sciendo International Conference KNOWLEDGE-BASED ORGANIZATION
Vol. XXXI No 1 2025

AUTOMATED TARGET PROFILING: LEVERAGING ARTIFICIAL INTELLIGENCE FOR OPEN-SOURCE INTELLIGENCE COLLECTION

Robert SOLER, Maurice DAWSON, Marc COLINA

Illinois Institute of Technology, Chicago, USA
rsoler@hawk.iit.edu

***Abstract**: The emergence of Artificial Intelligence (AI) and the growing complexity of cyber threats have created new prospects for automating cybersecurity functions. A pivotal domain that gains from this innovation is Open-Source Intelligence (OSINT), an essential element of threat intelligence. This paper details the design and execution of an AI-driven solution that automates the creation of extensive target profiles through the integration of social media data collecting and large language model (LLM) processing. The application uses RapidAPI to consolidate publicly accessible information from social media platforms, including Facebook, Twitter, and LinkedIn, and applies OpenAI's GPT-4o model to transform that data into organized intelligence reports. The system also integrates computer vision models, YOLOv8 and DeepFace, for facial recognition and verification. The platform, hosted on a secure AWS cloud instance, provides a scalable and interactive online interface for users to enter data and obtain downloadable reports. This research illustrates how the integration of AI with OSINT procedures can improve the efficiency, depth, and accessibility of operational threat intelligence collection, while also identifying existing deficiencies in AI incorporation within the cybersecurity domain.*

Keywords: AI, OSINT, Large Language Model, GPT-4o, Cybersecurity

1. Introduction

The federal government of the United States is experiencing transformations due to a new administration and security challenges arising from ongoing hostilities, like the Russo-Ukrainian War. The capacity to collect intelligence to comprehend the dynamic battlefield is essential for sustaining military dominance. Nevertheless, specific assessments indicate that OSINT generates substantial information that can be effectively utilized to inform critical strategic decisions and acquire intelligence for tactical superiority. OSINT and other technologies enable a nation to enhance its preparedness for the multidomain warfare concept, wherein cyber capabilities interlink the marine, space, ground, and air domains [1].

Furthermore, university courses have been designed to meet these requirements in response to national security challenges, as shown by the program at the Illinois Institute of Technology (Illinois Tech) [1-2]. The Illinois Tech course emphasized the utilization of Open-Source Software (OSS) and AI to develop tailored apps that execute essential activities for cybersecurity and intelligence professionals identified as vital for national security [3-4].

DOI: 10.2478/kbo-2025-0023

2. Background and Literature Review

A. Artificial Intelligence

Artificial Intelligence, or AI, is the technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity, and autonomy [5]. The advancement of AI over the past decade has been undeniable. From 2012 to 2022, the average computing power of AI models has increased by over 1,000,000 times [6]. Numerous AI models have emerged, and at the top of the popularity poll today are Large Language Models, or LLMs. These are a category of models that have been trained on so much textual data to the point that they have become capable of understanding and generating text in human language [7].

There are dozens of different LLMs usable to the public as of today. These include models like Claude, the GPT series, Grok, Gemini, and Copilot, among others. GPT-4o of the GPT series is the most popular one among this list, and will be the LLM used for this project. This model was created by OpenAI, an industry leader in AI and machine learning technologies. GPT-4o is a "large multimodal model capable of processing image and text inputs and producing text outputs" [8]. It is a highly advanced creation, as it exhibits human-level performance on different academic and professional benchmarks. This, combined with its abilities to process extremely large amounts of text in a very short time, makes it a very powerful tool for tasks that have to do with analyzing large amounts of textual information.

B. Threat Intelligence

In the past decade, AI and Machine Learning weren't the only tech fields that have advanced drastically; cybersecurity has too. The average detection, response, and attribution for cyber incidents have gotten faster by over 1800% since 2011 [9]. Furthermore, the annual global revenue for cybersecurity defenses against Advanced Persistent Threat (APT) groups, or hacker groups, has increased from $2B to $12.5B in the past decade [10].

