A Privacy-Preserving Architecture for the Protection of Adolescents in Online Social Networks

Markos Charalambous[1], Petros Papagiannis[1], Antonis Papasavva[1], Pantelitsa Leonidou[1],
Rafael Constaninou[2], Lia Terzidou[3], Theodoros Christophides[1], Pantelis Nicolaou[2]
Orfeas Theofanis[4], George Kalatzantonakis[4], Michael Sirivianos[1]
[1]Cyprus University of Technology, Limassol Cyprus
[2]Cyprus Research and Innovation Center, Nicosia, Cyprus
[3]Aristotle University of Thessaloniki, Thessaloniki, Greece
[4]LSTech LTD, Milton Keynes, United Kingdom
Email: {marcos.charalambous, petros.papagiannis, t.christophides, michael.sirivianos}@cut.ac.cy,
{as.papasavva, pl.leonidou}@edu.cut.ac.cy, {r.constantinou, p.nicolaou}@cyric.eu, lterz@csd.auth.gr,
{orfetheo, george}@lstech.io

***Abstract*—Online Social Networks (OSN) constitute an integral part of people's every day social activity. Specifically, mainstream OSNs, such as Twitter, YouTube, and Facebook are especially prominent in adolescents' lives for communicating with other people online, expressing and entertain themselves, and finding information. However, adolescents face a significant number of threats when using online platforms. Some of these threats include aggressive behavior and cyberbullying, sexual grooming, false news and fake activity, radicalization, and exposure of personal information and sensitive content. There is a pressing need for parental control tools and Internet content filtering techniques to protect the vulnerable groups that use online platforms. Existing parental control tools occasionally violate the privacy of adolescents, leading them to use other communication channels to avoid moderation. In this work, we design and implement a user-centric Cybersafety Family Advice Suite (CFAS) with Guardian Avatars aiming at preserving the privacy of the individuals towards their custodians and towards the advice tool itself. Moreover, we present a systematic process for designing and developing state of the art techniques and a system architecture to prevent minors' exposure to numerous risks and dangers while using Facebook, Twitter, and YouTube on a browser.**

Keywords–online social networks; online threats; cybersecurity risks; privacy; minors.

I. INTRODUCTION

The majority of teens (85%) use more than one social media site according to a Pew Research Center [1] survey ($N = 743$). A 2018 poll ($N = 1001$) [2] found that the average 5 to 15 year-olds spend about 15 hours online every week. Additionally, 90% of the 11 to 16 year-olds surveyed said that they have an online social network account. These numbers illustrate that the overwhelming majority of young people use OSNs, even if they are not old enough to legally register accounts for most mainstream OSNs, like Facebook, Instagram, Twitter, YouTube, and Snapchat. Alarmingly, there are many risks adolescents are exposed to when using OSNs. Specifically, a 2019 study [3] of 21.6K primary school children and 18.1K secondary school children found that 16% and 19%, accordingly, had seen content that encouraged people to hurt themselves. The same study reports that 11 to 18 year-olds reported seeing sexual content in the most popular OSNs. Last, reviews from over 2K young people aged 11 to 18, show that the 16% witnessed violence and hatred, 16% encountered sexual content, and the 18% witnessed others being victims of cyberbullying. A different study conducted in 2018 found that 59% of U.S. teens have been victims of cyberbullying or harassment online. Additionally, about a third (32%) of teens report that someone has spread false rumors about them on the Internet, while smaller shares (16%) have been the target of physical threats online. Notably, the majority of the victims tend to be females. The study concludes that 59% of the parents worry that their child might be getting bullied online, but most are confident they can teach their teen about acceptable online behavior [4].

Overall, the popularity of the Internet, and OSN usage in particular, is very high and with an increasing tendency among youngsters. Thus, the online risks for these sensitive age groups received increased awareness. To design an architecture for the protection of youngsters in OSNs, we list the most frequent dangers the young users might encounter. Existing literature [4]–[6] agrees to the following distinctive threats: i) cyberbullying; ii) cyberpredators; iii) sensitive information leakage; iv) manipulated content and pornography; and v) offensive images and messages.

Contributions. In summary, this work makes the following contributions:

1) The design and implementation of a privacy-preserving CFAS that utilizes machine learning classifiers and other filters to protect minors when using OSNs.
2) CFAS makes efforts to keep the minors fully aware of what their custodians and what the Family Advice Suite can monitor, filter, and analyze about their online activity.
3) CFAS employs fine-grained tools to spread awareness to the custodians and the minors about the various threats they face when using OSNs. It also utilizes the Guardian Avatar that interacts and advises the adolescents in a direct and user-friendly way.
4) The proposed architecture can accurately detect: (i) cyber-bullying; (ii) sexual grooming; (iii) abusive users; (iv) bot accounts; (v) personal information exposure; (vi) sensitive

 ISBN: 978-1-61208-818-1

content in pictures; (vii) hateful and racist memes; and (viii) disturbing videos.

Paper Organization. The rest of the paper is organized as follows. First, we provide a detailed demonstration of the proposed architecture in Section II, followed by our design principles in Section III. Then, we list and discuss how the classifiers hosted on the Intelligent Web-Proxy (IWP) work in Section IV. We also provide an early evaluation of the system via a virtual environment, and physical experiments with beta testers (Section V), before discussing existing related work on parental control tools in Section VI. Last, we conclude this work in Section VII.

II. Architectural Overview

In this section, we describe the main pillars of our architecture. This architecture comprises the following: 1) OSN Data Analytics Software Stack (Back-End); 2) Intelligent Web-Proxy; and 3) browser add-on. For the tool to work efficiently, all three components interact with each other, but none depends on the other to function. Figure 1 depicts the proposed architecture of the CFAS framework, including its main components and the interfaces that interconnects them. We describe the main purposes and functionalities of each component below.

A. OSN Data Analytics Software Stack

The first component of the CFAS architecture is the OSN Data Analytics Software Stack, referred to as the *Back-End* henceforth. This is a single machine, which is responsible to train machine learning algorithms for the detection of threats in OSNs. The trained classifiers and detection rules created on this machine are sent automatically to the registered Intelligent Web-Proxies (IWP) when available (see # in Figure 1). In addition, the Back-End stores anonymized OSN traffic data from the registered IWPs, *only* if both the custodian and the minor give their explicit consent (4* in the figure). These anonymized data are used to retrain the machine learning algorithms hosted in the Back-End to extract more accurate and intelligent classifiers, which are sent back to the IWPs to replace the existing classifiers, as shown in step # in Figure 1.

B. Intelligent Web-Proxy

The Intelligent Web-Proxy (IWP) is a small device that is connected to the router of the service provider in the house of the protected family. We note that every different network needs its own IWP to be protected as a single IWP supports only one network. The IWP consists of three modules that handle specific tasks, as described below.

1) DOM Tree Analysis: This part of the IWP captures all the incoming and outgoing traffic of the user (child). Note that the word *user* refers to the child protected by our architecture henceforth. First, the user requests a webpage using their browser (see 1 at Figure 1). The response of this request is sent to the IWP: the DOM Tree Analysis module, specifically (step 3 in the figure). After capturing the traffic, the DOM Tree Analysis module handles TLS connections and performs TLS termination to decrepit HTTPS websites (only Facebook and Twitter currently). Importantly, the IWP is tested to manage high network traffic load and extract the webpage content from the captured DOM tree. At the same time, the same data are sent to the Data Access Layer for analysis (see 4 in Figure 1). We describe how the Data Access Layer (DAL) works below.

