Generative Chatbot Framework for Cybergrooming Prevention

Pei Wang

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Jin-Hee Cho, Chair
Lifu Huang
Chang-Tien Lu

December 1, 2021
Blacksburg, Virginia

Keywords: Cybergrooming, Natural Language Processing, Chatbot

Generative Chatbot Framework for Cybergrooming Prevention

Pei Wang

ABSTRACT

Cybergrooming refers to the crime of establishing personal close relationships with potential victims, commonly teens, for the purpose of sexual exploitation or abuse via online social media platforms. Cybergrooming has been recognized as a serious social problem. However, there have been insufficient programs to provide proactive prevention to protect the youth users from cybergrooming. In this thesis, we present a generative chatbot framework, called SERI (Stop cybERgroomIng), that can generate simulated conversations between a perpetrator chatbot and a potential victim chatbot. To realize the simulation of authentic conversations in the context of cybergrooming, we take deep reinforcement learning (DRL)-based dialogue generation to simulate the authentic conversations between a perpetrator and a potential victim. The design and development of the SERI are motivated to provide a safe and authentic chatting environment to enhance the youth's precautionary awareness and sensitivity of cybergrooming while any unnecessary ethical issues (e.g., the potential misuse of the SERI) are removed or minimized. We developed the SERI as a preliminary platform that the perpetrator chatbot can be deployed in social media environments to interact with human users (i.e., youth) and observe the conversations that the youth users respond to strangers or acquaintances when they are asked for private or sensitive information by the perpetrator. We evaluated the quality of conversations generated by the SERI based on open-source, referenced, and unreferenced metrics as well as human evaluation. The evaluation results show that the SERI can generate authentic conversations between

two chatbots compared to the original conversations from the used datasets in perplexity and MaUde scores.

Generative Chatbot Framework for Cybergrooming Prevention

Pei Wang

GENERAL AUDIENCE ABSTRACT

Cybergrooming refers to the crime of building personal close relationships with potential victims, especially youth users such as children and teenagers, for the purpose of sexual exploitation or abuse via online social media platforms. Cybergrooming has been recognized as a serious social problem. However, there have been insufficient methods to provide proactive protection for the youth users from cybergrooming. In this thesis, we present a generative chatbot framework, called SERI (Stop cybERgroomIng), that can generate simulated authentic conversations between a perpetrator chatbot and a potential victim chatbot by applying advanced natural language generation models. The design and development of the SERI are motivated to ensure a safe and authentic environment to strengthen the youth's precautionary awareness and sensitivity of cybergrooming while any unnecessary ethical issues (e.g., the potential misuse of the SERI) are removed or minimized. We used different metrics and methods to evaluate the quality of conversations generated by the SERI. The evaluation results show that the SERI can generate authentic conversations between two chatbots compared to the original conversations from the used datasets.

I dedicate my thesis work to my family. A special feeling of gratitude to my parents, friends and professors.

Acknowledgments

First of all, I would like to express my sincere gratitude to my advisor, Dr. Jin-Hee Cho, for her guidance throughout this research. She has been accommodating all the way, guiding me patiently through every problem I face and nudging me in the right direction. Without her support and guidance, I could not have made progress with this research. I would also like to thank Dr. Lifu Huang as my co-advisor for his time and encouragement to go through the crucial part of my research and provide valuable guidance. I express my honest appreciation to Dr. Chang-Tien Lu. for his critical comments for revising this thesis and information for future work directions. Finally, I would like to thank my family and friends for all the moral support they have provided throughout this research. I also acknowledge that this research is partly supported by Virginia Tech's ISDA-ISERC (Integrated Security Destination Area-The Integrated Security Education and Research Center) Research Program.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

BLEU BiLingual Evaluation Understudy

ConvAI2 Conversational Intelligence Challenge 2

DDQ Deep Dyna-Q

DRL Deep Reinforcement Learning

ES-DDQ Emotion-SensitiveDeep Dyna-Q

GPT Generative Pre-Training

IRL Interactive Reinforcement Learning

MDP Markov Decision Process

ML Machine Learning

MLM Masked Language Modeling

MMI Maximum Mutual Information

MT5 Multilingual variant of T5

NLG Natural Language Generation

NLP Natural Language Processing

NLU Natural Language Understanding

NSP Next Sentence Prediction

PJ Perverted Justice

PTM Pre-trained Model

RNN Recurrent Neural Network

ROUGE Recall-Oriented Understudy for Gisting Evaluation

SARSA State-Action-Reward-State-Action

seq2seq Sequence-to-Sequence

SERI Stop cybERgroomIng

T5 Text-to-Text Transfer Transformer

TextCNN Convolutional Neural Network for text

Chapter 1

Introduction

Cybergrooming is defined as the crime of establishing a personal trust relationship with potential victims, commonly youth, via the Internet only for sexual exploitation or abuse [1]. Cybergrooming is one of the well-known online social deception attacks in online social media [2]. Developing an effective scheme to mitigate the threats of cybergrooming plays an important and practical role in online social media and computer science. A generative chatbot can be developed and deployed in an online social media environment to provide proactive prevention to protect potential youth victims from cybergrooming. This chapter introduces research motivation and challenges, followed by the research goal, the summary of the key contributions, and the structure of this thesis.

1.1 Motivation & Challenges

With the development of the Internet and computing devices, online social media technology has been evolved to a new advanced level. Today anybody can easily have a computer or a mobile device to actively access online resources and interact with other online people. Meanwhile, more and more young people have become the users of the online applications. As of 2017, approximately one-third of online users in the world are known young people below the age of 18 [3]. The Internet and social media technologies have brought people countless benefits in almost all areas of our life. On the other hand, the proliferation of

online social media has also introduced various social and security problems due to their easy access and deceptive exploitation. Cybergrooming is one of these problems in which a perpetrator uses the social media technology and environment to practice online sexual exploitation and abuse of children and teenagers [1, 4]. CyberTipline reported that about 60,000 cases were received on luring children for sexual purposes in cyberspace from 1998 to 2003 in the USA [5].

Due to the unpredictable harmfulness of cybergrooming, some researches have been conducted to study cybergrooming. The majority of these cybergrooming researches focused on investigating the special properties of cybergrooming and detecting online child sexual exploitation or predators by analyzing malicious conversations collected from online chatting rooms [6, 7, 8]. Detecting predators from collected online conversations is not a highly challenging work. It is a more reactive process. Because of the unique vulnerability of the youth as teenagers under puberty, a proactive approach is more effective and important to protect the youth from becoming victims in cybergrooming. A proactive prevention program can enhance the sensitivity and awareness of potential victims to the online cybergrooming incidents. Therefore, our research work is motivated to develop such a proactive cybergrooming prevention program that can help our youth to get alerts while the normal chatting atmosphere is gradually deviated to cybergrooming.

To develop a proactive cybergrooming prevention program, we first need to develop a generative chatbot that can be deployed in a chatting room to represent the perpetrator to interact with the potential victims, youth users. During the online conversation, the chatbot can simulate the perpetrator to chat with the youth user, observe the emotion changes of the user, and control the conversation to different stages accordingly. At the end, if the youth user becomes a victim, the chatbot can be based on the situation to provide different warning information to help the victim from cybergrooming.

This thesis aims to develop a generative chatbot framework that can provide authentic conversations between a perpetrator chatbot and a youth user to achieve stopping cybergrooming ultimately. This generative chatbot framework is named as 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 in a social media environment. 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. With the generative chatbot framework, various applications can also be developed on the framework to implement various protection details from different views.

While developing the generative chatbot framework, we have faced the following research challenges:

1. Unlike general conversations between a normal adult and a teen, the perpetrator's words are goal driven and tend to lead a conversation with a potential victim. Ultimately, the perpetrator aims to meet the potential victim in person and exploit the relationship to commit a serious, potential sexual crime. Thus, the perpetrator often takes multiple stages, such as establishing a trust relationship with a potential victim in the initial conversation, gradually escalating its stage to obtaining private information, and ultimately meeting up with the victim in person. However, no prior work has addressed such goal-oriented conversations in the context of cybergrooming.

2. A lack of proper datasets has been a non-trivial hurdle in developing a generative chatbot generating authentic conversations. Most related work to online sexual exploitation has used the Perverted Justice (PJ) dataset (Perverted Justice Foundation Inc. [9]), which is the only publicly available dataset that the chatbot can mimic.

The PJ dataset contains the chatlogs between cybergrooming perpetrators and professionally trained volunteers playing the role of potential youth victims. However, the limited volume of the PJ dataset (i.e., 100 sets of conversations) as well as a lot of noises, such as emojis, slangs, short abbreviations, unsegmented words, or URLs, have been a major challenge to generate high-quality conversations with high logical flows, fluency, and human-like languages.

1.2 Research Goal

Cybergrooming is an online threat to youth users of social media applications. Existing studies on cybergrooming mainly focused on detecting perpetrators or malicious conversations from the collected data. There have been insufficient programs to provide proactive prevention to protect potential youth victims from cybergrooming. To implement a proactive prevention, we need to have an agent that can be deployed in a social media environment to interact with human users and observe the chatting information during online conversations in real-time. A generative chatbot, if deployed in the chatting rooms of social media environments, can fulfill this fundamental need. To meet this need, this research aims to design and develop a generative chatbot framework that can build an authentic conversational platform using advanced natural language processing approaches. Different applications can be developed on the platform to implement different protection details for protecting youth users. With this framework, a safe cybergrooming protection environment could be established for youth users to make an authentic dialogue with a stranger or acquaintance and learn how to carefully respond to such a person asking for sensitive or private information.

To support the thesis research goal described above, we have the following objectives to achieve:

1. Design and develop a generative chatbot framework that can be deployed in a social media environment to implement a proactive cybergrooming prevention program.

