electronics

MDPI

Article

C^3-Sex: A Conversational Agent to Detect Online Sex Offenders

John Ibañez Rodríguez [1], Santiago Rocha Durán [1], Daniel Díaz-López [2,*], Javier Pastor-Galindo [3] and Félix Gómez Mármol [3]

[1] Faculty of Computer Engineering, Escuela Colombiana de Ingeniería Julio Garavito, AK.45 No.205-59, Bogotá 111166, Colombia; john.ibanez@mail.escuelaing.edu.co (J.I.R.); santiago.rocha@mail.escuelaing.edu.co (S.R.D.)

[2] School of Engineering, Science and Technology, Universidad del Rosario, Carrera 6 No. 12 C-16, Bogotá 111711, Colombia

[3] Faculty of Computer Science, Campus de Espinardo, University of Murcia, 30100 Murcia, Spain; javierpg@um.es (J.P.-G.); felixgm@um.es (F.G.M.)

* Correspondence: danielo.diaz@urosario.edu.co

Received: 8 August 2020; Accepted: 9 October 2020; Published: 27 October 2020

Abstract: Prevention of cybercrime is one of the missions of Law Enforcement Agencies (LEA) aiming to protect and guarantee sovereignty in the cyberspace. In this regard, online sex crimes are among the principal ones to prevent, especially those where a child is abused. The paper at hand proposes C^3-Sex, a smart chatbot that uses Natural Language Processing (NLP) to interact with suspects in order to profile their interest regarding online child sexual abuse. This solution is based on our Artificial Conversational Entity (ACE) that connects to different online chat services to start a conversation. The ACE is designed using generative and rule-based models in charge of generating the posts and replies that constitute the conversation from the chatbot side. The proposed solution also includes a module to analyze the conversations performed by the chatbot and calculate a set of 25 features that describes the suspect's behavior. After 50 days of experiments, the chatbot generated a dataset with 7199 profiling vectors with the features associated to each suspect. Afterward, we applied an unsupervised method to describe the results that differentiate three groups, which we categorize as indifferent, interested, and pervert. Exhaustive analysis is conducted to validate the applicability and advantages of our solution.

Keywords: chatbot; online child sexual abuse; criminal profiling; natural language processing; law enforcement agencies

1. Introduction

Human or drug trafficking, sexual abuse, child bullying, animal cruelty or terrorism are examples of realities that should not exist. Unfortunately, they are still present in our societies causing irreparable harm to innocent beings. Another problem that is also detected today and affects minors is the production, possession, and distribution of online child sexual abuse content, an offence clearly defined by the Luxembourg Guidelines (http://luxembourgguidelines.org) and supported by 18 organizations interested in the protection of children from sexual exploitation and sexual abuse.

During the last decades, this scourge has been magnified by the proliferation of forums, online communities, and social networks [1,2]. The content derivated from child sexual abuse is now more accessible than before and, in turn, more present than ever [3]. In this sense, the Child Sexual Exploitation database of INTERPOL registered almost three million images and videos containing child sexual abuse content to the date of July 2020. This database has been useful to identify 23,100 victims worldwide (more than 3800 in 2019), and to chase 10,579 criminals.

The current era of the information facilitates the formation of criminal groups that share and spread sexual content. As a result, many sex offenders can directly see or download such illegal multimedia that would once have been much harder to come by. According to [4], 75% of all exploitation cases were linked to the possession and distribution of child sexual content, 18% was related to child sex trafficking, and 10% to production itself.

Law Enforcement Agencies (LEAs) are making a great effort to pursue these criminals and seize child sexual content [5]. Analyzing P2P networks to measure the existing child sexual abuse content [6], blocking the access to specific illegal webpages [7] or automatically detecting suspect material with visual recognition techniques [8] are some of the dimensions that have been considered to uncover pedophiles and protect the rights of children. Nevertheless, despite those praiseworthy trials of chasing sexual offenders, there is still work to be done to have a sophisticated and effective system to avoid online child abuse [9]. Due to the problem severity, new approaches scaling to such magnitude are needed.

Fortunately, Artificial Intelligence (AI) is being developed by leaps and bounds and constitutes a new pillar on which to build more refined solutions. This paradigm is being implemented, for example, to assist in decision-making, to extract patterns in datasets, or to predict future events. Nonetheless, considering that child abuse content is mainly distributed through the immersion of the Internet and especially frequent in online communities, we are particularly interested in tools adaptable to World Wide Web scenarios.

Natural Language Processing (NLP) is one of the most used AI techniques for social media applications and aims to analyze and understand human language, or even to replicate it with empathetic responses. One of the most common NLP applications is the Artificial Conversational Entities (ACE), also known as chatbots, which do not require human intervention for maintaining a conversation with a user. On the other hand, it faces some challenges such as response blandness (tendency to generate uninformative responses), speaker consistency (possibility of making contradictory interventions), word repetitions, lack of grounding (responses out of context), empathy incorporation or explanation capacity [10]. According to the Hype Cycle for emerging technologies by Gartner [11], conversational AI platforms remain in the phases of "innovation trigger" and "peak of inflated expectations", meaning that they are currently getting substantial attention from the industry.

The paper at hand proposes C^3-Sex, an AI-based chatbot that automatically interacts with Omegle (https://www.omegle.com) users to analyze their behavior around the topic of child illegal content. The analysis of messages from a suspect that follow the tone of the conversation produces a vector with features that allow to describe the suspect behavior. This tool could be potentially useful for Law Enforcement Agencies (LEAs) who work in chasing and putting an end to child sexual abuse. In particular, this chatbot is an enhanced version of the one presented in [12], indeed including more sophisticated techniques, support for mobile apps and a more refined behavior.

The remainder of the paper is structured as follows. Section 3 describes some remarkable related works found in the literature. In Section 4, 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 5. Section 6 discusses the different features that support the profiling of a suspect based on his/her interaction with the C^3-Sex. Then, in Section 7 we perform an exhaustive evaluation of the proposal and analyze the obtained results. Finally, Section 8 contains some highlights derived from the work done and mentions some future research directions.

2. Background

A chatbot is generally built upon the following elements, as shown in Figure 1:

- Interaction channel, the front-end interface with the end-user in the form of an email agent, instant messaging service, web page or mobile app.
- Natural Language Processor (NLP), the component in charge of understanding the human peer.

- Natural Language Generator (NLG), responsible for responding to human interventions.
- Knowledge-based data, the element that brings the contextual information of the chatbot.
- Business logic, which defines how to interact with the end-user.
- Machine learning models, those parts of the system that enable automatic operating and simulates human behavior.

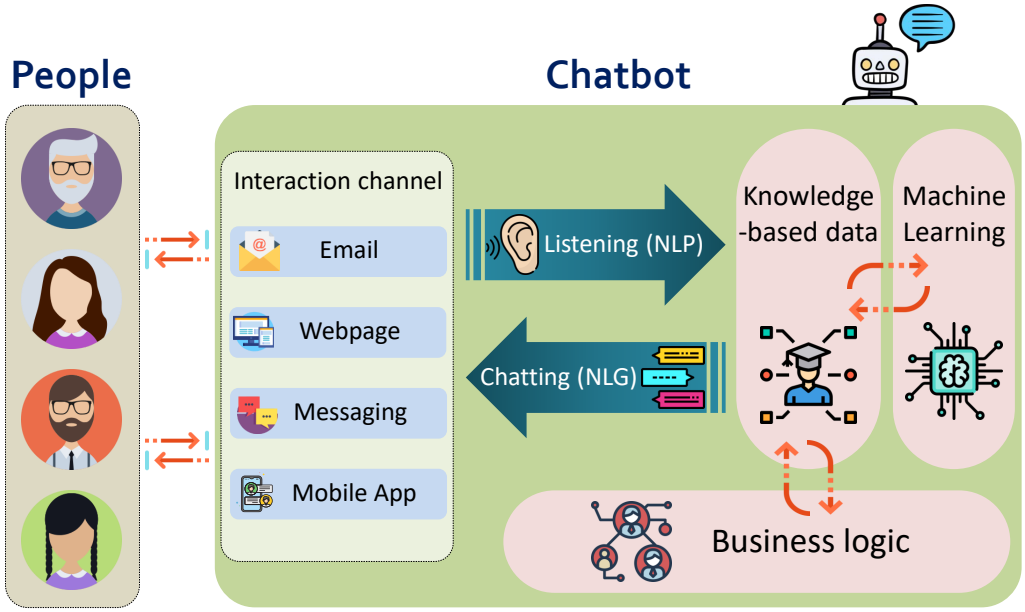


Figure 1. Anatomy of a chatbot.

Chatbots are used in a variety of fields for different purposes. The most common ones are support bots, which assist customer requests related to the delivery of a service or use of a product; financial bots, which resolve inquiries about financial services; or information bots that help new users on a page or disoriented people around a topic. Recently, they are also being applied to innovate in cybersecurity through the form of training bots or guide bots. The formers are a productive tool to educate end-users [13] and cyber analysts [14] in security awareness and incident response, whereas the latter informs the end-users about terms of security and privacy, such as Artemis [15].

