USING OSINT TO GATHER INFORMATION ABOUT A USER FROM MULTIPLE SOCIAL NETWORKS

Nilesh Sambhe[1], Piyush Varma[2], Arpan Adlakhiya[3*], Aditya Mahakalkar[4], Nihal Nakade[5] and Renuka Lakhe[6]

[1]Assistant Professor, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India
[2]Cyber Forensic Expert, FORnSEC Solutions, Nagpur, Maharashtra, India
[3]Student, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India
[4]Student, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India
[5]Student, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India
[6]Student, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

[1]nilesh.sambhe@gmail.com, [2]varmap662@gmail.com, [3]arpanadlakhiya.162@gmail.com, [4]adityamahakalkar49@gmail.com, [5]nihalnakade7@gmail.com, [6]renuka.lakhe@gmail.com

Abstract

With the widespread and easier access of the internet, many people have started to use various social networking sites each catering to their needs. It has been observed that most users prefer to use the same social media handle or username on multiple sites for easier management. This makes it possible to get a hold of the publicly available information about the user. But, with the increase in privacy protections and user restrictions, investigators often struggle to gather information about a user. We propose an automated software to perform this job which uses Open-Source Intelligence Gathering (OSINT) methods where all publicly available information of a user is gathered in an intelligently structured format all at one place. The software will search various social networking sites for the required user profile and gather all publicly available information. This information will then be available to investigators with the facility to export in various digital document formats.

Keywords: *social media, internet, user profile, information gathering, osint, social network, investigator, user interface*

1. Introduction

Social media when evolved in 1997 was merely just available as a service. But in today's modern era it is more than just that. In itself, a social network refers to a way in which people interact, and share information via virtual communities. Now it is just like our pseudo-online biodata whenever we meet any new person, we visit his/her social media profile and get an idea about almost everything like hobbies, interests, likes, dislikes, etc. Even today, there are various tools which do some data analysis using Artificial Intelligence (AI) / Machine Learning (ML) and show us about one's visual profile. This information is quite essential to people like investigators, who are assigned the task of gathering intel about a user. But they often struggle to gather this information about their target since it is a tedious and time-consuming process if done manually. Hence, we are introducing a system which focuses on automated data gathering from various social media platforms and makes it easy to access all in one place.

In this paper, we propose a system that makes use of OSINT to gather publicly available information about the users from various social networks. OSINT refers to the repurposing of public records for intelligence and investigations, including social media content not protected by privacy settings. While the investigating agent has to manually specify the search criteria such as username, tools such as search engines and web spiders will then automatically retrieve these data and as such are key elements in a process of constructing actionable intelligence from public records. As social media platforms maintain a presence in social life, their users continue to submit information, much of which is publicly accessible by default. [1]

* Corresponding Author

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Our software basically comprises two parts, a frontend User Interface (UI) and a backend server to run scripts and manage databases. The frontend is a user-friendly interface built in Electron that enables the user to interact with our software. The UI will enable the user to provide some input, and will display the output results in a structured manner. The backend is basically built in Python which uses the official Application Programming Interfaces (API) of different platforms to access the user's public data. For this the software needs the username of the person as input. Once the user enters a username, the software will execute appropriate scripts in the background and will fetch all the data from different platforms and will present it to the user in the UI. The software has the functionality of getting data in different output formats like PDF, Docx, etc.

The software will be structured in various modules. The first module is used to check whether a user exists in a given set of social media platforms. Based on its results it will execute site-specific modules to gather user information from those sites. This data will then be cached into a structured database system. The data will then be fetched from the database and rendered into the software UI. This software will be useful to various law enforcement agencies such as Police, Cyber Cell, Income Tax Department, Central Bureau of Investigation etc.

2. Literature Review

The emergence of social media dates back to nearly two decades ago. Since the social network has been rapidly growing over the past years, people have begun to adopt it as a valuable source of information. Social media became widely accepted by people beyond 2005. Suddenly, social media created a means through which people access information on every aspect of human life facilitating sharing of ideas, opinion, and interests among others increasingly comfortable with the world at large more than ever. This advent and outburst adoption of social media platforms is to a large extent facilitated as a result of advances in technology and the Internet.

