DEVELOPING AN OSINT TOOL FOR INVESTIGATING CYBERCRIME AND MISINFORMATION ON SOCIAL MEDIA PLATFORMS

A report submitted to

RAMAIAH INSTITUTE OF TECHNOLOGY

Bengaluru

ISP SENIOR PROJECT

as partial fulfillment of the requirement for the award of degree of

Bachelor of Engineering (B.E) in Information Science and Engineering

By

Eeshan Singh	1MS19IS037
Chirag G	1MS19IS038
Gaurav Sood	1MS19IS041
Jayesh Goyal	1MS19IS052

Under the Guidance of

Prashanth Kambli

Assistant Professor in ISE department of MSRIT

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, Affiliated to VTU)
BANGALORE - 54

MAY 2023

RAMAIAH
Institute of Technology

Department of Information Science and Engineering

Ramaiah Institute of Technology

Bengaluru - 54

CERTIFICATE

This is to certify that EESHAN SINGH (USN- 1MS19IS037), CHIRAG G (USN- 1MS19IS038), GAURAV SOOD (USN- 1MS19IS041) and JAYESH GOYAL (USN- 1MS19IS052) who were working for their ISP, SENIOR PROJECT under my guidance, have completed the work as per my satisfaction with the topic DEVELOPING AN OSINT TOOL FOR INVESTIGATING CYBERCRIME AND MISINFORMATION ON SOCIAL MEDIA PLATFORMS. To the best of my understanding, the work to be submitted in the dissertation does not contain any work, which has been previously carried out by others and submitted by the candidates for themselves for the award of any degree anywhere.

Name and Signature of the Guide   Signature of the HOD   Signature of the Principal

External Viva

Name of the examiners   Signature with date

1.

2.

RAMAIAH
Institute of Technology

DECLARATION

We hereby declare that the entire work embodied in this ISP SENIOR PROJECT report has been carried out by us at Ramaiah Institute of Technology under the supervision of Prashanth Kambli. This project report has not been submitted in part or full for the award of any diploma or degree of this or any other University.

Eeshan Singh	1MS19IS037
Chirag G	1MS19IS038
Gaurav Sood	1MS19IS041
Jayesh Goyal	1MS19IS052

Place: Bangalore
Date:

ABSTRACT

The open-source tool was designed to analyze multiple social media accounts and gather information such as profile data, followers, posts, similarity between different accounts, and various analysis from the metadata gathered. The analysis was targeted towards individuals of interest, and it was used to investigate malicious activities that occurred online, such as cyberbullying, spreading misinformation, and cybercrimes. The tool was implemented in the form of a CLI or a web application, and users were able to choose from multiple analysis techniques based on their requirements. Overall, the tool was very powerful and useful for social media analysis of an individual. The tool covered over 800+ websites.

ACKNOWLEDGMENT

We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We would like to express our profound gratitude to the Management and Dr. N.V.R Naidu Principal, M.S.R.I.T, Bengaluru for allowing us to explore our potential.

We sincerely thank our beloved Dr. Sanjay H A, HOD, Information Science and Engineering, for his constant support and guidance.

We wholeheartedly thank our project guide Prof. Prashanth Kambli, for providing us with the confidence and strength to overcome every obstacle at each step of the project and inspiring us to the best of our potential. We also thank him for his constant guidance, direction, and insight during the project.

Finally, we would like to express sincere gratitude to all the teaching and non-teaching faculty of ISE Department, our beloved parents, seniors, and my dear friends for their constant support during the course of work.

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Motivation and Scope

As the number of people using social networking and social media platforms continues to rise, there has been a significant increase in the use of these platforms for performing various malicious acts. As a result, social media analysis has become an essential field to keep these crimes in check by providing valuable insights into user behavior, preferences, and trends. This tool has been developed to facilitate easy analysis of social media, with all major platforms such as Instagram, Twitter, and Facebook included. The tool was mainly built using Python and offers a variety of analysis modules for users to choose from, including user profiling, network analysis, and hashtag analysis. Each module has an associated rating matrix with rate values ranging from 0 to 100, indicating the probability of a positive or negative match for the person being searched. The insights gained from the tool can be used by businesses and individuals to make informed decisions, conduct research projects, or pursue personal interests. The tool can be deployed as a CLI mainly as a Python or Nodejs script, but when deployed as a web application, it offers additional features beyond the functionality of the CLI.