In our context, a chatbot could be trained to pretend being an adult person interested in acquiring child sexual abuse material and then be deployed on the web to seek perverts, interact with them, and infer a possible possession of illicit content in an automatic and scalable manner. The software-driven agent(s) can automatically log in to communities, social networks, and forums to interact with large numbers of users. It is worth mentioning that LEAs would not be as effective in doing this manually, having the whole Internet at their disposal and few resources. As far as we know, the use of chatbots to profile suspects in an active way of child sexual content has been investigated little. Only a few approaches [16,17] employ them to emulate a victim, such as a child or a teenager.

3. State of the Art

Several scientific works have been conducted so far in the field of chatbots. Thus, for instance, Gapanyuk et al. [18] 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 that matches one of the lists, the corresponding answer is returned to the user. This work's main contribution is the implementation of a rule-based system that is encapsulated in a meta-graph as multiple agents.

Most early works about conversation systems are generally based on knowledge and 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, [19] 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.

The integration of an emotional module to chatbots is one way to engage users, i.e., to give the conversational system the ability to be friendly and kind depending on the user's current emotional state. Following this approach, Reference [20] builds a complex embedded conversational agent system, capable of doing high-quality natural language processing and sophisticated manipulation of emotions based on the Plutchik Model [21]. This chatbot analyzes and synthesizes the actual emotional status and the emotional expression depicted on the user messages so that a response can be generated in a customized way.

With the assumption that linguistic style can be an indicator of temperament, a chatbot with an explicit personality was proposed in [22]. 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, representing a specific reply profile suitable for a received user post.

Heller et al. [23] describe another related work, where a chatbot named Freudbot was constructed using the open-source architecture of Artificial Intelligence Markup Language (AIML). This chatbot aimed to improve the student–content interaction in online learning ecosystems. Explicitly, this chatbot technology is promising as a teaching and learning tool in online education.

In turn, Sabbagh et al. [14] present the HI^{2}P TOOL, which is focused on encouraging an information security culture and awareness between users. This tool incorporates different types of learning methods and topics like incidents, response, and security policies. The interaction with the user is based on the ALICE chatbot using the AIML, making the solutions efficient and straightforward.

Another case of a chatbot used for security training is presented in [13], where the chatbot Sally can interact with some groups of employees in a company with different education or experience on security. Sally was able to provide security training, which was evidenced by a growing knowledge of the target users.

Furthermore, the work presented in [24] investigates the behavior of people when they are aware that they are interacting with chatbots. The results show that the conversation can become pure and composed of short messages in such a situation, even if it can be extended in time. Conversely, 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) [25] is a proposal with 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 online child sexual abuse, Zambrano et al. [16] present BotHook, a chatbot to identify cyber pedophiles and catch cybercriminals. In this work, a module of attraction of pedophile interests and characterization was developed. Likewise, the work introduced in [17] 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.

In the same direction, a system to detect online child grooming is proposed in [26], which uses Bag of Words (BoW) to select words in the context of grooming, and Fuzzy-Rough Feature Selection (FRFS) for the selection of the most important features. Finally, a fuzzy twin Support Vector Machine was used for text classification using two training datasets: one from Perverted Justice (http://www.perverted-justice.com/) and another one from PAN13 https://pan.webis.de/data.html#pan13-author-profiling.

Likewise, an alternative method for detecting grooming conversations is proposed in [27] where a group of 17 characteristics associated with grooming conversations are identified and then used for text classification. Such proposed method exposes a low-computational cost and an accuracy (96.8%) close to the one obtained with more computational-demanding classifiers like Support Vector Machine (SVM) (98.6%) or K-nearest neighbor (KNN) (97.8%).

The authors of [28] propose a classifier for the detection of online predators, which employs Convolutional Neural Networks. Such proposal gets a classification performance (F1-score) that is 1.7% better than the one obtained with an SVM approach. Two datasets were used for the training of the model: PAN-2012 (https://pan.webis.de/data.html#pan12-sexual-predator-identification) and conversations gathered by the Sûreté du Quebec (Police for the Canadian province of Quebec).

A model that employs user reactions coming from different social networks to detect cyberbullying incidents is proposed in [29]. Such proposal argues that not all text with profanity addressed to a person may be actually considered as proof of bullying, as it depends on the reaction of the person. A dataset composed of 2200 posts were manually labeled as bullying or not-bullying and used to train a Support Vector Machine classifier.

A conversational agent pretending to be a child is proposed in [30], which aims to prevent cybercrimes associated with pedophilia, online child sexual abuse and sexual exploitation. This conversational agent uses game theory to manage seven chatter-bots that address a conversation strategy with the aim of identifying pedophile tendencies without making the chatter to suspect about the agent.

Grooming Attack Recognition System (GARS) is a system that calculates dynamically the risk of cybergrooming that a child is exposed to along a conversation [31]. The risk is calculated using fuzzy logic controllers, which consider the exchanged dialogs, the personality of the interlocutors, the conversations history of the child and the time that the child profile has been exposed on the Internet.

A study of different computational approaches to forecast social tension in social media (Twitter) is defined in [32], which is accompanied by a comparison of approaches based on precision, recall, F-measure, and accuracy. The considered approaches are tension analysis engine, which is their own proposal based on the conversation analysis method MCA (Membership, categorization, analysis) [33], machine learning approach with Naive Bayes classifier and sentiment analysis with the SentiStrength tool [34]. Conversational analysis and syntactic and lexicon-based text mining rules showed a better performance than machine learning approaches.

The proposal designed in [35] tries to detect whether an adult is pretending to be a child as part of an online grooming abuse. The proposal identifies a person as an adult or a child based on the writing style, and then it determines if a child is a fake child or not. That paper suggests that it is challenging to separate children from adults in informal texts (blogs or chat logs). Such proposal uses a set of 735 features gathered from the literature, which are used to build models based on algorithms like Adaboost, SVM, and Naive Bayes.

In rows 1–8 of Table 1 we found different literature efforts concerning the development of chatbots (conversational agents) that mainly uses rule-based and generative-based models, aimed to resolve different industry requirements or to improve the quality of bot conversation. Only works at rows 9–10 propose solutions that integrate conversational agents and classificators to detect cyber pedophiles. On the other hand, rows 11–18 contain different works that propose classificator models that employ different artificial intelligence techniques to detect sexual crimes, cyber bullying and harassment conversation. In the paper at hand, we propose a chatbot to face child abuse with a different approach that has not been considered before: in essence, our chatbot C^3-Sex emulates an individual interested in acquiring child sexual abuse material and then evaluates the responses and behavior to profile the suspect using 25 different features, which altogether provide essential information to LEA in the hunting of perverts.

Table 1. Comparative table of the analyzed related works.

	Related Work	Year	Type	Main Component(s)	Aim
1	Gapanyuk et al. [18]	2018	Conversational model	Question and Answering model using cosine distance, metagraph model	Develop a bot that provides information about goods
2	Serban et al. [19]	2017	Conversational model	Recurrent, sequence-to-sequence and latent variable NN	Develop a social bot
3	Tatai et al. [20]	2003	Conversational model	Emotional state generator, Plutchik emotional model	Develop an emotion sensitive social bot
4	Qian et al. [22]	2018	Conversational model	Identity coherent conversation machine	Develop a social bot with a predefined personality
5	Heller et al. [23]	2005	Conversational model	AIML Knowledge base	Develop a bot to support on-line education
6	Sabbagh et al. [14]	2012	Conversational model	AIML Knowledge base	Develop a bot for education in security
7	Kowalski et al. [13]	2013	Conversational model	Question and Answering model	Develop a bot for security training
8	Zhou et al. [25]	2018	Conversational model	Sequence-to-sequence model with emotion category embedding	Develop an emotion sensitive social bot
9	Zambrano et al. [16]	2017	Conversational model	Sequence-to-sequence model (Planned but not implemented)	Detect cyber pedophile
10	Laorder et al. [30]	2013	Conversational model	Question-answering based on AIML and game theory	Detect Paedophile, online child sexual abuse and Sexual exploitation
11	Anderson et al. [26]	2019	Classificator model	Fuzzy-Rough feature selection (FRFS) + Fuzzy Twin SVM	Detect grooming
12	Gunawan et al. [27]	2018	Classificator model	SVM and KNN	Detect grooming
13	Ebrahimi et al. [28]	2016	Classificator model	Convolutional Neural Network	Detect juvenile abuse, Sexual solicitation
14	Dadvar et al. [29]	2012	Classificator model	SVM	Detect cyberbullying
15	Michalopoulos et al. [31]	2014	Classificator model	Fuzzy logic controllers to evaluate risk	Detect grooming
16	Burnap et al. [32]	2015	Classificator model	Own tension analysis algorithm based on MCA (Membership, categorization, analysis), Naive Bayes and SentiStrength	Detect social tension
17	Meyer et al. [35]	2015	Classificator model	Adaboost, SVM and Naive Bayes	Detect grooming

4. Goals and Taxonomy of C^3-Sex

The implementation of an Artificial Conversational Agent (ACE) encompasses many challenges [10]. It is an arduous task bringing along a substantial technical complexity. On the other hand, it is crucial to define the context in which a chatbot will operate and the objectives it pursues.