2. Develop a perpetrator chatbot to manage and control the whole grooming conversations.

3. Develop a potential victim chatbot and use it to build the interactive conversations with the perpetrator chatbot.

4. Employ modern advanced natural language processing models to implement the simulation of authentic conversations between a perpetrator chatbot and a potential victim chatbot.

5. Make the perpetrator chatbot be able to identify and control grooming stages of the conversations with the potential victims, and achieve the ultimate attack goal.

6. Make the perpetrator chatbot be able to manage the authentic dialogue generation according to the grooming stages.

7. Train two chatbots with the public ConvAI2 dataset and the the grooming-domain-specific PJ dataset to make the generative chatbot framework be grooming-domain oriented.

8. Provide ethical statement so that the SERI will be used in the real social media environment properly.

1.3 Key Contributions

Through the developed generative chatbot framework, SERI, we made the following key contributions:

1. When training the perpetrator chatbot, we employed deep reinforcement learning (DRL) [10] to generate authentic, strategic dialogues where the perpetrator has a clear and ultimate attack goal to achieve offline sexual exploitation. By applying rewards matching a target stage and the corresponding occurrence generation, we augmented the quality of the dialogue model that can generate strategic conversation describing the cybergrooming attack behavior.

2. Based on the T5 (Text-to-Text Transfer Transformer) model [11], we applied a two-stage paradigm to train the SERI 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) [12]. Then the chatbots were fine-tuned on the domain-specific PJ dataset [9], which was pre-processed with a series of social text normalization tools to remove the informal slangs, abbreviations, unsegmented words, emojis, or URLs from those conversations.

3. We identified four grooming stages based on the existing conversational dataset and behavior changes. We applied these four grooming stages to model the perpetrator's conversational strategies to achieve the ultimate attack goal, which is meeting up with a victim in person. We predicted the grooming stage of each dialogue utterance by training a TextCNN [13]. Then the perpetrator chatbot was trained via the T5 model to provide a response in the target stage. The chatbot was supplied with the dialogues and the reward based on the corresponding grooming stage to guide the dialogue generation.

4. We defined and applied an attack stage-based grooming strategy to manage the dialogue generation based on the perpetrator chatbot and the four grooming stages. When the perpetrator leads the conversation and collects sufficient resources from the current stage, the perpetrator will switch to a next-stage by starting a trigger utter-

ance to continue its conversation. If the potential victim shows alertness to this switch towards cybergrooming and terminates the conversation, it indicates a failure of the cybergrooming attack by the perpetrator.

5. We evaluated the SERI by using both referenced metrics (i.e., BLEU [14], ROUGE [15], and BERTScores [16]) and unreferenced metrics (i.e., perplexity and MaUde scores [17]). The evaluation results demonstrated that the conversations generated by the SERI had higher quality than the ground truth conversations for all the above metrics. Particularly, the human evaluation verified that approximately 37% of dialogues generated by the SERI were preferred over the ground truth dialogues from the PJ dataset.

1.4 Structure of the Thesis

The thesis is organized as follows:

- Chapter 2 introduces the related work on cybergrooming detection, chatbot application tool, pre-training language models and DRL-based conversation generations.
- Chapter 3 describes all the preliminary models used in developing the SERI framework.
- Chapter 4 describes the design and implementation details of the proposed generative chatbot framework, SERI.
- Chapter 5 presents the experiment details that use different datasets to train the chatbots to generate authentic conversations.
- Chapter 6 provides an analysis on experimental results.
- Chapter 7 gives the conclusion and discusses the future work.

- Finally, an ethical statement is presented in Appendix A.

Chapter 2

Background & Related Work

2.1 Cybergrooming Detection

Because of the emergence of the Internet in 1990s, various applications, such as emails, web, news groups, and so on, were developed on the Internet to bring many benefits for our society. At the same time, misuse of the Internet for criminal behaviors was also started gradually. One of these behaviors was online grooming, in which pedophiles misused the Internet platform to seduce children for sexual exploitation. Some studies were started to explore the properties of online grooming and provide suggestions to avoid becoming a victim in online grooming. Durkin discussed the possible activities conducted by pedophiles and analyzed the implications for law enforcement [18]. His discussion revealed that the challenge work was how to detect and identify that an activity was a misuse of the Internet.

Since then many researches have been conducted to apply different technologies to detect the predator in online grooming or cybergrooming. Machine Learning (ML) became the main technology in this research activities. Different machine learning algorithms have been chosen to implement the detection of perpetrators and malicious conversations in online cybergrooming from the online forum or social media platforms by leveraging the lexical features as well as behavioral features, for example, Support Vector Machine (SVM) [6, 8, 19, 20], k-nearest neighbors (KNN) [20], Random Forest [8], Decision Tree [8], fuzzy logic [6], Naïve Bayes [7] and Neural Network (NN) classifiers [7, 8]. Existing studies have

also developed cybergrooming attack stages among perpetrators and the victims based on the conversational relationship [21]. Perpetrators usually build a relationship and evolve to a closer stage to realize the cybergrooming crime.

While most previous studies aimed at detecting potential perpetrators and analyzing features of cybergrooming from the collected data in social media platforms [22], there is no research on identifying the features of potential victims in cybergrooming scenarios.

2.2 Chatbot Application Tools

A chatbot is a computer program that can simulate and process human conversations in real-time [23]. When a chatbot is deployed in a computer application environment, it allows humans to interact with the environment as if they are communicating with a real person. Communications between chatbots and humans can be text, voice, and/or image.

Chatbots are built not only to mimic human conversations and entertain users, but also to be useful in various application areas such as education, information retrieval, business, e-commerce, health, and so on [24]. A chatbot may have its own knowledge domain. Chatbots that can answer nearly any user questions from whichever domain are called Generic chatbots. [25]. For example, Chorus for Google Hangouts is a generic chatbot [26]. Some chatbots that can work in multiple domains are called Open-Domain chatbots. Guardian is an example for open-domain chatbot [27]. A chatbot that only works for a specific or narrow knowledge domain is called a Domain-Specific chatbot [28]. SnapTravel is a specific domain chatbot used for booking hotels online [29].

Two essential technologies, namely, the pattern matching approach and the machine learning approach, are usually used to implement chatbots [25]. The pattern matching approach uses

pattern matching algorithms to match the user input to a rule pattern and choose an answer from a set of predefined responses. The chatbots that are implemented with the pattern matching approach are referred as role-based chatbots. Obviously, the machine learning approach is to use the Natural Language Processing (NLP) and related Artificial Intelligence (AI) technologies to extract the content from unstructured user input and generate the answer via various trained language models. To generate the response, the machine learning approach commonly considers the whole dialogue context, not only the current conversation. The chatbots that are developed with the machine learning approach are called generative chatbots.

Most NLP techniques for building chatbots consist of natural language understanding (NLU), and natural language generation (NLG) [30]. NLU is based on the context information to do intent classification and entity extraction. NLG is to use various trained language generation models to automatically generate response. Both NLU and NLG need an extensive set of data to train the language models to enhance its natural language generation quality [31]. If the training dataset is not enough, some grammatical errors may occur [32]. Some of NPL models are introduced in the following sections.

To mitigate the threat of cybergrooming, a chatbot can be developed and deployed in a social media environment to represent different chatting roles. The chatbot can be used to automatically interact with human users. Besides, it can also collect the conversational data, and analyze the data to provide different services. An early-stage chatbot, named Negobot, was developed to detect and analyze potential pedophiles in the social networks [33]. A game-theoretic reward can push the chatbot toward the next grooming stage or keep the current stage.

Holtzman et al. [34] introduced Temperature, Top-k Sampling and Nucleus Sampling mechanism which are widely used in dialogue generation.

Several chatbot programs have explored the pre-training language models for conversation generation. For example, the DialoGPT (i.e., dialogue generative pre-trained transformer) [35] fine-tuned GPT-2 on a large-scale conversation dataset and applied a max mutual information (MMI) method to generate coherent and diverse conversations. TransferTransfo [36] also extends GPT-2 with a multi-task objective, combining several unsupervised prediction tasks and shows strong improvements over the end-to-end conversational models.

However, no previous chatbots have been developed to avoid cybergroominng by simulating conversations between a cybergroomer and a victim.

2.3 Pre-trained Language Models

Natural Language Processing (NLP) is the general and powerful module that can be used to make various processes on natural languages in machine learning. NLP has been brought to a new level with the emergence of pre-trained models (PTMs). When pre-training models are applied on the large corpus, PTMs can learn universal language expressions, which are so beneficial for downstream NLP tasks that prevent training a new model from scratch.

Attention mechanism [37] is another technical approach in NLP to make the rapid development of language models. Mino et al. introduced Key-value Attention Mechanism [38]. Self attention and multi-head attention are introduced by Vaswani et al. [39] in 'Attention is All You Need."

Transformer [40] is one of the most symbolic encoder-decoder language framework with an innovative implementation of attention mechanism. Based on Transformer, BERT [41] developed a bidirectional encoder framework using Masked-Language Modeling and Next

Sentence Prediction to capture word-level or sentence-level representations. BERT is good at multiple downstream tasks like sentence classification and question answering. GPT, GPT2, GPT3 [42, 43, 44] achieved good performance in text generation tasks including article summarizing, dialogue generation, etc. using a "semi-supervised" approach. In recent years, pre-trained sequence-to-sequence models, such as BART [45] and T5 [11], have demonstrated their superior capabilities in natural language understanding and generation from the large-scale data training. MT5 [46], a multilingual variant of T5 that was pre-trained on a new Common Crawl-based dataset covering 101 languages.

We used the T5 model as the base model to train our chatbots.