Among the many different domains of cybersecurity, this paper will focus mainly on threat intel. Threat intel is defined as the set of data collected, assessed, and applied regarding security threats, threat actors, exploits, malware, vulnerabilities, and compromise indicators [11]. Though there are many types of threat intel out there, this paper will be focusing more on operational threat intel, which gives insight on the "who", "why", and "how" behind a past or potential attack [12]. More specifically, the tool being described in this paper is one that aids in gathering Open-Source Intelligence, or OSINT. OSINT is information about something or someone that can be legally obtained through public sources like newspapers, journals, and the internet [13], and this information is vital to threat intel.

Advances and investments in threat intel were brought about by cybercrimes and cyber-attacks. Over time, threats have grown in complexity and number. From 2010 to 2025, the number of APT groups has gone from 10 to 150 [14]. For this reason, the importance of threat intel has been recognized by both governments and private sectors alike. In fact, the United States has two federal laws, the Cybersecurity Act of 2015 and the Cybersecurity Enhancement Act of 2014 [15][16], that require private companies to share threat intel with each other.

Advancements in AI greatly complements cybersecurity, especially in threat intel. However, although integrating AI with threat intel has already caused breakthroughs, research shows that both the private sector and the government aren't utilizing AI in cybersecurity to its fullest potential yet [17]. In fact, there isn't much research done on the intersection between OSINT and AI [18].

3. System Components and Architecture

The Social Media Profiler isn't just one application that runs independently on its

own. Before diving into the project's architecture, it is important to first understand what an Application Programming Interface (API) is. An API is a software intermediary that allows applications to interact with each other. In other words, it lets different applications talk to each other.

This whole tool is composed of two main components: the back-end code and the front-end code. At the same time, it relies on two other third-party services as well: the RapidAPI hub and the OpenAI API.

A. RapidAPI

RapidAPI is an API hub owned by Nokia [19]. It is essentially an online platform where customers can subscribe to and use different sorts of API services. Many of these APIs are connected to social media platforms such as Facebook, Twitter, and LinkedIn [20],[21],[22],[23]. Using RapidAPI, the tool will communicate with these three social media platforms to gather data (and eventually OSINT) about the target.

B. OpenAI API

Unlike the APIs in RapidAPI, the OpenAI API doesn't connect its users to social media platforms. Rather, it allows its users to communicate with the GPT-4o API [8], which is the LLM that will be used to analyze all the provided OSINT and generate a profile report out of it.

C. Front-end Code

The front-end code is responsible for what the user will be directly interacting with. Using the React framework [24], this code builds the website where the user will input the social media URLs of their target. Then, it will display the generated report about the target. It also allows the user to download this report locally as a PDF file.

D. Back-end Code

This portion is the vast majority of the project. It's in charge of handling the user input that the website received, cleaning and processing the target's data gathered from RapidAPI, sending the processed information to and receiving the profile report from the OpenAI API, and sending that report to the website for the user to view. This back-end code runs a Flask web application [25] which hosts its own API for the website to communicate with. It also runs two computer vision models, Yolov8 [26] and DeepFace [27], to detect and recognize the face of the target.

E. Architecture

Figure 1 displays a high-level architecture of the tool. The program flow in this architecture will be discussed in the next section.