The primary purpose of this agent is to chase pedophiles on the web and the detection of sexual content. To this end, we intend to introduce the software-controlled instrument into suspicious chat

rooms to interact with other users. Therefore, for the proper functioning of the tool, C^3-Sex should comply with the following properties and directives:

1. Appropriateness: C^3-Sex should manage those situations in which the conversation is out of scope. It should be programmed to bring the subject back into our field of interest, that is, minors' sexual content. To this end, C^3-Sex has to emulate a human behavior accurately, being pragmatic in interpreting the suspect's intent, and responding accordingly. It is essential to be consistent with the created answer, depending on whether it is being asked, ordered, or affirmed. In essence, our software-controlled agent should not be revealed.
2. Platform flexibility: The ACE has to change scenarios from chat rooms to other environments dynamically. To date, C^3-Sex operates in Omegle rooms and is able to smoothly migrate at any time to Snapchat or Telegram sessions. The latter is a requirement directly extracted from our specific experience, where Omegle's suspects used to ask for a change of application. In this sense, Snapchat and Telegram are frequently used for sexting as it lets users transfer media files without limitations and without saving the conversation.
3. Illegal content holders hunting: C^3-Sex should exhibit the behavior of a human interested in acquiring online child sexual abuse content in order to pinpoint suspects possessing illegal content (such as images or videos) who are willing to share it with others.
4. Illegal content bidder hunting: The chatbot should also exhibit the behavior of a human interested in distributing online child sexual abuse content to identify suspects eager to obtain and consume this kind of illegal content.
5. Suspect profiling: The solution should produce a vector of features that characterize the conversation maintained between the chatbot and the suspect to help in the profiling of the latter.

C^3-Sex combines two main parts to fulfill these goals and achieve a functional conversational model, namely: the interactive module and the analysis module. The interactive module holds some interaction interfaces, a knowledge-based system, and a machine learning generative model. The analysis module, in turn, includes an emotional classifier and an opinion classifier. As a result of both modules execution, the chatbot finally outputs a vector of 25 features that characterize the ended conversation. The components and the overall functioning workflow of C^3-Sex can be observed in Figure 2 and they are further described next.

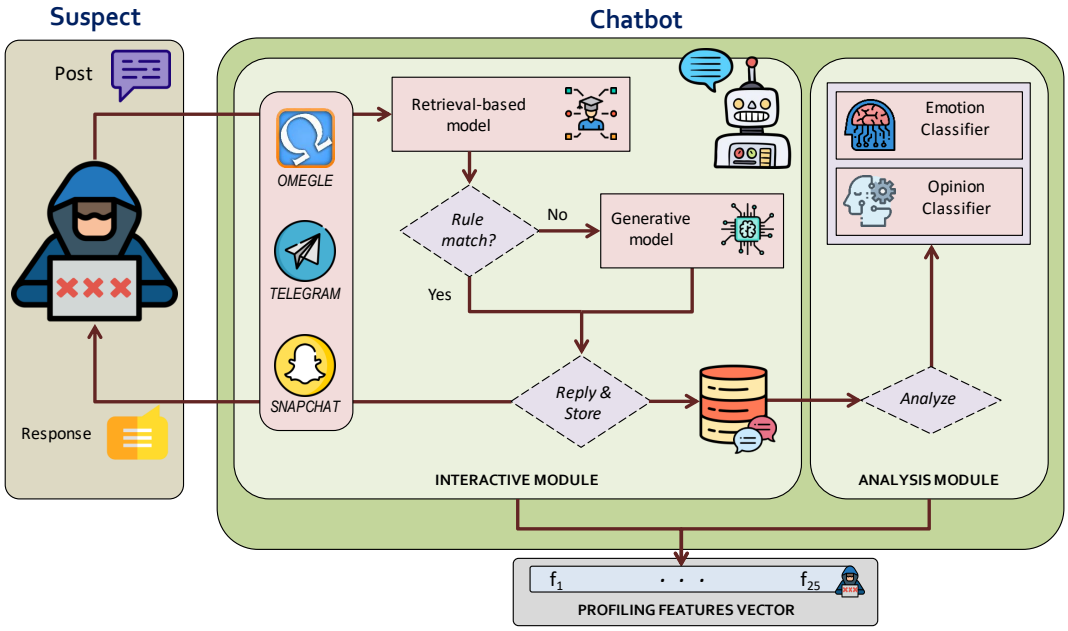


Figure 2. C^3-Sex overview depicting the workflow of posts and replies.

4.1. Interactive Module

The interactive module is active throughout the conversation with the suspects. Concretely, it is in charge of emulating a human behavior and consequently storing information of the suspect. This part is directly linked with the aforementioned properties 1, 2, 3, and 4. The main elements making this module work are explained below:

- Interaction channel: This is the front-end of the chatbot enabling the communication link with the users. This channel can dynamically switch to the different available interfaces. Thus, the default initial Omegle interface can migrate to Snapchat or Telegram interfaces. It is worth mentioning that this change is transparent to the rest of the system, which in turn continues running uninterruptedly. That is to say, the conversation continues naturally despite the change of platform.
- Retrieval-based model: It contains the context's information and leads the conversation when the theme of child content is in the air. In this model, the knowledge is expressed using a set of rules in AIML (Artificial Intelligence Markup Language). Therefore, the response to a question is selected according to the match with one rule. Next, Listing 1 shows an example of an AIML rule within our context, where the *pattern* is checked against the received post, and *template* contains the associated response to be sent if the post actually matches the *pattern*.

  This model is executed whenever the chatbot receives a post from the suspect. When a post matches a defined rule, it means that the conversation has reached the context of interest. As a result, the corresponding reply is provided to continue the proper flow of the interactions. More details are discussed in depth in Section 4.

  However, and despite the effective management of controlled posts by the retrieval-based model, this component is unfortunately not suitable for out-of-context situations. The knowledge base is not infinite, and there may be cases where no answer can be found (there is not a matching rule for a received post). As C^3-Sex should be able to send valid responses, even in the absence of matching rules, a second model comes into play to alleviate such a shortcoming.
- Generative model: This component is activated whenever the model mentioned above does not know how to respond in the first instance. In this case, the generative model can reply to requests out of our scope of online child sexual abuse content, allowing C^3-Sex to still send valid responses and not get blocked when the conversation goes sideways.

  This model uses memory to maintain the learned patterns, and when a question (post) comes from the chat, the algorithm replies with the patterns saved in memory. The specific configuration of the model is addressed in Section 5. Note that responses from a generative model may be more natural than those from a retrieval-based model, but may contain grammatical failures or sometimes have no sense. Some examples can be appreciated in Table 2.

Listing 1. Example of Artificial Intelligence Markup Language (AIML) Rule.

```
<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>
```

Table 2. Example of replies from the generative model of C^3-Sex.

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?*

As seen before, the three elements (interaction channel, retrieval-based model and generative model) are strongly linked to the chatbot functionality. Additionally, it is crucial to store the conversation maintained by the interactive module in a database so that the next module can sucessfully operate.

4.2. Analysis Module

The analysis module is activated once the conversation has finished, that is, when the suspect leaves the chatroom/app or C^3-Sex receives (but does not download) media content. This module is responsible for characterizing the conversation. Specifically, there are two models which process and analyze the stored messages:

- Emotional classification model: It assesses the underlying emotions within a given conversation. This model classifies each suspect's response into one of the following six emotions: anger, disgust, fear, joy, sadness, and surprise. The emotional classifier makes use of a Support Vector Machine (SVM) as classification and supervised learning paradigm (see Section 5).
- Opinion classification model: It extracts the opinion inclined towards (or against) specific topics. In this case, a message is categorized as a positive claim or as a negative statement. The opinion classification model is an excellent complement to the emotional classification model for comparing results and guaranteeing consistency. The opinion classifier uses a Multinomial Naïve Bayes algorithm, as specified in Section 5.

Once the conversation managed by the interactive module has finished, and the consequent emotional and opinion metrics have been obtained, the chatbot returns a vector of features that describes suspect's behavior. These features are detailed in Section 6. Nonetheless, before studying the profiling vector, we will detail exactly how each of the mentioned AI models has been built and developed.

5. Artificial Intelligence Models in C^3

Every AI-based model of our proposed chatbot has followed a generic data science life cycle to enable a custom, automatic, and intelligent Artificial Conversational Entity. Specifically, each model has gone through the following phases [36]:

1. Business understanding: Definition of the data context where the solution will be deployed and executed.
2. Data acquisition: Step focused on obtaining data that will be used and applied in the modeling phase.
3. Modeling: Specific technologies, algorithms, or paradigms are selected to achieve the defined goals. The selected models should be configured and trained to adapt them to the use case (defined in step 1) through useful information (collected in step 2).
4. Deployment: The AI elements are jointly embedded in the system to enhance the effectiveness and efficiency of the tool. The modeled tool (developed in step 3) is launched in real scenarios as a validated product that works according to its mission. This phase involves continuous updating and maintenance of the application.

Table 3 shows this data science life cycle for the C^3-Sex models.

Table 3. Data science life cycle for C^3-Sex models.

Model	Business Understanding	Data Acquisition	Modeling	Deployment
Retrieval-based model	Child illegal content	Context information	100 AIML rules	Context responses
Generative model	Child illegal content	Cornell Movies (https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html)	LSTM-NN	Out-of-context responses
Emotion classifier	Pedophiles	SemEval [37]	Support Vector Machine	Sentiment extraction
Opinion classifier	Pedophiles	Opinion dataset [38]	Multinomial Naive Bayes	Opinion extraction

5.1. Business Understanding

Social media has brought many advantages to our society, but it has also led to the generation of comfortable spaces for dishonest, beguiler, and deceitful people [39]. Apart from social networks [40], other services have been successful in recent years, such as the Deep Web or anonymous chats due to their private nature. These virtual environments mask users interested in certain crimes, such as child sexual abusers, protecting them from being arrested.

