2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress

C^3-Sex: a Chatbot to Chase Cyber perverts

Jossie Murcia Triviño, Sebastián Moreno Rodríguez,
Daniel Díaz López
Computer Science Faculty
Colombian School of Engineering Julio Garavito
AK.45 No.205-59, Bogotá, Colombia
{jossie.murcia, sebastian.moreno-r}@mail.escuelaing.edu.co
daniel.diaz@escuelaing.edu.co

Félix Gómez Mármol
Faculty of Computer Science
University of Murcia
Campus de Espinardo, s/n
30100, Murcia, Spain
felixgm@um.es

***Abstract*—Amongst the myriad of applications of Natural Language Processing (NLP), assisting Law Enforcement Agencies (LEA) in chasing cyber criminals is one of the most recent and promising ones. The paper at hand proposes C^3-Sex, a smart chatbot to interact with suspects in order to profile their interest regarding a given topic. This solution is based on our Artificial Conversational Entity (ACE) that connects to different online chat services to start a conversation regarding a specific matter, in our case child pornography, as this is one sensitive sexual crime that requires special efforts and contributions to be tackled. The ACE was designed using generative and rule-based models in charge of generating the posts and replies constituting the conversation from the chatbot side. The proposed solution also includes a module to analyze the conversations performed by the chatbot and to classify the suspects into three different profiles (indifferent, interested and pervert) according to the responses that they provide in the conversation. Exhaustive experiments were conducted obtaining an initial amount of 320 suspect chats from Omegle, which after filtering were reduced to 35 useful chats, that were classified by C^3-Sex as 26 indifferent, 4 interested and 5 pervert individuals.**

***Index Terms*—Natural Lenguage Processing, Chatbot, Criminal profiling, Law Enforcement Agencies, Child pornography.**

I. Introduction

Artificial Intelligence (AI) has recently captured the attention of many researchers worldwide, encompassing areas such as Machine Learning (ML), Computer Vision, Knowledge-Based Systems, Planning, Robotics, and Natural Language Processing (NLP), among others. Specifically, NLP aims to perceive and understand human language and to replicate it with empathetic responses. Some of the current NLP challenges include understanding complex structures of natural language, extensibility through syntax adaptation, adaptation of responses influenced by the interaction and extension of the conversation scope to an open context [1].

In turn, NLP entails the development of Artificial Conversational Entities (ACE), i.e. chatbots, defined as autonomous components interacting dynamically with humans. A chatbot is generally built upon: an interaction channel (e-mail, instant messaging, web page, mobile app, etc.), a Natural Language Processor (NLP), a Natural Language Generator (NLG), a knowledge-based data, one or more machine learning models and the business logic (see Figure 1).

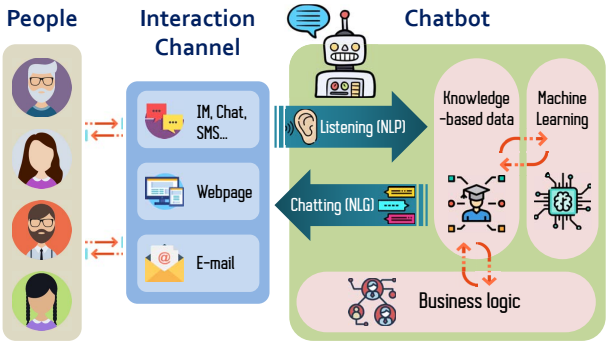

Fig. 1. Anatomy of a chatbot

Chatbots are used in a variety of fields for different purposes, such as i) Support bots, designed to solve customer requests related to the delivery of a service or use of a product, and ii) Financial bots, aimed to resolve inquiries about financial services. Chatbots may have some constraints regarding the requests that they can respond and the vocabulary that they can employ, which depends on the specific domain where they are serving on. Furthermore, according to the Hype Cycle for emerging technologies by Gartner[1], conversational AI platforms remain in the phases of "innovation trigger" and "peak of inflated expectations", meaning that they are getting substantial attention from the industry.

Besides the aforementioned use cases for chatbots, cybersecurity is one of the newest where to apply this technology [2]. Thus, there exist chatbots focused on training end-users [3] or cyber analysts [4] in security awareness and incident response. Further, there are also malicious chatbots devoted to malware distribution through a human-machine conversation [5]. In addition, there is software designed to guide the user in terms of security and privacy, such as Artemis [6], a conversational interface to perform precision-guided analytics on endpoint data. Most of these security chatbots are implemented in a question-answering context [7] using a post-reply technique. As far as we know, the use of chatbots to profile suspects in an

[1]https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018

978-1-7281-3024-8/19/$31.00 ©2019 IEEE
DOI 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00024

active way of child pornography has been little investigated, existing few approaches [8, 9] employing a chatbot to emulate a victim such as a child or a teenager. Likewise, our investigation aims to emulate a vulnerable person while the suspect offers him/her illegal content.