2.4 DRL-based Conversation Generation

Markov decision process (MDP) has been popularly employed in learning dialogue strategies for several decades. As a typical approach to learn efficient and effective dialogue strategies, RL has been commonly used [47, 48, 49, 50]. They mainly aimed to identify optimal dialogue strategies in providing high quality conversation while minimizing retrieval cost such as in an air travel information system. RL has been also used based on dialogue state transitions estimated from a corpus where a model-based RL is employed to find an optimal policy [49, 50] and automatically learn dialogue for maximizing system performance [51]. Also a simulation-based RL, such as Q-learning or SARSA (state-action-reward-state-action), has been used to generate training episodes [48] or to optimize dialogue management [52]. Recently, RL has been used with deep learning, called deep reinforcement learning (DRL), for dialogue generation. [53] considered future reward in chatbot dialogue in terms of informativity, coherence, and ease of answering and evaluated their model on diversity, length, and human judges. [54] used deep Q-learning for dialogue generation that can reflect both emotion and content.

[55] developed robots that can understand and help patients' requests under emergency situations. [56] used an interactive RL (IRL) method to train IRL agents by using simulated users, which offer affordable and faster evaluation, compared to using actual human users. [57] developed a world model and Deep Dyna-Q (DDQ) using simulated users to reduce cost of training RL agents. [58] further enhanced the DDQ by developing Emotion-Sensitive Deep Dyna-Q (ES-DDQ) model to present an emotional world considering emotion-related cues with simulated users.

However, no prior work has leveraged RL-based dialogue generation to model the behaviors of online social attackers (e.g., cybergroomers or sexual perpetrators) in terms of their attack goals and intents.

Chapter 3

Preliminaries

This chapter discusses all the preliminaries used for the SERI's development. It covers the common Sequence-to-Sequence models, and transformer-based models.

3.1 Sequence-to-Sequence Mechanism

Sequence-to-Sequence (seq2seq) [59] is a kind of machine learning tasks which aims to map a fixed-length input with a fixed-length output where the length of the input and output may differ. Seq2seq tasks are popular in natural language processing including machine translation, text generation, text summarization, etc. **Encoder-decoder** is a widely used structure for seq2seq tasks.

3.1.1 Recurrent Neural Network Encoder-Decoder.

Recurrent neural network (RNN) [60] Encoder-Decoder is one simple Encoder-Decoder framework. The RNN encoder encodes the input sequence X into a fixed-length vector C (context vector), and the RNN decoder decodes C into the target sequence Y. As shown in Figure3.5, to generate a target sequence, there are two inputs which are the input sequence and the previous generated sequence.

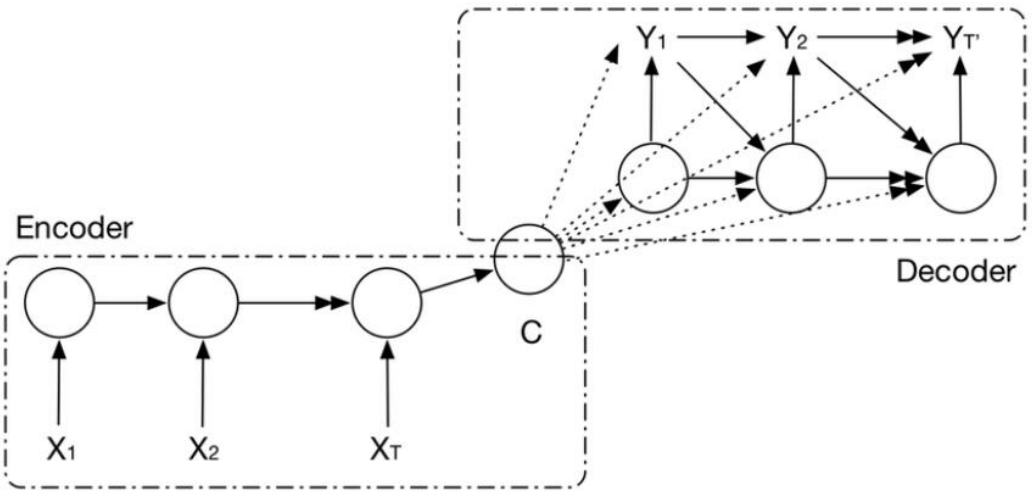

Figure 3.1: RNN encoder-decoder architecture

3.1.2 Encoder

The **encoder** is a stack of several recurrent units that takes two input vectors. One input is the vector from the embedding table. Another input vector is the previous hidden state. Each word is represented as x_t where t is the order of the input word. The calculation of a hidden state h_t is given by:

$$h_t = f(W^{(hh)}h_{t-1} + W^{(hx)}x_t) \tag{3.1}$$

where $W^{(hh)}$ is the learnable weights connecting current hidden state and previous hidden state, while $W^{(hx)}$ is the learnable weights connecting current hidden state and input word.

3.1.3 Context Vector

The **context vector** is transformed from the hidden states by a customized function q:

$$c = q(h_1, \ldots, h_T). \tag{3.2}$$

Here, when selecting h, the context variable is just the hidden state h_T of the input sequence at the final time step. The context vector aims to encapsulate the information for all input elements in order to help the decoder make accurate predictions.

3.1.4 Decoder

The **decoder** is a stack of recurrent units where each predicts an output y_t at a time step t. The decoder hidden state $s_{t'}$ is calculated using a function g, the context vector c, the previous hidden state $s_{t'-1}$, and the previous output state $y_{t'-1}$ by:

$$s_{t'} = g(W^s s_{t'-1} + W^c c + W^y y_{t'-1}). \tag{3.3}$$

After obtaining the hidden state of the decoder, we can use an output layer and the softmax operation to compute the conditional probability distribution $P(y_{t'}|y_1, ..., y_{t'-1}, c)$ for the output at time step t' .

3.2 Attention Mechanism

Since the context vector C in Figure 3.1 has a fixed-length, the capabilities of feature extraction and expression are limited. The attention mechanism [61] solves this problem by

enhancing the important parts of the input data and fading out the rest.

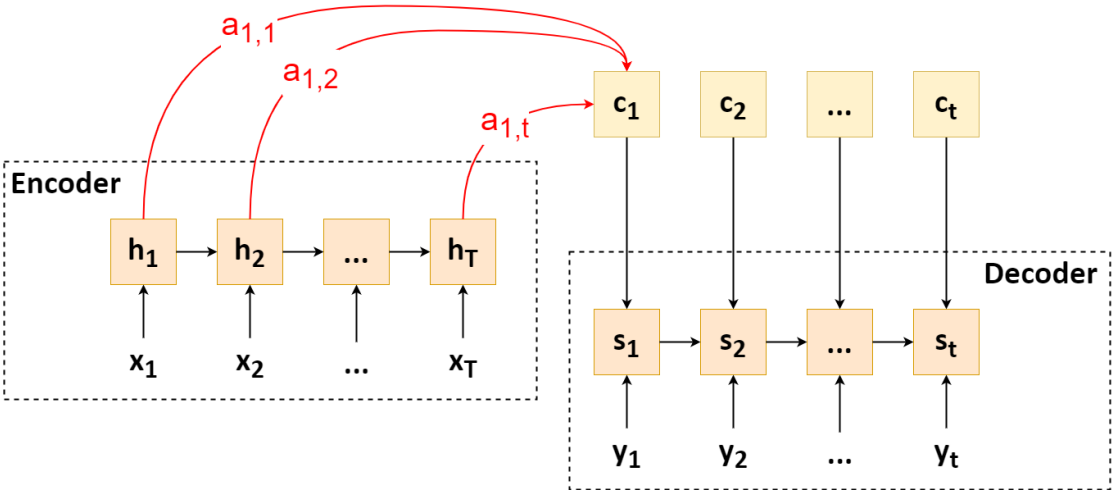

Figure 3.2: Attention mechanism in Seq2seq model

3.2.1 Soft Attention

Soft Attention is when we calculate the context vector as a weighted sum of the encoder hidden states as we had seen in the Figure 3.2. Here we introduce the concept of **(query)** q which usually denotes **the subset of previous decoder hidden states** according to the task. We also define the **input** of attention as x which is the hidden states of the encoder.

Each context vector is calculated according to the formula below:

$$c_i = \sum_{i=1}^{N} a_i x_i \tag{3.4}$$

where N is the length of input X, a is the attention distribution which is computed by:

$$a_i = \frac{\exp(s(x_i, q))}{\sum_{j=1}^{N} s(x_j, q)}. \tag{3.5}$$

A softmax is applied to get the normalized alignment scores.

The **scoring functions** $s(x_i, q)$ is the output score of a feedforward neural network to capture the alignment between input at i and query q. There are two common-used forms of function $s(x_i, q)$, **additive attention**:

$$s(x_i, q) = v^\top \tanh(W x_i + U q), \tag{3.6}$$

where v, W and U are all learnable weight matrices.

Scaled Dot-Product Attention:

$$s(x_i, q) = \frac{x_i^\top \cdot q}{\sqrt{\mathrm{d}}}, \tag{3.7}$$

where $\sqrt{\mathrm{d}}$ is a scaling factor which is to prevent gradient vanishing or exploding of softmax's backpropagation. As shown below:

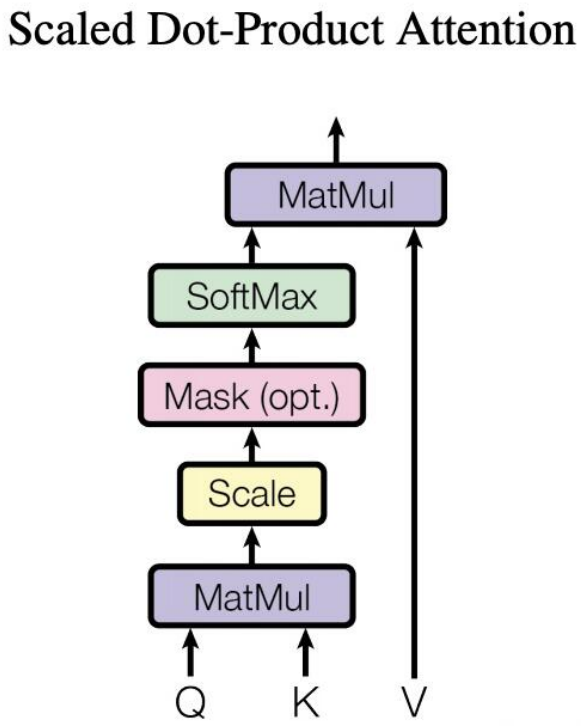

Figure 3.3: Scaled Dot-Product Attention

3.2.2 Key-Value Pair Attention

Key-value Pair Attention is developed to fix the problem that the input attention (hidden state) h participates in the process of calculating both attention a and output y.

