

DARK WEB CRAWLER

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

N HARSHA VARDHAN REDDY (40110821)

M V SAI KUMAR REDDY (40110755)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

CATEGORY -1 UNIVERSITY BY UGC

Accredited with Grade "A++" by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHISALAI, CHENNAI - 600119

APRIL - 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Category - I University by UGC

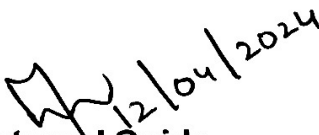
Accredited "A++" by NAAC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

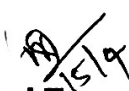
This is to certify that this Project Report is the bonafide work of **N HARSHA VARDHAN REDDY(40110821)** and **M V SAI KUMAR REDDY(40110755)** who carried out the Project entitled "**DARK WEB CRAWLER**" under my supervision from November 2023 to April 2024.

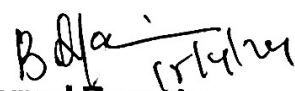

12/04/2024
Internal Guide

Dr. MINU SUSAN JACOB, Ph.D.


Head of the Department
HEAD OF THE DEPARTMENT
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SATHYABAMA Ph.D.
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Jeppiaar Nagar, Rajiv Gandhi Salai,
Chennai - 600 119.

Submitted for Viva voce Examination held on 15-04-2024


15/4
Internal Examiner


15/4/24
External Examiner

DECLARATION

I, N HARSHA VARDHAN REDDY(40110821), hereby declare that the Project Report entitled **DARK WEB CRAWLER** done by us under the guidance of **Dr. Minu Susan Jacob Ph.D.** at sathyabama institute of science and technologies is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

DATE: 15/4/24

PLACE: Chennai

N. Harsha Vardhan Reddy
SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to The **Board of Management of Sathyabama Institute of Science and Technology** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph. D, Dean**, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.**, Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.Minu Susan Jacob Ph.D**, for him valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Computer Science and Engineering who were helpful in many ways for the completion of the project.

ABSTRACT

The Dark Web is a hidden part of the internet that is not accessible through regular search engines. It is known for its anonymity and is often associated with illegal activities such as drug trafficking, cyber-attacks, and the sale of stolen data. In order to explore this secretive realm, researchers and law enforcement agencies rely on tools known as Dark Web Crawlers. A Dark Web Crawler is a specialized software program designed to navigate through the Dark Web and collect information on websites, forums, and other online platforms. These crawlers employ advanced techniques to discover and index content that is not easily accessible through conventional means. They assist in identifying potential threats, monitoring illegal activities, and gathering intelligence for investigative purposes. Dark Web Crawlers play a crucial role in maintaining cyber-security and combating criminal activities on the Dark Web. By employing advanced algorithms and intelligent search techniques, these powerful tools provide valuable insights and help protect individuals, organizations, and society as a whole. A portion of the internet that is hidden and inaccessible using standard search engines is known as the "Dark Web." Because of its anonymity, it is frequently linked to unlawful operations like drug trafficking, computer hacking, and the sale of stolen data. Researchers and law enforcement organizations use tools known as Dark Web Crawlers to investigate this covert domain. A specific piece of software called a "Dark Web Crawler" is intended to explore the Dark Web and gather data from blogs, forums, and other online resources. These crawlers use cutting-edge methods to find and index content that is difficult to reach through standard channels. They aid in spotting potential dangers, keeping an eye on illicit activity, and gathering information for forensic purposes. In order to protect cyber security and stop criminal activity on the Dark Web, Dark Web Crawlers are essential. These potent technologies offer useful insights and aid in the protection of people, companies, and society at large by utilizing cutting-edge algorithms and intelligent search strategies