2) Data Access Layer: The Data Access Layer hosts all the trained classifiers and detection rules generated from the Back-End that are used to check all the received captured traffic.

Figure 2 demonstrates the functionality of the Data Access Layer, which is the main storage unit hosted in the IWP and the Back-End of the CFAS infrastructure. First, the data captured by the DOM Tree Analysis are sent to the *Decision Mechanism* of DAL (step 1 in Figure 2). Every bit of information (Facebook chat, Facebook news-feed pictures, Facebook posts created by the user, Facebook pictures uploaded by the user, visited YouTube videos, and visited Twitter user profiles) is sent individually. Upon reception of this data, the Decision mechanism creates a unique Execution ID (ExecID), see step 2 in the figure. This unique string is used by the Decision mechanism to define the job number of the trained classifier, which is used to analyze the data.

Then, the Decision mechanism requests the Data Access API to store this data in the database: a MongoDB (step 3). Once the data are stored, the Data Access API binds them with a unique number, which is used as a primary key to identify these data: DataID. The DataID is sent back to the Decision mechanism (step 4), which is combined with the ExecID to call the suitably trained classifier to detect suspicious behavior (see step 5). Once the trained classifier receives the ExecID and the DataID, it sends the DataID to the Data Access API to request the retrieval of data for analysis (step 6), which in return are sent back to the trained classifier (step 7). Once the trained classifier finished the analysis of the data, it sends its results to the Data Access API, along with the ExecID and DataID to be stored in the database (step 8). Then, the trained classifier sends the ExecID and DataID back to the Decision Mechanism to inform it that the analysis finished (step 9).

In response, the Decision Mechanism requests the results of the job from the Data Access API (step 10), and the Data Access API responds with the results of the analysis (step 11). Last, based on the results of the trained classifier, and thresholds set in the Decision mechanism, the Decision mechanism is responsible to decide whether a notification needs to be sent to the user via the CFAS browser add-on, and to the custodian of the user, via the Parental Console. If this is the case, the Decision Mechanism triggers an event via the Notification Module (step 12). Note that step 12 in Figure 2 is the same as step 5 and step 5* in Figure 1.

3) Parental Console: The last component hosted in the Intelligent Web-Proxy is the Parental Console. The Parental Console is a fine-grained web-based platform that enables the custodian of the user to manage which data of the user (child) he/she and the IWP can see. Also, via the Parental Console, the custodian can choose what the IWP filters, protects, and blocks. Additionally, custodians can set the level of the child's cybersafety. To set these options in operation, the child receives notifications on their browser add-on through the Notification Module, informing them that their custodian has made some changes in the options.

We highlight that for these options to operate, the child needs to approve them via their browser add-on. This way, we ensure that the child gave their consent about what the IWP captures, analyzes, filters, and blocks. At the same time, this functionality ensures that the child knows exactly what notifications their custodian will be receiving about the

 ISBN: 978-1-61208-818-1

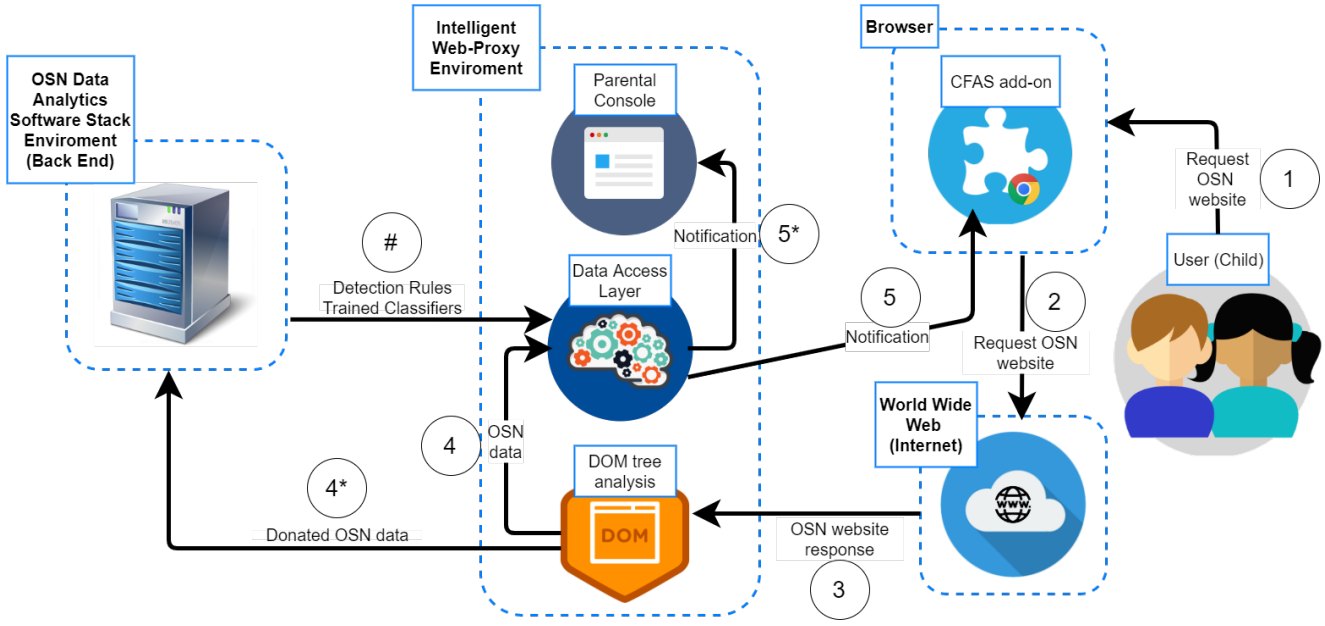

Figure 1. Cybersafety Family Advice Suite Architecture

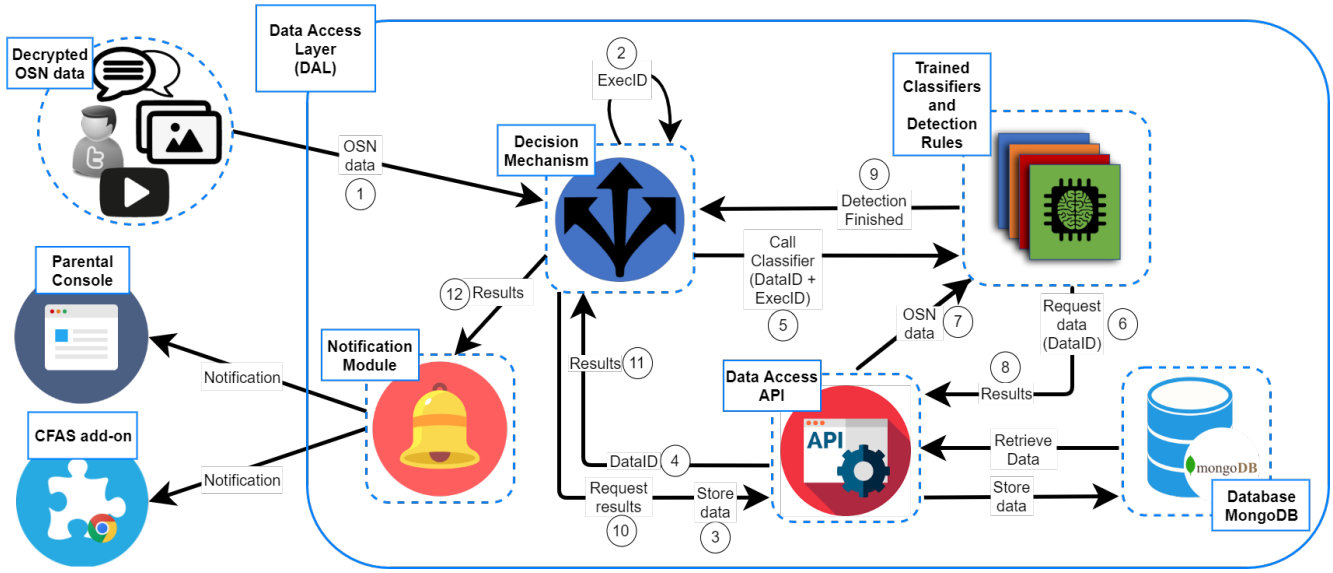

Figure 2. Data Access Layer (DAL) processes. DAL is the main storage unit of the IWP and the Back-End of the CFAS infrastructure.