For Key-value Pair Attention, the hidden state h are decomposed into two respective parts, which are a key and a value, The key is used for calculating the attention distribution, and the value is used for encoding the context representation. And the rest is similar to Soft Attention.

3.2.3 Self Attention

Self Attention is developed based on Key-value pair attention. In Self Attention, all query, key and value are linear-transformed by input x:

$$Q = W_q X; K = W_k X; V = W_v X. \tag{3.8}$$

Compared with using a single input X only as in Key-value Pair Attention, this transformation makes query, key and value all learnable to improve the fitting performance of the model.

3.2.4 Multi-head Attention

Multi-head Attention is to linearly project the queries, keys and values h times with different, learned linear projections to d_k, d_k and d_v dimensions. The single attention function is perform in parallel on each of these projected versions of queries, keys and values. d_v-dimensional output values are concatenated and once again projected, resulting in the final

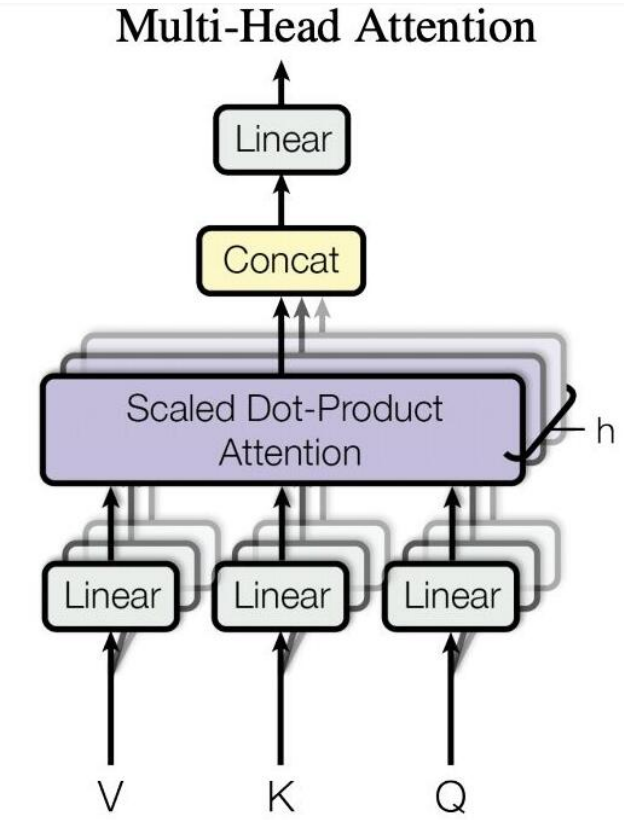

Figure 3.4: Multi-Head Attention

values, as depicted in Equation 3.9 and Figure 3.4. The multi-head attention is formulated by:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \\ \text{where head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \end{aligned} \tag{3.9}$$

3.3 Transformers

A transformer [39] is an encoder-decoder model that adopts the mechanism of multi-head self-attention. It has **two inputs**: input sequence $= (i_1, i_2, \ldots, i_p, i_N)$, output sequence $= (t_1, t_2, \ldots, t_q, t_M)$ and **one output**: output_probabilities $= (o_1, o_2, \ldots, o_q, o_M) = \text{Transformer}(inputs, outp$ as shown in Figure 3.5, where i_p, t_q and o_q are the token indexes in the dictionary.

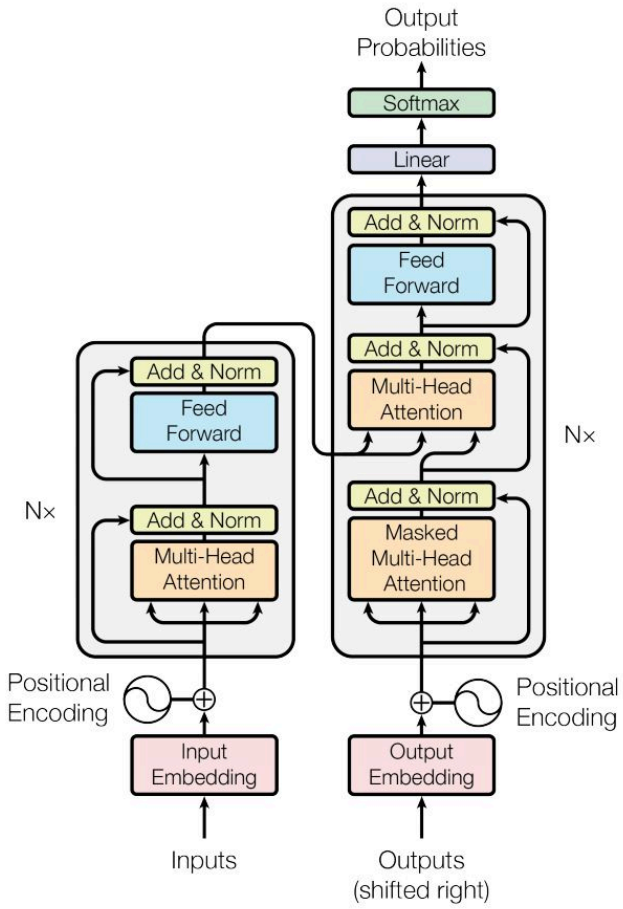

Figure 3.5: Transformers encoder-decoder architecture

Next, the inputs are vectorized with **Embedding** method. All inputs are limited with same length which means longer inputs are truncated , while shorter inputs are padded to required length.

One of the drawbacks of Attention mechanism is missing the positional information of word when calculating attention. However, the **Positional Encoding** is applied to extract and pass the position information of each element inside the inputs.

The **transformer encoder** is a stack of multiple identical layers, where each layer has two sub-players. The first is a multi-head self-attention pooling and the second is a position-

wise feed-forward network. Specifically, in the encoder self-attention, queries, keys, and values are all from the the outputs of the previous encoder layer. Inspired by the ResNet, a residual connection is applied around both subplayers. The residual connection is immediately followed by layer normalization. As a result, the transformer encoder outputs a vector representation for each position of the *input sequence*.

The **transformer decoder** is also a stack of multiple identical layers with residual connections and layer normalizations. Besides the two sub-layers described in the encoder above, the decoder inserts a third sub-layer, known as the encoder-decoder attention. In the encoder-decoder attention, queries are gain from the outputs of the previous decoder layer, keys and values are gain from the output of encoder. In the decoder self-attention, queries, keys, and values are all from the the outputs of the previous decoder layer. However, there is a limitation that each position in the decoder is allowed to only attend to all positions in the decoder before. So, a masked attention is applied to ensure that the prediction only depends on those output tokens that have already been generated.

Multiple language models are developed based on Transformer, we will discuss BERT, GPT2 and T5 below.

3.3.1 Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) encodes context bidirectionally and requires minimal architecture changes for a wide range of natural language processing tasks [41].

There are two unsupervised tasks for BERT Pre-training, that are Masked Language Model(MLM) and Next Sentence Prediction (NSP). MLM learns the context before and after a word rather than only afterwards which makes BERT able to better capture the representation of word,

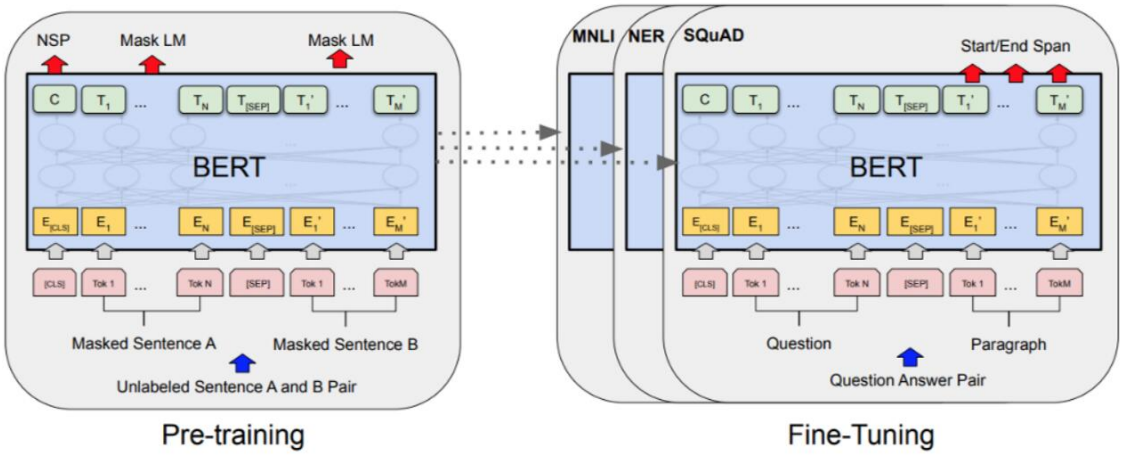

Figure 3.6: Pre-training and Fine-tuning architecture of BERT

both semantically and syntactically. NSP is helpful because many important downstream tasks such as Sentence Classification are based on the relationship of sentences, which cannot be directly captured by language modeling.

During supervised learning of downstream tasks, BERT representations will be fed into an added output layer, with minimal changes to the model architecture depending on nature of tasks, such as predicting for every token vs. predicting for the entire sequence. All the parameters of the pretrained transformer encoder are also fine-tuned, while the additional output layer will be trained from scratch.

3.3.2 The Generative Pre-Training 2 Model

Generative Pre-Training 2 (GPT-2) is a large transformer-based language model, with generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task.