TABLE OF CONTENTS

Chapter No	TITLE	Page No.
	ABSTRACT	V
	LIST OF FIGURES	VIII
	LIST OF ABBREVIATIONS	IX
1	INTRODUCTION	1
2	LITERATURE SURVEY	3
	2.1 Inferences from Literature Survey	
	2.2 Existing System in Proposed System	5
	2.3 Open problems in Existing System	6
3	REQUIREMENTS ANALYSIS	10
	3.1 Risk Analysis of the Project	10
	3.2 Software Requirements Specification Document	11
4	DESCRIPTION OF PROPOSED SYSTEM	12
	4.1 Flow chart of process using machine learning	13
	4.2 Selected Methodology or process model	13
	4.3 Architecture / Overall Design of Proposed System	15
	4.4 Description of Software for Implementation and Testing plan of the Proposed Model/System	16
	4.5 Project Management Plan	20
5	IMPLEMENTATION DETAILS	
6	RESULTS AND DISCUSSIONS	25

7	CONCLUSION	31
	7.1 Conclusion	31
	7.2 Future Work	31
	REFERENCES	32
	APPENDIX	34
	A. SOURCE CODE	34
	B. SCREENSHOTS	40
	C. RESEARCH PAPER	45
	D. CERTIFICATE	52

LIST OF FIGURES

FIGURE NO	FIGURE NAME	Page No.
4.1	Flow chart	13
4.2	System Architecture	15
4.5	Project Management Plan	20
6.1	Result For First Crawl	26
6.2	Result For Second Crawl	28
6.3	Result For Third Crawl	30
B.1	Main Page	40
B.2	Token Generation Page	40
B.3	Keys Page	41
B.4	Links Targeting Page	41
B.5	Dash Board For Link Extraction	42
B.6	Information Extraction	42
B.7	Classification Extraction	43
B.8	Pie Chart For Lables	43
B.9	Crawler For Database	44

LIST OF ABBREVIATIONS

S NO	ABBREVIATION	EXPANSION
01	AI	Artificial Intelligence
02	API	Application Programming Interface
04	UI	User Interface
05	UX	User Experience

CHAPTER 1

INTRODUCTION

The Dark Web Crawler, an innovative tool in the realm of digital exploration, has emerged as a powerful assistant in navigating the hidden corners of the internet. With its advanced capabilities, this cutting-edge technology enables users to delve into the mysterious depths of the dark web, uncovering hidden information, and shedding light on obscured activities and communities.

Designed to operate in the anonymous and encrypted networks of the dark web, the Dark Web Crawler utilizes specialized algorithms and techniques to scan and index the vast expanse of this concealed landscape. Unlike traditional search engines that operate on the surface web, this tool is specifically tailored to comprehend and navigate the complex network of Tor and I2P networks, unearthing hidden websites, forums, marketplaces, and other digital entities.

One of the primary objectives of the Dark Web Crawler is to provide unparalleled access to information that is typically unattainable through conventional means. With a comprehensive and ever-expanding database, it empowers users to obtain valuable insights into the hidden side of the internet, facilitating research, investigations, and intelligence gathering. Whether it be for law enforcement agencies combating cybercrime, journalists investigating illicit activities, or security professionals seeking vulnerabilities, this tool grants unprecedented visibility into the underbelly of the online world.

Moreover, the Dark Web Crawler ensures user anonymity and security in accordance with the principles of the dark web. By leveraging encryption and privacy-focused techniques, it shields users from unwanted surveillance, protects sensitive data, and safeguards the integrity of the exploration process. This commitment to maintaining user privacy is instrumental in building trust and encouraging widespread adoption of this invaluable tool.

While exploring the dark web can be a daunting and unfamiliar experience for many, the Dark Web Crawler aims to simplify the process and make it more approachable. Its user-friendly interface and intuitive controls allow users, regardless of their technical expertise, to seamlessly navigate through the dark web's labyrinthine web of hidden content. By offering a streamlined and efficient search experience, this tool eliminates the barriers to entry often associated with delving into the dark web, enabling.

In conclusion, the Dark Web Crawler represents a groundbreaking advancement in the field of digital exploration. Its ability to comprehend the hidden layers of the internet facilitates access to vital information, empowers various sectors, and protects user privacy. As the dark web continues to evolve, this indispensable tool serves as a guiding light, illuminating the depths of this mysterious realm and unveiling the secrets that lay within.

