separately on the ConvAI2 dataset. We observe that the perpetrator chatbot tends to generate more leading dialogues while the potential victim chatbot generates response messages more consistently.

Fine-tuning the chatbots on the PJ dataset. A perpetrator usually follows the four grooming stages, as shown in Table 1, to gradually obtain trust from the potential victim and achieve the final cybergrooming goal progressively. To model the perpetrator's responses at the four stages, we fine-tune the four subchatbots for the perpetrator based on the in-domain PJ dataset. To obtain the messages for each stage, we cut conversations in the PJ dataset into several blocks and assign a stage for each block based on the criteria in Table 2.

The connection strength of a block from the previous utterances is crucial to determine each block locus and improve the quality of training of each stage. To split the conversations into blocks, we estimate two types of connectivity from the pre-trained BERT next sentence prediction model (Devlin et al., 2019): (1) The connectivity score, g_1, between each utterance and the last utterance from the perpetrator; and (2) The connectivity score, g_2, between each utterance and the last utterance from the victim. Thus, the connectivity between each utterance and the previous contexts is represented by $g_1 + g_2$. Furthermore, the beginning of each block is refined by comparing the connectivity scores to three utterances: The first utterance of the current block and its two previous utterances from the perpetrator. We use the utterance with the minimum of $g_1 + g_2$ as the new beginning of the block. This way allows us to refine the beginning of all the blocks and obtain four groups of blocks for the four stages. We fine-tune the four perpetrator subchatbots on the four groups of blocks separately. Further, we fine-tune a victim chatbot based on the victim utterances from the PJ dataset.

Finally, to generate consistent and high-quality (i.e., human-like) conversations, we allow each chatbot to generate five candidate messages at each time and select the best one based on their connec-

Stages	Label Distribution of Each Block
$\tilde{s}_1$	More than 80% utterances are labeled as $\tilde{s}_1$
$\tilde{s}_2$	More than 60% utterances are labeled as $\tilde{s}_2$
$\tilde{s}_3$	More than 50% utterances are labeled as $\tilde{s}_3$
$\tilde{s}_4$	More than 40% utterances are labeled as $\tilde{s}_4$

Table 2: Conversation segmentation criteria for the four relationship stages.

tivity scores to the previous message. The connectivity scores are computed based on the pre-trained BERT next sentence prediction model and used to ensure the consistency of a generated message with the context earlier.

Stage evolution of the perpetrator subchatbot. We design a cybergrooming stage evolution for the chatbots by observing whether the conversation of each stage maintains a certain number of rounds (e.g., 20). If the conversation of stage $\tilde{s}_1$ lasts 20 rounds between the perpetrator and victim chatbots, the perpetrator will move to stage $\tilde{s}_2$. Once the victim detects the perpetrator's grooming intent, he/she will leave the chat conversation immediately and the current stage lasts less than 20 rounds.

Parameter	Value	Parameter	Value
Learning rate (lr)	$5e^{-5}$	Epochs	4
Epsilon (ε)	$1.0e^{-6}$	Batch size	8
Warmup steps	500	GPU	Yes
Early stopping	0	Vocabulary	T5-base

Table 3: Parameters and their default values used for the SERI framework.

4 Experiment Setup

Datasets. We trained our chatbots using two chatlog datasets. The ConvAI2 dataset (Dinan et al., 2019) is a two-person casual chat dataset in JSON format with several different repeated labels. The sentences with the "history" label fit best for our task. Hence, we use the 2,000 dialogues with more than 60K utterances from the ConvAI2. We manually downloaded the PJ dataset from the PJ website [1] in HTML format. It contains 100 dialogues with more than 100K chat records between perpetrators and professionally trained volunteering undercover police officers mimicking potential victims [2]. We randomly divided the PJ dataset into train set, valid set, and test set with a ratio of 8:1:1. Table 3 summarizes the key parameters of the T5 model.

[1] http://www.perverted-justice.com/?archive=byUserVotes

[2] http://www.perverted-justice.com/index.php?pg=policeinfo

Role	BLEU	ROUGE	BERTScore
	Max:100	Max:1	Max:1
Perpetrator	2.9906	0.0970	0.8311
Victim	2.6884	0.1063	0.8274

Table 4: BLEU, ROUGE, and BERTScore-based analysis for the conversations generated by the SERI.

Data cleaning. The ConvAI2 dataset is well-organized and ready for our chatbots training. However, the PJ dataset contains a lot of noises, such as URLs, Hashtags, Mentions, or Emojis. We removed the noises by regular expressions in Python library 'Preprocessor.' There are repeated occurrences of informal languages, such as lexical slangs and consecutive words without spaces. To segment consecutive words with spaces, we applied 'wordsegment' library in Python. Lexical slangs can be normalized with a state-of-the-art lexical normalization model, called MoNoise (van der Goot, 2019).

