AI-assisted Online Child Protection

Bassel Walid, STEM high school for boys – 6th of October

Saif Ahmed, STEM high school for boys – 6th of October

Abstract

No one can deny the internet's great effect on the progress that we see around us. Despite all the benefits of the internet, it has its dark side, when misused the internet represents a threat, especially for children. Lacking child internet safety is a hazardous problem that people usually ignore or simply underestimate. According to the Center for Cyber Safety and Education, 53% of kids in grades 4-8 revealed their number to a stranger, and about 11% have met a stranger[1]. These statistics are skyrocketing, which is scary. In this research paper, a prototype for an online child safety software that utilizes AI to identify obscene photos. The reasoning behind writing this research is to explore one of the ways that a child protection software can be created and to explore and empathize the importance of maintaining child privacy while protecting them. The software here will mainly focus on limiting the threats from these sources. The algorithm is designed to block all unwanted websites with any unfriendly or aggressive content that does not suit children. Unblocked websites are continuously analyzed, where users receive the results of these updates periodically. As for the videos, AI can recognize and analyze videos to ensure the safest online experience for the child. The chatrooms are the last approach we are taking into consideration in our project. We are scanning the chats to recognize any obscene or immoral content, when any prohibited content is detected the parents are alerted at once.

I. Introduction

In recent years children's usage of the internet has skyrocketed now that 80-98% of American children aged 3-18 have internet access in their homes. This presents parents with a problem that has been getting more dangerous over the years, child internet safety.

Child internet safety has been listed as the 4th "biggest problem" according to a survey done by the C.S Mott children hospital in Michigan, going up from the 8th most prominent problem in the year prior, and this makes sense, as more and more children get their hands-on electronic devices that can access the internet.

As a result of the high percentage of children with access to the internet exposure to sexual content has also risen. It was shown in two surveys conducted in the UK that the percentage of children exposed to sexual content has risen across the years as shown in [Table 1].

	Overall	Gender		Age (years)					
	Total	Boys	Girls	11	12	13	14	15	16
Study 1 (EU, 2010)									
Images or video of someone naked, percent	13.7	15.8	11.5	5.3	7.2	10.7	14.7	20.9	23.5
Images or video of someone's "private parts," percent	9.3	11.2	7.3	3.0	3.7	6.7	10.3	14.8	17.2
Images or video of people having sex, percent	10.1	13.6	6.7	3.1	4.6	6.9	11.2	16.2	19.1
Images or video of movies that show sex in a violent way, percent	2.9	3.8	2.1	1.0	1.2	1.9	3.6	4.5	5.4
Images or video of any of the above, percent	20.5	24.4	23.4	7.5	10.8	15.8	23.4	32.6	35.4
Study 2 (UK, 2018)									
Images or video of someone naked, percent	33.2	35.7	30.2	—	—	—	30.6	35.7	—
Images or video of someone's "private parts," percent	25.0	27.6	21.9	—	—	—	22.4	27.4	—
Images or video of people having sex, percent	20.8	23.5	17.8	—	—	—	20.2	21.4	—
Images or video of movies that show sex in a violent way, percent	9.9	11.3	8.4	—	—	—	8.1	11.6	—
Images or video of any of the above, percent	38.0	40.4	35.1	—	—	—	35.4	40.6	—

TABLE 1: THE PERCENTAGE OF CHILDREN EXPOSED TO SEXUAL CONTENT

This shows the need to both start educating parents on the dangers of unsupervised internet usage and to start creating and implementing online child protection solutions.

The main challenge that would face any developer trying to achieve such a goal is to maintain the privacy of the children while allowing the parents to easily supervise the children's usage. As a result, a low information-to-results ratio had to be achieved to make sure not too much of the child's information is accessed and still gets the required results. After that, the second challenge is maintaining the overall privacy of the information, something which would require the software to be either disrupted according to a Software as a service model or to be locally hosted, a combination of both approaches was selected for this prototype. The third problem would be the number of places that a child's safety can be compromised, this problem comes as a result of the many places a child can interact with strangers, from anonymous chatrooms to online game lobbies, this limits the places this A.I can help. The last problem is the many platforms that children use, this problem is not as big as the others, but it still increases the work required to integrate the A.I with different platforms like Android, IOS, and Windows**.**

II. Implementation methods

In the creation of the software, already existing ideas and code modules were used in combination with new ideas to maximize efficiency. The project is separated into three main parts, general internet content moderation, video caption analysis, and predatory behavior chat scanning.

III. Main programming languages and programs used.

In the project, Python was used as the primary programming language due to its versatility when it comes to A.I and algorithms, quick integration with other programming languages, and the wide range of already existing modules and programs. This allows for more development time to be dedicated to adding features and optimizing them instead of making custom versions of these programs (i.e. the image detection algorithm). In professionally developed software, it is of course preferred to use custom-made software though to allow for deeper integration and optimization.

IV. General Internet Content Moderation

This is the first and simplest part of this project, the moderation algorithm is very simple but quite effective.

On the first run, the program presents a list of other child safety measures that they can deploy to enhance their child's safety, for example, it recommends users to use a DNS that offers a family plan or a child safety blocklist (i.e. OpenDNS Family Shield, Cloudflare 1.1.1.1 For Families, etc.) which simplifies the work needed for the project because they already block a majority of adult content, the program also recommends users to make sure they are using Google SafeSearch to prevent explicit images from popping out in Google's image search.