The Dark Web Crawler is a game-changing development in the area of digital exploration, to sum up. Its capacity to comprehend the internet's secret layers makes it easier to obtain important data, gives various sectors more power, and safeguards user privacy. This vital tool serves as a guiding light, illuminating the depths of this enigmatic world and revealing the secrets that lay within as the dark web continues to develop.

The Dark Web Crawler uses specific algorithms and techniques to scan and index the huge expanse of this hidden world. It is designed to operate in the anonymous and encrypted networks of the dark web.

This tool is specifically designed to comprehend and navigate the intricate network of Tor and I2P networks, uncovering secret websites, forums, marketplaces, and other digital entities. In contrast to conventional search engines that operate on the surface web.

CHAPTER 2

LITERATURE SURVEY

2.1 INFERENCES FROM LITREATURE SURVEY

1. Dark web illegal activities crawling and classifying using data mining techniques.

- Authors: Alaidi, A. H. M., Roa'a, M., ALRikabi, H. T. S., Aljazaery, I. A., & Abbood, S. H.
- Published in: iJIM (International Journal of Interactive Mobile Technologies) (2022)
- Explanation: This paper discusses the use of data mining techniques for crawling and classifying illegal activities on the dark web. It explores methods to identify and categorize illicit content.

2. Exploring Dark Web Crawlers: A systematic literature review of dark web crawlers and their implementation.

- Authors: Bergman, J., & Popov, O. B.
- Published in: IEEE Access (2023)
- Explanation: This literature review provides insights into the various dark web crawlers and their implementations. It explores the state of the art in crawling the dark web.

3. SpyDark: Surface and Dark Web Crawler.

- Authors: Dalvi, A., Paranjpe, S., Amale, R., Kurumkar, S., Kazi, F., & Bhirud, S. G.
- Published in: 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC) (2021)
- Explanation: This paper presents SpyDark, a crawler designed for both surface and dark web crawling. It discusses its capabilities and features for cybersecurity purposes.

4. A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence.

- Authors: Koloveas, P., Chantzios, T., Tryfonopoulos, C., & Skiadopoulos, S.
- Published in: 2019 IEEE World Congress on Services (SERVICES) (2019)
- Explanation: This research presents a crawler architecture focused on harvesting content from the clear web, social web, and dark web. It particularly targets cyber-threat intelligence related to IoT (Internet of Things).

5. Crawling the dark web: A conceptual perspective, challenges and implementation.

- Authors: AlKhatib, B., & Basheer, R.
- Published in: Journal of Digital Information Management (2019)
- Explanation: This paper offers a conceptual perspective on crawling the dark web, highlighting the challenges and implementation considerations involved in this task.

6. Dark Web Crawling for Cybersecurity: Insights into Vulnerabilities and Ransomware Discussions.

- Authors: Dalvi, A., Kulkarni, P., Kore, A., & Bhirud, S. G.
- Published in: 2023 2nd International Conference for Innovation in Technology (INOCON) (2023)
- Explanation: This research focuses on dark web crawling for cybersecurity purposes. It provides insights into vulnerabilities and discussions related to ransomware on the dark web.

7. A Review of Dark Web: Trends and Future Directions.

- Authors: Sobhan, S., Williams, T., Faruk, M. J. H., Rodriguez, J., Tasnim, M., Mathew, E., & Shahriar, H.
- Published in: 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC) (2022)
- Explanation: This review paper discusses the trends and future directions in the context of the dark web. It provides insights into the evolving landscape of the dark web.

8. Using dark web crawler to uncover suspicious and malicious websites.

- Authors: Pannu, M., Kay, I., & Harris, D.
- Published in: Advances in Human Factors in Cybersecurity (2019)
- Explanation: This paper explores the use of dark web crawlers to uncover suspicious and malicious websites. It discusses the potential applications of crawling for cybersecurity.

9. Dark web—onion hidden service discovery and crawling for profiling morphing, unstructured crime and vulnerabilities prediction.

- Authors: Rawat, R., Rajawat, A. S., Mahor, V., Shaw, R. N., & Ghosh, A.
- Published in: Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021 (2021)
- Explanation: This work focuses on dark web crawling for discovering onion hidden services. It aims to profile morphing and unstructured crime while predicting vulnerabilities.