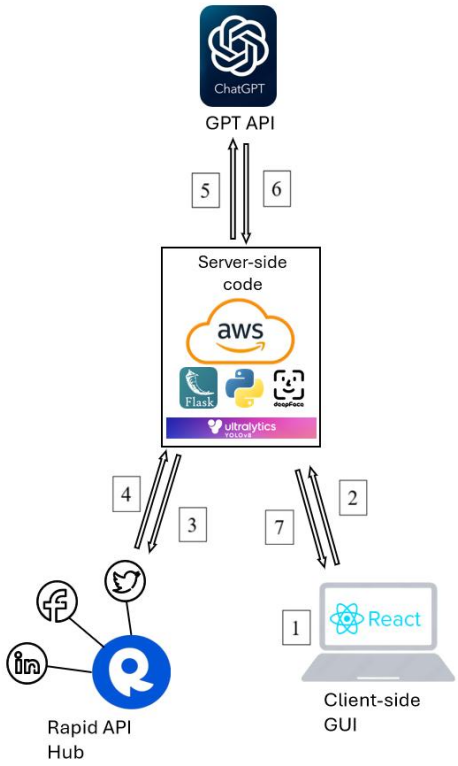

Figure 1: Software architecture of the Social Media Profiler

4. Implementation

Below is the program flow of the architecture in Figure 1. Note that each step is referenced in Figure 1 using its respective number.

1. User inputs the social media URLs of the target individual.
2. Website sends those URLs to the server-side code.
3. Server-side code uses those URLs to query RapidAPI for the target's social media data.

4. RapidAPI responds to the server-side code with the target's social media data.
5. Server-side code cleans and processes the data into organized information, then sends it to the OpenAI API.
6. GPT-4o processes all the information given by the server-side code and generates a profile report out of that, then sends it back to the server-side code.
7. Server-side code formats the report and then sends it over to the website so that the user may view the report.

As stated in the previous section, two computer vision models are used to detect the face of the target. This happens in step 5 of the program flow. Yolov8 detects if there is a human in the target's profile pictures, and if there is, then DeepFace analyzes each human's face and picks which face is the target's based on the rest of the OSINT gathered from their social media. This picture will then be included in the report.

The entire tool was then deployed in the cloud. An AWS EC2 instance was used to host the back-end and the front-end code (the website). This was done to make the tool scalable and available to anyone with internet access. Additionally, SSL was used to serve the site via HTTPS, for security. However, the SSL certificate used for this was just self-signed, as it only served as a proof-of-concept and production certificates require payment.

5. Results and Discussions

Upon engineering the tool to follow the discussed system architecture, the OSINT tool was finished. Booting up the AWS EC2 instance and running the programs for the tool, the website in Figure 2 can be visited:

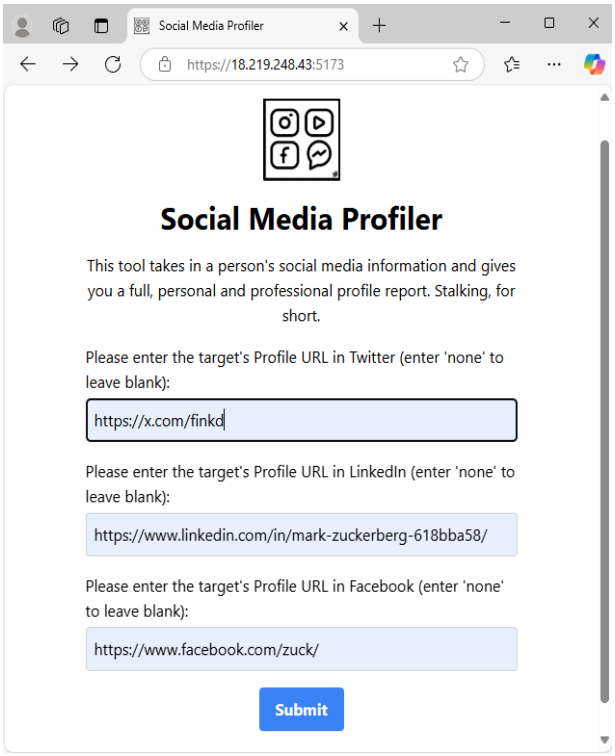

Figure 2: website GUI for the Social Media Profiler

Figure 3 illustrates a profile report that provides a comprehensive intelligence study of the target. This information encompasses details such as the entity's race, complete name, date of birth, educational history, and other publicly accessible data available on the Internet. This information facilitates the development of a link analysis [29-30]. Future research will further integrate AI by not only automating jobs but also training the LLM to deliver analyses comparable to, if not superior to, those of individuals with extensive expertise in the Intelligence Community (IC). This would be coupled with cybersecurity technologies that assess system vulnerabilities and exploits, thereby offering enhanced understanding [2].

