CASPER-2.0: An AI-based ghost protecting children from online threats

Summary

The main aim of the CASPER-2 project is to develop an artificial intelligence-based application-agnostic agent operating on user-interface human-computer interaction level to detect & prevent inappropriate content for children.

Keywords

Application-agnostic solution, artificial intelligence, children protection, human-computer interaction, software, online child protection.

Actors involved in the project

- Serbia (School of electrical engineering): Aleksandar Jevremovic, Milan Cabarkapa, Marko Krstic, Mladen Veinovic, and Milos Stojmenovic

  The School of Electrical Engineering in Belgrade has over a century-long history of education and research in the fields of electrical engineering, telecommunications, and (later) computer science. Over 1.000 students enrol every year.

- North Macedonia (Faculty of Computer Science & Engineering): Ivan Chorbev, Ivica Dimitrovski, Petre Lameski, Eftim Zdravevski

  The Faculty of Computer Science and Engineering (FCSE) at UKIM is among the largest and most prestigious faculties in the field of computer science and technologies in North Macedonia. It started to work in 1985 under the name "Institute of informatics".

- Portugal (O Mundo da Carolina): Nuno Garcia, Nuno Pombo

  O Mundo da Carolina is a non-profit association that aims to support children with chronic diseases and who are in an unfavourable socioeconomic condition.

The NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618

The project

Our project started in a less developed area, with the main objective of protecting children and young people using the Internet.

At first, we just discussed the possibilities of using AI on the human-computer interaction (HCI) level to create an application-agnostic solution for filtering inappropriate content. Our team members - scholars from Serbia, North Macedonia, and Portugal - already had a great experience in domains like cybersecurity, human-computer interaction, and artificial intelligence. Then, we found a NGI Trust call and we decided to apply, sending our proposed solution. As the proposal was accepted, we obtained funding and we also got the chance to work with great people and field experts like Mr. Alasdair Reid, Sigita Jurkynaitė, Javier Nesofsky, Raffaele Buompane, Christian Schunck, and Casper Dreef.

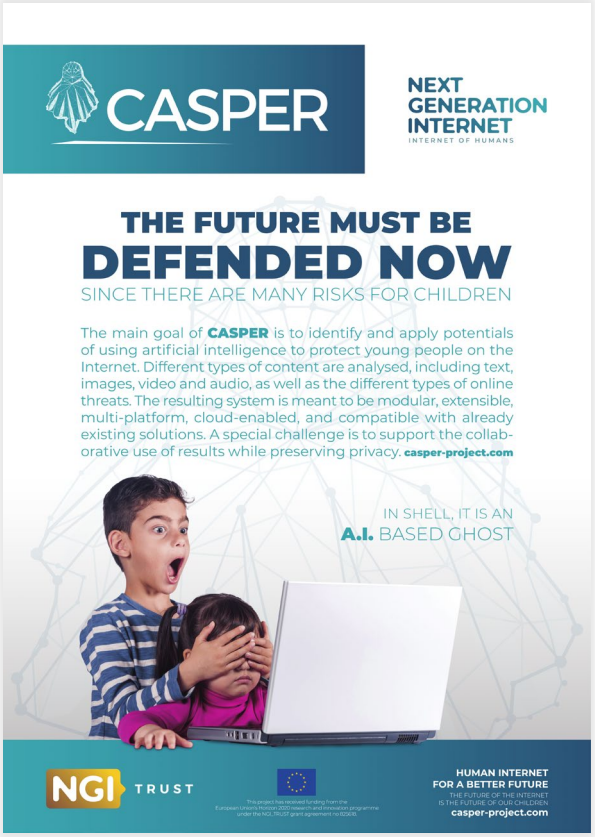

Figure 1. CASPER poster

The problem

Our project aims to provide an application-agnostic solution for filtering inappropriate content from online communications. This means to detect the inappropriate content and, where applicable, to prevent exposure of this content to end-users.

Figure 2. CASPER project members at the kick-off meeting in Belgrade, Sept. 2019

At the very beginning, since most of our researchers are parents of young children and adolescents, they tried to apply existing solutions to protect them when using the Internet. It was surprising to discover that there was no effective solution available to cover different applications and different threat types. So, our original idea and our research started somehow to scratch this personal itch. Later, the COVID-19 pandemic showed us how necessary and critical this project was, since people - including the most vulnerable groups - were spending even more time online.

The NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618

Originally, within the **Casper 1.0** project, we mostly focused on protecting young children from being exposed to nudity, pornography, and online cyberbullying. In other words, we developed and tested algorithms to be integrated into a compact product.

After this step, we received a grant for **CASPER 2.0** in order to implement the results from CASPER 1.0 viability study. The main focus of the second project was to achieve real-time performance and complete a Minimum Viable Product.

During the project we faced many challenges, including selecting effective and efficient algorithms, optimizing to achieve real-time performance, protecting users' privacy in distributed classification scenarios, getting relevant training datasets, etc. The bright side is that overcoming some specific problem usually gave us additional ideas on how to expand or improve the project.

In fact, after testing different algorithms and developing prototypes, we understood that the potential of our original approach was much bigger. As such, within the Casper 2.0 project we expanded the scope to support other languages other than English, and to protect another vulnerable user profile: elderly people. Of course, from completely different types of threats - fake news and online frauds.

The solution

The Casper 1.0 project started in August 2019 and ended in July 2020. Casper 2.0 started in August 2020 and ended at the end of April 2021. Our main objective was to develop an application agnostic AI-based solution on HCI/UI level to detect/prevent inappropriate content for children.

During the Casper 1.0 project, we made the first steps in recording HCI and then post-processing it to select optimal algorithms. Within the Casper 2.0 project instead, we focused on achieving real-time performance, by optimizing architecture, using dedicated hardware, or using edge-computing architecture.

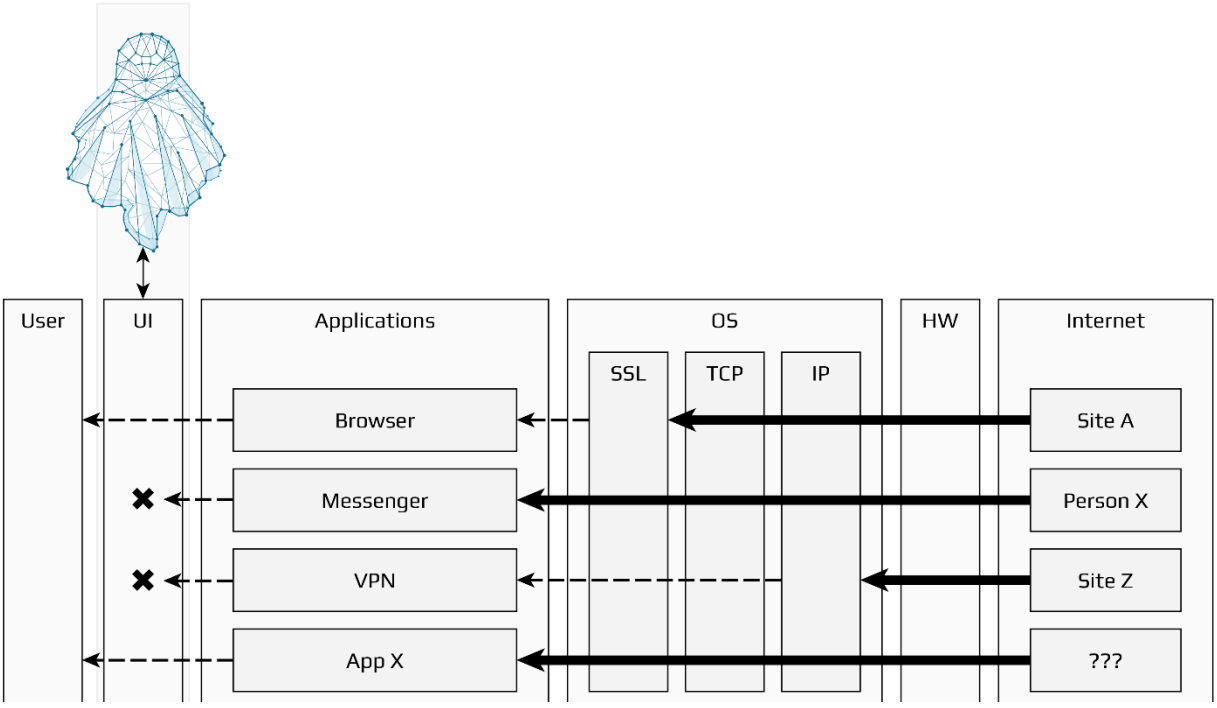

Figure 3. Example of content filtering via CASPER

The NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618

While working on our solution, we used different programming languages (Python mostly) to develop new algorithms and the platform, and different existing algorithms for OCR, text/image classification, etc. We also used WireGuard as a VPN solution for edge-computing scenarios. The MVP product is developed for Windows OS. Some features (for instance screenshot capturing and UI reaction) need to be adapted for different desktop and mobile platforms. We plan to support Android and Linux in the future.