10. Hadoop-based dark web threat intelligence analysis framework.

- Authors: Yang, Y., Yu, H., Yang, L., Yang, M., Chen, L., Zhu, G., & Wen, L.
- Published in: 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) (2019)
- Explanation: This paper introduces a Hadoop-based framework for analyzing dark web threat intelligence. It explores the use of big data technologies for understanding dark web threats.

2.2 EXISTING SYSTEM AND PROPOSED SYSTEM

- **Existing System:**

The existing system for Dark Web crawlers is riddled with several disadvantages that hinder their efficiency and effectiveness. One major drawback is the lack of scalability. Current Dark Web crawlers face challenges when it comes to handling the vast and continuously expanding size of the Dark Web. With each passing day, new websites and content are added, making it difficult for crawlers to keep up and ensure coverage.

Another disadvantage is the issue of accessibility. The nature of the Dark Web, with its hidden services and anonymity, poses challenges for crawlers to access and retrieve information. Many websites on the Dark Web require specialized access tools, such as Tor, which can complicate the crawling process. Additionally, a significant portion of the Dark Web is protected by encryption and authentication methods, which further limits the crawler's ability to extract meaningful data.

- **Proposed System**

A Dark Web crawler is a software tool specifically designed to navigate the Dark Web, a part of the internet that is not indexed by traditional search engines and is known for its non-indexed content and anonymity. The proposed work for a Dark Web crawler involves several key steps and considerations.

Firstly, the crawler would need to be developed with robust capabilities to handle the unique challenges of the Dark Web. This includes the ability to navigate through hidden services, utilize Tor or similar anonymity networks, and effectively handle the encryption and secure communication protocols commonly used in the Dark Web.

Next, the crawler would need to implement different crawling strategies to efficiently explore the Dark Web. This could involve focusing on specific keywords or topics, following links from known sources, or dynamically adjusting the crawl depth based on the discovered content. The crawler should also be able to handle the large volume of data typically found in the Dark Web, by employing techniques like distributed crawling or data prioritization based on relevance or importance.

2.3 OPEN PROBLEMS IN EXISTING SYSTEM

- **Anonymity and Privacy Preservation:**

- Problem: Dark web users rely on anonymity to protect their identity. Crawlers must respect this anonymity while still collecting relevant data.
- Solution: Develop techniques for ethical and privacy-preserving crawling that minimize intrusion into user privacy.

- **Content Verification and Trustworthiness:**

- Problem: Much of the information on the dark web is unverified or potentially fraudulent. Crawl results need mechanisms to assess content reliability.
- Solution: Implement content verification algorithms or integrate with reputation systems to assess the trustworthiness of sources.

- **Dynamic Content and Rapid Changes:**

- Problem: Dark web content is highly dynamic and frequently changes addresses. Keeping up-to-date with new content and avoiding broken links is a challenge.
- Solution: Develop strategies for real-time monitoring of dark web sources and adapt to rapid changes in content and URLs.

- **Hidden Services Discovery:**

- Problem: Discovering new hidden services and onion links is crucial for comprehensive crawling but is challenging due to their secretive nature.
- Solution: Investigate techniques for automated discovery of hidden services or collaboration with trusted sources for link sharing.

- **Multilingual Content and Translation:**

- Problem: Dark web content is often multilingual, making it difficult to understand and analyze. Language barriers can hinder effective crawling.
- Solution: Implement language detection and translation capabilities to make multilingual content accessible for analysis.

- **Deep Web vs. Dark Web Differentiation:**

- Problem: Distinguishing between the deep web (legitimate, unindexed content) and the dark web (illicit, intentionally hidden content) is challenging.
- Solution: Develop techniques to differentiate between deep and dark web content to focus crawling efforts effectively.

- **Content Classification and Filtering:**

- Problem: Dark web content varies widely, from illegal activities to forums for

legitimate discussions. Filtering relevant content while avoiding illegal material is crucial.

- Solution: Implement content classification algorithms to filter and categorize crawled data while avoiding illegal content.