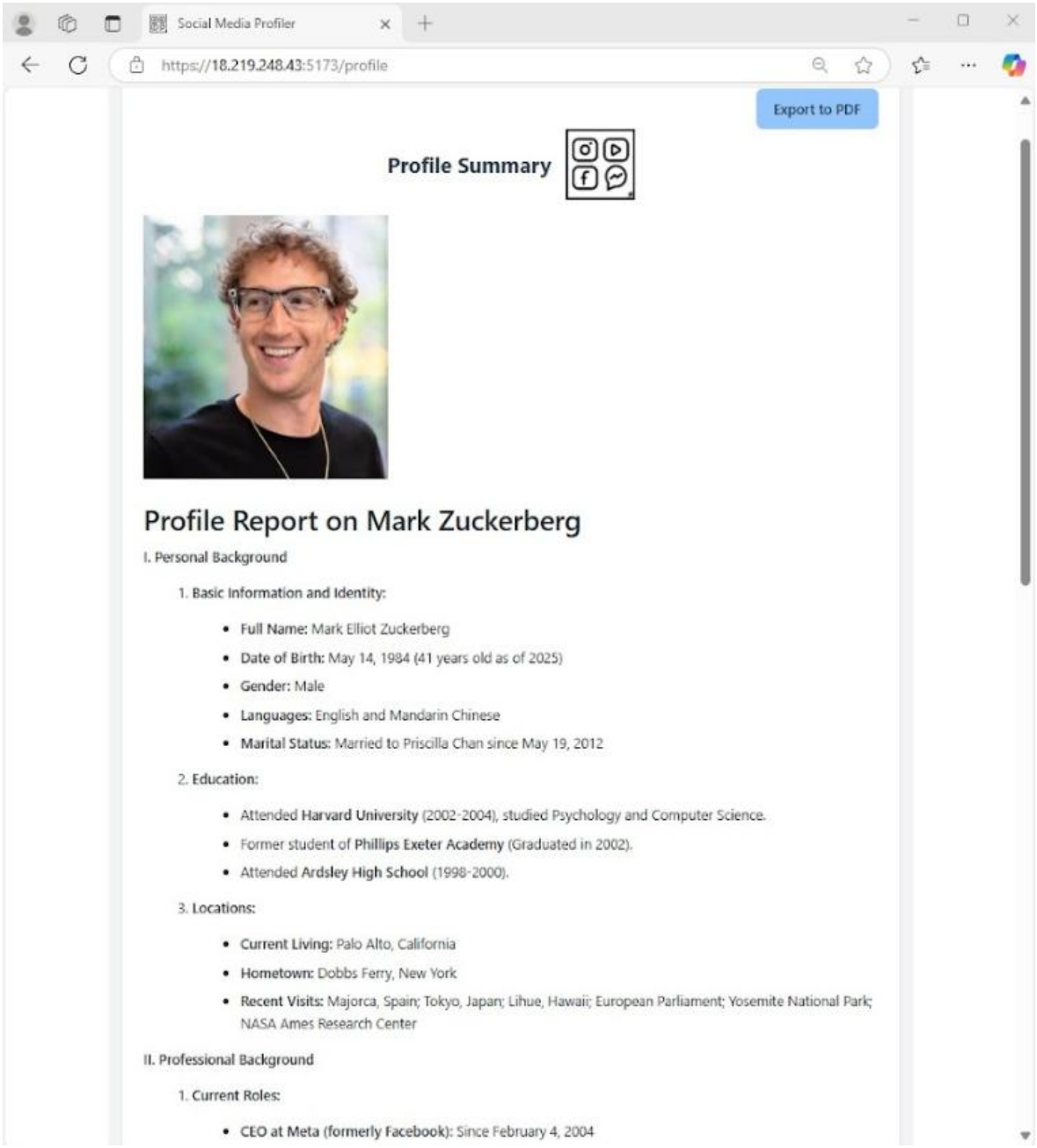

Figure 3: Generated profile report for the target

6. Conclusions

This research presents an innovative method for automating OSINT-driven threat intelligence through a layered architecture that integrates AI capabilities with contemporary web technologies. The system efficiently converts unstructured public information into actionable profiles by integrating social media data extraction via RapidAPI with sophisticated processing from OpenAI's GPT-4o model. The incorporation of computer vision tools enhances the data by validating and scrutinizing target imagery. The deployment to AWS illustrates the tool's scalability and practical relevance. The initiative highlights the significant potential of AI in enhancing cyber threat intelligence, especially for identifying and characterizing individuals of interest in investigative and defensive contexts. The prototype confirms the viability of automated OSINT profiling

while highlighting the necessity for additional investigation into ethical implications, accuracy standards, and the resilience of AI-driven decision-making in cybersecurity scenarios. This study enhances the expanding body of studies promoting the comprehensive incorporation of AI into national and business threat intelligence systems.

Reference List

[1] Dawson ME Jr. Cyber Warfare: Threats and Opportunities [Internet]. Porto, Portugal: Universidade Fernando Pessoa; 2020 [cited 2025 Apr 20]. Available from: https://bdigital.ufp.pt/entities/publication/5e93c236-588a-4564-bfb6-982d33fe11b4.

[2] Dawson M. Integrating intelligence paradigms into cyber security curriculum for advanced threat mitigation. In: Information Technology – New Generations. Cham, Switzerland: Springer Nature; 2024. pp. 77–81.