One of the main CASPER use-cases is related to the use of messaging apps. Although many social network platforms implement mechanisms for protection from inappropriate content, their messaging part is still vulnerable. One of the reasons for that is that the same mechanisms cannot be applied without endangering users' privacy and bypassing end-to-end encryption used in messaging apps. However, CASPER agents could provide children protection even for these applications by analysing content completely locally on end devices. This way both privacy and encryption issues can be overridden, providing application-agnostic protection to our children.

Our idea is to make this software available to anyone who needs this kind of protection, regardless of their economic situation. On the other side, it is critical to prevent this software from becoming mass surveillance software. So, going open source turned out to be the only possible choice. Thanks to NGI we were able to develop a functional prototype. When the funding from NGI stops, we plan to continue our work as a foundation, and we expect our primary source of income to be donations.

Our project is competitive in relation to other alternatives mostly because of its holistic and application-agnostic approach. Practically, it means that even if someone manages to overcome the DNS protection or to inject malware to the victim's computer, our software will still manage to protect them from revealing personal data or being exposed to inappropriate content.

We strongly believe that protecting children and other vulnerable categories when using the Internet is one of the critical tasks for creating a better society and better Internet. Going open source is also a critical aspect of that belief.

Privacy means protection from threats that we are (still) unable to fight effectively. That's why vulnerable groups of Internet users, including children and elderly people, need their privacy to be respected. In some cases, it even means preventing them to perform some actions - actions that could be harmful to them, but that they are not in a position to recognise.

Figure 4. Aleksandar Jevremovic, Professor at Singidunum University in Serbia, researcher at the School of Electrical Engineering in Belgrade.

Results

Since it is a "mission-critical" software, we are very cautious about releasing it in public. At this moment, we have the software running on some personal computers that our researchers can control. However, we have to do some additional optimisations, in terms

of efficacy and efficiency, before we are sure that it can provide the required level of protection to end-users.

Our current metrics mostly depend on the quality of datasets we use for training the algorithms. Because this is a very sensitive topic, we can't do testing with real users. When we release the software in public, we'll also use false positives/negatives as relevant metrics.

We presented our project at different conferences, and the reactions were mostly very positive. Especially in cases where young parents were eager to download and test the solution. We consider this to be a success we were striving for since the beginning of the project. On the other side, since our goal is to make this software freely available, if we manage to continue our work based on donations, we'll consider that to be another big win.

Considering its open-source nature, our project can also be seen as a research framework in this field. This means that other researchers can use the project as a base for their research. Preserving privacy is one of the critical goals of this project, and the open-source license is a way to insist on that.

Testimonial

Projects like ours, that are oriented towards the common good but are not profitable, are not usually very interesting for investors. Regardless of that, the NGI Initiative decided to support our work on the project, not only by funding researchers but also by providing mentors, coaches, consulting, access to relevant events, training, webinars, connections, and much more.

We received great help from our coaches and mentors. Since our team is mostly made of researchers and scientists that lack relevant business experience, they helped us to find business models that would fit our needs. Also, they helped us to consider other aspects, especially those relevant for providing project funding in the future.

We had some very basic questions in some domains, but the coaches were more than patient to guide us and give us advice on how to overcome some difficulties we faced during the project.

Future plans

We will try to continue our work as a foundation, and we already had some interesting talks with UNICEF, the WeProtect Global Alliance, and others. As soon as we have a functional prototype, it will be much easier to attract funding.

Additionally, we already have some ideas to apply a future CASPER foundation in other domains, but we are currently too busy working on this project. Even if we had more developers and researchers, we would use them on this project. There is a lot of work to be done in the next 5-10 years.

The NGI_Trust project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825618