The everyday use of online social networks (OSN) such as Twitter, LinkedIn, and Facebook have seen a steady rise in adoption since 2005. Online social networks mirror a subset of our everyday social interactions and possess information that crosses geographic borders.

As consumers of social media, we are starting to see an emergence of specialized social networking sites such as Twitter and LinkedIn that provide datasets that are both overlapping and disjoint from datasets of more general-purpose social networking sites such as Facebook. While Facebook and Twitter tend to have casual social interactions, LinkedIn has specialized in professional networking. A survey by the PEW Internet and American Life project, a non-profit think-tank dedicated to uncovering trends in American life, found that more than 50% of online social media users have two or more online profiles. [2]

Online social networks have a lot of information to offer for OSINT such as social contacts, activities, and personal details of an individual of interest, but on the contrary to what many might believe, not all information on the web is easily accessible. [2] Investigators are met with several obstacles including privacy and platform restrictions as well as data availability and longevity.

With growing privacy concerns, many social networking platforms have continued to add privacy control mechanisms to restrict access to private information. Information flowing into online social networks is collected on a massive scale, but is tightly controlled by the social media platform regarding how it flows out. Social networking platforms generally control information flowing out based on social relationships, user-based privacy settings, as well as rate limiting, activity monitoring, and IP address-based restrictions. The social graph is far from static, relationship dynamics change frequently and profiles are updated constantly. While many social networking platforms such as Facebook disallow the use of screen scrapers and other data mining tools through their terms of service agreements, the legal enforceability of these terms remains unclear. [2]

Investigators commonly need to verify and expand on the known associates of a target for vetting a security clearance or general target profiling. Online social networks provide a cost-effective means to gather contacts of an individual, but with a target's privacy protections enabled many valuable contacts may be obscured from an investigator. The lack of an all-in-one software which can ease the tedious task of gathering such information facilitates the idea of designing such a system.

3. Implementation

This section gives the overall explanation of the proposed system, including the flowchart of the proposed system, procedure of data gathering, data caching and displaying the results into the UI. For simplicity of our system, we will be using the five most popular social media platforms to gather data as follows – Facebook, Instagram, Twitter, LinkedIn, and Reddit.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

3.1. Flowgraph of proposed system

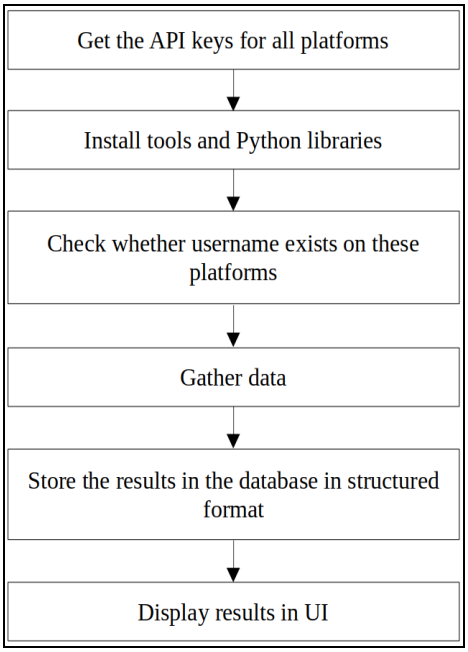

Figure 1. Flowgraph of the Proposed System

The flow graph in Figure 1. shows the process of data gathering and displaying the results. It is explained in brief below:

a. Getting the API keys

To connect to the social media platforms, we need to get the API keys in order to gather information from their websites. This step is essential since websites do not allow direct access to user data for security purposes. This can be done by creating a developer account on these platforms, and authenticating using the API keys.

b. Installation of tools and Python libraries

We are using official Python libraries to access the APIs for different websites. To check whether a username exists we are using a widely used open-source tool named Sherlock, which is also written in Python. After installing these tools and libraries we are all set to gather data.