[3] Oerlemans JJ, Langenhuijzen S. Balancing national security and privacy: examining the use of commercially available information in OSINT practices. Int J Intell CounterIntell. 2025;38(2):579–597.

[4] Browne TO, Abedin M, Chowdhury MJM. A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications. Int J Inf Secur. 2024;23(4):2911–2938.

[5] Stryker C, Kavlakoglu E. What is Artificial Intelligence (AI)? [Internet]. Armonk, NY: IBM; 2024 Aug 9 [cited 2025 Apr 20]. Available from: https://www.ibm.com/think/topics/artificial-intelligence.

[6] Sevilla J, Heim L, Ho A, Besiroglu T, Hobbhahn M, Villalobos P. Compute trends across three eras of machine learning. In: International Joint Conference on Neural Networks (IJCNN). [Place unknown]: IEEE; 2022. pp. 1–8.

[7] IBM. What are large language models (LLMs)? [Internet]. Armonk, NY: IBM; 2023 Nov 2 [cited 2025 Apr 20]. Available from: https://www.ibm.com/think/topics/large-language-models.

[8] OpenAI. GPT-4 technical report [Internet]. [Place unknown]: arXiv; 2023 [cited 2025 Apr 20]. Available from: https://arxiv.org/abs/2303.08774.

[9] EuRepoC. Dashboard - EuRepoC: European Repository of Cyber Incidents [Internet]. [Place unknown]: EuRepoC; 2023 Oct 13 [cited 2025 Apr 20]. Available from: https://eurepoc.eu/dashboard/.

[10] Kost E. What is an advanced persistent threat (APT)? [Internet]. Mountain View, CA: UpGuard; 2024 Nov 17 [cited 2025 Apr 20]. Available from: https://www.upguard.com/blog/what-is-an-advanced-persistent-threat.

[11] Conti M, Dargahi T, Dehghantanha A. Cyber threat intelligence: Challenges and opportunities. Adv Inf Secur. 2018;1–6.