1.2 Methodology

The utility tool is written in Python and is designed to work with popular social media networks such as Instagram, Twitter, and Facebook. It includes various features and functions, such as username enumeration, hashtag analysis, and gathering post information and other metadata. The tool follows a three-phase approach to data collection and analysis: data collection, data processing, and data analysis.

1.3 Data Collection

The tool utilizes publicly available information from targeted social media platforms to analyze a person's required information. For example, the tool employs the Instagram Graph API to access multiple types of user data, such as user profiles, posts, and followers. Similarly, the tool uses the Twitter API to access user profiles, tweets, and followers for Twitter data. After gathering data from these websites, it is stored in a database for further processing and analysis.

1.4 Data Examination

After completing the data collection process, the tool employs various algorithms and statistical methods to perform data analysis and derive insights. The tool offers several data analysis features, including username enumeration, hashtag analysis, and post metadata collection.

1.5 Enumeration of Usernames

One of the features of the tool is username enumeration, which enables users to gather information about an individual or a group of target audience using their usernames on various social media platforms. This method aims to collect the target user's publicly available information, such as their full name, biography, and location. Additionally, it can also gather data on the user's followers and those they follow, which is useful for personality analysis.

1.6 Analysis of Hashtags

The tool also includes a feature for hashtag analysis, which enables users to study the usage of hashtags related to a specific topic on a particular social media platform. This feature can assist in identifying popular or trending hashtags, as well as the users who are utilizing them. Additionally, it can be used to analyze the level of engagement associated with different hashtags.

1.7 Collection of Post-Metadata

The tool feature known as post metadata collection enables users to gather information about posts made by specific social media accounts. This function can extract data such as the post's date and time, the number of likes, comments, and shares, and the content of the post.

1.8 Plugin Creation

The tool's functionality can be extended through the use of plugins, which allow users to add their own custom features. Creating plugins is an integral part of the tool's development process because it enables users to tailor the tool to their specific needs.

To create a plugin, users need to write code that interacts with the tool's API and performs specific functions. For instance, a plugin could be designed to gather data on users who have liked a particular post or to identify the most engaged users on a specific social media platform. Once created, plugins can be shared with other users, allowing the wider tool community to benefit from the collective knowledge.

1.9 Outcome

The the tool comprises several key elements such as data collection, analysis, and plugin development that work together to provide users with a comprehensive solution for analyzing social media profiles. The program is user-friendly, featuring a simple command-line interface that allows for quick and easy access to important information on social network accounts. Additionally, the tool's functionality can be expanded through the use of plugins, enabling users to customize the tool to their specific needs. All in all, the tool is a powerful and flexible solution for a wide range of social media analysis tasks.

The following section provides a detailed explanation of the research conducted in this field and how frameworks and standard methods can be applied to our project based on insights and ideas gathered from various research papers worldwide. It outlines the development of a rough system design and framework for the tool, highlighting the system layout and components envisioned for its construction. This information is presented and explained in Chapter 3 of the report.

Chapter 4 delves into the specific implementation methods employed in our project and the diverse technology and system designs utilized in developing the tool. This chapter covers descriptions of JavaScript web engines, data transfers, and analysis techniques featured in the report. The significance of using Python and JavaScript in the project is also elucidated in this chapter.

Subsequently, Chapter 5 presents the final results obtained from analyzing various names across all available social media platforms. The outcomes of this analysis are described and displayed.

Finally, Chapter 6 discusses the gaps identified in this project and presents the conclusions drawn from the research.

Chapter 2

LITERATURE REVIEW

Social media analysis has become an increasingly popular research area in recent years due to the massive amount of data available on various platforms. Several tools and platforms have been developed to facilitate social media analysis, including both open-source and proprietary software. In this review, we will compare some of the existing tools with our tool, highlighting the strengths and weaknesses of each.

One popular social media analysis tool is Gephi, an open-source software platform that provides advanced network analysis capabilities. Gephi is mainly used for visualizing and analyzing large networks of data, such as social media networks, and provides a range of features for exploring, manipulating, and exporting network data. While Gephi is a powerful tool for network analysis, it may not provide the same level of user profiling and hashtag analysis that our tool offers.