Unlike original transformer architecture, GPT-2 applied the transformer decoder model discarding the encoder part, so there is only one single input sentence rather than two separate

source and target sequences. GPT2 has a uni-directional framework which is trained to predict the future left-to-right context.

We first trained our chatbot with GPT2 model.

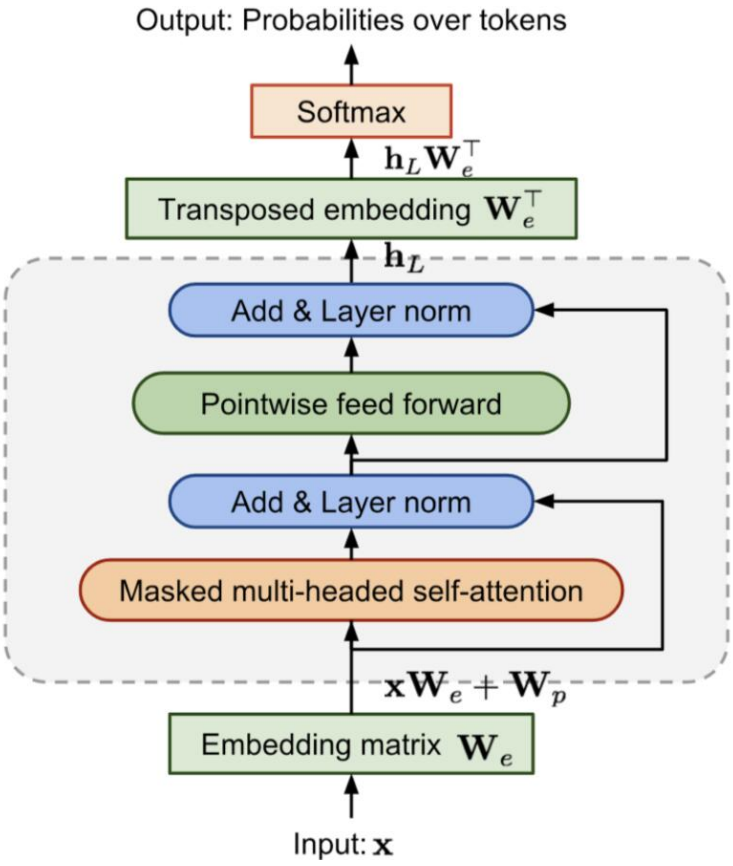

Figure 3.7: GPT2 architecture

3.3.3 Text-To-Text Transfer Transformer

Text-To-Text Transfer Transformer (T5) is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format.

For the high-level self-supervised pre-training method, T5 takes BERT-style compared with

Language Model style and Deshuffling style. For the noise adding strategy, T5 finds out that replacing spans is the best after comparing with masking tokens and dropping tokens.

Overall, T5 merges the advantages of several mechanisms like Encoder-Decoder, Fully visible and Bert-span. T5 model reaches the SOTA in many tasks.

We finally decide to use T5 model for chatbot training.

The comparison between BERT, GPT2 and T5 are shown in the table below:

	BERT	**GPT2**	**T5**
Base model	Transformer encoder	Transformer decoder	Transformer encoder-decoder
Bi-directional	✓	✗	✓
Pre-training corpus	BooksCorpus (800M) & Wikipedia (2500M)	BooksCorpus (800M)	C4(750 GB)
Parameter size (base)	110M	117M	220M

Table 3.1: Comparison between BERT, GPT2 and T5.

Chapter 4

The Proposed Generative Chatbot Framework

The proposed generative chatbot framework, named as SERI, is the fundamental platform to implement the proactive prevention to protect the potential victims from cybergrooming. Figure 4.1 provides an architectural overview of the proposed SERI framework. The SERI contains two chatbots with the four components as follows: (1) Training a cybergrooming stage classifier to assign a stage to each perpetrator's utterance in the PJ dataset; (2) Pre-training both chatbots for a perpetrator and a potential victim on the large-scale ConvAI2 dataset; (3) Fine-tuning the two chatbots on the preprocessed PJ dataset, and specifically, the perpetrator chatbot is trained with a DRL policy and a reward that measures how likely the generated utterance is from the target grooming stage; and (4) Advancing the perpetrator chatbot to a higher-level stage to continue the dialogues.

4.1 Classification of perpetrators' Messages per Stage.

The previous study [22] defined six stages from the perpetrator perspective to indicate the evolution of the cybergrooming conversations, based on which, we train a TextCNN [13] to predict a stage label for each utterance from the perpetrators. Specifically, given each utterance u, we encode it by the T5 encoder and further feed the contextual representations

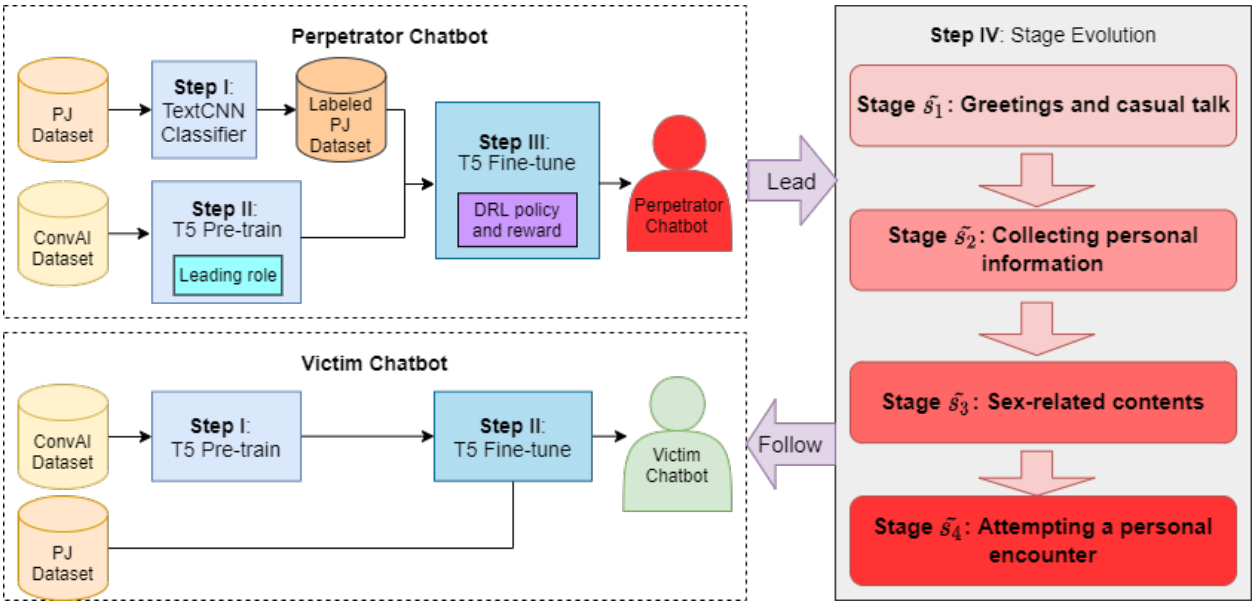

Figure 4.1: Architecture of the proposed SERI framework.

Stages	Conversation Content
$\tilde{s}_1$	Greetings, casual talks for initiation of a trust relationship
$\tilde{s}_2$	Private information collection, such as identity as name, age, gender; social relationship as family, school, location; or interests and schedule
$\tilde{s}_3$	Sexual questions or conversations, or sending/ requesting sexual pictures/videos
$\tilde{s}_4$	Attempts of in-person contact or requesting online or in-person meeting

Table 4.1: Cybergrooming stages

into a TextCNN model. The output of the convolutional layer after dropout, denoted as $\boldsymbol{u}$, is the contextual representation of u. Then we apply a linear function to classify u to one of the six stages with the softmax function. This stage classifier is optimized by minimizing $\mathcal{L}_C$, which is a categorical cross-entropy loss defined as follows:

$$\mathcal{L}_C = -\frac{1}{|U|}\sum_{u\in U}\sum_{s\in S} y_{u,s}\cdot\log(\tilde{y}_{u,s}), \tag{4.1}$$

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

where U is the set of utterances in a dialogue and S is the set of all the target stages. The $\tilde{\boldsymbol{y}}_u$ denotes a vector of probabilities over all stages for u and $\tilde{y}_{u,s}$ is the probability of predicting

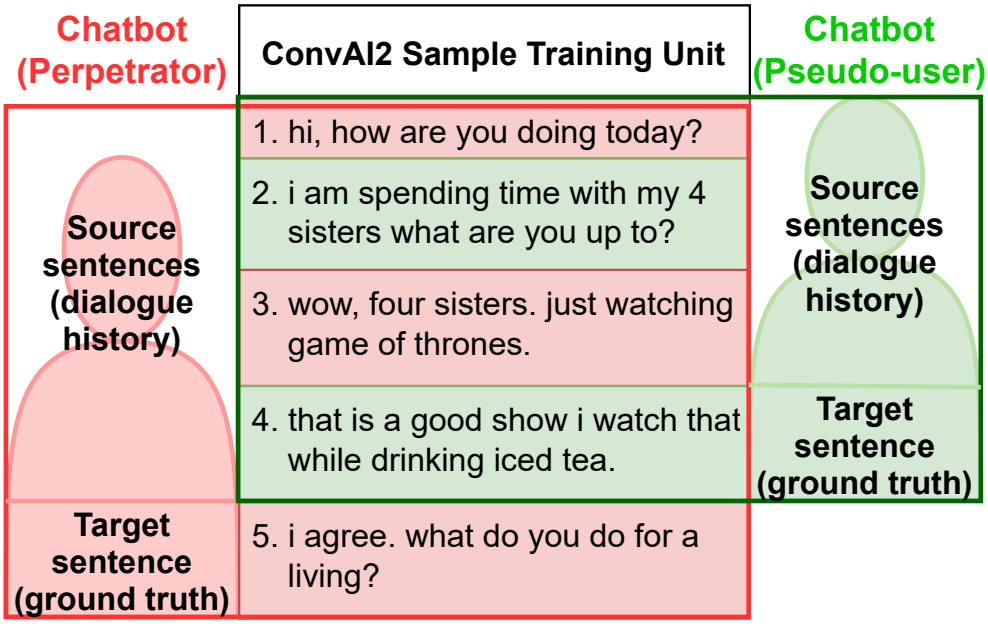

Figure 4.2: A sample training unit for the perpetrator and pseudo-user (i.e., potential victim) chatbots.

u with stage s. $y_{u,s}$ indicates whether s is the true stage label of u by $y_{u,s} = 1$ or not by $y_{u,s} = 0$. The parameters $\boldsymbol{W}$ and $\boldsymbol{b}$ from the dense layer are learnable.