Given that C^3-Sex is designed to explicitly detect criminals related to child sexual content, it is interesting to deploy it in those websites or scenarios where it is most likely to interact with suspects. In this regard, our context is defined by the Omegle platform where there are chatrooms under the 'sex' topic, with the possibility of migrating to other anonymous applications such as Telegram or Snapchat. Therefore, our software-driven agent should operate correctly in holding conversations in those platforms, considering their jargon, common patterns, and specific properties.

The retrieval-based model and the generative model should be prepared to manage conversations around sex and minors, whereas the emotion classifier and opinion classifier should work adequately to categorize the behavior and mood of potential pedophiles. The combination of the four models should guarantee:

- Maintenance of a fluid conversation in an anonymous chatroom that is primarily about sex, even if a change of platform is required.
- Analysis of the stored messages to characterize the suspect's interest in obtaining or distributing illegal child content.
- Profile the suspects depending on their level of interest towards online child sexual abuse content.

5.2. Data Acquisition

AI models should be trained with datasets to configure their operation. Given those datasets, AI models, and particularly machine learning algorithms, learn from observations and extract patterns. Once the model is configured, validated, and ready, the data of our context feeds those technologies to produce the desired outputs.

The retrieval-based model is used for knowledge representation and is built upon a series of rules focused on maintaining a conversation pivoting around the topic of online child sexual abuse. It is worth noting that this model is built with handmade rules, so it is not explicitly trained. However, a previous data acquisition is essential before creating rules for the context.

The more sophisticated the rules, the better the agent's simulation. In this regard, it is vital to understand the use case and analyze conversations in order to create custom, effective and valid rules.

Secondly, the generative model has been trained with the Cornell Movie Dialogs dataset [41]. This dataset contains conversations between characters from more than 600 movies. In this sense, the model gets configured with a naturalness attractive for the bot to respond in casual or jovial situations, for which a very elaborated response is not expected.

The emotion classifier is a model trained with the SemEval dataset [37]. This dataset contains news headlines from major newspapers (e.g., The New York Times, CNN, BBC News), each of them labeled through a manual annotation with one of the following six emotions: anger, disgust, fear, joy, sadness, and surprise. Therefore, this classifier is able to categorize a sentence or text in one of those six possible feelings.

In turn, the opinion classifier was trained with a dataset [38] containing reviews of movies, restaurants, and other products labeled as positive (1) or negative (0). Thanks to this classifier, we can categorize chat interactions.

5.3. Modeling

Once the data has been selected, found, or collected, the AI models should be configured, trained, and validated for the use case. To this extent, it is necessary to define the hyperparameters of the models (their configuration) to determine their behavior. Subsequently, they are trained with labeled datasets so that they learn to predict or classify. In the production environment, each model responds following the patterns learned from the training phase.

The retrieval-based model has been build with 100 AIML rules. These rules are used for different phases of the interaction: the formation of the friendly relationship (focused on capturing suspect's interest), the establishment of the sexual relationship (which already conveys an interest in the exchange of sexual material) and the assessment of risk (to gain the trust of the suspect). The old version of C^3-Sex [12] used a set of 60 AIML rules, so it has been augmented from a deeper review of experienced conversations between the C^3-Sex and suspects. Particularly, 40 additional AIML rules that consider new terms used in sexual-related conversations were created, giving the C^3-Sex the capacity to answer properly in more situations. Also, this new set of AIML rules hinders an easy identification of C^3-Sex as a chatbot by the suspects. Some examples of new AIML rules are shown in Listing 2.

The generative model responds according to the message sent by the suspect. This model was built using a Long Short Term Memory (LSTM) Neural Network (NN) [42] and trained with the Cornell Movie Dialogs dataset, where each instance of data is preprocessed by eliminating additional blank spaces, numbers and special characters, and treated as a Post-Reply message. Thus, to train this model, 50 epochs (top-down analysis of the datasets) and approximately four days of continuous execution were necessary. In addition, different manual validations were performed over the chatbot to analyze the error generated in the training of the LSTM-NN, avoiding both overfitting and underfitting.

Listing 2. Example of added AIML Rules.

```
<aiml version="1.0">
    <category>
        <pattern>* FEEL *</pattern>
        <template>cool, u know whats more exciting? pics haha</template>
    </category>
    ...
    <category>
        <pattern>* link to * child *</pattern>
        <template>
        i d like but i prefer in private add me snap, my nickname
        is p\_ramirezxxx
        <template>
    </category>
    ...
    <category>
        <pattern> * bored * bit horny </pattern>
        <template>
        I m horney and bored, do u wanna shared some content
        </template>
    </category>
</aiml>
```

The emotion classifier is based on the supervised method of Support Vector Machine (SVM). Given the SemEval dataset of sentences labeled with associated emotions (among the six possible), SVM is trained to seek those hyperplanes that separate the groups of sentences of each class. Consequently, the emotion classifier learns to interpret the emotion of future inputs with a rate of 0.5, which indicates how much the model changes in response to the estimated error each time that the model weights are updated [43]. The model can map a message within a specific hyperplane, that is, within an emotion.

The opinion classifier uses a Multinomial Naïve Bayes solution [44] with a simple preprocessing (steaming, removing stopwords, among other techniques) and an alpha of 1. It calculates the sentiment of texts considering the joint probabilities between the appearing words and existing sentiments (in our case, negative or positive). The representation of messages is based on a matrix with the words as columns.

5.4. Deployment

The deployment of C^3-Sex (https://github.com/CrkJohn/Snapchat, https://github.com/Santiago-Rocha/PGR) may be done through a connection with an online chat service where it can be possible to interact with different suspects. The four designed models should work together within the same system as a unique entity. In this manner, suspects directly interact with C^3-Sex through the interaction channel, remaining unaware that the communication is actually conducted with its AI-powered software modules.

In the interactive module, the communication with Omegle is done through a Chrome driver handled by functions of the Python library Selenium (https://pypi.org/project/selenium), which allows manipulating the DOM tree of the Omegle interface. On the other hand, the Snapchat management was implemented in the Snapbot (https://github.com/CrkJohn/Snapchat) module with the usage of Appium (http://appium.io/), Selenium, and Android Studio (https://developer.android.com/studio). Appium and Selenium allow us to manipulate the application as if automation tests or unit tests were conducted, and Android Studio allows us to know the identifiers of each component like login button, text entry, or send button. Therefore, we could achieve smooth and error-free navigation on the Snapchat. A physical mobile device was employed to install the Snapbot module and manage Snapchat conversations.

The implementation of the retrieval-based model was based on AIML (Artificial Intelligence Mark-up Language). Such model makes use of the library python-aiml (https://pypi.org/project/

python-aiml/), which serves as an inference engine to read XML files in the AIML way, which compose the entire knowledge base of the retrieval-based conversational module. Additionally, the generative model of C^3-Sex was implemented using Python 3 with a variety of libraries: (i) TensorFlow (https://www.tensorflow.org) as Machine Learning library used to build the recurrent neural network (RNN) which simulates the generative conversational agent, (ii) Sklearn (https://pypi.org/project/scikit-learn/) for the adoption of ML algorithms, such as the Bayesian network in the opinion classification model, and (iii) Pandas (https://pandas.pydata.org), a data analysis library used to manage and read data structures, such as CSV and dataframes.

Concerning the analysis module, the opinion classifier has been developed with Pandas Section 5.4 and the emotion classifier has been developed with R (https://www.r-project.org/).

The chatbot should be coherent, cohesive, and tolerant against unforeseen situations. To this end, C^3-Sex interventions are handled by the generative and retrieval-based models, generating trust with the suspect and subtly giving direction to the conversation to finally guess the possession of or expectation about possessing illegal content. In those cases where there is an imminent delivery of multimedia, a change of online chat service should be suggested in the conversation since Omegle does not allow to exchange images or videos. In fact, the change of platform must not affect the operation of the modules.

At the final step, the chatbot should be able to analyze the suspect's mood in his/her messages and his/her opinion on specific topics. For this purpose, C^3-Sex uses the emotion and opinion classification models to analyze the conversation once it has ended. This analysis of the suspect's behavior in conjunction with other metrics will be used to definitively build the profiling vector of the suspect.

6. Profiling Vector of the Suspect

Once a conversation is completed in both Omegle and Snapchat platforms, the entire record of such conversation is thoroughlyanalyzed to extract metrics about the the suspect's interest in child sexual abuse content. Finally, the profiling vector is completed with 25 features extracted and calculated from the interaction with the suspect. These features are shown in Table 4.

Features f_{1-6} refer to metrics of time for the conversation maintained between the C^3-Sex and the suspects. These metrics are important to identify how much a suspect is interested in a conversation. This due to the conversation time could be considered as an indicator of the interest of the suspect in the conversation topic. Features f_{8-11} aims to identify the activity of the suspect, and the subsequent activation of the generative or retrieval models to generate a response from the C^3-Sex to the suspect. Analyzing the number of responses generated by each model is critical to understand the behavior of the C^3-Sex.