Another tool worth mentioning is Social Mention, a free web-based application that allows users to monitor social media platforms for mentions of particular keywords or phrases. Social Mention provides real-time analysis of data from various platforms, including Twitter, Facebook, and YouTube, and provides a range of metrics and analysis tools for users to explore. However, Social Mention may not provide the same level of in-depth analysis as your tool, which offers user profiling, network analysis, and hashtag analysis.

A third tool that is worth considering is Brandwatch, a proprietary social media analysis platform that provides real-time data insights for businesses and individuals. Brandwatch offers a range of features, including sentiment analysis, competitive analysis, and audience insights, which can be used to inform marketing and business strategies. However, Brandwatch is a commercial product, which may not be suitable for individuals or small businesses with limited budgets.

Compared to these existing tools, this tool offers a more comprehensive range of analysis modules, including user profiling, network analysis, and hashtag analysis. Additionally, the rating matrix that tool uses to indicate the probability of a positive or negative match provides a unique and useful feature for users to evaluate the insights gained from the analysis. The fact that this tool can be deployed both as a CLI script and as a web application also makes it a versatile and accessible tool for users with varying levels of technical expertise.

Overall, this tool offers a valuable contribution to the field of social media analysis, providing users with a powerful and user-friendly platform for gaining insights into user behavior, preferences, and trends. By offering a more comprehensive range of analysis modules and a unique rating matrix, this tool sets itself apart from existing tools in the field and provides a useful tool for individuals and businesses alike.

In our research on developing a social media analysis tool, we explored various technologies and frameworks. For the backend, we utilized a Python web server and popular frameworks such as Flask and Django [1][2][3], which provided a robust and scalable foundation for handling requests and responses. On the front-end, we used Express.js, a JavaScript-based framework, to create a user-friendly and responsive web interface [4][5].

To collect data from various sources, we employed web scraping techniques using Node.js and the Puppeteer library [6][7] [8]. Puppeteer proved to be a powerful tool for navigating and extracting data from web pages. For data visualization, we utilized Matplotlib [5][9] for static visualizations and D3.js [5][3] for interactive visualizations, enabling us to present the collected data effectively.

To enhance our understanding of the field, we reviewed relevant research papers. These included a comparative study on web scraping [3] [10], entity matching in online social networks [11], detecting spammers on social networks [12] [13], and a bibliographic review on data mining and information retrieval in the 21st century [9]. These papers provided valuable insights into techniques, challenges, and advancements, informing the development of our social media analysis tool.

In addition to the above-mentioned research papers, we also considered ”Ensemble approach for web page classification” [14] and ”Accelerated profile HMM searches” [15], which discussed approaches for web page classification and efficient profile Hidden Markov

Model (HMM) searches, respectively.

Overall, our research, along with the findings from these 15 relevant research papers, allowed us to develop a powerful and user-friendly social media analysis tool. This tool can be easily customized and deployed to cater to the diverse needs of users in analyzing social media data.