The paper at hand proposes C^3-Sex, a chatbot based on the application of Machine Learning and Knowledge-Based Systems, able to interact with suspects around topics related to child pornography. Once the conversation has finished, some additional machine learning algorithms are employed to analyze the chat logs and make a profile of the suspect within three different categories (indifferent, interested and pervert). The collected chats, joint with the values for each of the defined metrics, could be used for a Law Enforcement Agency (LEA) to identify and process a suspect of child pornography.

The remainder of the paper is structured as follows. Section II describes some remarkable related works found in the literature. In Section III, the key goals and components of C^3-Sex are introduced, while the main aspects of the data science lifecycle and the achieved proposal are presented in Section IV. Section V discusses the different user profiles that can be deduced from the interaction between the suspect and C^3-Sex. Then, in Section VI we perform an exhaustive evaluation of the proposal and analyze the obtained results. At last, Section VII contains some highlights derived from the work done and mentions some future research directions.

II. State of the art

Several scientific works have been conducted so far in the field of chatbots. For instance, Gapanyuk et al. [10] propose a hybrid chatbot model composed of a question-answering module and a knowledge-based scheme. The question-answering module contains a list of pairs of questions and answers so, when a user asks a question matching one of the lists, the corresponding answer is returned to the user. Their main contribution is the implementation of a rule-based system that is encapsulated in a meta-graph as multiple agents.

Most of the early works about conversation systems are generally based on knowledge and are designed for specific domains. These knowledge-based approaches require no data to be built, but instead they require much more manual effort and expert knowledge to build the model, which is usually expensive. Thus, [11] proposes a deep learning hybrid chatbot model which is generative-based. This proposal is composed of 22 response models, including retrieval-based neural networks, generation-based neural networks, knowledge-based question-answering systems, and template-based systems. In addition, it develops a reinforcement learning module based on estimating a Markov decision process.

Integration of an emotional module within chatbots is a way to engage users, i.e., to give the conversational system the ability to be friendly and kind depending on the current emotional state of the user. To this end, [12] builds a complex embedded conversational agent system, capable of processing high-quality natural language as well as sophisticated manipulation of emotions based on the Plutchik Model [13]. This chatbot analyzes and synthesizes the actual emotional status and the emotional expression depicted on the user messages, so a response can be generated in a customized way.

Assuming that linguistic style can be an indicator of temperament, a chatbot with an explicit personality was proposed in [14]. The objective of this chatbot is to generate responses that are coherent to a pre-specified disposition or profile. The proposal uses generic conversation data from social media to generate profile-coherent responses, which represent a specific response profile suitable for a received user post.

Heller et al. [15] describe another related work, where a chatbot named "Freudbot" was constructed using the open source architecture of Artificial Intelligence Markup Language (AIML). The aim of this chatbot was to improve the student-content interaction in distance education. Explicitly, this chatbot technology is promising as a teaching and learning tool in distance and online education.

In turn, Sabbagh et al. [4] present the HI^{2}P TOOL, focused on encouraging an information security culture between users. HI^{2}P incorporates different types of learning methods and topics like incidents response and security policies. The interaction with the user is based on the ALICE[2] chatbot using the AIML, making the solutions simple and efficient.

Another case of chatbot used for security training is presented in [3], where the chatbot Sally is able to interact with some groups of employees in a company who have different education or experience on security. Sally was able to provide security training, which was evidenced in a grow up in the knowledge of the target users.

Furthermore, the work presented in [16] investigates on the behavior of people when they are aware that they are interacting with chatbots. The results show that in such a situation, the conversation can become simple and composed of short messages, even if it can be extended in time. On the opposite, conversations with a human can become complex and composed of long messages, but shorter in time. Additionally, the same research found that language skills such as vocabulary and expression are easily transferred to a machine.

Emotional Chatting Machine (ECM) [17] is a machine learning approach that considers the emotional state of the conversation to generate appropriate responses in content (relevant and grammatical) and in emotion (emotionally consistent).

Particularly related to the topic of sexual harassment and child pornography, Zambrano et al. [8] present BotHook, a chatbot to identify cyber pedophile and catch cyber criminals. In this work, a module of attraction of pedophile interests and characterization was developed. Likewise, the work introduced in [9] discusses the efficiency of current methods of cyber perverts detection and proposes some futuristic methods such as metadata and content data analysis of VoIP communications, as well as the application of fully automated chatbots for undercover operations.