Using a simple extension that grabs the URLs of the webpages that the child visits and sends it to the python extension so it can look to see if this website is part of a trusted website list that includes sites like Wikipedia, government websites, etc. which we are sure don't include adult content, if the website isn't a part of that list the program it'll search through the website's content to see if it has any of the blacklisted words that exist in a list defined by the child's parent and this list includes terms like references of pornography, swear words, etc., if it doesn't find any matches it deletes the entry and waits for another link, if it finds any of these words the program it fetches the URLs of all images in the website using the Beautiful Soup module for python and sends them to the explicit image detection algorithm to scan these photos to check if the website has any explicit photos, after it checks the photos on the website it saves all the results of this scan (all blacklist matches, explicit photo scan results) to a file sent with the alert that is sent to the parent when the website the scan is done.

The program also saves these URLs to an online open-source database that can be community checked to improve the program.

V. Video Caption Analysis

When the HTML parser finds that the domain name is YouTube it instead calls for the video caption analysis, the video caption analysis program first gets the unique video ID given to all videos uploaded to YouTube, it then uses the YouTube API to fetch the captions for the video which it then sends to Google's Natural Language Processing API and it calls its content classification to classify the themes in the captions to figure out whether the video contains any age-inappropriate subjects (i.e. gambling), after it receives the themes from the API it saves them in a text file with the video title and ID.

This is not an arguably bad way to analyze YouTube videos, but it has a flaw that is being slowly fixed. auto-generated captions, while a very useful feature to have, it makes mistakes all the time which can affect the results for videos that do not contain human-written captions and while a lot more channels are starting to add captions to their videos, we still include the name of the video so the parents can check the watch history to review the video again.

VI. Predatory Behavior Chat Scanning

Now that this is the hardest part of the program due to how there are no existing resources on this except ChildSafe.ai which is still not even in beta, but this also means that there is a whole new field of child safety that is barely explored. While currently there is no open-source solution available, one can be made with enough resources as is going to be discussed here.

First, we propose training an NLP API to detect specific behaviors that connect with online exploitation (i.e. persuasion, manipulation, deception, etc.) and that gets easier due to the abundance of chat logs that the Pervert-Justice foundation has compiled over 15 years of decoy operations that allowed them to achieve over 623 convictions, only one of which was from research, that means that we have over 600 full chat logs of actual convicted online predators which when paired with the multiple research papers available on the techniques that online predators utilize to lure young children one can train a very reliable model by feeding these chat logs and tagging them with appropriate tags, secondly the proposed A.I have to have a predator identification and police reporting algorithm which tries to look up the predator using the given username, altering the parents and reporting the case to local authorities immediately and giving them easy access to the results of the lookup if there are any, finally, there has to be a database that records these chatlogs anonymously with the parent's permission to be available to further help development in this area of child safety, the database should be available to researchers, physiologists to help research newer behaviors and techniques that these predators are using, it should also be possible to A.I researchers so they can feed these into their models to improve them.

VII. Conclusion

As internet use is extending to younger children, there is an increasing need for research focusing on the risks young users are experiencing, as well as the opportunities, and how they should cope. The Internet represents a significant threat to children because many children lack the simplest protection ways. Approximately 34% of students report experiencing cyberbullying during their lifetime. Over 60% of students who experience cyberbullying reported that it immensely impacted their ability to learn and feel safe while at school. We chose our approach to solve the problem, which is using AI to limit the threats caused by the internet, especially Chatrooms, websites, and videos. We use algorithms to recognize any spam and any unfriendly or immoral content. Any source that was found to fulfill these conditions will be prohibited immediately, and parents shall be alarmed. By working on decreasing these threats, we are helping in solving this terrible problem. We are providing a safer climate for children to use the internet without any fear. There are 71% of teens have hidden their online behavior from their parents so we provide parents with a program that will make them feel comfortable and rest assured that their children are in safe hands. We are working for a better future.

VIII. References

[1] Przybylski, A. K., & Nash, V. (2018). Internet Filtering and Adolescent Exposure to Online Sexual Material. Cyberpsychology, Behavior, and Social Networking, 21(7), 405-410. doi:10.1089/cyber.2017.0466

[2] Marcum, C. D. (2007). Interpreting the Intentions of Internet Predators: An Examination of Online Predatory Behavior. Journal of Child Sexual Abuse, 16(4), 99-114. doi:10.1300/j070v16n04_06

[3] Teimouri, M., Benrazavi, S.R., Griffiths, M.D., et al. (2018). A Model of Online Protection to Reduce Children's Online Risk Exposure: Empirical Evidence from Asia. Sexuality & Culture 22, 1205–1229. doi:10.1007/s12119-018-9522-6

[4] Savirimuthu J. (2012) Online Child Safety, Civil Society and the Private Sector: Alternative Strategies. In: Online Child Safety. Palgrave Macmillan, London. doi:10.1057/9780230361003_6

[5] Mathiesen, K. (2013). The Internet, children, and privacy: the case against parental monitoring. Ethics Inf Technol 15, 263–274. Doi:10.1007/s10676-013-9323-4

[6] C. (n.d.). CHILDREN'S INTERNET USAGE STUDY. Retrieved January 3, 2021, from https://iamcybersafe.org/s/parent-research

[7] Davis, M. M. (2015, August 10). Top 10 Child Health Problems: More Concern for Sexting, Internet Safety (Rep.). Retrieved http://mottnpch.org/sites/default/files/documents/081015_top10.pdf11

[8] K., Roberts, A., Cui, J., Smith, M., AIR; Bullock Mann, F., Barmer, A., and Dilig, R., RTI, H. (n.d.). (Rep. No. 2020144)