If the proportion of replies from the retrieval model f_9 is bigger than the proportion of replies from the generative model f_{10}, for a total suspect posts (f_8), it is possible to deduce that the suspect has an affinity with child sexual abuse content, as the retrieval model only gets activated when a message coming from the suspect matches an AIML rule built specifically for the child sexual abuse context.

Features f_{12-13} refer to the amount of links to external sites that the suspect shares, which is useful to quantify the grade of confidence that the suspect has gained, as the existence of more external URLs implies a bigger willing to share content.

In turn, features f_{14-19} and f_{24} allow to determine how pleasant or enjoyable a conversation was for the suspect according to the emotions identified in the suspect posts by the emotion classifier. Positive emotions (Joy, Surprise) imply that the suspect is comfortable with the conversation topic.

Features f_{20-23} represent the general opinion of the suspect about the conversation topic maintained with the C^3-Sex . This also helps to identify the affinity of the suspect with a specific topic.

At last, f_{25} allows to identify the response time between a message of the C^3-Sex and a reply from the suspect, which is also an important aspect to determine the interest of the suspect in a conversation topic, due to the fact that a shorter response time may imply a greater interest.

Table 4. Suspect profiling features.

Feature	Name	Domain	Description
f_1	Initial timestamp in Snapchat	*Time*	Full timestamp (year, month, day, hour, minute, second) of the start of the conversation in Snapchat
f_2	Initial timestamp in Omegle	*Time*	Full timestamp (year, month, day, hour, minute, second) of the start of the conversation in Omegle
f_3	Final timestamp in Snapchat	*Time*	Full timestamp (year, month, day, hour, minute, second) of the end of the conversation in Snapchat
f_4	Final timestamp in Omegle	*Time*	Full timestamp (year, month, day, hour, minute, second) of the end of the conversation in Omegle
f_5	Snapchat duration	$\mathbb{Q}^+$	Total time on the Snapchat website in seconds, subtraction between metrics f_1 and f_3
f_6	Omegle duration	$\mathbb{Q}^+$	Total time on Omegle in seconds, subtraction between metrics f_2 and f_4
f_7	Conversation length	$\mathbb{N}^+$	Number of interactions by suspect and bot.
f_8	Suspect posts	$\mathbb{N}^+$	Number of times that the suspect interacted with our bot
f_9	Suspect posts about sex	[0–1]	Proportion of times that child sexual abuse rules (from the retrieval-based model) matched the total of suspect posts. (f_8)
f_{10}	Suspect posts not about sex	[0–1]	Proportion of times that the generative model responded because the retrieval model could not respond over the total of suspect posts (f_8)
f_{11}	Received bytes	$\mathbb{N}^+$	Total number of bytes of text received by the suspect
f_{12}	URLs from Snapchat	$\mathbb{N}^+$	Number of hyperlinks received by the suspect in Snapchat
f_{13}	URLs from Omegle	$\mathbb{N}^+$	Number of hyperlinks received by the suspect at Omegle
f_{14}	Anger	[0–1]	Rate between the number of times that the anger emotion was identified in the suspect posts and f_8
f_{15}	Disgust	[0–1]	Rate between the number of times that the disgust emotion was identified in the suspect posts and f_8
f_{16}	Fear	[0–1]	Rate between the number of times that the fear emotion was identified in the suspect posts and f_8
f_{17}	Sadness	[0–1]	Rate between the number of times that the sadness emotion was identified in the suspect posts and f_8
f_{18}	Joy	[0–1]	Rate between the number of times that the joy emotion was identified in the suspect posts and f_8
f_{19}	Surprise	[0–1]	Rate between the number of times that the surprise emotion was identified in the suspect posts and f_8
f_{20}	Negativity	[0–1]	Rate between the number of times that a negative post was identified by the opinion model and f_8
f_{21}	Positivity	[0–1]	Rate between the number of times that a positive post was identified by the opinion model and f_8
f_{22}	Neutrality	[0–1]	Rate between the number of times that a neutral post was identified by the opinion model and f_8
f_{23}	Opinion about sex	[0–1]	Average of the opinions found in the suspect's posts about sex, where 0 is a negative opinion and 1 is a positive opinion. Not neutral opinions are considered.
f_{24}	Emotion about sex	[0–1]	Average of the emotions found in the suspect's posts about sex, where 0 is a negative emotion (anger, disgust, fear, sadness) and 1 is a positive emotion (joy, surprise)
f_{25}	Response speed	$\mathbb{Q}^+$	Average time between C^3-Sex messages and suspect responses

7. Experiments

Our proposal's suitability has been validated through different experiments that aim to verify the chatbot's effectiveness in (i) holding a conversation with a human, (ii) characterizing the suspect's behavior with features, and (iii) automatically contacting a big number of users. Figure 3 shows the design of the experiments.

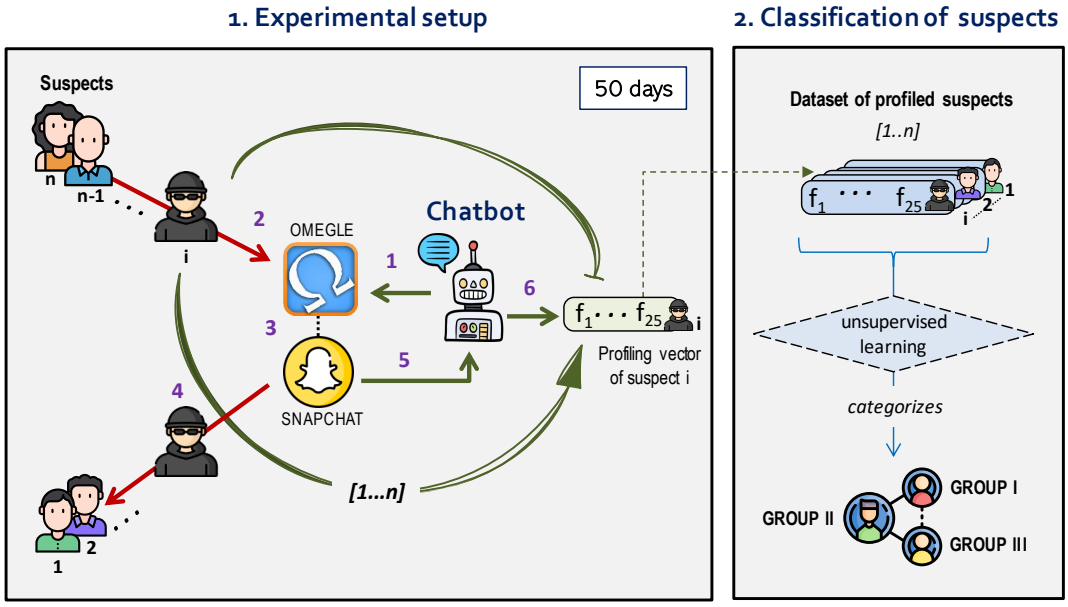


Figure 3. Experiments launched with C^3-Sex.

The settings of the experiment are described in Section 7.1, while the analysis of results is carried out in Section 7.2.

7.1. Experimental Setup

Experiments were conducted by running an instance of a C^3-Sex that connects to the online chat platform Omegle. Omegle was chosen because it allows to easily start communication with a peer in some place of the word around a common interesting topic. For these experiments, the interesting topic that was configured was "sex". In this way, the C^3-Sex may start a conversation with a suspect, who has also typed "sex" as interesting topic during the access to the Omegle chat service.

In the experiments, C^3-Sex simulated the behavior of a person interested in online child sexual abuse material, building trust for the suspects to transfer content and allowing profiling them according to the previously defined metrics based on the conversation performed.

The chatbot was executed for 50 days to contact and profile suspects through Omegle automatically. The workflow of each interaction of the C^3-Sex with a suspect is detailed in the left of Figure 3:

1. C^3-Sex joins into an Omegle chat room typing sex as an interesting topic.
2. Suspect *i* also logs in to the chat room.
3. The conversation starts between the user and the chatbot. In case the suspect is willing to exchange multimedia content, the C^3-Sex suggests to use Snapchat to exchange the multimedia content.
4. At some point, suspect *i* leaves the chat room.
5. Accordingly, C^3-Sex closes the conversation.
6. Finally, C^3-Sex analyzes the conversation using the profiling metrics and AI models (emotional classification and opinion classification model), calculates the features, and saves the associated profiling vector in a dataset for further analysis. Then, the C^3-Sexreturns to step 1 to start a new conversation with another suspect.

As a result of 50 days of experimentation, the chatbot has built a dataset of *n* profiled suspects composed by the associated *n* vectors of features. In the following section, we will study in-depth the results of the experiments, executed for 50 days (from 19-04-2020 to 07-06-2020), time in which the C^3-Sex was able to contact 7199 suspects.

7.2. Classification of Suspects

After completing the 50 days of activity, we had a database containing the information about 7199 suspects, including the 25 features from f_1 to f_{25} calculated for each of them. The operation of the chatbot has finished, and the investigator should inspect the dataset to detect pedophiles. In our case, the high complexity of analyzing these unlabeled users by hand led us to propose an unsupervised learning method to provide a first aggregated description of the results. This analytical procedure may support the competent authorities with the manual exploration and identification of sex offenders.