c. Check whether a username exists on these platforms

We will ask the user to input the username, this username will then be passed as an argument to the tool Sherlock. This tool will then run scripts to check whether a username exists on each of the platforms, and will return the results.

d. Gathering data

From the results returned by Sherlock, we will start gathering the data from the platforms on which the username exists. Using the API keys, we will get access to all the publicly available data about the user. This data will then be structured into JSON format for ease of use.

e. Storing the results in database

The results gathered will then be cached into the database so as to reduce the overhead of fetching the same data repeatedly. This will help lower the required processing time and increase the efficiency of our software.

f. Displaying the results in UI

After the results are stored in the database, they are fetched and displayed in the software UI in proper format. The results can then be exported to various file formats such as PDF, Docx, CSV, etc.

3.2. Procedure for gathering data

For the implementation of the proposed system, we are using Python programming language. Python is an interpreted, high-level and general-purpose programming language. It has been gaining popularity among software developers and data scientists who manage and analyze user data. Python has a huge set of available libraries for gathering and analyzing data, hence making it very useful is our system.

For developing the software UI, we are using the Electron framework. Electron is an open-source framework developed and maintained by GitHub. It allows for development of cross-platform desktop

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Graphical User Interface (GUI) applications using web technologies.

The generalized block diagram of our software is as follows:

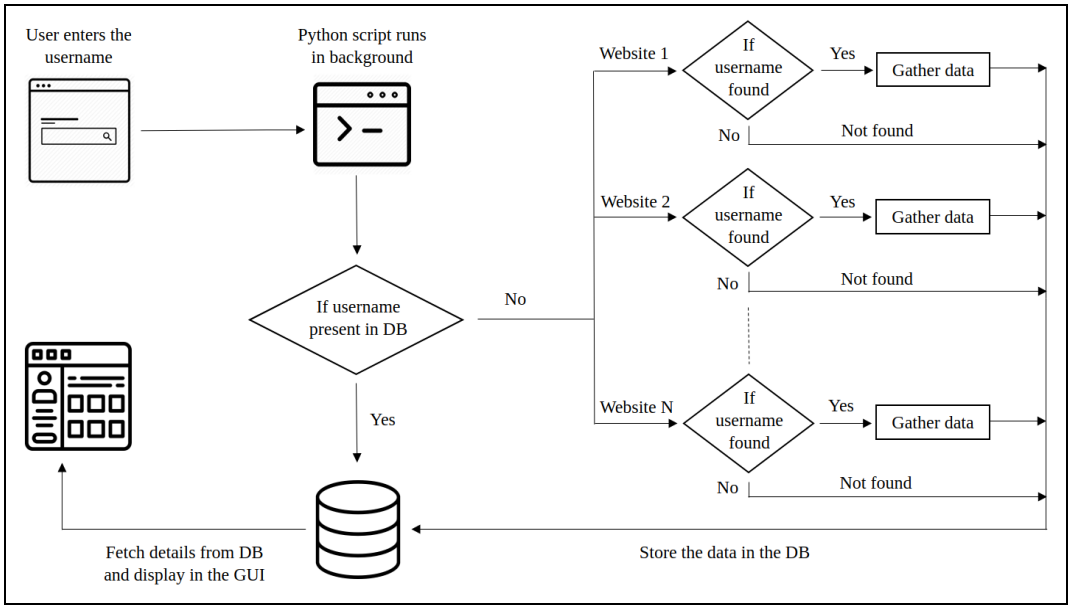

Figure 2. Generalized Block Diagram

The entire procedure for gathering and displaying data is described in a series of steps below:

Step 1: Creating a developer account on each platform in order to gain access to API keys.

a) Facebook – Create an account on Facebook and login to Facebook for Developers. From the "My Apps" section we can navigate to the "Graph API Explorer" and get the access tokens.

b) Instagram – Create an Instagram account, we also need the Facebook for Developers account, login to Facebook for Developers. From the "Instagram Basic Display API" page create a new app and grab the access tokens.

c) Twitter – Create an account on Twitter and login to Twitter Developers. We need to submit a simple request to get the API keys. Once the request is approved, we can get the access tokens.

d) LinkedIn – Create an account on LinkedIn and login to LinkedIn Developers. Create a new app which will give us the access tokens.

e) Reddit – Append the user account's URL with ".json" and retrieve information in JSON format.