As described above, we found different efforts in the literature w.r.t. the development of chatbots using different

[2]http://www.alicebot.org

approaches like: knowledge base schemes, generative and retrieval neural networks or question answering systems, amongst others. Even if the use of chatbots to face sexual crimes is emerging, these are still an immature application stage which requires more research to beat challenges like handling the particular behavior of a sexual crime suspect, fetching a knowledge database applicable for child pornography domain, generating trust with a suspect to achieve the exchange of illegal data and even using the conversation as a digital evidence that can be used by LEAs. In the paper at hand, we propose the use of a chatbot to face child abuse with a different approach that has not been considered before, i.e. the combined use of a retrieval-based and generative-based models that allow to build specific domain conversations that can also be spontaneous. The retrieval-based model allows us to guide the conversation, while the generative-based model allows us to handle situations where an unexpected post is received from the suspect. In essence, our chatbot C^3-Sex emulates an individual interested in the topic of child pornography. Additionally, our proposal has an important component which is focused on the profiling of the suspect using 6 different metrics which altogether contribute important information to LEA in the hunting of perverts.

III. Goals and key components of C^3-Sex

For our proposal C^3-Sex to achieve its overall target of chasing and spotting perverts by interacting with them in certain chat rooms, the corresponding Artificial Conversational Agent (ACE) should, in turn, enforce the following goals:

1) **Illegal content holders hunting**: C^3-Sex should exhibit the behavior of a human interested in acquiring child pornography, in order to pinpoint suspects possessing illegal content (such as images or videos) and are willing to share it with others.
2) **Illegal content bidder hunting**: Our chatbot should also exhibit the behavior of a human interested in distributing child pornography, so to identify suspects eager to obtain and consume this kind of illegal content, even if a payment is required.
3) **Appropriateness**: C^3-Sex should be able as well to manage situations where the conversation evolves towards topics out of the main one for which the chatbot is intended, i.e., child pornography. This can be provoked by suspects who intend to unveil the bot. An appropriate response should be generated for a question within the same context where the conversation is flowing.
4) **Suspect profiling**: Our solution should perform an analysis of the conversation maintained between the chatbot and the suspect with the purpose of profiling the latter and assigning him to some category (indifferent, interested and pervert).

To fulfill these goals and achieve a functional conversational model, our chatbot combines two main approaches, namely: Knowledge-based systems (represented in the retrieval based model) and Machine learning (represented in the generative models). Additionally, the proposal includes a sentiment module consisting of two models: emotional classification and opinion classification models. The overall functioning workflow of the proposal can be observed in Figure 2.

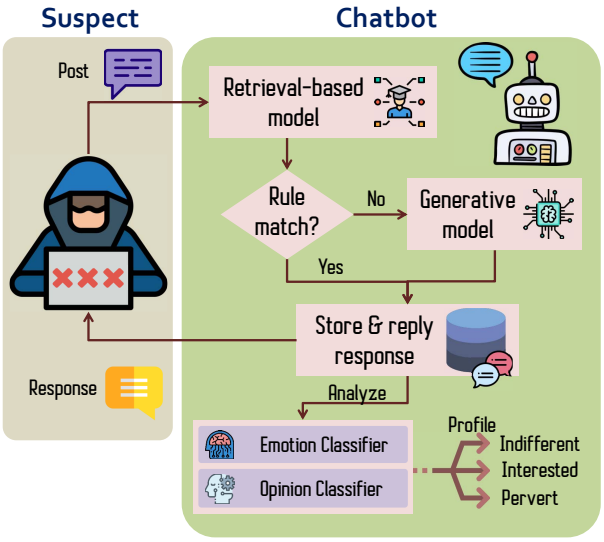

Fig. 2. C^3-Sex overview depicting the workflow of posts & replies

Next we further describe each of the aforementioned key components of C^3-Sex:

1) **Retrieval-based model**: It is a model where the expertness knowledge is represented, saved and inferred from a set of rules expressed using an Artificial Intelligence Markup Language (AIML). A response to a question is selected according to the match of the input with one rule. While this model guarantees an effective implementation even if its knowledge depends on the number and richness of the rules, the response naturalness may be poor.
   Next we show an example of an AIML rule within the context of child pornography:

```
<aiml version="1.0">
  <category>
    <pattern>* DO * WANNA * WITH ME
    </pattern>
    <template>actually im looking
    for some hot content, do you
    have?</template>
  </category>
  ...
</aiml>
```

   This model is the first one to be applied whenever a post from the interlocutor of the chatbot arrives. In case such post matches with some of the rules of the retrieval-based model (as we have seen above), then the corresponding reply is provided. Yet, in case the post does not match any rule from the ones existing in the knowledge base, then the generative-based response model outputs a reply, as we will see next.

2) **Generative model**: It is a model trained by means of a Long Short Term Memory (LSTM) Neural Network (NN) [18] using specific training datasets (as we will see later). The algorithm uses memory to maintain the learned patterns and when a question (post) comes from the chat, the algorithm replies with the patters saved in memory. Responses from a generative model may be more natural than the ones from a retrieval based model, but may contain grammatical failures or sometimes have no sense, we can see some examples in TABLE I.