Since our ultimate goal is to detect perverts, we individually pursued those users who demonstrated suspicious patterns in obtaining or transmitting sexual content. With this aim, a possible solution was to differentiate various groups within our dataset, which shared common ways of thinking, behaving, and acting, that is, with similar metrics. Thanks to our subsequent expert analysis of the groups that emerged, we could infer different types of suspects according to their features. One or more groups could share aspects related to pedophilia.

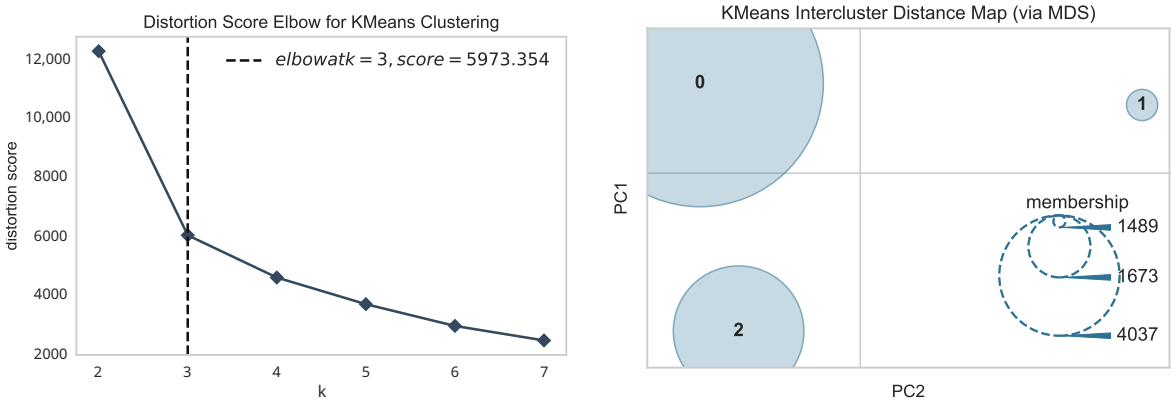
In this context, we chose the K-Means clustering method [45] due to its simplicity and efficiency, linear computational complexity, and convergence in detecting clusters. Our dataset did not present particularities to force us to use a specific algorithm in this sense, and K-Means maintains an excellent accuracy–complexity ratio.

K-Means algorithm classified each instance of the dataset in the most similar cluster among the existing K groups. In particular, this method is based on the distances between instances' features, so the cluster assigned was the one in which the centroid was closest to the instance in the feature multidimensional space. Due to this algorithm's nature, it operates better as there are fewer dimensions (features).

In our specific use case, the original set of 25 features would have produced a large multidimensional space and subsequent inaccurate clusters. For this reason, we performed a feature selection process in which we tested several combinations among the 25 features, always standardizing them and reducing the dimensionality with a Principal Component Analysis (PCA, limited to 95% of explained variance). As a result, we obtained that the configuration with lower values of distortion was formed by the features f_5 (Snapchat duration),f_6 (Omegle duration), f_9 (Suspect posts about sex), f_{23} (Opinion about sex), f_{24} (Emotion about sex) and f_{25} (Speed of response). Fortunately, those metrics aggregate other (discarded) features and are, from our intuition, truly relevant for the search of perverts.

Finally, considering only the selected features, we calculated the optimal number of clusters K with the distortion score elbow method [45]. Assuming that a larger number of clusters always groups more precisely (in its maximum case, each instance would belong to its own cluster), the idea is to find the lowest possible K that guarantees the best balance between the average distance of the cluster points to its centroid (distortion) and the total number of clusters. Figure 4a shows that $K = 3$ is the optimal number of groups in our dataset to divide the suspects faithfully. On the other side, Figure 4b represents the inter distances between the cluster centers as a result of a Multidimensional Scaling (MDS). The clusters' size is proportional to the number of elements they contain, in particular 1489, 1673 and 4037 instances.

Based on the formed groups, in the following sections, we characterize and compare each group of suspects according to their features.



(**a**) Distortion scores per number of possible clusters (**b**) Clusters with the best accuracy and performance

Figure 4. Application of K-Means to generate groups of suspects.

7.3. Comparative of Detected Groups

Once we detected groups of similar suspects, we compared them according to the selected features. To provide a clear and homogeneous view of the differences, we applied a quantile-based scaler to the features (where values go from 0 to 1). Note that this transformation is applied to each feature independently and maps the original values to a uniform distribution using quantiles information. Another type of transformation could significantly offer bad visualizations for comparison, alter the data distribution, or suffer a lot from outliers. Conversely, the quantile transform respects the essence of the variability of the data and maintains the so valuable outliers. Despite some distribution distortion, the conversion facilitated a great comparison. Note that we did not want to remove anomalies because those specific antipatterns might potentially expose pedophiles.

In this regard, Figure 5 shows the high-level comparative of each group according to the scaled features. For this particular experiment we have deducted that the three found groups refer to the categories of indifferent, interested, and pervert as we will explain next. These assigned names for each cluster come from the analysis of the features of the included conversations, but it is not a judgment 25 statement, and must just be considered as a way to qualitative describe the results.

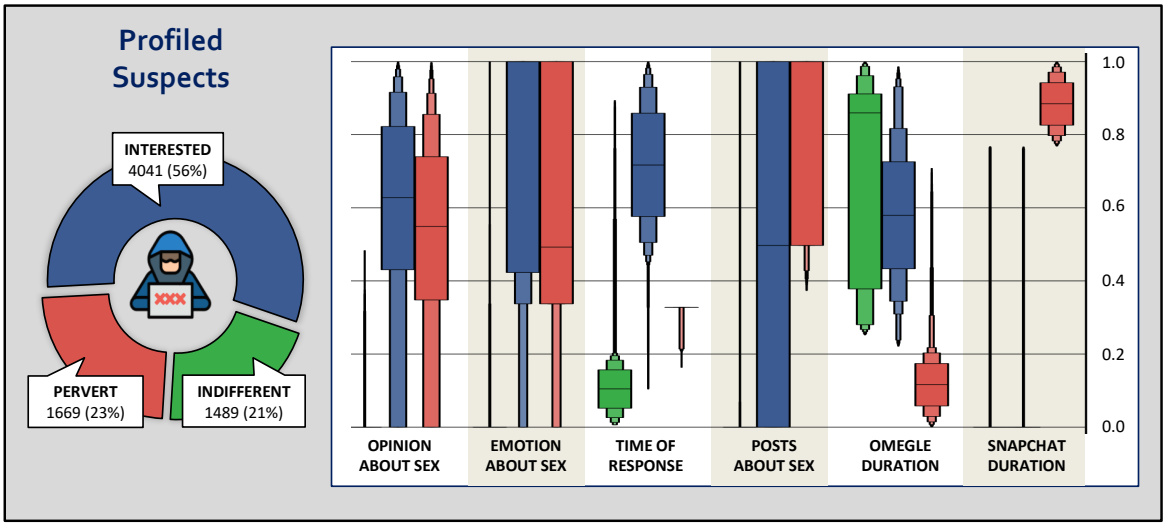


Figure 5. Comparative of the different types of profiled suspects.

7.3.1. Indifferent

We have considered the green group, the cluster with 1489 elements, to be the indifferent users. In terms of online child sexual abuse, they have a bad opinion about it, demonstrate bad emotions

when dealing with child abuse topics, and do not send sexual messages. Moreover, they are also the quickest to respond, which could demonstrate an apparent devotion to rejecting the transmission of compromised content. Another proof of the latter is that they spend all the conversation on Omegle, the platform that does not allow sending multimedia, and when the change to Snapchat is made, they leave.

7.3.2. Interested

The cluster denoted in blue, composed of 4041 suspects, demonstrates a higher interest. In general, they have a favorable opinion about child abuse proposals and present positive emotions in their sexual interactions. Nevertheless, in this group, there are users with different frequencies in posting sexual content. In this sense, one could say that these suspects might be interested in reading or talking about sex, without welcoming the transmission of content. Therefore, they are not fast to respond, which might indicate skepticism when talking about sex, although they do not reject it. Also, they focus their activity in Omegle, where no photos or videos are allowed.

7.3.3. Pervert

Finally, the red-colored group is formed by 1669 suspects and presents a similar interest in child sex content as interested users. However, the main difference with the rest of the groups is their predisposition for the transmission of sexual content. They demonstrate a receptive attitude to having a conversation about sex as the interested group, but they take little time to respond, they spend a short time in Omegle and register a high activity in Snapchat, the platform to which users move to exchange multimedia.

Therefore, having described the different types of profiled suspects, in the following section, we analyze in-depth the values of the different metrics by groups to precisely characterize their behavior.

7.4. Behavior of Profiled Suspects

Considering the principal differences among the distinguished groups, it is worth analyzing each group's specific behavior in detail. Figure 6 compares the distribution of values per detected group for the selected features. Note that it is a more accurate perspective than Figure 5, which served us to categorize each group.