[12] Baker K. What is Cyber Threat Intelligence? [Internet]. Austin, TX: CrowdStrike; 2025 Mar 4 [cited 2025 Apr 20]. Available from: https://www.crowdstrike.com/en-us/cybersecurity-101/threat-intelligence/.

[13] Hwang YW, Lee IY, Kim H, Lee H, Kim D. Current status and security trend of OSINT. Wirel Commun Mob Comput. 2022;2022:1–14.

[14] MITRE ATT&CK®. APT Groups [Internet]. [Place unknown]: MITRE; [cited 2025 Apr 20]. Available from: https://attack.mitre.org/groups/.

[15] United States. Cybersecurity Act of 2015, 6 U.S.C. §§ 1501–1533, 2015.

[16] United States. Cybersecurity Enhancement Act of 2014, 15 U.S.C. §§ 7421–7430, 2014.

[17] Sun N, Ding M, Jiang J, Xu W, Mo X, Tai Y, Zhang J. Cyber Threat Intelligence Mining for Proactive Cybersecurity defense: A survey and new perspectives. IEEE Commun Surv Tutor. 2023;25(3):1748–1774.
[18] Evangelista JR, Sassi RJ, Romero M, Napolitano D. Systematic literature review to investigate the application of Open Source Intelligence (OSINT) with Artificial Intelligence. J Appl Secur Res. 2020;16(3):345–369.
[19] RapidAPI. RapidAPI [Internet]. [Place unknown]: RapidAPI; [cited 2025 Apr 20]. Available from: https://rapidapi.com/.
[20] RealtimeAPI. Facebook RealtimeAPI [Internet]. [Place unknown]: RapidAPI; [cited 2025 Apr 20]. Available from: https://rapidapi.com/realtimeapi-realtimeapi-default/api/facebook-realtimeapi.
[21] RockApis. LinkedIn API8 [Internet]. [Place unknown]: RapidAPI; [cited 2025 Apr 20]. Available from: https://rapidapi.com/rockapis-rockapis-default/api/linkedin-api8.
[22] AhnKiet N. Facebook Data API2 [Internet]. [Place unknown]: RapidAPI; [cited 2025 Apr 20]. Available from: https://rapidapi.com/tml25092004/api/facebook-data-api2.
[23] davethebeast. Twitter241 [Internet]. [Place unknown]: RapidAPI; [cited 2025 Apr 20]. Available from: https://rapidapi.com/davethebeast/api/twitter241.
[24] Meta. React [Internet]. [Place unknown]: Meta; [cited 2025 Apr 20]. Available from: https://reactjs.org/.
[25] Grinberg M. Flask [Internet]. [Place unknown]: [Publisher unknown]; [cited 2025 Apr 20]. Available from: https://flask.palletsprojects.com/.
[26] Jocher G, et al. YOLOv8 [Internet]. [Place unknown]: Ultralytics; [cited 2025 Apr 20]. Available from: https://docs.ultralytics.com/.
[27] Serengil SI. DeepFace [Internet]. [Place unknown]: [Publisher unknown]; [cited 2025 Apr 20]. Available from: https://github.com/serengil/deepface.
[28] Amazon Web Services, Inc. Amazon Web Services (AWS) [Internet]. [Place unknown]: AWS; [cited 2025 Apr 20]. Available from: https://aws.amazon.com/.
[29] Bier EA, Card SK, Bodnar JW. Entity-based collaboration tools for intelligence analysis. In: IEEE Symposium on Visual Analytics Science and Technology. Columbus, OH, USA: IEEE; 2008. pp. 99–106.
[30] Dawson M, Lieble M, Adeboje A. Open source intelligence: Performing data mining and link analysis to track terrorist activities. In: Information Technology – New Generations. Cham, Switzerland: Springer International Publishing; 2018. pp. 159–163.