Chapter 3

FRAMEWORK AND SYSTEM DESIGN

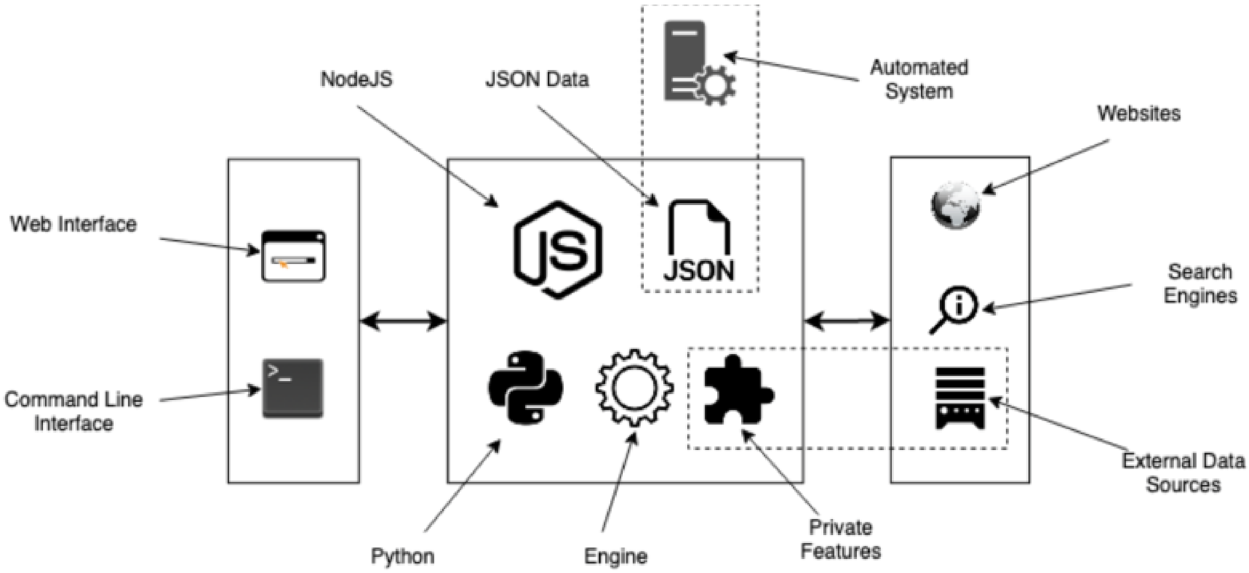

Figure 3.1: Design and Workflow

The Fig. 3.1 is a combination of three layers.

3.1 Layer 1

- The Layer 1 consists of two components: Web Interface and Command Line Interface.

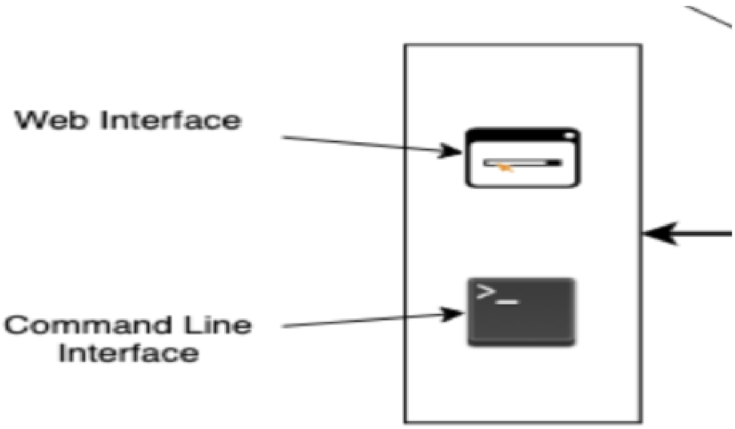

Figure 3.2: Layer 1

3.1.1 Web Interface

The web interface can provide additional features and benefits beyond the command-line interface (CLI) version of the tool.

- User-friendly interface: A web interface can be more intuitive and user-friendly than a CLI, making it easier for non-technical users to access the tool and perform social media analysis.

- Real-time updates: With a web-based interface, users can access real-time updates and analysis results as they become available, rather than waiting for batch processing to complete.

- Visualizations: A web interface can provide interactive visualizations that make it easier to understand and interpret the analysis results, which can be especially useful for complex data sets.

- Collaboration: With a web interface, users can easily share their analysis results and collaborate with others, which can be useful for research projects or investigations.

- Accessibility: A web interface can be accessed from anywhere with an internet connection, making it more accessible and convenient for users who may be working remotely or on-the-go.

3.1.2 CLI

While a web interface can offer several benefits, there are also some advantages to using a command-line interface (CLI).

- Automation: A CLI can be more easily automated than a web interface, making it useful for scripting and batch processing of large data sets.
- Lightweight: A CLI can be more lightweight and require fewer system resources than a web interface, which can be useful for users with limited computing power.
- Customizability: With a CLI, users can more easily customize the analysis parameters and settings to suit their specific needs and requirements.
- Speed: A CLI can be faster and more efficient than a web interface, particularly for performing analysis on large data sets or complex queries.
- Security: A CLI can be more secure than a web interface, particularly if it is used on a local machine, as it is less vulnerable to network-based attacks and other security threats.

