How AI-Powered Web Scraping Tools Fight Crime

Vaidotas Šedys

According to data published by Statista, monetary damage of cybercrime reported to the Internet Crime Complaint Center (IC3) increased significantly between 2001 and 2022, more than doubling in the last two years.

Such a massive surge cannot be plausibly explained by the rising rates of reporting without assuming that cybercrime itself is on the rise. Of course, a proportion of criminal activity, we do not know how sizable, still remains unreported. And then there is a bleak side of cybercrime that would not show up in such statistics as it doesn't cause direct monetary damage, for example, sharing child pornography.

All this speaks to the desperate need for better means to detect and counter online criminal activity. As a method to efficiently collect threat intelligence at scale, web scraping is a clear candidate for being such a means. However, this method has its practical limitation, which require improving web scrapers with cutting-edge artificial intelligence (AI) and machine learning (ML) technology.

What is web scraping?

Web scraping is the process of extracting publicly available web data using automated tools known a scrapers and crawlers. A web crawler, or spider, crawls the internet to make a list of URLs according to pre-programmed criteria. A web scraper then extracts the needed information from those pages.

Public data collection serves a myriad of use cases in business, academia, and other fields due to its ability to extract large volumes of information efficiently and, if needed, in real time. The following are some of the common usages:

- Investigative journalism, for example, scraping historical archives, documents, or news articles to detect patterns that lead to or support a story.
- Price aggregation to provide price comparison service that allows users to get the best available deals in online stores.
- Brand protection, which involves automatically checking websites to see how the company's ads are displayed and how its trademark is used.
- Various academic and business research, such as labor market research through online job posting data.

Automated big data collection is also crucial for developing ML models. Algorithms are usually trained on large volumes of various data that can only be effectively acquired with the help of scrapers.

However, this symbiotic relationship between AI and data scraping has been challenged by recent lawsuits against big tech companies like OpenAI, Microsoft, Google, and Midjourney. The legal concerns over feeding ML algorithms with massive web data sets are related to presumed breaches of data privacy, copyright, and open-source software licensing laws.

As we wait for the case law to answer these understandable concerns, it is important to see the other side of this technology - how the combination of web intelligence and AI/ML helps fight in the name of the law in cyberspace.

Utility and limitations of web scraping

Web intelligence is used by both cybercrime researchers aiming to advance our knowledge and theoretical base and law enforcement and cyber security practitioners.

For researchers, web scraping provides data on various online crimes, like cyberbullying or the way it is presented in online media,

allowing them to investigate public sentiment. With such insights, scientists can develop ideas for crime-countering solutions and advise policymakers.

Online security specialists scrape thousands of websites for early identification of threats and data leaks. Cyber threat intelligence is also used to train AI-based tools that protect companies against security breachers.

For law enforcement officers, web crawling and scraping is the way to get clues for investigations as well as collect evidence. For example, the dark web, which is inaccessible by common web browsers and search engines, is ridden with illegal e-commerce trading in everything from counterfeited goods to drugs and child abuse content. Web crawlers that are built to crawl the dark web, identify illegal activities, and gather data that can become leads for investigations and, ultimately, proof of guilt.

However, due to the many requests web scrapers have to make to the servers when monitoring cyber threats, they are often misidentified as malicious actors. This results in IP bans that break the scraping process and significantly delay any research or investigation. Residential and datacenter proxies are used to rotate IPs and avoid triggering website protection against being flooded. Unfortunately, they are not enough to convincingly mimic organic user behavior and can still be blocked.

Other online data collection challenges are related to the fact that websites have different HTML structures and dynamic layouts. Manually adapting web scrapers to the changing website layout and building custom tools for every differently-built website might take away the speed and efficiency benefits of this method of threat intelligence collection.

Integrating AI and ML into web data collection

Investigation of keywords used in cybercrime research papers has shown that this field's interest in machine learning has been rising rapidly in recent years. The reasons for that are easy to see.

Machine learning can help overcome previously mentioned limitations that specialists encounter when using web crawling and web scraping to tackle cybercrime. AI-powered proxy tools, such as Web Unblocker, are better at preventing IP bans. They are capable of dynamic fingerprinting, that is, providing user-related information such as browser type, location, and window resolution, most appropriate for the particular website that is being scraped.

Additionally, ML-based adaptive scrapers are trained on data from many websites with different layouts and underlying structures. Thus, one such scraper can be used for multiple sites and adapt to dynamic website structures.

AI and ML-powered scraping solutions can also automatically collect and classify data based on previously annotated web content while removing redundant and irrelevant information to ensure that the collected data meets the pre-set standard.

Ultimately, all this saves time and manual labor for cybercrime investigators and cybersecurity officers while producing high-quality datasets.

In the future, such AI-enhanced scraping solutions might become even better at predicting cyber threats or even trapping cyber criminals. For example, trained on communications with online fraudsters or sexual predators, ML-based chatbots might be developed to convincingly simulate victims' responses and help law authorities extract information. Additionally, AI-powered scrapers will get better at automatically collecting and analyzing visual content to gather evidence and spearhead investigations.

Fighting cybercrime with AI and web scraping

The combination of AI and automated data gathering solutions is already actively used and being further developed for the future of countering criminals online.

1. Partnerships with regulatory authorities to combat cybercrime

Oxylabs has launched the 4β (4beta) project to enable researchers, academic institutions, public sector, and non-profit organizations to make use of big data. The initiative has already attracted numerous partnerships, some of which tackle cybercrime.

The first one involved creating an AI-powered scraper for RRT — The Communications Regulatory Authority of the Republic of Lithuania. In the first couple months of use, the tool scanned over 280,000 Lithuanian websites to detect illegal content, mainly related to child sexual abuse, and report it to RRT specialists for further investigation.

Another partnership is with the Environmental Protection Department of Lithuania. The project aims at combating illegal online advertisements related to environmentally harmful activity. Previously, the department's employees had to manually look for ads that offer illegal hunting or fishing equipment or advertise other illicit activities. Oxylabs helped to develop an AI-powered web scraping solution to automate these procedures and offer more accurate identification of illicit online proposals.

2. The future of fighting cybercrime in the UK

In 2021, the Government Communications Headquarters (GCHQ) of the United Kingdom outlined how they plan to develop AI systems for analyzing web data to combat cybercrime. The report shines a light on the rarely openly advertised ways in which AI and web data collection are used by government agencies. The agency's examples include using AI to uncover connections between the individuals that form weapon, drug, and human trafficking networks. Additionally, AI models analyze data to predict cyber attacks and trace malware back to its origins.

GCHQ points out another important benefit of using visual data collection and AI analysis for countering child sexual exploitation — automated procedures prevent human analysts from being unnecessarily exposed to traumatically disturbing content.

In conclusion

Increasing digitalization and online anonymity provide perfect conditions for cybercrime to proliferate. The solution to this problem must come from improving the technological capabilities of law enforcement.

AI and, particularly, machine-learning-based web scraping and data analysis might help detect, report, and investigate online crime, chasing the scale and proficiency of criminal activity itself. However, tipping the scales in the law's favor calls for purposeful cooperation between businesses, NGOs, and governmental law enforcement agencies when utilizing such technology.

About the Author

Vaidotas Šedys is the Head of the Risk Management Department at Oxylabs, a market-leading web intelligence solutions provider. Having extensive experience in payment and digital risk management, Vaidotas established himself as an influential force in the online web data gathering industry, employing innovative methods to ensure the most ethical and secure SaaS business processes. Currently, Vaidotas is leading a team of 9 professionals that is successfully overseeing risk-vulnerable areas of business operations and countering emerging risks.