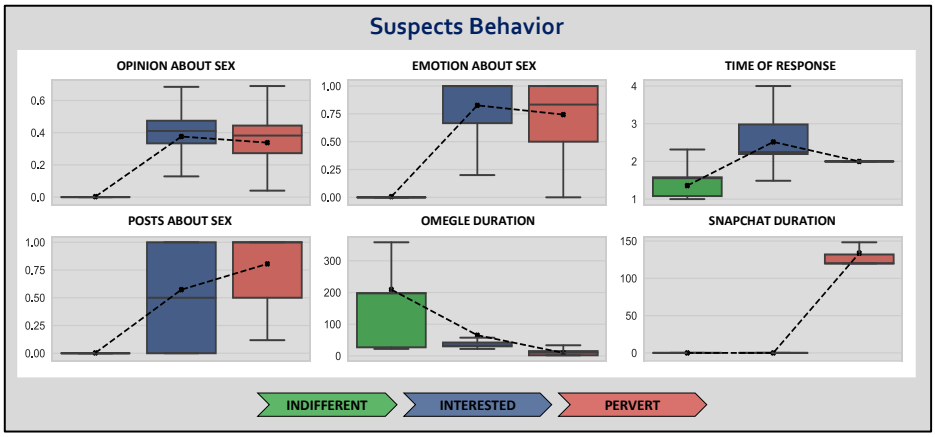


Figure 6. Behavior of profiled groups.

First of all, interested and pervert suspects maintain a positive tendency around 0.4, reaching even a 0.7 of positivity, in child abuse interactions. On the other hand, as indifferent users practically do not send messages about child sexual aspects, it demonstrates a totally negative opinion against minor contents.

Moreover, interested and pervert suspects experience very positive emotions within sexual messages, more than 0.75 of emotional positivism for the former in most cases, and more than 0.50 in the latter. On the contrary, indifferent users also show strong negative emotions since they do not post about sex.

In terms of response time, indifferent users generally take less than 2 s to respond. On the other hand, interested users typically require 2 and 3 s to respond, eventually reaching 4 s. Surprisingly, perverts spend exactly two seconds, on average, to interact with our chatbot. This non-variable value makes us suspect that these users, apart from being interested in sex, were actually programmed to interact automatically, that is, they were probably bots controlled by software. Nevertheless, these numbers demonstrate that all users using Omegle are interested in chatting and go away from keyboard.

As mentioned before, indifferent users do no usually send messages with sexual content. Regarding interested users, there is a continuous uniform distribution at all levels, from suspects who barely post about sex to others who mention sexual topics within almost 100% of their messages. Additionally, most perverts send sexual messages in over 50% of their interactions.

Finally, the differences in chatroom duration are significantly disparate. In the Omegle conversation, indifferent users spend up to 200 s (3–4 min) in most cases, while interested and pervert suspects do not usually reach 50 s in the best cases. When moving to Snapchat, both the indifferent and interested users definitely leave the conversation. Yet, perverts remain online in Snapchat for a considerable period of 2 to 4 min in predisposition to broadcast or receive multimedia.

With these results, C^3-Sex is proven to be a powerful tool for profiling suspects based on premeditated metrics. The following sections show how long the chatbot has been running in order to profile 7199 users effectively.

7.5. Experiments Overview and Limitations

C^3-Sex is an automatic and software-controlled conversational agent deployed on the web that interacts with users autonomously. Figure 7 represents the weekly temporal activity of our chatbot in the experiments, showing its high capacity to analyze the network. On average, C^3-Sex can interact with 900 suspects weekly, a value that is already high compared to a manual intervention replicated by a human. In particular, in the seventh week, the chatbot chatted with more than 500 users, and in weeks 2, 3, and 5, it surpassed the incredible number of 1500 suspects. Moreover, in the second week C^3-Sex was able to maintain contact with nearly 2500 network users.

The diagram helps us to evaluate the chatbot very positively, which was able to stay online throughout the eight weeks of the experiment, with a total of 7199 users contacted. Therefore, this tool has is a valuable support for profiling people on a large scale and with descriptive statistics of behavior.

However, C^3-Sex also has some limitations. Although it profiles suspects through a wide range of features, it is unfeasible for the current version of the C^3-Sex to infer the potential intention of the suspect or the probability of being a pedophile. Such determination is the job of the cyber intelligence analyst who employs C^3-Sex and who must analyze the generated clusters to determine the type of users allocated in each one. A coming step in the evolution of C^3-Sex could be to use the 7199 conversations, which were categorized with an unsupervised method in Section 7 as belonging to indifferent, interested or pervert suspects, to train a supervised model that may be able to make an automatic categorization of the new suspects that interact with C^3-Sex. On the other hand, artificial intelligence models included in Table 3 could be trained with more data and thus achieve a more intelligent chatbot when interacting and analyzing.

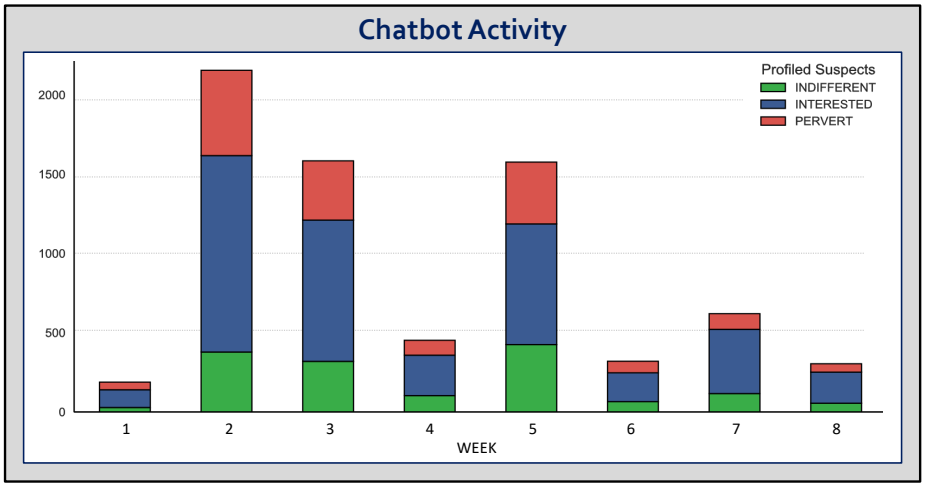


Figure 7. Chatbot activity.

8. Conclusions and Future Work

With the aim of humbly contributing to the honorable task of prosecuting sexual crimes, specifically online child sexual abuse, our conversational agent C^3-Sex has been proposed along with this paper. C^3-Sex is composed of four models, namely: Retrieval-based, Generative, Emotional classification, and Opinion classification. Altogether, these models constitute a solution able to keep conversations with suspects and profile them with up to 25 features related to the conversation's content and the behavior in the chatroom. The proposed analytic classification task reveals that these profiling features segregate different kinds of users and behaviors. The final goal of C^3-Sex is to profile users and expose potential holders and bidders of illegal content related to online child sexual abuse, who can later be investigated by a Law Enforcement Agency.

As future work, we are willing to investigate and compile real chats of sex offenders to implement a supervised model and provide a robust categorization mechanism. The latter would also allow the validation of the tool with a specific numeric accuracy. In fact, in the next release, the chatbot could include the percentage of the dangerousness of the other peer, having predictive functions on-the-fly as the conversation goes on. We also 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 by generating more specific AIML rules for the retrieval model and training the generative model with a dataset associated with 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. Some of these new types of sexual crimes would require C^3-Sex to keep more complex conversations for a longer time.

Author Contributions: Conceptualization, D.D.-L.; methodology, D.D.-L. and F.G.M.; software, J.I.R. and S.R.D.; validation, J.P.-G.; formal analysis, J.I.R., S.R.D. and D.D.-L.; investigation, J.I.R., S.R.D. and J.P.-G.; resources, D.D.-L.; data curation, J.I.R., S.R.D. and Javier Pastor-Galindo; writing—original draft preparation, J.I.R., S.R.D. and D.D.-L.; writing—review and editing, J.P.-G. and F.G.M.; visualization, J.P.-G. and F.G.M.; supervision, D.D.-L. and F.G.M.; project administration, D.D.-L. and F.G.M.; funding acquisition, D.D.-L. and F.G.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by the Colombian School of Engineering Julio Garavito (Colombia), by the Escuela de Ingeniería, Ciencia y Tecnología and the Dirección de Investigación e Innovación at the Universidad del Rosario (Colombia), by an FPU predoctoral contract (FPU18/00304) granted by the Spanish Ministry of Science, Innovation and Universities, 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.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACE	Artificial Conversational Entity
AI	Artificial Intelligence
AIML	Artificial Intelligence Markup Language
ECM	Emotional Chatting Machine
KNN	K-Nearest Neighbor
LEA	Law Enforcement Agency
LSTM-NN	Long Short Term Memory Neural Network
MCA	Membership, Categorization, Analysis
MDS	Multidimensional Scaling
NLG	Natural Language Generator
NLP	Natural Language Processor
P2P	Peer to Peer
PCA	Principal Component Analysis
SVM	Support Vector Machine
VoIP	Voice over IP