However, we found the cybergrooming stages proposed by Zambrano et al. [22] were not clearly defined because some perpetrators' utterances could fall to multiple stages according to their definitions. Based on our understanding of the grooming stages, we proposed four stages by restructuring the six stages proposed by Zambrano et al. [22]. To be specific, the new stage $\tilde{s}_1$ is combined from s_1 and s_4, $\tilde{s}_2$ is combined from s_2 and s_3, $\tilde{s}_3$ is the same as s_5, and $\tilde{s}_4$ is the same as s_6. The key conversation contents and topics for the perpetrators covered by each new stage are summarized in Table 4.1. In the end, through the TextCNN stage classifier and stage consolidation, each utterance in the PJ dataset is assigned a stage label.

Further, we will use this classifier to obtain the stage of a given message from perpetrator in section 4.3.

4.2 Pre-training the Chatbots on the ConvAI2 Dataset.

We build two chatbots from the T5 model implemented by PyTorch to play the roles of the perpetrator and potential victim respectively. Due to the limited size of the in-domain PJ dataset, we first pre-train the T5 model with a large-scale ConvAI2 dataset, which contains broad topic dialogue turns, to improve the fluency of the generated conversations from both chatbots.

We noticed that in the cybergrooming conversations, the perpetrators mostly lead the conversations. A similar pattern is also recognized in the ConvAI2 dataset where the conversation is usually between two persons and the leading person is usually the one who starts the conversation. Thus, to train the perpetrator's chatbot with the ability to lead the conversation, we use the leading person's dialogues in the ConvAI2 to train the perpetrator chatbot and use the other one's responses to train the victim chatbot.

The chatbots training needs to consider the dialogue history from both sides to predict the next utterance response. Then we concatenate four or five consecutive utterances as a training unit[1] and set the last utterance as the prediction target (i.e., ground truth response). The three or four preceding sentences are the training input (i.e., dialogue history) for the victim and perpetrator chatbots, respectively, because we allow the perpetrator's content as the beginning of both chatbots. Figure 4.2 shows an example of two training units for a perpetrator (red box) and a victim chatbot (green box).

In the pre-training phase, given a source dialogue history as x, we have the T5 model [11]

[1]For training perpetrator chatbot, we take five consecutive utterances as a training unit, while for training the victim chatbot, we use four.

to generate a response by optimizing the following objective:

$$\mathcal{L} = -\sum_{i} \log P(y_i | y_{i-k}, \ldots, y_{i-1}; x; \Theta), \tag{4.2}$$

where Θ is the set of the T5 model parameters and y_i is the i-th token of the response utterance.

In the next section we will fine-tune the chatbot on the model we pre-trained in this section.

4.3 Fine-tuning the Chatbot on the PJ Dataset.

The goal of fine-tuning is to shift the chatbot to generate cybergrooming responses. The fine-tuning of the potential victim chatbot follows the same procedure of the pre-training phase but on the domain-specific PJ dataset. We take two methods for fine-tuning the perpetrator chatbot discussed below.

4.3.1 Rule-based Method: Fine-tuning the Perpetrator Chatbots on the PJ Dataset

A perpetrator usually follows the four grooming stages, as shown in Table 4.1, 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 4.2.

The connection strength of a block from the previous utterances is crucial to determine each

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 4.2: Conversation segmentation criteria for the four relationship stages.

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 [41]: (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 connectivity 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 be-

tween 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.

4.3.2 DRL Policy Method: Fine-tuning the Perpetrator Chatbots on the PJ Dataset

The perpetrator will take strategies to gradually level up the grooming stages defined in Table 4.1 to build a trust relationship with a victim and complete the ultimate grooming goal. As a result, the perpetrator chatbot is able to generate stage-related conversations during the fine-tuning on the PJ dataset. We optimize a DRL policy to generate the utterances closer to the intended stage.

State. A state is denoted by the two previous dialogue turns to contain four consecutive utterances $[u_1, u_2, u_3, u_4]$. The dialogue history is further vectorized by feeding the concatenation of u_1 to u_4 into a T5 encoder.

Action. An action is a dialogue utterance to generate. The action space can be considered unlimited since any length sequences within the max-length hyper-parameters can be generated.

Reward. We implement a classification confidence based reward to encourage the chatbot to follow the expected grooming states. We train the stage classifier TextCNN in the previous section and use it to evaluate how well the generated sentence $\mathbf{y}'$ matches the target stage. The confidence of the stage classifier is estimated by:

$$p(s|\mathbf{y}') = \text{softmax}(\text{TextCNN}(\mathbf{y}', \theta)), \tag{4.3}$$

where $\mathbf{y}'$ represents the generated sentence, $p(s|\mathbf{y}')$ denotes the probability distribution over all the target stage labels, and θ are the parameters of the stage classifier, fixed during fine-tuning. The reward is obtained by:

$$R = [p(s_i|\mathbf{y}')], \tag{4.4}$$

where $\mathbf{y}'$ is the generated target sentence sampled from the model's distribution at every time step in decoding and s_i is the correct stage from the ground truth.

Gradients and objectives. The reward is used for learning a policy. The policy gradient is given by:

$$\nabla_\Theta J(\Theta) = E[R \cdot \nabla_\Theta \log P(\mathbf{y}^\mathbf{s}|\mathbf{x}; \Theta)], \tag{4.5}$$

where R is the stage classifier reward, $\mathbf{y}^\mathbf{s}$ is sampled from the distribution of model outputs at every decoding time step, and Θ are the parameters of the model.

The overall objectives for ϕ are the combination of the loss of the T5 model (Eq. (4.2)) and the policy gradient of the reward in Eq. (4.5). We tested multiple candidate ratios between the two items and identified that 1:0.3 is the best ratio between the loss of the T5 model (Eq. (4.2)) and the policy gradient of the reward regarding our metrics. Figure 4.3 summarizes the procedures of estimating the loss after integrating the DRL into our model.

Output filtering. After the fine-tuning, to assure the generation of consistent and logically smooth (i.e., human-like) conversations, each chatbot is allowed to produce five candidate utterances every time. We can choose the best utterance based on the *connectivity* scores of the five occurrences to the previous utterance and the *similarity* score to the previous utterance. The *connectivity* scores are computed based on the pre-trained BERT next sentence prediction function and can ensure the consistency of a response utterance with the previous contexts. The *similarity* scores are computed from the Semantic Textual Similarity

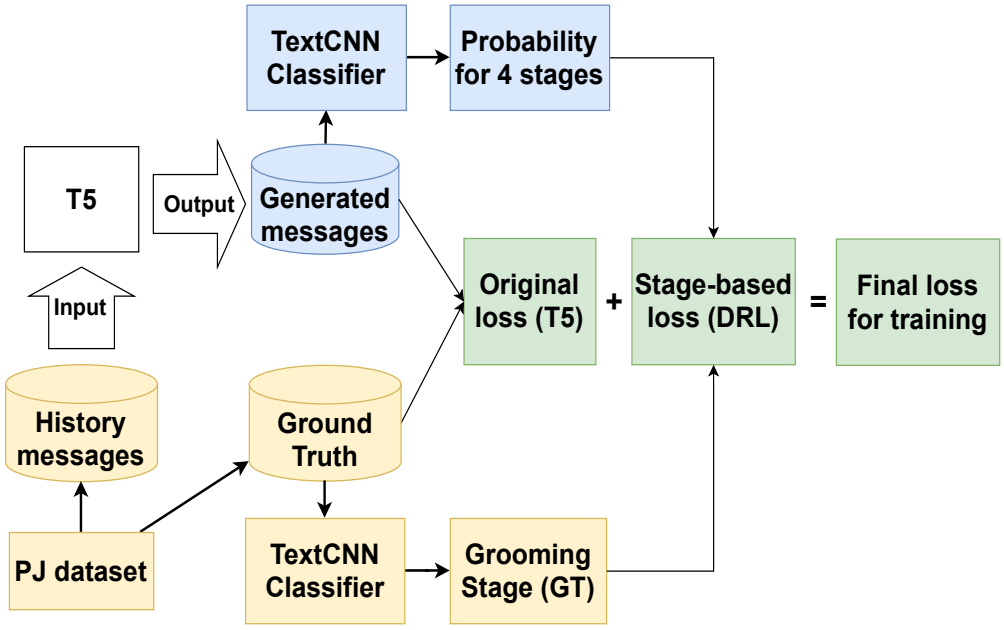

Figure 4.3: Calculation of loss after integrating DRL.

Stages	Trigger Sentence of Each Stage
$\tilde{s}_1$	hi , how are you doing today ?
$\tilde{s}_2$	you parents know you be chatting with me ?
$\tilde{s}_3$	how many pictures you have , any sexy ?
$\tilde{s}_4$	what will we do if you meet me ?

Table 4.3: Trigger sentences of the four relationship stages.

model [62] and can maintain the diversity of a generated utterance to prevent duplicate generation. Given a different scale of the connectivity score and the similarity score, we find that 1:3 is the best ratio between them for output filtering.