Metrics. To evaluate the performance of our chatbot, we use both referenced and unreferenced metrics for evaluating automatic dialogues (Finch and Choi, 2020). For referenced metrics, we use BLEU (Post, 2018), ROUGE (Lin, 2004), and BERTScore (Zhang* et al., 2020) to evaluate the quality of the chatbot generated utterances by comparing them against the ground truth from the PJ dataset. For unreferenced metrics, we use perplexity and MaUde scores (Sinha et al., 2020). Perplexity is to measure how easily a sentence can be understood and lower perplexity indicates higher fluency. MaUde measures multiple aspects of quality in languages in terms of fluency, reasonableness (i.e., logical flow), or avoiding repetition. We compare the scores of perplexity and MaUde under both the ground truth utterances from the PJ dataset and the conversations generated by our proposed SERI.

We also conducted human evaluation by randomly selecting 200 conversation samples where each sample contains 4 history utterances and 2 target utterances (i.e., the original utterance from the PJ dataset and an utterance generated by the SERI). For each sample, we ask two graduate students and one NLP expert to compare the two target utterances and select which one is more valid and consistent with the history utterances than the other.

	Perpetrator	Victim
Ground truth dialogues	357.06	477.82
Generated dialogues	139.46	188.97

Table 5: Perplexity score-based analysis.

	Perpetrator	Victim
Ground truth dialogues	0.8442	0.8625
Generated dialogues	0.8662	0.8641

Table 6: MaUde score-based analysis based on PJ evaluation dataset.

5 Experimental Results & Analysis

Referenced metrics-based analysis. Table 4 shows the BLEU, ROUGE, and BERTScore evaluation results. A higher score (with the max value specified) means a higher similarity between the generated and ground truth dialogues. As shown in Table 4, BLEU and ROUGE scores instead reflect a low similarity between the generated and ground truth dialogues because online chatting languages are often informal and do not follow strict grammar or fluency rules. The BERTScore is relatively high due to: (1) Most of the words are functional and uninformative, such as *yes*, *haha*, or *why*, which makes it difficult for the BERT to learn meaningful contextual representations; and (2) The BERTScore is highly sensitive to some particular word pairs which do not capture any meaningful semantics of very short messages.

Unreferenced metrics-based analysis. In Table 5, we demonstrate perplexity scores of the ground truth dialogues and the generated dialogues by the SERI. For the PJ evaluation dataset, we observe higher perplexity scores in ground truth dialogues than the generated dialogues by our SERI. The perplexity measures how easily sentences can be understood in terms of grammatical correctness or logical flows. Therefore, it is reasonable to observe lower perplexity scores based on our generated dialogues than the ground truth dialogues as the original dialogues are very informal and contain a lot of grammatical errors.

In Table 6, we also demonstrate the MaUde scores of ground truth and our SERI's generated dialogues by each chatbot (i.e., perpetrator and victim) under the PJ evaluation dataset. This MaUde score represents the general reasonableness of the language being used. We observed higher MaUde scores in our generated dialogues than the original conversations. This implies that our chatbots can effectively mitigate the adverse effects of informal languages used in the PJ dataset.

	Utterance
Context	1: nutting , you miss me 2: ya 3: you better 4: what if i don't ? , lol , jk 5: i'll get you 6: can't get me through the competition duh , i'm not scared of you
Original response	lol, how much you miss me
Generated response	i'm scared of you right now

Table 7: Inter-agreement sample of human evaluation.

Human evaluation analysis. Based on the human evaluation, we find that the utterances generated by the SERI are chosen over human-written utterances by at least two annotators for 74 out of 200 samples, reaching a 37% passing rate for this Turing test (Turing, 2009), showing the promising performance of the SERI. Table 7 shows an example where the utterance generated by the SERI was preferred by the three annotators.

6 Conclusions & Future Work

This study gives the following **key findings**: (1) Training the dialogue model with accurate context utterances and target utterance can help distinguish the role of chatbots; (2) Segmenting a training corpus according to each stage can help control the output of a dialogue model; and (3) Our human evaluation also shows the promising performance of the SERI by reaching a 37% passing rate for the Turing test.

As discussed in Section 5, we found some **limitations** because the PJ dataset has poor readability. Even if we cleaned the PJ dataset using data cleaning methods, the inherently poor quality of the languages used in the PJ dataset was a big challenge to improve the performance of the generated dialogues by our SERI. This implies that the languages used by cybergrooming perpetrators and potential victims are not really intelligent compared to those of normal people.

For the **future research**, we plan to: (1) Conduct deeper data cleaning to find more effective ways to normalize social slangs; and (2) Investigate how deep reinforcement learning can optimize the current generation model to introduce a perpetrator's strategic conversations.