Post	Reply
I need some good content	*I will make this one more!*
I want weird porn material	*I don't think so. Weird?*
What are you doing now?	*I am chatting with you.*
I want to offer you something, however it is a little weird and illegal	*Is it weird that I want to be the case?*

TABLE I
EXAMPLE OF REPLIES FROM THE GENERATIVE MODEL OF C^3-SEX

3) **Emotional classification model**: In order to assess the underlying emotions within a given conversation, this model makes use of a Support Vector Machine (SVM), trained with the Semantic Evaluation (SemEval) 2007 dataset [19] (one of the most traditional datasets for emotion classification), as a classification and supervised learning paradigm, with a learning rate of 0.5. The SVM algorithm establishes the optimal hyperplanes in a multidimensional space that distinctly separate the six labeled emotions on the SemEval dataset. Suspects responses will be treated as new observations for the model, where each response is classified as one of the six SemEval emotions, namely: anger, disgust, fear, joy, sadness and surprise.
4) **Opinion classification model**: In order to discern whether a given post yields an opinion inclined towards (or against) a specific topic, we have developed the opinion classification model leveraging a Multinomial Naive Bayes with a simple pre-processing (steamming, removing stopwords, etc.) and an alpha of 1. The dataset used [20] consists of 2000 samples of positive and negative reviews of movies, restaurants and other products. The representation of characteristics was based on a document-term matrix taking into account the frequency of words. After training with 90% of the samples, we tested our model with the remaining 10%, achieving an 80% of accuracy.

The opinion classification model is a good complement for the emotional classification model, as the results from one of them could be compared with the results from the other one to guarantee consistency in the prediction of the suspect with regard to child pornography that is done by C^3-Sex.

IV. DATA SCIENCE LIFECYCLE FOR C^3-SEX

Our proposed chatbot follows a generic data science life cycle encompassing the following phases [21]: i) Business understanding, ii) Data acquisition, iii) Modeling and iv) Deployment. This data science life cycle supports each and every activity developed and gives a high-level perspective of how data science must be structured to build a functional Artificial Conversational Entity.

A. Business understanding

Business understanding entails the definition of the data context where the solution will be deployed and executed. In this regard, social media has a growing impact in our lives, allowing people to get access to new interactive services such as anonymous chats. Likewise, it has allowed to find out, in a relatively easy fashion, certain interests of its users, as well as to start some sort of interaction with them. And such sharing of personal information and accessibility to interactive services is what defines our data context. However, social media might also bring concerns when such data and services are employed by dishonest, beguiler and deceitful people, especially those perverts interested in child pornography.

Actually, perverts are using social media platforms today to communicate with each other, aiming at sharing their child pornography material. Furthermore, some of them even chase innocent, naive and sometimes reckless children in the Internet, bamboozling them to obtain further child pornography material from their defenceless victims.

In our solution, the chatbot C^3-Sex simulates an undercover agent from a Law Enforcement Agency aiming at preventing child pornography activities that can start in social media interactions. The goal of this agent is to chat with a suspect (determined by a previous investigation that tags such user as a potential pervert) over a topic where he/she can express his/her thoughts regarding some selected matters. Once a conversation has finished or a determined time has elapsed, the objective is to perform an analysis to classify and profile the suspect.

B. Data acquisition

Data acquisition refers to the collection of data from the context in order to analyze and pre-process it, so these can be used later in the modeling phase. With regard to the **generative model**, different data sets were reviewed and finally the natural conversational data set PAPAYA[3] was selected to further train the LSTM-NN model and validate the abstraction of text patterns. PAPAYA dataset was chose due it groups conversations around different topics (politics, religion, society), giving the bot versatility in situations where the suspect introduce in the conversarion no sexual-related topics.

With regard to the knowledge base for the **retrieval based model**, several manual conversations with potential suspects were carried out. Some of the conversations led to the identification of suspects that after an initial conversation requested a change in the communication channel to another social media platform like Snapchap, Kik or Telegram where they actually transferred illegal content to our chatbot. Based on the previous conversations, multiple AIML rules were created

[3]https://github.com/bshao001/ChatLearner

to give C^3-Sex enough guidelines about the interaction with the suspects and how to deal with them.

As outlined above, in order to train the **emotional classification model** we employed the dataset (SemEval 2007) [19], consisting of news headlines from major newspapers (e.g., The New York Times, CNN, BBC News), each of them labeled through a manual annotation following a scheme of six emotions: anger, disgust, fear, joy, sadness, and surprise.

In turn, the dataset used to train the **opinion classification model** [20] contains sentences labelled as positive (1) or negative (0). All data come from reviews of movies, restaurants and other products.