Stage evolution of the perpetrator chatbot. The perpetrator chatbot not only generates utterances close to a stage but also evolves the grooming stage to a higher level after maintaining a sufficient number of dialogue runs (e.g., 20). For example, if the chatbot stays at stage $\tilde{s}_1$ for 20 rounds, including 10 perpetrator responses and 10 victim responses, this perpetrator will move forward to stage $\tilde{s}_2$. Each stage will start with a trigger sentence

(see Table 4.3), a trigger sentence can direct the conversation into the topic of a specific stage. Each grooming stage evolves based on a trade-off between the risk and benefit to the perpetrator. That is, if the perpetrator is too aggressive, the victim may be aware of the malicious intents and terminate the conversation to cause a failure of the cybergrooming attack. Otherwise, the user can continue the conversation while the perpetrator may not be able to make good grooming progress.

Chapter 5

Experiment Setup

Parameter	Value	Parameter	Value
Learning rate (γ)	$5e^{-5}$	Epochs	4
Epsilon (ε)	$1.0e^{-6}$	Batch size	8
Warmup steps	500	GPU	Yes
Early stopping	0	Vocabulary	T5-base
T5 loss:DRL loss	1:0.3	Diversity coef	3

Table 5.1: Parameters and their default values used for the SERI framework.

5.1 Datasets

Two chatlog datasets are used in our project. The ConvAI2 dataset [12] is a two-person casual chat dataset in the JSON format with topic labels. We collect 2,000 dialogues with more than 60,000 utterances under the "history" label from the ConvAI2. We manually downloaded each piece of the PJ dataset from the official PJ website [9] in HTML format. It consists of 100 grooming conversations with more than 100,000 chat records between real perpetrators and the professionally trained volunteers acting as potential victims [63]. The PJ dataset is split into the training, validation, and testing sets randomly following the ratio of 8:1:1. All other key parameters in our SERI framework are listed in Table 5.1.

5.2 Data Cleaning

Due to the well-organized structure, the ConvAI2 dataset is directly suitable for the pre-training steps of our chatbots. However, the PJ dataset is noisy with Emojis, Mentions, URLs, or Hashtags. As preparation, we removed the noises by a regular expression-based Python library 'Preprocessor.' This dataset also has plenty of informal languages, such as lexical slangs, and has long consecutive words omitting space separators. We applied 'wordsegment' library to segment those consecutive words by adding essential spaces in Python. Similarly, the lexical slangs can be normalized with MoNoise [64], a state-of-the-art lexical normalization model.

5.3 Metrics

The performances of the SERI chatbots are evaluated in terms of the quality of automatic dialogues [65]. We conduct evaluations by referenced metrics, unreferenced metrics, and human evaluations.

5.3.1 Referenced Metrics

The referenced metrics are commonly known as BLEU [14], ROUGE-L [15], and BERTScore [16].

BLEU. BLEU calculates penalty based on the length of the generated sentence and precision of n-gram between generated sentence and references. The calculation of BLEU is given by:

$$BLEU = BP \cdot exp\left(\sum_{n-1}^{N} w_n log P_n\right) \tag{5.1}$$

where n denotes n-gram where N is the max n-gram order (default 4), w_n is the weight for different n-gram, P_n is the precision of the generation for different n-gram ,and BP denotes Brevity Penalty. The Brevity Penalty is to prevent short generation whose calculation is given by:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \tag{5.2}$$

where c is the length of the target sentence, r is the length of the reference sentence which has the closest length to the target sentence.

ROUGE-L. ROUGE-L calculates the longest common subsequence between the target and reference. The calculation of precision, recall and F-score are shown as:

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{5.3}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{5.4}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{5.5}$$

where m is the length of reference sentence, n is the length of target sentence, and β is the weight coefficient for balancing recall and precision. In our calculate, we use the F-score with β equals to 1.

BERTScore. BERTScore is a metric based on the pre-trained BERT model, computing BERT embeddings and pairwise cosine similarity between generated sentence and reference.

For a candidate $\hat{x}$ and a reference x, the recall, precision and F-scores are:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} max x_i^\top \hat{x}_j \tag{5.6}$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{x_i \in x} max x_i^\top \hat{x}_j \tag{5.7}$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{5.8}$$

We compare the three metrics of the utterances of the proposed SERI against the ground truth utterances in the PJ dataset where the higher measures are better.

5.3.2 Unreferenced Metrics

The unreferenced metrics are perplexity and MaUde scores [17]. The perplexity score is an indicator of how to easily understand a given sentence while a lower perplexity score represents higher fluency. The MaUde score can judge the language quality in multiple aspects, such as fluency, reasonableness (i.e., logical flow), or repetition avoidance.

5.3.3 Human Evaluation

Human evaluation is conducted by two graduate students and one NLP expert. Each participating person completes all the evaluations questions. The questions are prepared by randomly selecting 200 conversation snippets with four dialogue histories and two candidate utterance responses. For the two candidate responses, one response is from the ground truth

PJ dialogue while the other response is generated by the SERI. Humans can decide which one is more consistent and fluent as the response of the history dialogues.

Chapter 6

Experimental Results & Analysis

6.1 Referenced Metrics-based Analysis

The results of three referenced metrics BLEU (Bi-Lingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), and BERTScore are shown in Table 6.1. The scores indicate that the perpetrator chatbot has lower BLEU and ROUGE scores compared to the victim chatbot, reflecting a lower similarity between the SERI's dialogues and the ground truth dialogues. This is because most online chats are informal without strict grammar or fluency rules along with the distinctiveness of dialogue generation task. The higher BERTScore of the perpetrator can be explained by: (1) BERT failed to learn informative contextual representations from many of the functional and uninformative words, such as *yes*, *haha*, or *why*; and (2) The BERTScore is highly sensitive to certain word pairs which fail to capture any meaningful semantics of very short messages.

6.2 Unreferenced Metrics-based Analysis

In The results of perplexity scores from the ground truth dialogues and the SERI generated ones are shown in Table 6.2. The ground truth dialogues from the PJ dataset show much

Role	BLEU	ROUGE	BERTScore
	Max:100	Max:1	Max:1
Perpetrator	2.556	0.091	0.830
Victim	2.688	0.106	0.827

Table 6.1: BLEU, ROUGE, and BERTScore-based analysis for the conversations generated by the SERI.

	Perpetrator	Victim
Ground truth dialogues	315.93	477.82
Generated dialogues	124.82	188.97

Table 6.2: Perplexity score-based analysis.

higher perplexity scores than the SERI's generated dialogues. Since the perplexity score measures the level of easy understanding, the lower perplexity score from the SERI's dialogues means that the original PJ dialogues have more informal expressions, and grammar or logical errors than the SERI's.

Table 6.3 shows MaUde scores from the ground truth dialogues and the SERI's dialogues under the PJ dataset. Since MaUde score measures the reasonableness of the dialogues, the SERI's dialogues demonstrate a slightly higher MaUde scores than the original PJ dataset. This implies that our SERI chatbots can be significantly resistant to some negative effects caused by the informal languages.

6.3 Human Evaluation Analysis

From the human annotators' responses, at least two annotators agreed 74 SERI produced sample responses out of the total 200. This count achieves a 37% success rate of the Turing test [66], which demonstrates the SERI's promising role in dialogue generation. We provide the SERI response that received unanimously positive response by annotators in Table 6.4

	Perpetrator	Victim
Ground truth dialogues	0.844	0.862
Generated dialogues	0.853	0.864

Table 6.3: MaUde score-based analysis based on PJ evaluation dataset.

showing the inter-agreement sample response of human evaluation.

6.4 Impact of Pre-training and DRL

As shown in Table 6.5, the dialogue generated by the model with pre-training on the ConvAI2 dataset shows a better performance with all five metrics, compared to the model without pre-training. As shown in Table 6.6, we observed that overall the model with DRL outperforms the model without DRL. The model with DRL reaches a higher score of BLEU and ROUGE. The aim of using DRL is to simulate the stage strategy of a real perpetrator, which can lead to a higher similarity between the ground truth and generated conversations. Although the model with DRL shows less performance in perplexity score, it was not significant compared to the perplexity performance in the model without DRL. Although the model with DRL showed lower performance in the perplexity score, it outperformed the model without DRL in the MaUde score. This proves that overall using DRL does not introduce a significant quality loss in the generated text while introducing the perpetrator's goal-driven conversation.

6.5 Challenges

We found that there are still some challenges with the existing metrics for dialogue evaluation. First, the existing metrics cannot effectively reflect the logical fluency between one utterance and its history utterances. We observed that if conversations are free from grammar errors,

	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 6.4: Inter-agreement sample of human evaluation.

	With Pre-training	W/O Pre-training
BLEU	2.556	2.505
ROUGE	0.091	0.081
BERTScore	0.830	0.829
Perplexity	124.82	140.33
MaUde	0.853	0.850

Table 6.5: Impact of pre-training on the ConvAI2 dataset.

the existing metrics simply give high scores without considering logical flows. Second, the existing metrics cannot show the performance of the domain-specific application, such as our chatbot. That is, any existing metrics could not provide meaningful measures to indicate the grooming effect of the perpetrator's utterances on the vulnerability of the victim to cybergrooming.

	With DRL	W/O DRL
BLEU	2.556	2.472
ROUGE	0.091	0.084
BERTScore	0.830	0.829
Perplexity	124.82	118.93
MaUde	0.853	0.8333

Table 6.6: Influence of DRL.

Chapter 7

Conclusions & Future Work

7.1 Summary of Key Findings

We discover several **key findings** from this SERI framework: (1) Pre-training the seq-to-seq dialogue model on a high-quality general conversation corpus first (i.e., ConvAI dataset) and then fine-tuning it on a target corpus (i.e., PJ dataset) enhanced the performance of the proposed SERI compared to training on the target corpus directly; (2) Preparing and segmenting the ConvAI2 data to train the two chatbots with different training data units can match the role of leading conversations by the perpetrator chatbot; (3) Implementing a grooming stage-based deep reinforcement learning method can encourage the chatbot to generate dialogues in accordance with the evolving stages from the perpetrator perspective; and (4) Evaluating the chatbot by the human evaluation demonstrates a promising Turing test rate of 37% to pick the utterances generated by the SERI.