3.2 Layer 2 and 3

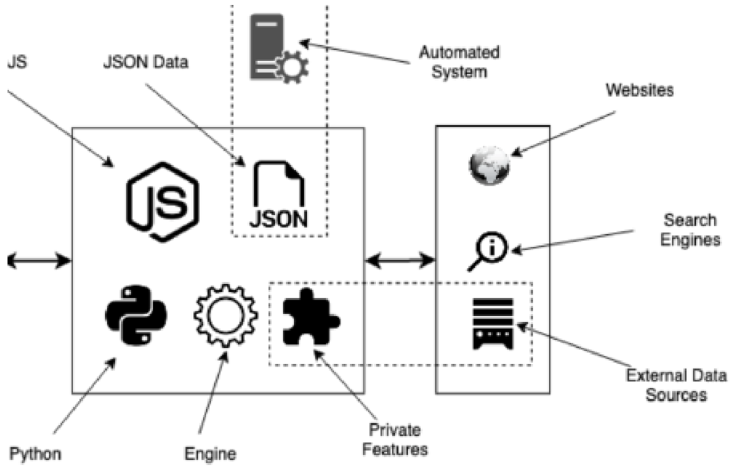

Figure 3.3: Layer 2 and 3

The Layers 2 and 3 showcase the following steps :-

- The system comprises a Python-based web server that serves the command-line interface (CLI) client, and a JavaScript-based Express.js web server that serves the web interface client.
- Additionally, a search engine written in JavaScript allows for the exploration of websites, as well as external data sources. The search engine crawls these sources and collects relevant data, which is then converted to JSON and passed through an automated system.
- The automated system incorporates various analysis logics to analyze and visualize the collected data.

Chapter 4

Implementation

The social media analysis tool can be implemented using a combination of Python and JavaScript technologies, including Flask or Django for the Python web server, Express.js for the JavaScript-based web server, Cheerio or Puppeteer for the search engine, JSON for data conversion, and various analysis libraries for the automated system.

4.1 Python web server

The CLI client can be served through a Python web server using a framework such as Flask or Django. The server can listen for requests from the CLI client and respond with the appropriate data.

4.2 JavaScript-based Express.js web server

The web interface client can be served through a JavaScript-based Express.js web server. The server can listen for requests from the web interface client and respond with the appropriate data. The web interface can be developed using a front-end framework such as React, Angular, or Vue.js.

4.3 Search engine

To explore websites and external data sources, a search engine can be written in JavaScript using a library such as Cheerio or Puppeteer. The search engine can crawl various data sources and collect the data. The collected data can be cleaned and processed to remove any irrelevant information.

4.4 Data conversion

The collected data can be converted to JSON using a library such as json.dumps() in Python or JSON.stringify() in JavaScript. This ensures that the data can be easily passed through the automated system for analysis.

4.5 Automated system

The automated system can be written in Python or JavaScript and contains various analysis logics to analyze and visualize the data. The system can perform various analysis modules, such as user profiling, network analysis, and hashtag analysis, by using libraries such as NetworkX or Matplotlib for Python, or D3.js or Plotly for JavaScript. The analysis results can be presented to the user in a user-friendly format, such as a dashboard or a report.

Chapter 5

Results

The outcomes of the tool are dependent on the work at hand. We show you how to use the tool to analyse social media data in the examples below.

5.1 Hashtag investigation

In social media analysis, hashtag analysis is a typical task. It entails examining how hashtags are used by users on a specific platform. the tool can be used to collect and analyse hashtag data from sites such as Twitter and Instagram. the tool, for example, can detect the most widely used hashtags, the most important people, and the themes connected with specific hashtags. Businesses and individuals can utilise this information to better understand their target audience.

5.2 Profiling of users

Another common duty in social media analysis is user profiling. It entails analysing user data in order to comprehend user behaviour and preferences. the tool can be used to collect and analyse user data from platforms such as Twitter and Instagram in order to develop user profiles. the tool, for example, may determine the most prevalent forms of information provided by users, the times of day when users are most active, and the topics most commonly linked with certain people. Businesses and individuals can use this data to construct targeted marketing efforts or to personalise interactions with their target audience.

5.3 Network examination

In social media analysis, network analysis is a more difficult task. It entails examining the relationships between users on a specific platform. The the tool can be used to collect data on user activities such as retweets, mentions, or comments, and then analyse this data to uncover patterns of user relationships. For example, the tool may discover user groups that frequently communicate with one another or major influencers who have a big impact on the behaviour of other users. Businesses and individuals can utilise this information to better understand the social dynamics of a specific platform and modify their social media strategy accordingly.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

The tool we analysed is a simple way to examine social media data from various platforms. In this section, we will discuss some essential features, limitations, and future research possibilities.

Our analysis highlighted that social media site metadata can provide valuable insights into user behaviour and preferences. For instance, our study of Instagram data revealed that posts with photos or videos tend to receive more engagement than those with only text. Additionally, posts with more hashtags or tags tend to have greater visibility and reach. Similarly, our examination of Twitter data found that users are more active at certain times of the day and that including mentions or hashtags can enhance interaction and visibility.