online activity of the child, and what OSN traffic activity the custodian can see. We note that our proposed architecture promotes a conversation and close communication between the custodian and the child. This way, the family protected by CFAS can agree on what online activity of the child the custodians need to monitor, and what are the main risks and threats involved in using OSNs. Moreover, this architecture promotes OSN threat awareness, hence enforcing a culture of safe OSN usage. To achieve this, we introduce specific Parental and Back-End visibility options and Cybersafety options.

1) Parental Visibility Options: These options define what the custodian of the user can see, while enabling various levels of monitoring for the custodians, always with the explicit consent of the user. We define three Visibility Levels:

- Level 1: This is the lowest level of parental visibility, meaning that the custodian cannot see any data regarding the OSN traffic of the user. We note that the custodian still receives notifications regarding the threats detected by the trained classifiers hosted in the IWP, without mentioning the name of the perpetrator or revealing any OSN data. For the sake of the following examples, we assume that the protected child's name is *John*: "John might be a victim of cyberbullying."
- Level 2: This level of visibility allows the custodian to select some of the following OSN activity of the child to be visible to them: suspicious Twitter usernames the child

visited, disturbing YouTube videos the child watched, Facebook wall, photos, and friends of the child. Once the user gives their consent via their browser add-on for this data to be visible to the custodian, the visibility option is operational. A notification example: "John might be a victim of cyberbullying by Eve", where John is the protected child, and Eve is the perpetrator.

- Level 3: This is the default and highest level of parental visibility. When this option is selected, it adds all the options from Level 2, along with data regarding the user's Facebook chat. So at this level, the custodian of the child can see all the incoming and outgoing traffic of the child's Facebook wall, photos, notifications, friends, and chat, *only* in case of an incident. A notification example: "John might be a victim of cyberbullying by Eve. Click here to see the suspicious chat". This way, the custodian can see *portions* of the chat between the user and the perpetrator that show signs of cyberbullying.

We note that these options expire once every six months, so the custodian and the child can reset them as they wish. All the above levels of visibility can be set up after a mutual agreement between the custodian and the user while keeping the user fully aware of what their custodian can see.

2) Back-End Visibility options: Through the Back-End Visibility options, the Cybersafety Family Advice Suite offers options regarding which OSN traffic data is sent to the Back-End. OSN data sent to the Back-End are used to retrain the machine learning algorithms and detection rules hosted there to make them more accurate in future predictions. The custodian can choose among the child's Facebook wall, photos, notifications, friends, and chat. We note that the user needs to give their consent for the data to be sent to the Back-End. We define the following Back-End Visibility Levels:

- Level 1: This is the lowest level of Back-End visibility. If this option is set, no data is sent to the Back-End.
- Level 2: In this level, the custodian allows the IWP to send data to the Back-End regarding the child's Facebook wall, friend's Facebook wall, and the child's Facebook friends profiles. The custodian may select one or all of the above. Also, these data may be sent anonymized or not.
- Level 3: This is the highest level of Back-End visibility. When this option is set, it allows the IWP to send all the data from level 2, in addition to the child's Facebook chats. Once again, these data may be sent anonymized or not, and always with the consent of both the custodian and the child.

3) Cybersafety Options: Last, the Parental Console allows the custodian to choose the child's level of Cybersafety. These options define how aggressive the IWP can be, regarding the protection of the user: what the IWP can filter, protect, block, replace, encrypt, or watermark. This options can be configured at two different levels:

- Level 1: This is the lowest level of cybersafety. If set, the IWP only pushes notifications to the user explaining that certain suspicious or malicious activity is detected. This means that the IWP still detects suspicious activity, but it does not hide, protect, encrypt, blocks, or watermarks any content. Via the Parental Console, the custodian can choose the notifications they wish for the child to receive for each detection mechanism. The detection mechanisms include: a) cyber grooming; b) hate or inappropriate speech (cyberbullying); c) distressed behavior (when the child is suicidal, scared, depressed); d) fake activity (fake OSN profiles); e) personal information exposure (when the child is about to publish personal information); f) hateful memes; g) inappropriate YouTube videos; and h) sensitive content in pictures (when the child is about to share a benign picture that includes nudity without protection, like a picture in a swimsuit).
- Level 2: At this level, the custodian may choose any of the above IWP detection mechanisms to take action and filter, replace, protect, encrypt, or block content before it reaches the browser of the protected child. The detection mechanisms remain the same as level 1, but the custodian needs to select at least one to be operational for this level to hold.

Overall, the IWP is responsible for capturing the incoming and outgoing traffic of Facebook, Twitter, and YouTube of the user and send it to the locally hosted trained classifiers to detect malicious activity. In case the suspicious activity is detected by one or more trained classifiers, the IWP pushes a notification to the browser add-on of the user to inform them about the imminent threat detected. At the same time, the suspicious malicious content is blocked or filtered by the browser add-on to protect the minor, given that the Cybersafety Option Level 2 is set by the custodian and the user. The IWP hosts trained classifiers and detection rules to perform the following actions:

1) detect nudity in images included in the captured traffic;
2) encrypt sensitive images with steganography;
3) detect and warn the minor in case they are about to share personal information;
4) detect cyberbullying in Facebook conversations;
5) detect sexual grooming in Facebook conversations;
6) detect hateful and racist memes in Facebook feed;
7) detect bot, aggressive, bully, and spam Twitter users;
8) detect inappropriate videos for children on YouTube;
9) provide sentiment analysis of the chat of the minor;
10) generate informative notifications to the minor;
11) push notifications to the custodian about an incidence (e.g., sexual grooming);
12) push notifications to the child via the browser add-on;
13) submit data to the Back-End through a secure tunnel; and
14) block adult, or any other site, defined by the custodian.

C. Browser add-on

The last component of our architecture is the browser add-on (CFAS add-on in Figure 1). The browser add-on is the gateway between the IWP and the user, responsible to inform the user about the threats detected from the IWP, and the Visibility and Cybersafety options set by their custodian.

Importantly, our browser add-on operates as a Guardian Avatar that the child may interact with to ask for advice. Our avatar operates as the *guardian angel* of the user while using different OSN platforms (Facebook, Twitter, and YouTube only, currently). By following the Guardian Avatar approach as a gamification feature [7], CFAS aims to encourage the users to use it and interact with it because of its extended usability and improved user experience functionalities.