During the implementation and evaluation in Chapter 6, we also discussed **limitations** from the current development stage of the chatbot mostly from the domain-specific PJ dataset that has informal styles and poor readability. Although the raw PJ data were cleaned by the automatic text processing tools, it hinders the improvement of the quality of SERI's dialogue generation. Based on the observation of the PJ dataset, the languages used by the perpetrators and the victims have inherently poor quality, which means their use of informal languages are the key features that distinguish their conversation from other normal

conversations.

7.2 Future Work Directions

We have plans of the following **future research directions**:

- Consider advanced and more intelligent data cleaning methods to deal with social slangs.
- Investigate how game theory can optimize the current seq-to-seq model to introduce a perpetrator's strategic conversations.
- Develop new metrics that can capture the effectiveness of the perpetrator's occurrences on the vulnerability or resilience of the victim to cybergrooming.

7.3 Publications

From this research, we have the following papers accepted or to be submitted:

- **P. Wang**, Z. Guo, L. Huang, and J.H. Cho "SERI: Generative Chatbot Framework for Cybergrooming Prevention," *The First Workshop on Evaluations and Assessments of Neural Conversation Systems (EMNLP-EANCS 2021)*, Nov. 2021.
- **P. Wang**, Z. Guo, L. Huang, and J.H. Cho "Deep Reinforcement Learning-based Authentic Dialogue Generation for a Cybergrooming Prevention Program," to be submitted to *The 60th Conference of Association of Computational Linguistics (ACL)*, Dec. 1, 2021.

- Z. Guo, **P. Wang**, J.H. Cho, L. Huang, and K. Kim, "Insights and Lessons Learned from Text Mining-based Social-Psychological Feature Analysis: Cybergrooming Case Study," to be submitted to *IEEE Intelligent Systems*, Dec. 5, 2021.

Appendices

Appendix A

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[1], 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[2].

[1] https://www.cs.cmu.edu/~biglou/resources/
[2] https://gdpr-info.eu/

Bibliography

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

[2] Z. Guo, J.-H. Cho, I.-R. Chen, S. Sengupta, M. Hong, and T. Mitra, "Online social deception and its countermeasures: A survey," *IEEE Access*, vol. 9, pp. 1770–1806, 2021.

[3] UNICEF, "The state of the world's chidren 2017: Children in a digital world," 2017. Available at https://reliefweb.int/report/world/state-worlds-children-2017-children-digital-world-enar , Accessed: 12-01-2021.

[4] S. Marchenko, "Web of darkness: Groomed, manipulated, coerced, and abused in minutes," 2017. Available at https://www.biometrica.com/icmec-online-grooming/ , Accessed: 12-01-2021.

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

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

[7] P. Bours and H. Kulsrud, "Detection of cyber grooming in online conversation," in *2019*

IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6, 2019.

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

[9] Perverted Justice Foundation Inc., "Perverted-justice.com archives," 2020. Available at http://www.perverted-justice.com/?archive=byUserVotes , Accessed: 12-01-2021.

[10] H. Lai, A. Toral, and M. Nissim, "Thank you BART! Rewarding pre-trained models improves formality style transfer," *arXiv preprint arXiv:2105.06947*, 2021.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

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

[13] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[14] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, (Belgium, Brussels), pp. 186–191, Association for Computational Linguistics, Oct. 2018.

[15] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summa-*

rization Branches Out, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, Jul. 2004.

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

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

[18] K. F. Durkin, "Misuse of the internet by pedophiles: Implications for law enforcement and probation practice," *Federal Probation*, vol. 61, no. 3, pp. 14–18, 1997.

[19] Z. Dhouioui and J. Akaichi, "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, 2016.

[20] F. E. Gunawan, L. Ashianti, and N. Sekishita, "A simple classifier for detecting online child grooming conversation," *TELKOMNIKA*, vol. 16, pp. 1239–1248, June 2018.

[21] G. Winters and E. Jeglic, "Stages of sexual grooming: Recognizing potentially predatory behaviors of child molesters," *Deviant Behavior*, pp. 1–10, Sep. 2016.

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

[23] Oracle Cloud Infrastructure, "What is a chatbot?," 2021. Available at https://www.oracle.com/chatbots/what-is-a-chatbot, Accessed: 12-01-2021.

[24] B. A. Shawar and E. Atwell, "Chatbots: Are they really useful?," *LDV Forum*, vol. 22, pp. 29–49, 2007.

[25] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, 11 2020.

[26] Google Hangouts, "Chorus - a conversational agent powered by crowdsourcing," 2021. Available at http://talkingtothecrowd.org/, Accessed: 12-01-2021.

[27] N. Good and C. Wilk, "Introducing the guardian chatbot," 2021. Available at https://www.theguardian.com/help/insideguardian/2016/nov/07/introducing-the-guardian-chatbot, Accessed: 12-01-2021.

[28] P. Kucherbaev, A. Bozzon, and G.-J. Houben, "Human-aided bots," *IEEE Internet Computing*, vol. 22, no. 6, pp. 36–43, 2018.

[29] R. V. Hooijdonk, "Chatbots are changing the nature of customer service," 2019. Available at https://www.toolbox.com/marketing/customer-experience/guest-article/chatbots-are-changing-the-nature-of-customer-service/, Accessed: 12-01-2021.

[30] R. Perera and P. Nand, "Recent advances in natural language generation: A survey and classification of the empirical literature," *Computing and Informatics*, vol. 36, no. 1, pp. 1–31, 2017.

[31] H. T. Hien, P.-N. Cuong, L. N. H. Nam, H. L. T. K. Nhung, and L. D. Thang, "Intelligent assistants in higher-education environments: The fit-ebot, a chatbot for administrative and learning support," in *Proceedings of the Ninth International Symposium on Information and Communication Technology (SoICT 2018)*, (New York, NY, USA), pp. 69–76, ACM, December 2018.

[32] J. Kim, H.-G. Lee, H. Kim, Y. Lee, and Y.-G. Kim, "Two-step training and mixed encoding-decoding for implementing a generative chatbot with a small dialogue corpus," in *Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG)*, (Tilburg, the Netherlands), pp. 31–35, Association for Computational Linguistics, 2018.

[33] C. Laorden, P. Galán-García, I. Santos, B. Sanz, J. M. Hidalgo, and P. G. Bringas, "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*, pp. 261–270, Springer, 2013.

[34] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," 2020.

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

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

[37] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, p. 4291–4308, Oct 2021.

[38] H. Mino, M. Utiyama, E. Sumita, and T. Tokunaga, "Key-value attention mechanism for neural machine translation," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, (Taipei, Taiwan), pp. 290–295, Asian Federation of Natural Language Processing, Nov. 2017.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[40] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020.

[41] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

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

[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

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

[45] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[46] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," 2021.

[47] E. Levin, R. Pieraccini, and W. Eckert, "Learning dialogue strategies within the markov decision process framework," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 72–79, 1997.

[48] E. Levin, R. Pieraccini, and W. Eckert, "Using markov decision process for learning dialogue strategies," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 1, pp. 201–204 vol.1, 1998.

[49] S. Singh, M. Kearns, D. Litman, and M. Walker, "Reinforcement learning for spoken dialogue systems," *Advances in Neural Information Processing Systems (NIPS)*, vol. 12, pp. 956–962, 1999.

[50] N. Roy, J. Pineau, and S. Thrun, "Spoken dialogue management using probabilistic reasoning," in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 93–100, 2000.

[51] S. Singh, D. Litman, M. Kearns, and M. Walker, "Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system," *Journal of Artificial Intelligence Research*, vol. 16, pp. 105–133, 2002.

[52] L. Daubigney, M. Geist, S. Chandramohan, and O. Pietquin, "A comprehensive reinforcement learning framework for dialogue management optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 8, pp. 891–902, 2012.

[53] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (Austin, Texas), pp. 1192–1202, Association for Computational Linguistics, Nov. 2016.

[54] R. Lan, J. Wang, W. Huang, Z. Deng, X. Sun, Z. Chen, and X. Luo, "Chinese emotional dialogue response generation via reinforcement learning," *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 4, pp. 1–17, 2021.

[55] B. Rofi'ah, H. Fakhrurroja, and C. Machbub, "Dialogue management using reinforcement learning.," *Telkomnika*, vol. 19, no. 3, 2021.

[56] A. Bignold, F. Cruz, R. Dazeley, P. Vamplew, and C. Foale, "An evaluation methodology for interactive reinforcement learning with simulated users," *Biomimetics*, vol. 6, no. 1, p. 13, 2021.

[57] B. Peng, X. Li, J. Gao, J. Liu, and K.-F. Wong, "Deep Dyna-Q: Integrating planning for task-completion dialogue policy learning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2182–2192, 01 2018.

[58] R. Zhang, Z. Wang, M. Zheng, Y. Zhao, and Z. Huang, "Emotion-sensitive deep dyna-Q learning for task-completion dialogue policy learning," *Neurocomputing*, vol. 459, pp. 122–130, 2021.

[59] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014.

[60] B. Quast, "rnn: a recurrent neural network in r," *Working Papers*, 2016.

[61] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2015.

[62] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[63] Perverted Justice Foundation Inc., "Perverted-justice.com info for police," 2020. Available at http://www.perverted-justice.com/index.php?pg=policeinfo , Accessed: 12-01-2021.

[64] R. van der Goot, "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*, (Florence, Italy), pp. 201–206, Association for Computational Linguistics, Jul. 2019.

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

[66] A. M. Turing, "Computing machinery and intelligence," in *Parsing the Turing Test*, pp. 23–65, Springer, 2009.