Ethical Statement

Our goal in developing the SERI is to simulate the authentic conversations between perpetrators and potential victims, especially human youth users. A general approach to ensure proper rather than malicious application should incorporate ethical considerations as the first order principles in each step of the system design. In this paper, we focus on developing a chatbot approach to educate youth users by increasing their awareness and sensitivity to cybergrooming and its consequence and accordingly protect them from cybergrooming. We acknowledge the pros and cons of releasing details of the SERI. Here we provide some example scenarios where the SERI should or should not be used:

- **Should-Do:** Educational parties use the SERI to develop curricula to educate youth in terms of how to respond to online abusive messages and avoid cybergrooming when a youth has a chance to have online conversations with a stranger or acquaintance talking about sexually sensitive or private information.
- **Should-Do:** Parents who want to learn grooming conversations to educate their children to be resistant and resilient against the potential risk of encountering sexual predators.
- **Should-Not-Do:** Anyone using the SERI as a tool for online sexual exploitation or abuse of children.

Besides the above regulations that we will use to ensure the properly and ethically use of SERI, we will also design several strategies to prevent the misuse and its adverse influence:

- First, part of the adverse influence and ethical concerns of SERI lies in the sensitive and inappropriate languages used by the chatbots. To mitigate this issue, we will design approaches and leverage linguistic resources, such as the profane lexicons[3], to replace filthy words in the training dataset with moderate ones and balance between simulating a realistic cybergrooming scenario and avoiding any potential ethical issues or bad influence to youths.
- Instead of releasing the source code and models of SERI to the public, we will make them to be accessible only to parties for research purposes by request.
- When delivering SERI as an education program, we will only include the perpetrator chatbot and allow youths to chat with it. We will design approaches to monitor the language generated by the chatbot and stop the conversation by the monitoring system or the users whenever filthy language is detected. This will prevent the SERI from being misused by a bad party as the SERI will stop working when the user is detected as an adult or potential perpetrator. Finally, the conversational data will be encrypted and stored under the regulations and standards stated in the legal frameworks, such as GDPR[4].

[3] https://www.cs.cmu.edu/~biglou/resources/

[4] https://gdpr-info.eu/

References

P. Anderson, Z. Zuo, L. Yang, and Y. Qu. 2019. An intelligent online grooming detection system using AI technologies. In 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pages 1–6.

P. Bours and H. Kulsrud. 2019. Detection of cyber grooming in online conversation. In 2019 IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–6.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

K. R. Choo. 2009. Online child grooming: A literature review on the misuse of social networking sites for grooming children for sexual offences, volume 103. Canberra: Australian Institute of Criminology.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186.

Z. Dhouioui and J. Akaichi. 2016. Privacy protection protocol in social networks based on sexual predators detection. In Proceedings of the International Conference on Internet of Things and Cloud Computing, ICC'16, New York, NY, USA. Association for Computing Machinery.

E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, R. Lowe, et al. 2019. The second conversational intelligence challenge (ConvAI2). arXiv preprint arXiv:1902.00098.

M. A. Fauzi and P. Bours. 2020. Ensemble method for sexual predators identification in online chats. In 2020 8th International Workshop on Biometrics and Forensics (IWBF), pages 1–6. IEEE.

S. E. Finch and J. D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. CoRR, abs/2006.06110.

F. E. Gunawan, L. Ashianti, and N. Sekishita. 2018. A simple classifier for detecting online child grooming conversation. TELKOMNIKA, 16(3):1239–1248.

C. Laorden, P. Galán-García, I. Santos, B. Sanz, J. M. Hidalgo, and P. G. Bringas. 2013. Negobot: A conversational agent based on game theory for the detection of paedophile behaviour. In International Joint Conference CISIS'12-ICEUTE 12-SOCO 12 Special Sessions, pages 261–270. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

C. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

S. Marchenko. 2017. Web of darkness: Groomed, manipulated, coerced, and abused in minutes.

Koons Family Institute on International Law and Policy. 2017. Online Grooming of Children for Sexual Purposes: Model Legislation & Global Review. International Centre for Missing and Exploited Children.

Perverted Justice Foundation Inc. 2020. Perverted-justice.com archives.

M. Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

K. Sinha, P. Parthasarathi, J. Wang, R. Lowe, W. L. Hamilton, and J. Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. ACL.

A. M. Turing. 2009. Computing machinery and intelligence. In Parsing the Turing Test, pages 23–65. Springer.

UNICEF. 2017. The state of the world's chidren 2017: Children in a digital world.

R. van der Goot. 2019. MoNoise: A multi-lingual and easy-to-use lexical normalization tool. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 201–206, Florence, Italy. Association for Computational Linguistics.

G. Winters and E. Jeglic. 2016. Stages of sexual grooming: Recognizing potentially predatory behaviors of child molesters. Deviant Behavior, pages 1–10.

T. Wolf, V. Sanh, J. Chaumond, and C. Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. CoRR, abs/1901.08149.

P. Zambrano, J. Torres, L. Tello-Oquendo, R. Jácome, M. E. Benalcázar, R. Andrade, and W. Fuertes. 2019. Technical mapping of the grooming anatomy using machine learning paradigms: An information security approach. IEEE Access, 7:142129–142146.

T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In International Conference on Learning Representations.

Y. Zhang, S. Sun, M. Galley, Y. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. 2019. DialoGPT: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536.