C. Modeling

Our chatbot (see Figure 2) is based on four models: Retrieval-based, Generative, Emotional classification and Opinion classification.

The Artificial Conversational Entity is composed of both generative and retrieval based models. An implementation that uses only generative models could make mistakes in the coherence of the responses, but it may look flexible and natural. On the other hand, an implementation including only retrieval models could generate coherent responses, but it could also have a limited domain knowledge. Therefore, a smart combination of generative and retrieval based models could be useful to output coherent and natural responses to the suspect.

In our proposal, the **generative model** was built with an LSTM-NN, which is a recurrent neural network trained with the logs of those conversations and interactions made with a suspect. The LSTM-NN algorithm treats each instance of the training data set as a Post-Reply message. This component is intended to generate a dynamic response based on the training data set, which may include general responses given that the knowledge contained in the training data set is general and not specific for a domain.

The development of the generative-based model followed three stages:

- **Pre-processing**: This stage involves cleaning every post received from the suspect to remove extra blank spaces, numbers, special characters and parsing the text to lower case. Additionally, a format change is applied given that conversations in the training data set may be structured under an HTML format and a simpler format is required to compose the data.
- **Training**: It was done using the natural conversational data set PAPAYA and involved 60 epochs, 2 layers, 1024 units, and an initial weight of 0.1, taking approximately 4 days to train.
- **Testing**: Different tests were performed over the ACE to analyze the error generated in the training of the LSTM-NN, so to avoid both overfitting and underfitting.

The **retrieval model**, in turn, contains all the rules that define how an interaction within a child pornography domain should be maintained. These rules are defined under an AIML syntax and are included in some phase of the interaction, such as friendship forming (to make a brief personal introduction to the other peer of the communication), sexual relationship forming (to express a specific interest in sexual-related topics), sexual content offering, transaction (to request or offer illegal content), and risk assessment (to get confidence from the suspect).

- **Modeling:** Based on ALICE and deleting some unnecessary rules, our team interacted with some potential suspects, adding different rules in order to contribute to each phase of the interaction.
- **Testing:** We have used incremental development to test our model and improve the response given by the Chatbot. The number of rules increased with the progressive interaction with the suspects.

D. Deployment

C^3-Sex is intended to be integrated in a real context through the development of a software component that emulates a user that logs in an online chat platform, e.g., Omegle[4], selects "sex" as a topic of interest, and starts a conversation with the peer randomly selected by the platform. After a short introduction, our chatbot subtly requests content related to child pornography. Moreover, it also manages the situation where the suspect requests changing the communication platform, e.g. towards Kik, Snapchat or Telegram. The communication will immediately end as soon as the suspect transfers at least one illegal content (image or video). In case the suspect does not offer sexual content, the chatbot will deceitfully offer child pornography content, requesting a payment for it. In case the suspect accepts the deal, i.e., we are clearly facing with a pervert, the communication is also concluded abruptly. After a conversation is ended up, C^3-Sex will make the analysis of the conversations to profile the suspect and, in case it is required, a LEA can be informed so to possibly prosecute the identified criminal.

V. Profiling of suspects

Once a conversation between our C^3-Sex and a suspect is terminated, the whole log of such conversation is comprehensively analyzed in order to determine the affinity of the suspect with regards to child pornography. To this end, the following metrics are utilized:

1) **Average response time** ($\tau \in \mathbb{R}^+$): It measures the elapsed time between the generation of a reply by the ACE and the next interaction from the suspect, considered as the response to the previous reply. The purpose of τ is to quantify the interest of the suspect, assuming that when suspects have a high interest they will usually respond within a short period of time.
2) **Child pornography matched rules** ($R \in \mathbb{N}$): It measures the number of times that child pornography rules from the **retrieval based model** matched, allowing C^3-Sex to determine the affinity of the suspect with the topics included in such rules. It is assumed that a

[4]https://www.omegle.com

conversation with a big amount of child pornography rules matches, implies that the suspect has an affinity with some or various child pornography topics.

3) **Recognized emotions** ($E \in \{0,1\}$): The emotional module identifies in each post from the suspect that fired a rule within the retrieval-based model an emotion belonging to the scheme of six emotions (anger, disgust, fear, joy, sadness, and surprise) obtained from the **emotional classification model**. We define $E = \frac{\sum_{i=0}^{N} E_i}{N}$, where N is the number of posts identified by our retrieval-based model from the rules fired, and $E_i \in \{0,1\}$ is the classification for the i-th post, where 0 indicates that the post contains a negative emotion (anger, disgust, fear, sadness) and 1 a positive one (joy, surprise).
4) **Opinion classification** ($O \in [0,1]$): The opinion module classifies those posts from the suspect that fired a rule within the retrieval-based model. We define $O = \frac{\sum_{i=0}^{N} O_i}{N}$, where N is the number of posts identified by our retrieval-based model from the rules fired, and $O_i \in \{0,1\}$ is the prediction for the i-th post, where 0 indicates a negative post and 1 a positive one.