In addition, the user can select their favorite avatar icon from a list of icons. The Guardian Avatar "follows" the user

in their online-activities as a virtual friend. When the IWP detects any malicious behavior or incidents, the notifications (warnings, advice, etc.) appear as chat bubbles of the avatar, in a friendly and encouraging text. An example of the avatar notifying the minor about a detected incident is depicted in Figure 3. With the addition of the avatar, it is expected that the CFAS warnings and advice will be less disturbing for children (especially for the adolescents) and will make users more willing to use it.

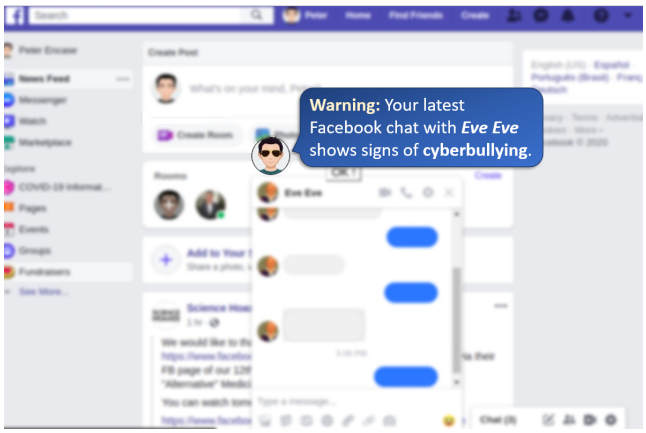

Figure 3. Guardian Avatar notifies the minor of any detected incidents

The browser add-on can:

1) notify the user about the activity detected by the IWP;
2) notify the user about what their custodian can see based on the preferences (Parental Visibility options) applied;
3) notify the user about what data is sent to the Back-End to aid the machine learning classifiers to become more accurate (Back-End Visibility options);
4) let the user change the options about what OSN traffic activity their custodian can see;
5) let the user change the options about what data is sent to the Back-End;
6) let the user flag content/text as cyberbullying activity, sexual cyber grooming activity, aggressive behavior activity, fake identity activity, and false information activity in case the IWP failed to detect so;
7) let the user flag sensitive or nudity content in case the IWP failed to detect so; and
8) let the user flag content/text as an incorrect sensitive content, cyberbullying, sexual grooming, aggressive behavior, fake identity, and false information activity in case the IWP detected so.

Overall, we propose a fully privacy-preserving architecture for the protection of minors when they use OSNs, both towards their custodians and towards the system itself. First, the minor is empowered to choose the online activity and warnings that their custodian receives in case a threat is detected by the IWP. This can be done via the Parental Visibility Options. Second, the user can choose which online activity the IWP filters, captures, and protects via the Cybersafety Options. Also, the IWP, the device that is responsible for capturing and analyzing the online activity of the minor to detect online threats, is connected and physically exists within the network of the user. Thus, the online activity of the minor is captured and analyzed locally and is isolated within the network of the user. In addition, the IWP never makes any data visible to the rest of the system (Back-End or other IWPs) without the explicit authorization and consent of both the user and their custodian via the Back-End Visibility Options.

III. DESIGN

We now detail the design of the proposed architecture. Instead of simple rule-based filters, our architecture utilizes advanced machine learning algorithms. The downside of having rule-based filters is that they are blunt. There are situations where there is a particular piece of content that technically does not violate the specified policies, but when this content is analyzed with advanced machine learning techniques, it might turn out to be hate speech, sarcasm, sexual grooming, etc. Such techniques allow us to detect bullies or predators that are close to the line. To sum up, the aim is to have these granular standards so that our design can control for bias. Our design approach is based on the following design principles:

1) We place all functionalities (filters, text replacement, notifications, data submission to the Back-End, etc.) in the IWP instead of the browser add-on when it can be correctly and efficiently implemented. This way, we prevent a minor from modifying or disabling the system's functionality through the browser add-on. For example, in case a minor accidentally or willingly disables the browser add-on, the IWP does not get affected, and all the processes and functionalities can continue their operation normally. We assume that the device of the minor is still configured to route social network services through the IWP and that the child does not have the permission, knowledge, or access to alter the configuration of the IWP or their personal device. Also, the IWP can notify the custodian through the Parental Console that the browser add-on of the minor is not responding anymore.

This architecture aims to provide the ability to seamlessly support multiple types of clients (desktop browsers, mobile apps, etc.) with a minimal client or client platform configurations or modifications. Moreover, the browser add-on does not support complex functionalities other than javascript and HTML scripts. For example, functionalities, like text replacement, picture encryption, filtering, etc., are too complex to be implemented and run on a browser add-on.

In case the IWP is down, the browser add-on calls REST API requests from the Back-End, and the Back-End DAL is employed to identify suspicious content. This means that the OSN traffic activity of the user is sent outside of the network, to the Back-End, for analysis. Whether a suspicious activity is detected by the Back-End or not, all the user OSN traffic data is automatically deleted from the Back-End. Having some functionalities on the IWP prevents it from calling REST API requests from the Back-End every time it needs to analyze OSN traffic activity. In addition, placing some functionalities on the IWP, solves the potential problem of the whole system being down in case of Back-End unavailability, thus solving the problem of single-point failure. Examples: i) The IWP can push notification to the browser add-on without the need of the Back-End. ii) Before any content reaches the minor's device, the IWP can replace cyberbullying content without calling REST API requests from the Back-End, using the functionality installed on it already.

2) Rules and trained classifiers are generated in the Back-End. Trained classifiers are placed in the IWP only if they can

 ISBN: 978-1-61208-818-1

run efficiently. The Back-End collects data from all the IWPs to generate detection rules and trained classifiers. Data collected from the IWPs are used to generate cyberbullying, sexual cyber grooming, distressed behavior, aggressive behavior, fake identity, and false information detection rules.

3) Warning, flagging, and feedback functionality is placed on the browser add-on. The Guardian Avatar displays notifications in dialogue boxes after the IWP detects suspicious behavior and pushes a notification to the browser add-on. The user can flag content as cyberbullying activity, sexual cyber grooming activity, aggressive behavior activity, fake identity, false information, and sensitive picture through the browser add-on in case the IWP failed to detect so. The user can also give feedback based on the activity detected by the IWP. For example, in case the IWP detects cyberbullying, it pushes a notification to the browser add-on. The Guardian Avatar shows the notification/warning to the user explaining that cyberbullying was detected (Figure 3). Then, the user can provide feedback on whether this detection is accurate or not.

4) The minor can check the content their custodian, the IWP, and the Back-End can see. The custodian can set up the Visibility settings in a fine-grained way and always with the consent of the minor. This way, we enable various levels of monitoring for parents and the Back-End with the child's consent, while keeping the child fully aware of what their custodians and the Back-End can see, e.g., chat messages.

Overall, we propose a system that eases the tension of ensuring the safety of minors while respecting their privacy with respect to what their custodians and third parties can see. By automating the detection of malicious communication, we enable custodians to be continuously aware of their child's safety. This is achieved without the parent having to go through the minor's online communication manually, thus, without having to invade the minor's privacy. Our approach aims to warn the custodians about the suspicious online activity that was detected, without violating the privacy of the minor. For example, if the minor has a Facebook online conversation with sexual content with somebody, the custodian of the minor will receive a warning that such a conversation is taking place, once the IWP captures it. Still, the parent won't be able to see the actual content because that would violate the teenager's privacy. Instead, the parent can only see the actual conversation through their Parental Console once the explicit consent of the child has been granted. To sum up, our design principles intend to encourage custodians to have a conversation with the minor; thus, bringing families closer and spreading awareness about the numerous threats that exist in contemporary OSNs.