- **Robustness Against Adversarial Measures:**

- Problem: Dark web operators may employ countermeasures against crawlers, such as IP blocking or misinformation. Crawlers need to be resilient.
- Solution: Enhance crawler robustness by implementing techniques like IP rotation, CAPTCHA solving, and fingerprinting resistance.

- **Ethical Considerations and Legal Compliance:**

- Problem: Crawling the dark web raises ethical and legal concerns. Ensuring compliance with relevant laws and ethical guidelines is essential.
- Solution: Establish clear ethical guidelines for dark web crawling and ensure compliance with applicable laws and regulations.

- **Data Volume and Storage:**

- Problem: Dark web data can be vast and challenging to store and process. Scalable storage and data management solutions are needed.
- Solution: Explore distributed storage and processing solutions to handle large volumes of dark web data effectively.

- **User Behavior and Interaction:**

- Problem: Understanding user behavior and interactions on the dark web is essential for threat analysis and profiling.
- Solution: Develop techniques to capture and analyze user interactions and behavior within dark web forums and marketplaces.

- **Integration with Threat Intelligence:**

- Problem: Dark web crawling generates valuable threat intelligence data. Integrating this data with broader cybersecurity threat intelligence systems is crucial.
- Solution: Develop standard data formats and APIs for sharing dark web threat

intelligence with broader cybersecurity communities.

- **Evaluation Metrics:**

- Problem: Measuring the effectiveness of dark web crawlers is challenging due to the lack of ground truth data and the hidden nature of the content.
- Solution: Define comprehensive evaluation metrics and benchmarks for assessing the performance and coverage of dark web crawlers.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 RISK ANALYSIS OF THE PROJECT

FEASIBILITY STUDY

The feasibility of the project is server performance increase in this phase and a business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis, the feasibility study of the proposed system is to be carried out. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- **ECONOMICAL FEASIBILITY**

This study is carried out to check the economic impact that the system will have on the organization. The amount of funds that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

- **TECHNICAL FEASIBILITY**

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands being placed on the client. The developed system must have modest requirements, as only minimal or null changes are required for implementing this system.

- **OPERATIONAL FEASIBILITY**

The aspect of the study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. User must

not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system

3.2 PYTHON:

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python can be used on a server to create web applications.
- Python can be used alongside software to create workflows.
- Python can connect to database systems. It can also read and modify files.

3.3 Streamlit Frame Work:

Streamlit is an open-source app framework in python language. It helps us create beautiful web apps for data science and machine learning in a little time. It is compatible with major python libraries such as scikit-learn, keras, PyTorch, latex, numpy, pandas, matplotlib, etc. Syntax for installing this library is shown below.

EXECUTION:

1. Set up your python development environment.
2. pip install streamlit.
3. Then run using this query: streamlit run module.py

3.4 SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

- **Hardware specifications:**

- Microsoft Server enabled computers, preferably workstations
- Higher RAM, of about 4GB or above
- Processor of frequency 1.5GHz or above

- **Software specifications:**

- Python 3.6 and higher
- Anaconda software

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

A Dark Web crawler is a software tool specifically designed to navigate the Dark Web, a part of the internet that is not indexed by traditional search engines and is known for its non-indexed content and anonymity. The proposed work for a Dark Web crawler involves several key steps and considerations.

Firstly, the crawler would need to be developed with robust capabilities to handle the unique challenges of the Dark Web. This includes the ability to navigate through hidden services, utilize Tor or similar anonymity networks, and effectively handle the encryption and secure communication protocols commonly used in the Dark Web.

Next, the crawler would need to implement different crawling strategies to efficiently explore the Dark Web. This could involve focusing on specific keywords or topics, following links from known sources, or dynamically adjusting the crawl depth based on the discovered content. The crawler should also be able to handle the large volume of data typically found in the Dark Web, by employing techniques like distributed crawling or data prioritization based on relevance or importance.

In order to efficiently browse and gather vital information from the secretive and anonymous world of the Dark Web, the proposed work for a Dark Web crawler calls for a combination of cutting-edge technical abilities, strong security measures, and clever crawling tactics. The development of a Dark Web crawler requires a comprehensive understanding of the unique challenges and complexities inherent in this clandestine corner of the internet. Let's delve deeper into the various processes and considerations involved in this ambitious endeavor.