With the aim of classifying the suspects into different categories (indifferent, interested, pervert) according to the results obtained in the previous 4 metrics derived from the interaction between a suspect and our chatbot, equation (1) is defined (see Fig. 3).

$$\varphi = \frac{1}{1 + \exp\left(-\frac{R \cdot E \cdot O}{\tau \cdot \delta_1} + \delta_2\right)} \tag{1}$$

where $\varphi \in [0,1]$ represents the likelihood of being a pervert and $\delta_1, \delta_2 \in \mathbb{N}$ are discretionary parameters to make $\varphi \to 0$ when $\frac{R \cdot E \cdot O}{\tau} \to 0$. It is straightforward to check that equation (1) fulfills the following conditions:

- $R \uparrow\uparrow \vee\ E \uparrow\uparrow \vee\ O \uparrow\uparrow \vee\ \tau \downarrow\downarrow \Leftrightarrow \varphi \to 1$
- $R \downarrow\downarrow \vee\ E \downarrow\downarrow \vee\ O \downarrow\downarrow \vee\ \tau \uparrow\uparrow \Leftrightarrow \varphi \to 0$

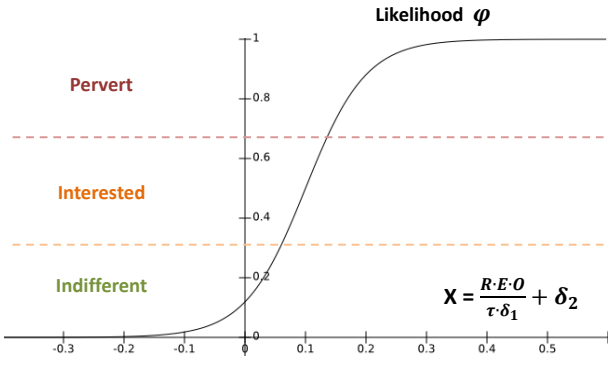

Fig. 3. Graphic representation of likelihood φ

Using the linear ranges defined by the heuristic value of φ, it is possible to show the suspects tendency to a criminal profile, which could vary from someone who is in total disagreement to someone who consumes and distributes content. Below we describe in further detail each of the considered suspect profiles.

1) Indifferent: An indifferent suspect is recognized as an interlocutor that disapproves or has no affinity with child pornography, reflected in no or few matched child pornography rules. Additionally, such a suspect usually has a high average response time showing little interest in the topic. Finally, the recognized emotions in the conversation with an indifferent suspect should indicate negative emotions, reflecting disagreement with the topic. This suspect profile would exhibit a likelihood of being a pervert $\varphi < \frac{1}{3}$.

2) Interested: This case is considered to allow the identification of suspects that show certain interest about child pornography, expressing enthusiasm in the topic, or even a tendency to become a potential consumer or one that already is. This profile includes suspects that have some knowledge about child pornography which can be determined by a medium number of matched child pornography rules. The average response time is medium in this case and the conversation could pose an equal amount of posts with positive and negative emotions. This suspect profile would exhibit a likelihood $\frac{1}{3} \leqslant \varphi < \frac{2}{3}$.

3) Pervert: A pervert suspect would trigger a high number of child pornography rules and the average response time should be low. Regarding the recognized emotions, it is expected that the conversations with a pervert contains a large amount of posts with identified positive emotions. This suspect profile would exhibit a likelihood of being a pervert of $\varphi \geqslant \frac{2}{3}$. These suspects are especially interesting for Law Enforcement Agencies as the results can project the profile of a consumer and distributor of illegal content, and there exists a high probability that they have actually committed a crime..

VI. Experiments

Several experiments were conducted to validate the suitability of our chatbot C^3-Sex in a real context, aiming to identify the suspect profile (indifferent, interested or perverted) behind a conversation using for this purpose the metrics defined in Section V. For the ease of reading, the settings of the experiment are reported in Section VI-A, while a significant analysis of the results is carried out in Section VI-B. In order to validate our model, we only use a quantitative approach due to the nature of the problem. K-fold validation could be used to ensure the quality of our results; however, there were very few data to perform the partitions and it was difficult to get more.