IV. IMPLEMENTATION

We implement all the architecture components, and integration's that we describe in Sections II and III. In this section, we provide the details of the prototype implementation. Note that we employ classifiers created in previous work for the detection of threats in OSNs. We note that these classifiers are generated on the Back-End and hosted on the IWP. In case the classifiers detect suspicious activity, the IWP pushes notifications to the browser add-on of the user, and the Parental Console.

A. Detection of Abusive Users on Twitter

When the minor visits a Twitter user account, the IWP captures the username of the visited user, and it calls the Twitter API to collect the last 20 tweets (including retweets) of that user [8]. This information is then sent to a classifier developed by Chatzakou et al. [9] for analysis. The developed classifier is trained with Twitter annotated data [10] [11] and analyzes the last 20 tweets of the visited Twitter user to detect whether it is an aggressive, bully, spam, or normal account.

B. Fake and Bot user detection on Twitter

When the minor visits an account on Twitter, the IWP captures the username of the Twitter account and sends it for analysis via a REST API call developed by [17] and Echeverria et al. [18]. This API returns True if the Twitter user account is a bot, and False otherwise. In case of the former, the IWP pushes a notification to the browser add-on of the minor, and to the Parental Console of the custodian (based on the Parental Visibility options).

C. Detection of Hateful and Racist memes on Facebook

The IWP captures the Facebook incoming and outgoing traffic of the minor and performs TLS termination of the DOM tree. All the images that are extracted from the DOM tree are sent to the classifier developed by Zannettou et al. [12] to be labeled as a hateful meme or not. This classifier is trained using images from Twitter, Reddit, 4chan's Politically Incorrect board [13], and Gab [14]. In case the detection is positive, the picture will be automatically replaced by the IWP with a static image to inform the minor.

Similarly, when the minor uploads an image on Facebook, the picture is analyzed by the aforementioned classifier to detect whether that image is hateful or racist. If so, then the IWP pushes a notification to the guardian avatar to advise the minor that the image they try to upload contains hateful content, and they shouldn't upload it.

D. Sexual Predator Detection on Facebook

When the minor is chatting with a friend on Facebook, the conversation is captured by the IWP and is sent to the classifier developed by Partaourides et al. [15] for analysis. A previous version of this classifier was trained with data from Perverted Justice website [16] to recognize patterns similar to the ones from convicted sexual predators. Upon positive detection, the IWP pushes a notification to the browser add-on of the minor, notifying them that signs of sexual predator have been detected. The custodian can see only portions of the chat between the minor and the predator via the Parental Console, only if the minor consents so via the Parental Visibility options explained in Section II. We note that the custodian can only see portions of the chat that the classifier detects as a sexual grooming pattern.

E. Cyberbullying Detection on Facebook

Similar to the Sexual Predator detection, when the minor is chatting with a friend on Facebook, the conversation is captured by the IWP and is sent to the classifier developed by Partaourides et al. [15] for analysis. This classifier returns percentages of how angry, frustrated, and sad the minor is during the Facebook chat conversation, using sentiment analysis. If any of these three feelings exceed 65%, the IWP pushes a

 ISBN: 978-1-61208-818-1

notification to the browser add-on of the child to warn them that the Facebook chat they are having seems to be toxic for them. Similar to the sexual predator detection above, the custodian is only able to see portions of the suspicious chat, only if the minor gave their consent beforehand.

F. Personal Information Leakage Detection on Facebook

When the user tries to make a post on Facebook, the IWP captures the text written by the user and analyzes it to detect dates, times, phone numbers with or without extensions, links, emails, IP and IPv6 addresses, prices, credit card numbers, street addresses, and zip codes. We implement this detection technique using existing Python libraries [19]. In case any of the above personal information is detected, the IWP pushes a warning to the minor to remove the sensitive information from their post. In case the minor dismiss these warnings, a notification is sent to the Parental Console of the custodian (in accordance with the Parental Visibility options).

G. Watermarking and Steganography

For the purposes of this detection mechanism, we consider any image that includes nudity (topless images of boys, or swimsuit images) as sensitive content images. When the minor tries to send a sensitive image to a friend over Facebook chat, the image first passes in the IWP for analysis. We followed similar techniques to Ghazali et al. [20] and Kolkur et al. [21] to develop our skin and nudity detection techniques. In case the image contains sensitive content, the IWP watermarks it [22]. Then, the IWP hides the original image in another static image using steganography. This way, only the person that the picture was sent to is allowed to see the hidden original image. We note that for this to work, the receiver needs to be part of the Cybersafety Family Advice Suite network as decryption keys hosted on the Back-End are requested from the IWP to decrypt the image. Similarly, if the minor tries to post an image that contains sensitive content on their Facebook wall, the IWP watermarks and performs steganography techniques to the image before posting it on Facebook. The minor, using the browser add-on, can set who is able to see (decrypt) this picture (family members, friends, classmates, etc.). For this scenario, we assume that the minor allows the image to be visible to family members only, and that their family members are registered CFAS members and have their own IWP set up at home. When a family member of the minor scrolls Facebook, their IWP captures that image and communicates with the CFAS Back-End to check if they have permission to see this image. If this is the case, then the IWP decrypts the image automatically. In case the image does not contain sensitive content, the IWP only applies watermarking on it before posting it. The receivers that are not part of the CFAS network can only see the static encrypted image.

H. Disturbing videos on YouTube

Our architecture also detects disturbing YouTube videos for young children, using the developed classifier by Papadamou et al. [23]. This classifier was trained using YouTube videos [24] and can discern inappropriate content with 84.3% accuracy. When a minor visits a YouTube video, the IWP captures the YouTube link, which includes the YouTube video ID, and it calls the YouTube API to collect the video features [25]. These features include the video upload date, likes, tag, title, thumbnail, etc. The IWP then sends these video features to the developed classifier for analysis. In case the classifier returns positive detection (inappropriate), then it warns the minor that the video they are watching is not suitable for them via the browser add-on.

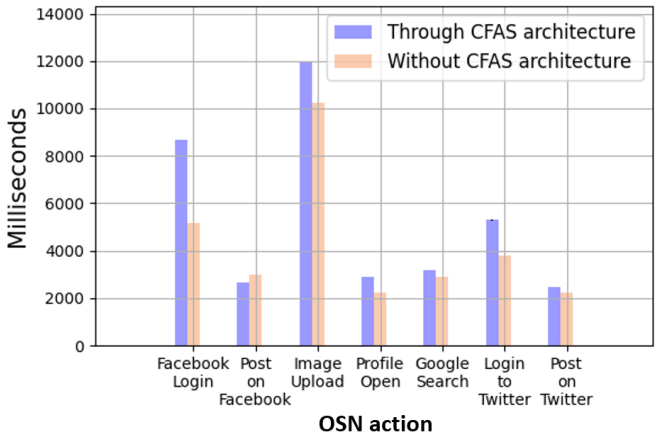

Figure 4. OSN actions with CFAS & without CFAS

V. Evaluation

In this section, we evaluate the performance of the prototype implementation of the Cybersafety Family Advice Suite.