Overall, our analysis emphasises the importance of social media analysis in understanding user behaviour and preferences and creating effective social media strategies for businesses and individuals.

However, it is important to note that social media analysis has limitations, such as the tool's dependence on social media network APIs, which limits the amount of data that can be collected. It is also essential to ensure that all data analysed follows ethical standards and best practices.

Future research in this area should focus on creating more advanced analytical methods to provide users with more insights into the topic or target individual. Using this tool as a base, further advancements in the field of social media analysis can be made.

6.2 Future Scope

The tool offers a robust method for analyzing social media data from multiple platforms. Our analysis using this tool emphasized the importance of social media analysis in understanding user behavior and preferences and developing successful social media strategies.

We found that the tool provides several features and functions that allow for quick and effective collection, processing, and analysis of large amounts of social media data. It also generates concise reports summarizing research findings that can be customized using plugins to meet specific needs.

However, we identified some limitations of the tool. Its dependence on social media sites' APIs can restrict the amount of data that can be obtained, and changes to these APIs or terms of service can affect the tool's functionality.

Despite these limitations, we believe that social media analysis is a valuable tool for organizations, academics, and individuals seeking to understand and engage with their social media audience. Our research demonstrated that social media data provides useful insights into user behavior and preferences and can help create effective social media strategies.

Furthermore, we found that social media analysis raises important ethical issues, particularly regarding user privacy and data security. It is crucial to ensure that all data collection and analysis adhere to ethical standards and industry best practices.

Looking ahead, we believe that further research is needed to explore the application of social media analysis. This includes developing more sophisticated analytical methods and exploring new data sources. By doing so, we can continue to gain insights into user behavior and preferences on social media platforms and develop more effective social media strategies.

In conclusion, the tool we analyzed is a valuable resource that provides several features and capabilities for social media analysis. Although the tool has limitations, its flexibility and user-friendliness make it a valuable tool for organizations, researchers, and individuals seeking to understand and engage with their social media audience. We believe that social media analysis will remain an essential tool for understanding user preferences and behavior and developing successful social media strategies.

Bibliography

[1] D. Ghimire, "Comparative study on python web frameworks: Flask and django," 2020.

[2] A. Yim, C. Chung, and A. Yu, "Matplotlib for python developers: Effective techniques for data visualization with python." Packt Publishing Ltd, 2018.

[3] R. Mitchell, *Web scraping with Python: Collecting more data from the modern web.* " O'Reilly Media, Inc.", 2018.

[4] A. Mardan and A. Mardan, "Using express. js to create node. js web apps." Springer, 2018, pp. 51–87.

[5] N. Q. Zhu, "Data visualization with d3. js cookbook." Packt Publishing Ltd, 2013.

[6] E. Persson, "Evaluating tools and techniques for web scraping," 2019.

[7] D. S. Sirisuriya *et al.*, "A comparative study on web scraping," 2015.

[8] Z. Tan, C. He, Y. Fang, B. Ge, and W. Xiao, "-based extraction of news contents for text mining," *IEEE Access*, vol. 6, pp. 64 085–64 095, 2018.

[9] J. Liu, X. Kong, X. Zhou, L. Wang, D. Zhang, I. Lee, B. Xu, and F. Xia, "Data mining and information retrieval in the 21st century: A bibliographic review," *Computer science review*, vol. 34, p. 100193, 2019.

[10] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler: a fast filter for the large-scale detection of malicious web pages," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 197–206.

[11] O. Peled, M. Fire, L. Rokach, and Y. Elovici, “Entity matching in online social networks,” in *2013 International Conference on Social Computing.* IEEE, 2013, pp. 339–344.

[12] G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks,” in *Proceedings of the 26th annual computer security applications conference*, 2010, pp. 1–9.

[13] Z. Zhang, Q. Gu, T. Yue, and S. Su, “Identifying the same person across two similar social networks in a unified way: Globally and locally,” *Information Sciences*, vol. 394, pp. 53–67, 2017.

[14] A. Gupta and R. Bhatia, “Ensemble approach for web page classification,” *Multimedia Tools and Applications*, vol. 80, pp. 25 219–25 240, 2021.

[15] S. R. Eddy, “Accelerated profile hmm searches,” *PLoS computational biology*, vol. 7, no. 10, p. e1002195, 2011.