A. Settings

The experiments were conducted by running an instance of C^3-Sex on a real context which emulates a user that logs in the online chat platform Omegle, selects "sex" as a topic of interest, and starts a conversation with the peer randomly selected by Omegle. The interaction with Omegle is done through a Chrome driver handled by functions of the

python library Selenium[5]. As the conversation evolves, one of the two previously developed models (generative or retrieval-based model) is employed to answer each post of the suspect. In case the suspect suggests using another communication channel to make the exchange of content, C^3-Sex will send a Telegram username and will be able to listen posts through the web version of Telegram. And if the suspect does not offer sexual material, our chatbot will deceitfully offer child pornography content. After a conversation is ended up, C^3-Sex will make the analysis of the conversations using the previously developed models (emotional classification and opinion classification model) and consequently will calculate the 4 defined metrics. In order to train the generative model, we have used the Tensor Flow library and for the calculus of the likelihood φ (equation 1), the discretionary parameters $\delta_1 = 0.05$ and $\delta_2 = 2.0$ were selected.

B. Results

C^3-Sex was executed for 15 hours, gathering a total of 320 chats. These chats were filtered to exclude the ones that lasted just a few seconds and therefore are not suitable to be analyzed. Chats with bots (e.g., bots publishing porn services) were also excluded from the analysis. After this filtering, 35 chats remained, which were deeply reviewed and analyzed as shown next.

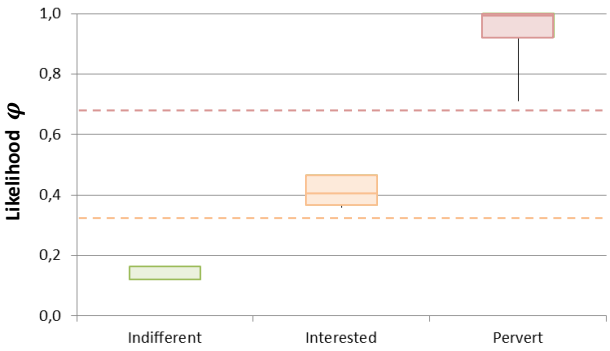

Fig. 4. Distribution of suspect profiles for a set of 35 chats

Fig. 4 shows the distribution of suspect profiles for the set of 35 chats. From these, 26 chats were classified in the suspect profile "indifferent" as φ was less than $\frac{1}{3}$ and are concentrated around the median value $\varphi \approx 0.12$. On the other hand, 4 chats were classified in the suspect profile "interested" with a median of likelihood $\varphi \approx 0.41$, having values more concentrated in the third quartile. At last, 5 chats were classified in the suspect profile "pervert", having a $\varphi > \frac{2}{3}$ with a median of likelihood $\varphi \approx 0.99$.

In turn, Fig. 5 shows the distribution of the 4 analyzed metrics over the 5 chats that were categorized as belonging to the suspect profile "pervert". As we can observe, the metric with a bigger dispersion is the average response time τ, with a median value of 2 seconds. Additionally, the metric E for

[5]https://pypi.org/project/selenium

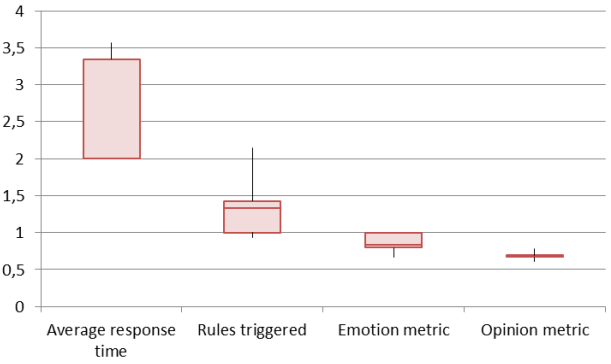

Fig. 5. Distribution of metrics for pervert chats

recognized emotions and O for opinion classification had a similar distribution with an interquartile range $IQR \approx 0.17$.

The results obtained from our metrics demonstrate the suitability and efficiency of our the proposal, since the chatbot is able to maintain long conversations without the suspect realizing its true nature thanks to the generative model. Simultaneously, C^3-Sex knowledge-based module builds a friendly relationship, giving enough trust to the suspect, who eventually will express his/her tastes, emotions, and opinions regarding child pornography. Furthermore, the sentiment module (emotional and opinion classifiers) allows us to create several metrics, which are assigned to a heuristic function that correctly dimensions the proposed profiles and guarantees the identification of illegal consumers and distributors.

In addition, as demonstrated by these experiments, our chatbot solution C^3-Sex exhibits a behavior that can be considered suitable for supporting the labors of LEA or any other organization fighting against child pornography. C^3-Sex was able to successfully identify chats conducted by suspects with a pervert profile and with a known username. A deeper review of such chats allows also to identify special features from those chats, such as communication patterns (metric R) and expressiveness (metrics E and O).