A. Performance Evaluation

To test the performance in regard to the number of concurrent users, we set a small home cluster using a laptop with 4GB Ram, a quad-core Intel Core i5 processor that is running Ubuntu 18.04 64bit and Google Chrome Version 80.0.3987.162 (64 Bit), which is used as the minor's laptop that hosts the browser add-on. In addition, we set up two virtual machines with 2GB RAM each, and one tablet of 3GB RAM: 4 users in total. The IWP is a virtual machine hosted on the Google cloud, configured with 4GB RAM, a dual-core Intel Xeon CPU, running Centos 7 (64 Bit), and it is using the mitmproxy [26]: the HTTPS proxy. Also, the IWP hosts a MongoDB for Data storage and Python3 for the API Calls. We run the experiments with a downlink of $\sim$20 Mbps and an uplink of $\sim$5 Mbps.

Figure 4 depicts the time in milliseconds needed for OSN actions to be executed with and without CFAS. Each machine executes the OSN actions using a JavaScript automated method in a serial manner. Then, we calculate the average time that each machine needed to finish each action using the start time and end time of each action. We observe that with CFAS, there are reasonable delays regarding the execution of some actions (e.g., Facebook Login, Image Upload, Twitter Login). This delay is acceptable since extra processing is needed to load and execute the CFAS tools. Other actions' delay is negligent ($\sim$1 second).

B. User Experience

In this section, we present the results of a user experience evaluation questionnaire given to minors and custodians after interacting with CFAS. The participation of minors required their custodians' consent. The sample consists of 30 minors and 12 custodians that had no knowledge or experience of the CFAS tools. The questionnaires were GDPR-compliant and

 ISBN: 978-1-61208-818-1

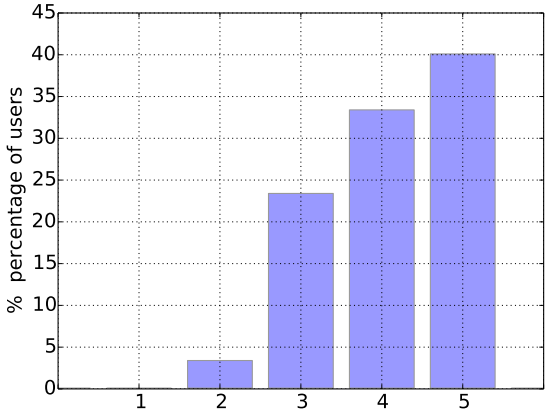

Figure 5. (Minors) Would you allow CFAS to send notifications to your custodian regarding suspicious detection? (1: Totally Disagree, 5: Totally Agree)

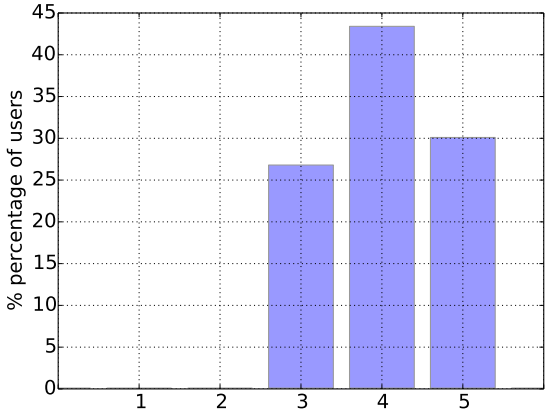

Figure 6. (Minors) Do you believe CFAS would improve your safety when using OSNs? (1: Totally Disagree, 5: Totally Agree)

anonymous. The study has received data protection approvals by the Ethics Committee of the Cyprus University of Technology, and by the Office of the Commissioner for Personal Data Protection of the Republic of Cyprus.

To evaluate our tools, the minors had to answer a variety of questions regarding their usability, accessibility, and performance. The minors were between 12 to 16 years old and reported using the Internet daily for entertainment and education purposes. The percentages of minors in our sample that have a registered Facebook, Instagram, and YouTube account are 53.3%, 33.3%, and 13.3%, respectively.

We report some of the results we obtained from the questionnaires given to minors and their custodians after they used the CFAS tools. When minors asked whether they would allow CFAS to send notifications to their custodians, the majority reported high, and complete agreement (Figure 5). In addition, the majority of minors believe that these tools could improve their safety when using OSNs, as depicted in Figure 6. Importantly, all of the minors report being very happy with the capabilities of CFAS (Figure 7). Alarmingly, Figure 8 depicts that many minors had their personal data (24%) and photos (7%) misused, being a victim of cyberbullying (7%), and witnessing inappropriate speech and racism (37%) on social networks. Note that the minors could select any that applied to them for this question.

On the other hand, the overwhelming majority of the custodians report that their child never complained of being a victim or a spectator of such threats online (Figure 9). Although this is a small number of participants, it depicts that it is usually the case that minors don't report the threats they face on OSNs to their custodians. Last, all of the custodians agree that CFAS could improve the safety of minors online (Figure 10), and the overwhelming majority of custodians report that they would install CFAS at home (Figure 11).

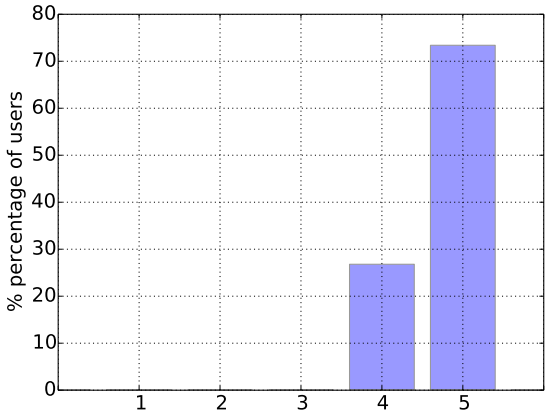

Figure 7. (Minors) Are you satisfied with CFAS capabilities? (1: Totally Disagree, 5: Totally Agree)

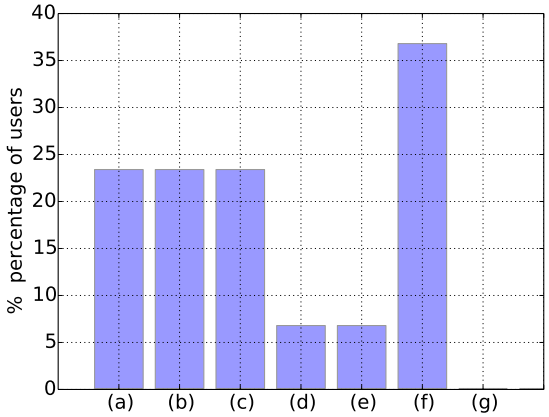

Figure 8. (Minors) Have you ever experienced the following online-threats? Select all that apply to you: (a) I prefer not to say; (b) None; (c) Personal data misused; (d) Personal photo misused; (e) Cyberbullying; (f) Inappropriate speech and racism; and (g) Sexual grooming

VI. Related Work

This section reviews some web-based and mobile applications that try to protect adolescents on the Internet and OSNs. We list the ones most relevant to the concepts of CFAS.