References

1. Díaz López, D.O.; Dólera Tormo, G.; Gómez Mármol, F.; Alcaraz Calero, J.M.; Martínez Pérez, G. Live Digital, Remember Digital: State of the Art and Research Challenges. *Comput. Electr. Eng.* **2014**, *40*, 109–120. [CrossRef]
2. Pina Ros, S.; Pina Canelles, A.; Gil Pérez, M.; Gómez Mármol, F.; Martínez Pérez, G. Chasing offensive conducts in social networks: A reputation-based practical approach for Frisber. *ACM Trans. Internet Technol.* **2015**, *15*, 1–20. [CrossRef]
3. Taylor, M.; Quayle, E. *Child Pornography: An Internet Crime*; Psychology Press: London, UK, 2003.
4. Adams, W.; Flynn, A. *Federal Prosecution of Commercial Sexual Exploitation of Children Cases, 2004–2013*; US Department of Justice, Office of Justice Programs, Bureau of Justice: Washington, DC, USA, 2017.
5. Krone, T. International Police Operations Against Online Child Pornography. In *Trends and Issues in Crime and Criminal Justice*; Australian Institute of Criminology: Canberra, Australia, 2005; Volume 2005.
6. Wolak, J.; Liberatore, M.; Levine, B.N. Measuring a year of child pornography trafficking by U.S. computers on a peer-to-peer network. *Child Abus. Negl.* **2014**, *38*, 347–356. [CrossRef] [PubMed]
7. McIntyre, T.J. Blocking child pornography on the Internet: European Union developments. *Int. Rev. Law Comput. Technol.* **2010**, *24*, 209–221. [CrossRef]
8. Ulges, A.; Stahl, A. Automatic detection of child pornography using color visual words. In Proceedings of the 2011 IEEE International Conference on Multimedia and Expo, Barcelona, Spain, 11–15 July 2011; pp. 1–6. [CrossRef]
9. Gottfried, E.D.; Shier, E.K.; Mulay, A.L. Child Pornography and Online Sexual Solicitation. *Curr. Psychiatry Rep.* **2020**, *22*, 10. [CrossRef] [PubMed]
10. Gao, J.; Galley, M.; Li, L. Neural Approaches to Conversational AI. *Found. Trends Inf. Retr.* **2019**, *13*, 127–298. [CrossRef]
11. Walker, M. Hype Cycle for Emerging Technologies, 2018. In *2018 Hype Cycles: Riding the Innovation Wave*; Gartner: Stamford, CT, USA, 2018.
12. Murcia Triviño, J.; Moreno Rodríguez, S.S.; Díaz López, D.O.; Gómez Mármol, F. C3-Sex: A Chatbot to Chase Cyber perverts. In Proceedings of the 4th IEEE Cyber Science and Technology Congress, Fukuoka, Japan, 5–8 August 2019; pp. 50–57. [CrossRef]
13. Kowalski, S.; Pavlovska, K.; Goldstein, M. Two Case Studies in Using Chatbots for Security Training. In *Information Assurance and Security Education and Training*; Dodge, R.C., Futcher, L., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 265–272.

14. Sabbagh, B.A.; Ameen, M.; Wätterstam, T.; Kowalski, S. A prototype For HI2Ping information security culture and awareness training. In Proceedings of the 2012 International Conference on E-Learning and E-Technologies in Education (ICEEE), Lodz, Poland, 24–26 September 2012; pp. 32–36. [CrossRef]
15. Filar, B.; Seymour, R.; Park, M. Ask Me Anything: A Conversational Interface to Augment Information Security Workers. In Proceedings of the SOUPS, Santa Clara, CA, USA, 12–14 July 2017.
16. Zambrano, P.; Sanchez, M.; Torres, J.; Fuertes, W. BotHook: An option against Cyberpedophilia. In Proceedings of the 2017 1st Cyber Security in Networking Conference (CSNet), Rio de Janeiro, Brazil, 18–20 October 2017; pp. 1–3. [CrossRef]
17. Açar, K.V. Webcam Child Prostitution: An Exploration of Current and Futuristic Methods of Detection. *Int. J. Cyber Criminol.* **2017**, *11*, 98–109. [CrossRef]
18. Gapanyuk, Y.; Chernobrovkin, S.; Leontiev, A.; Latkin, I.; Belyanova, M.; Morozenkov, O. The Hybrid Chatbot System Combining Q&A and Knowledge-base Approaches. In Proceedings of the 7th International Conference on Analysis of Images, Social Networks and Texts (AIST 2018), Moscow, Russia, 5–7 July 2018; pp. 42–53.
19. Serban, I.V.; Sankar, C.; Germain, M.; Zhang, S.; Lin, Z.; Subramanian, S.; Kim, T.; Pieper, M.; Chandar, S.; Ke, N.R.; et al. A deep reinforcement learning chatbot. *arXiv* **2017**, arXiv:1709.02349.
20. Tatai, G.; Csordás, A.; Kiss, Á.; Laufer, L.; Szaló, A. The chatbot who loved me. In Proceedings of the ECA Workshop of AAMAS, Budapest, Hungary, 10–15 May 2003.
21. Plutchik, R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* **2001**, *89*, 344–350. [CrossRef]
22. Qian, Q.; Huang, M.; Zhao, H.; Xu, J.; Zhu, X. Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; pp. 4279–4285.
23. Heller, B.; Proctor, M.; Mah, D.; Jewell, L.; Cheung, B. Freudbot: An investigation of chatbot technology in distance education. In Proceedings of the EdMedia: World Conference on Educational Media and Technology. Association for the Advancement of Computing in Education (AACE), Montreal, QC, Canada, 27 June 2005; pp. 3913–3918.
24. Hill, J.; Ford, W.R.; Farreras, I.G. Real conversations with artificial intelligence: A comparison between human—human online conversations and human–chatbot conversations. *Comput. Hum. Behav.* **2015**, *49*, 245–250. [CrossRef]
25. Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
26. Anderson, P.; Zuo, Z.; Yang, L.; Qu, Y. An Intelligent Online Grooming Detection System Using AI Technologies. In Proceedings of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 23–26 June 2019; pp. 1–6. [CrossRef]
27. Gunawan, F.E.; Ashianti, L.; Sekishita, N. A simple classifier for detecting online child grooming conversation. *Telkomnika Telecommun. Comput. Electron. Control* **2018**, *16*, 1239–1248. [CrossRef]
28. Ebrahimi, M. Automatic Identification of Online Predators in Chat Logs by Anomaly Detection and Deep Learning. Master's Thesis, Concordia University, Montreal, QC, Canada, 2016.
29. Dadvar, M.; de Jong, F. Cyberbullying Detection: A Step toward a Safer Internet Yard. In Proceedings of the 21st International Conference on World Wide Web, Association for Computing Machinery, New York, NY, USA, 16 April 2012; pp. 121–126. [CrossRef]
30. Laorden, C.; Galán-García, P.; Santos, I.; Sanz, B.; Hidalgo, J.M.G.; Bringas, P.G. 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*; Herrero, Á., Snášel, V., Abraham, A., Zelinka, I., Baruque, B., Quintián, H., Calvo, J.L., Sedano, J., Corchado, E., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 261–270.
31. Michalopoulos, D.; Mavridis, I.; Jankovic, M. GARS: Real-time system for identification, assessment and control of cyber grooming attacks. *Comput. Secur.* **2014**, *42*, 177–190. [CrossRef]
32. Burnap, P.; Rana, O.F.; Avis, N.; Williams, M.; Housley, W.; Edwards, A.; Morgan, J.; Sloan, L. Detecting tension in online communities with computational Twitter analysis. *Technol. Forecast. Soc. Chang.* **2015**, *95*, 96–108. [CrossRef]

33. Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2544–2558. [CrossRef]
34. Sacks, H.; Jefferson, G.; Schegloff, E. *Lectures on Conversation*; Wiley-Blackwell: Hoboken, NJ, USA, 2010.
35. Meyer, M. Machine Learning to Detect Online Grooming. Master's Thesis, Department of Information Technology, Uppsala University, Uppsala, Sweden, 2015.
36. Ericson, G.; Rohm, W.; Martens, J.; Sharkey, K.; Casey, C.; Harvey, B.; Nevil, T.; Gilley, S.; Schonning, N. Team Data Science Process Documentation. *Retrieved April* **2017**, *11*, 2019.
37. Strapparava, C.; Mihalcea, R. Semeval-2007 task 14: Affective text. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 23–24 June 2007; pp. 70–74.
38. Kotzias, D.; Denil, M.; De Freitas, N.; Smyth, P. From group to individual labels using deep features. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, USA, 12–15 August 2015; pp. 597–606.
39. Gómez Mármol, F.; Gil Pérez, M.; Martínez Pérez, G. Reporting Offensive Content in Social Networks: Toward a Reputation-based Assessment Approach. *IEEE Internet Comput.* **2014**, *18*, 32–40. [CrossRef]
40. Pastor Galindo, J.; Nespoli, P.; Gómez Mármol, F.; Martínez Pérez, G. The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. *IEEE Access* **2020**, *8*, 10282–10304. [CrossRef]
41. Danescu-Niculescu-Mizil, C.; Lee, L. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, Portland, OR, USA, 23 June 2011.
42. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
43. Brownlee, J. Understand the Impact of Learning Rate on Neural Network Performance. 2019. Available online: https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks (accessed on 12 December 2019).
44. Gupte, A.; Joshi, S.; Gadgul, P.; Kadam, A.; Gupte, A. Comparative study of classification algorithms used in sentiment analysis. *Int. J. Comput. Sci. Inf. Technol.* **2014**, *5*, 6261–6264.
45. Kodinariya, T.M.; Makwana, P.R. Review on determining number of Cluster in K-Means Clustering. *Int. J.* **2013**, *1*, 90–95.