VII. Conclusions and Future Work

With the aim of humbling contributing to the honorable task of prosecuting sexual crimes, specifically child pornography, C^3-Sex has been proposed along this paper. C^3-Sex is composed of four models: Retrieval-based, Generative, Emotional classification and Opinion classification. All together, these models constitute a solution able to keep conversations with suspects and profile them to identify perverts. The final goal of C^3-Sex is to hunt holders and bidders of illegal content related to child pornography, who can later be investigated by a Law Enforcement Agency.

As future work we plan to improve the models that compose our chatbot, so a more human-like interaction between the chatbot and the suspects can be performed, reducing the probability that the suspect can unveil C^3-Sex. This should be achieved through the generation of more specific AIML rules

for the retrieval model, and the training of the generative model with a data set associated to a context of sexual conversations. Additionally, in the future we expect to address other types of sexual crimes related to children, like grooming, sexual exploitation, sexting, sextortion, sex scam or sex trafficking, among others. Some of these new types of sexual crimes would require C^3-Sex to be able to keep more complex conversations for a longer time.

ACKNOWLEDGMENT

This work has been partially supported by the Colombian School of Engineering Julio Garavito (Colombia) through the project "Developing secure and resilient architectures for Smart Sustainable Cities" approved by the Internal Research Opening 2018 and by the project "Strengthening Governance Capacity for Smart Sustainable Cities" (grant number 2018-3538/001-001) co-funded by the Erasmus+ Programme of the European Union, as well as by a Ramón y Cajal research contract (RYC-2015-18210) granted by the MINECO (Spain) and co-funded by the European Social Fund.

REFERENCES

[1] Al Rahman, Abdullah Al Mamun, and Alma Islam. "Programming challenges of chatbot: Current and future prospective". In: *IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. Dec. 2017, pp. 75–78.

[2] Félix Gómez Mármol, Manuel Gil Pérez, and Gregorio Martínez Pérez. "I dont trust ICT: Research challenges in cyber security". In: *IFIP International Conference on Trust Management*. Springer. 2016, pp. 129–136.

[3] Stewart Kowalski, Katarina Pavlovska, and Mikael Goldstein. "Two Case Studies in Using Chatbots for Security Training". In: *Information Assurance and Security Education and Training*. Ed. by Ronald C. Dodge and Lynn Futcher. 2013, pp. 265–272.

[4] B. A. Sabbagh et al. "A prototype For HI2Ping information security culture and awareness training". In: *International Conference on E-Learning and E-Technologies in Education (ICEEE)*. 2012, pp. 32–36.

[5] Pan Juin Yang Jonathan, Chun Che Fung, and Kok Wai Wong. "Devious Chatbots - Interactive Malware with a Plot". In: *Progress in Robotics*. 2009, pp. 110–118.

[6] Bobby Filar, Richard Seymour, and Matthew Park. "Ask Me Anything: A Conversational Interface to Augment Information Security Workers". In: *13th Symposium on Usable Privacy and Security, SOUPS*. 2017.

[7] Simon Keizer and Harry Bunt. "Multidimensional Dialogue Management". In: *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. 2006, pp. 37–45.

[8] P. Zambrano et al. "BotHook: An option against Cyberpedophilia". In: *2017 1st Cyber Security in Networking Conference (CSNet)*. 2017, pp. 1–3.

[9] Kemal Veli Açar. "Webcam Child Prostitution: An Exploration of Current and Futuristic Methods of Detection". In: *International Journal of Cyber Criminology* 11.1 (2017), pp. 98–109.

[10] Yuriy Gapanyuk et al. "The Hybrid Chatbot System Combining Q&A and Knowledge-base Approaches". In: *7th International Conference on Analysis of Images, Social Networks and Texts*. 2018, pp. 42–53.

[11] Iulian V Serban et al. "A deep reinforcement learning chatbot". In: *arXiv preprint arXiv:1709.02349* (2017).

[12] Gábor Tatai et al. "The chatbot who loved me". In: *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2003.

[13] Robert Plutchik. "The nature of emotions". In: *American scientist* 89.4 (2001), pp. 344–350.

[14] Qiao Qian et al. "Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation". In: *27th International Joint Conference on Artificial Intelligence*. 2018, pp. 4279–4285.

[15] Bob Heller et al. "Freudbot: An investigation of chatbot technology in distance education". In: *EdMedia: World Conference on Educational Media and Technology*. 2005, pp. 3913–3918.

[16] Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. "Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations". In: *Computers in Human Behavior* 49 (2015), pp. 245–250.

[17] Hao Zhou et al. "Emotional chatting machine: Emotional conversation generation with internal and external memory". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[18] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[19] Carlo Strapparava and Rada Mihalcea. "Semeval-2007 task 14: Affective text". In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. 2007, pp. 70–74.

[20] Dimitrios Kotzias et al. "From group to individual labels using deep features". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 597–606.

[21] Gary Ericson et al. *Team Data Science Process Documentation*. Tech. rep. Microsoft Azure, 2017, p. 456.