Qustodio is a parental control software [27] that enables parents to monitor and manage their kids' web and offline activity on their devices. It also tracks with whom the child communicates on various OSNs and can be used as sensitive content detection and protection tool (using filters). Last, it monitors messages, calls, and the location of the minor's device. Kidlogger allows custodians to monitor what their children are doing on their computer or smartphone [28]. It performs keystroke logging, keeps a schedule of which websites the minors visit and what applications they use, and with whom they are communicating on Facebook. Also,

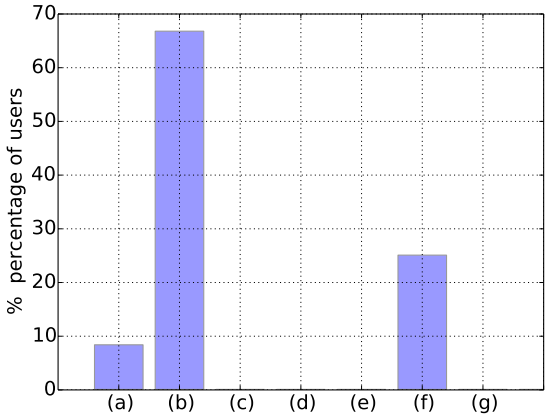

Figure 9. (Custodians) Has your child ever reported to you being a victim of the following? (a) I prefer not to say; (b) None; (c) Personal data misused; (d) Personal photo misused; (e) Cyberbullying; (f) Inappropriate speech and racism; and (g) Sexual grooming

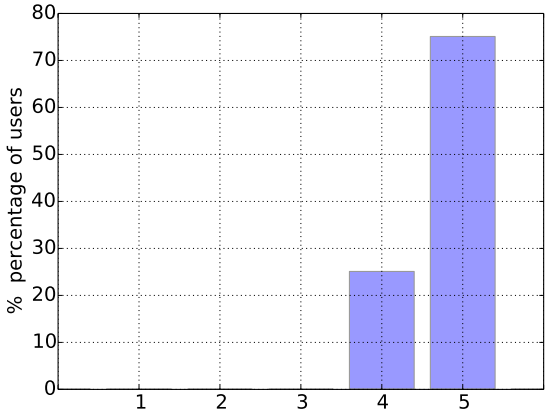

Figure 10. (Custodians) Do you think that CFAS would improve the safety of minors when using OSNs? (1: Totally Disagree, 5: Totally Agree)

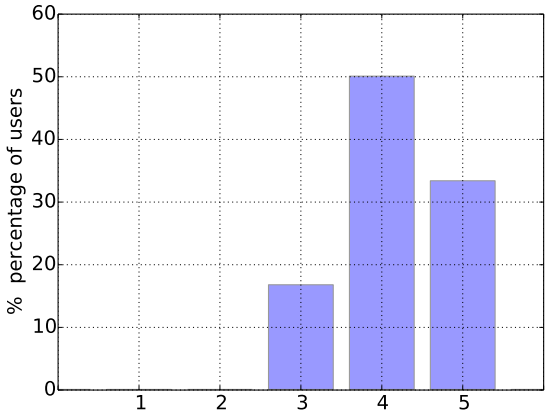

Figure 11. (Custodians) Would you install CFAS at home? (1: Totally Disagree, 5: Totally Agree)

Kidlogger offers sound recording of phone and online calls, smartphone location tracking, and photo capture monitoring. Web of Trust (WoT) is a browser add-on and smartphone application for website reputation rating that warns users about whether to trust a website or not [29].

Mspy is a smartphone application that monitors almost all the applications and activities on the smartphone of the minor [30]. Alarmingly, the application may be installed on the smartphone of the minor by the custodian and remain hidden, so the minor cannot know they are being monitored. Syfer [31] is a device, still in production, that can be plugged into the router of the house network and analyses the traffic activity for possible threats. It protects against cyber threats in realtime, stops invasive data collection, offers a VPN, has artificial intelligence for enhanced security, and blocks advertisements. It doesn't log any information, and it offers encrypted activity. It restricts inappropriate content with real-time website analysis provided by their AI engine. Bark [32] monitors text messages, YouTube, emails, and 24 different social networks for potential safety concerns. Bark looks for activity that may indicate online predators, adult content, cyberbullying, drug use, suicidal thoughts, and more. In case anything suspicious is detected, the custodians receive automatic alerts along with expert recommendations from child psychologists for addressing the issue. They offer an application for iOS, Android, Kindle, browser add-ons for Google chrome on PC and Safari on Mac, and Kindle. The user has to allow the Bark application to send all the traffic data to Bark's Back-End for analysis and detection.

The majority of the existing applications follows a more traditional approach (monitoring, restrictions over online activities). Most applications consider parents or custodians as the end-users, instead of the children [33] [34]. Many of the applications do not have interfaces for children but are just installed as services running in the background [35]. A new notion suggests designing and developing tools and software that is more "children-aware" and "children-friendly". Online safety applications should consider the child as the major user and try to enrich children's self-regulation and their risk coping skills in cases of online dangers [36]. By enforcing this child-friendly approach, we achieve a collaboration where parents and children need to communicate and discuss online risks and behavior in contrast with the approach of restriction and monitoring. We aim to teach children how to cope with online threats and use social media with responsibility and self-awareness. CFAS follows this approach by involving the child in the process of setting the filters, and parental and Back-End visibility options. In addition, the cybersafety tools require the child's consent to be activated. Last, we note that this work is a follow up of the work presented by Papasavva [37].

VII. Conclusion

In this paper, we present the architecture of a user-centric privacy-preserving advanced family advice suite for the protection of minors on OSNs. The architecture comprises three main components, namely, the Data Analytics Software Stack, the Intelligent Web-Proxy, and a browser add-on, which operates as a guardian angel of the child while using OSNs. This architecture aims to protect minors when using OSNs while preserving their privacy. We propose Guardian Avatars that interact with, warn, and advise adolescences when they face threats on OSNs. Also, the custodian of the adolescent receives notifications on their Parental Console in case a malicious activity is detected by the classifiers hosted on the IWP to be aware of the threats their child was exposed to. Importantly, the custodian can only see the relevant content, which indicated to be suspicious, only if the minor had previously given their explicit consent.

Blocking content from the minors or thoroughly monitoring

 ISBN: 978-1-61208-818-1

their every online-move should not be the solution as it violates the privacy of the adolescents. The proposed architecture advertises the collaboration between parents and children and aims at bringing the family to work together to protect the vulnerable groups of the Internet while using OSNs.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie ENCASE project (Grant Agreement No. 691025), and the CYberSafety II project (Grant Agreement No. 1614254). This work reflects only the authors' views.

REFERENCES

[1] M. Anderson and J. Jiang, "Teens, Social Media & Technology 2018," *Pew Research Center: Internet, Science & Tech*, vol. 31, 2018.

[2] "Children and parents media use and attitudes: annex 1." 2019, URL: https://bit.ly/2JIshIk [accessed: 2020-08-25].

[3] "Online Abuse - How safe are our children?" 2019, URL: https://bit.ly/390zOhO [accessed: 2020-08-25].

[4] "Pew Research Center. A Majority of Teens Have Experienced Some Form of Cyberbullying." 2018, URL: https://pewrsr.ch/32o2AHY [accessed: 2020-08-25].

[5] "EU Kids Online II Dataset: A cross-national study of children's use of the Internet and its associated opportunities and risks," 2017, URL: https://ab.co/30dr3NB [accessed: 2020-08-26].