Step 2: When the user will download and install the software, all the dependencies will be downloaded and installed by the installer automatically. Users don't have to explicitly install these dependencies.

Some of the dependencies include:

a) sherlock - A tool used to hunt down social media accounts by username across social networks

b) tweepy - To use the official Twitter API to access public information about a user

c) facebook-sdk - Official Python SDK to access the Facebook Graph API

d) instagram-basic-display-python – Python wrapper to access the Instagram Basic Display API

e) python-linkedin - LinkedIn API Wrapper for Python to access user information

Step 3: After all the dependencies are installed, the first step is to get the username from the user. This username will be passed to Sherlock tool. Sherlock will parse and check whether the user exists on that platform. If the user exists, then Sherlock returns the name of that platform as the result.

Step 4: If the user is found in the Sherlock tool's output, the respective site modules will be used to gather user data. Each site module will have the interface with the API's respective methods. The site modules will use these methods to get the complete user data that is publicly made available by the user. This data will be returned in a JSON

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

format which is iterable so as to ease the retrieval of data.

Step 5: From the results returned by the site modules, the required data about the user is stored into the database. This database also acts as a cache storing all the search results from all previous searches. If the user searches for the same username again, then this data is directly fetched from the database, hence reducing the time of execution of background scripts.

Step 6: This is the final step in the process in which all the results stored in the database are fetched and displayed in the software UI. The software UI made in Electron will display the data in a user-friendly format, and will allow the user to export the results to a variety of file formats such as PDF, Docx, CSV, etc.

4. Conclusion and Future Scope

In summary, we examined the methods of gathering data from various social media networks. We analyzed how the publicly available data can be accessed using different tools and interfaces, how the social media platforms restrict the gathering of private data in accordance to the user permissions. While there is little we can do to change the availability of information, our software does address data longevity issues by serving as a mechanism to enable efficient and automatic information gathering from social networks.

While previously investigators would have found it infeasible to search for a target's information, we have proposed a practical and efficient system to search for information. By doing so we have enabled investigators and other researchers to examine data that was previously tedious and time-consuming to find all in one place.

The software has an immense amount of scope in the field of digital investigations and forensics. People working in cyber cells and crime investigations can use this software to gather intel and whereabouts about their targets without wasting much time by manually visiting social media sites. The software can also be used by normal users to find contact information with ease.

As future work, the software can be greatly enhanced with the use of Artificial Intelligence and Machine Learning, which can dynamically gather data from a plethora of social networking websites. The information collected by the software can further be used for data analysis which can serve as a tool to predict a user's behavior by analyzing the content posed by users on different social media platforms and use it to predict the user's mood, likes-dislikes, thoughts, etc. Currently we are only limited to a few popular social networks, so in future we can add support for other social networks to increase the amount and accuracy of the data gathered. To improve the performance of our software, we can use distributed computing to run scripts for multiple websites parallelly or search for more than one username simultaneously.

Acknowledgments

We would like to express our special thanks of gratitude to our industry mentor as well as our teachers for sharing their pearls of wisdom and gave us this wonderful opportunity to work on our research project, which helped us to learn many new concepts and methods about gathering publicly available information.

References

6.1. Journal Article

[1] Trottier, D., "Open-Source Intelligence, Social Media and Law Enforcement: Visions, Constraints and Critiques.", European Journal of Cultural Studies 18, no. 4–5 **(2015)**, pp. 530-47.

6.2. Online Theses

[2] Holland, B. R., "Enabling Open-Source Intelligence (OSINT) in private social networks", Graduate Theses and Dissertations, Iowa State University **(2012)**. Available: https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=3354&context=etd

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731