4.1 Flow Chart of process using ML:

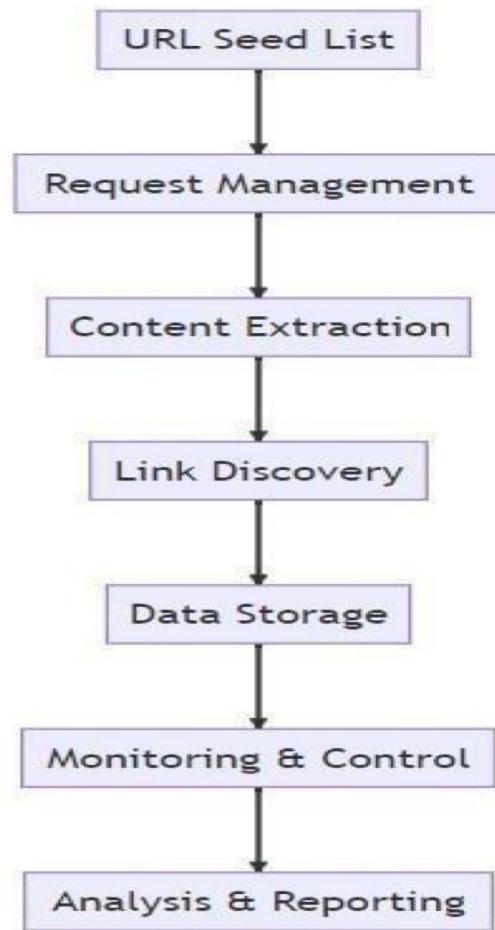


Fig 4.1 FLOW CHART

4.2 SELECTED METHODOLOGY OR PROCESS MODEL

4.2.1 Web Crawling Module:

The web crawling module is the core component of the dark web crawler system. This module is responsible for automatically fetching data from the dark web. It employs advanced crawling techniques to navigate through the hidden layers of the internet and extract relevant information. The module utilizes specialized algorithms and heuristics to efficiently discover and index dark web sites. It explores different routes, including Tor networks and other anonymizing technologies, to reach hidden services. The web crawling module also incorporates advanced techniques to bypass security measures and anti-crawling mechanisms, ensuring that as much data as possible is collected. It can handle a large volume of data and is designed to continuously crawl the dark web to keep the system up-to-date with the latest

information.

4.2.2 Data Processing Module:

The data processing module plays a crucial role in extracting meaningful information from the raw data collected by the web crawling module. This module applies various data mining and text analysis techniques to process, filter, and categorize the crawled data. It employs natural language processing algorithms to extract entities, keywords, and sentiment from the text. It also utilizes machine learning algorithms to recognize patterns and detect anomalies in the collected data. The processed data is then stored in a structured format, facilitating efficient retrieval and analysis. This module also ensures data integrity and takes measures to anonymize any personal or sensitive information that may be present in the collected data, while still retaining its usefulness for further analysis.

4.2.3 User Interface Module:

The user interface module provides an intuitive and user-friendly interface for interacting with the dark web crawler system. It allows users to input search queries and specify search criteria to retrieve relevant results from the collected data. The module presents the results in a clear and organized manner, allowing users to browse through the retrieved information easily. It also includes advanced search and filtering options to narrow down the results based on various parameters such as date, source, or relevance. The user interface module may incorporate visualizations, such as charts or graphs, to present aggregated data and insights in a more understandable format. It also provides functionality for exporting and sharing the retrieved data and analysis results. Additionally, this module may include features for user management, system configuration, and customization, ensuring a personalized and efficient user experience for both novice and expert users.