[6] T. Andreas, T. Nicolas, S. Makis, P. Kwstantinos, and S. Michael, "Cyber Security Risks for Minors: A Taxonomy and a Software Architecture," in *11^{th} International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)) July 12–14, 2013, Thessaloniki, Greece*. SMAP, Nov. 2016, pp. 93 – 99, ISBN: 978-1-5090-5246-2, URL: https://ieeexplore.ieee.org/abstract/document/7753391 [accessed: 2020-08-26].

[7] S. Deterding, M. Sicart, L. Nacke, K. O'Hara, and D. Dixon, "Gamification: Using game design elements in non-gaming contexts," *ACM CHI*, vol. 125, pp. 2425–2428, 2011, ISBN: 9781450302685.

[8] "Twitter API," 2020, URL: https://developer.twitter.com/en/docs [accessed: 2020-08-25].

[9] D. Chatzakou *et al.*, "Mean Birds: Detecting Aggression And Bullying On Twitter," in *Proceedings of the 2017 ACM on Web Science Conference (WebSci) June, 2017, New York, NY, United States*. ACM, Jun. 2017, pp. 13–22, ISBN: 9781450348966,URL: https://dl.acm.org/doi/pdf/10.1145/3091478.3091487 [accessed: 2020-08-26].

[10] "Restricted Dataset for "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior"," 2020, URL: https://zenodo.org/record/3706866 [accessed: 2020-08-25].

[11] "Dataset for "Mean Birds: Detecting Aggression and Bullying on Twitter"," 2018, URL: https://zenodo.org/record/1184178 [accessed: 2020-08-25].

[12] S. Zannettou *et al.*, "On The Origins Of Memes By Means Of Fringe Web Communities," in *Proceedings of the Internet Measurement Conference 2018 (IMC) October, 2018, New York, NY, United States*. ACM IMC, Oct. 2018, pp. 188—-202, ISBN: 9781450356190, URL: https://dl.acm.org/doi/pdf/10.1145/3278532.3278550 [accessed: 2020-08-26].

[13] A. Papasavva, S. Zannettou, E. De Cristofaro, G. Stringhini, and J. Blackburn, "Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 8-11 June, 2020, Atlanta, Georgia, US*, vol. 14, Jun. 2020, pp. 885–894, URL: https://www.aaai.org/ojs/index.php/ICWSM/article/view/7354 [accessed: 2020-08-26].

[14] "Dataset for "On the Origins of Memes by Means of Fringe Web Communities"," 2018, URL: https://zenodo.org/record/3699670 [accessed: 2020-08-25].

[15] H. Partaourides, K. Papadamou, N. Kourtellis, I. Leontiades, and S. Chatzis, "A Self-Attentive Emotion Recognition Network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 4-8 May, 2020, Barcelona, Spain*. IEEE, May 2020, pp. 7199–7203, ISBN: 978-1-5090-6631-5, ISSN: 2379-190X, URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9054762 [accessed: 2020-08-26].

[16] "Perverted Justice Data," 2019, URL: http://www.perverted-justice.com/ [accessed: 2020-08-25].

[17] "Astroscreen," 2019, URL: https://www.astroscreen.com/ [accessed: 2020-08-25].

[18] J. Echeverria *et al.*, "LOBO: Evaluation Of Generalization Deficiencies In Twitter Bot Classifiers," in *Proceedings of the 34^{th} Annual Computer Security Applications Conference (ACSAC) December, 2018, New York, NY, United States*. ACM, Dec. 2018, pp. 137–146, ISBN: 9781450365697, URL: https://dl.acm.org/doi/pdf/10.1145/3278532.3278550 [accessed: 2020-08-26].

[19] "GitHub - madisonmay/CommonRegex: A collection of common regular expressions bundled with an easy to use interface," 2019, URL: https://bit.ly/2Zu4gh8 [accessed: 2020-08-26].

[20] G. Osman, M. S. Hitam, and M. N. Ismail, "Enhanced skin colour classifier using RGB ratio model," *arXiv*, 2012.

[21] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia, "Human skin detection using RGB, HSV and YCbCr color models," *arXiv*, 2017.

[22] "Watermark with PIL," 2005, URL: http://code.activestate.com/recipes/362879/ [accessed: 2020-08-25].

[23] K. Papadamou *et al.*, "Disturbed YouTube For Kids: Characterizing And Detecting Inappropriate Videos Targeting Young Children," in *Proceedings of the International AAAI Conference on Web and Social Media 26 May, 2020, Palo Alto, California USA*. AAAI, May 2020, pp. 522–533, ISBN: 978-1-57735-823-7, ISSN: 2334-0770, URL: https://www.aaai.org/ojs/index.php/ICWSM/article/view/7320 [accessed: 2020-08-26].

[24] "Dataset: "Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children"," 2020, URL: https://zenodo.org/record/3632781 [accessed: 2020-08-25].

[25] "YouTube API," 2020, URL: https://developers.google.com/youtube/v3 [accessed: 2020-08-25].

[26] "mitmproxy HTTPS proxy," 2020, URL: https://mitmproxy.org [accessed: 2020-08-25].

[27] "Qustodio," 2020, URL: https://www.qustodio.com/en/ [accessed: 2020-08-26].

[28] "KidLogger parental control," 2016, URL: http://kidlogger.net [accessed: 2020-08-26].

[29] "Web of Trust," 2020, URL: https://www.mywot.com [accessed: 2020-08-26].

[30] "mSpy," 2020, URL: http://www.spyrix.com/android-monitor.php [accessed: 2020-08-26].

[31] "SYFER Complete Cybersecurity," 2020, URL: https://mysyfer.com [accessed: 2020-08-26].

[32] "Bark," 2020, URL: https://www.bark.us [accessed: 2020-08-26].

[33] K. Badillo-Urquiola *et al.*, ""Stranger Danger!" Social media app features co-designed with children to keep them safe online," in *Proceedings of the 18^{th} ACM International Conference on Interaction Design and Children (IDC)) July 19, 2019, New York, NY, United States*. ACM, Jun. 2019, p. 394–406, ISBN: 9781450366908, URL: https://dl.acm.org/doi/pdf/10.1145/3311927.3323133 [accessed: 2020-08-26].

[34] B. McNally *et al.*, "Co-designing mobile online safety applications with children," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems(CHI)) July 19, 2018, New York, NY, United States*. ACM, Apr. 2018, p. 523, ISBN: 9781450356206, URL: https://dl.acm.org/doi/pdf/10.1145/3173574.3174097 [accessed: 2020-08-26].

[35] P. Wisniewski, A. K. Ghosh, H. Xu, M. B. Rosson, and J. M. Carroll, "Parental control vs. teen self-regulation: Is there a middle ground for mobile online safety?" in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)) February 25, 2017, New York, NY, United States*. ACM, Feb. 2017, pp. 51—-69, ISBN: 9781450343350, URL: https://dl.acm.org/doi/pdf/10.1145/2998181.2998352 [accessed: 2020-08-26].

[36] A. K. Ghosh, C. E. Hughes, M. B. Wisniewski, Pamela J, and J. M. Carroll, "Circle of Trust: A New Approach to Mobile Online Safety for Families," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)) April 21, 2020, New York, NY, United States*. ACM, Apr. 2020, p. 1–14, ISBN: 9781450367080, URL: https://dl.acm.org/doi/pdf/10.1145/3313831.3376747 [accessed: 2020-08-26].

[37] A. S. Papasavva, "A Privacy-preserving Architecture for Parental Control Tools for the Protection of Minors on Online Social Networks," 2019.