4.3 ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

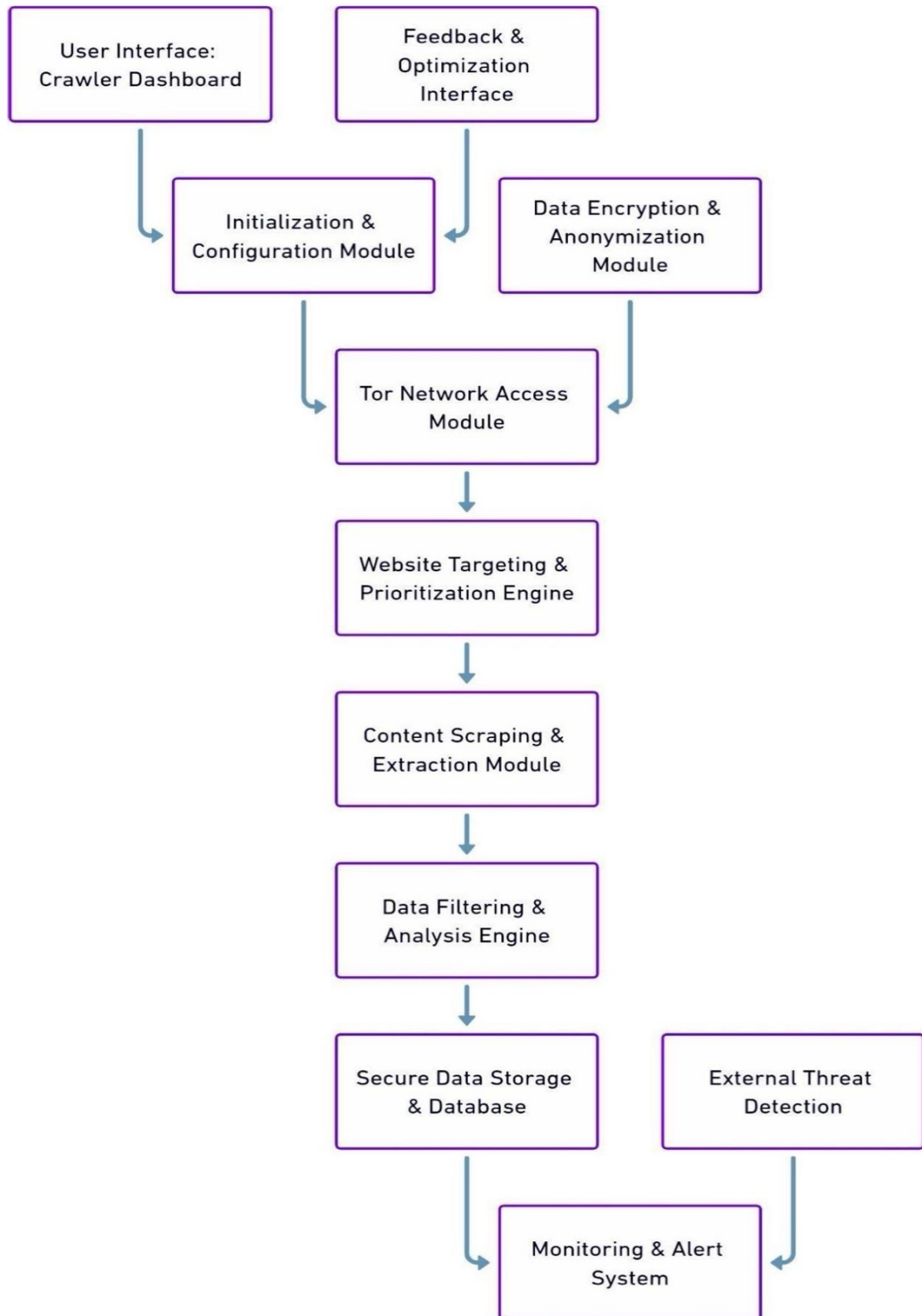


Fig 4.2 SYSTEM ARCHITECTURE

4.4 DESCRIPTION OF SOFTWARE FOR IMPLEMENTATION AND TESTING PLAN OF THE PROPOSED MODEL/SYSTEM

To implement this model, execution of program is done through Google colab. Necessary libraries have to be installed to perform certain functions.

DESCRIPTION OF PROGRAMMING LANGUAGE AND SOFTWARE

PYTHON

Among programmers, Python is a favourite because to its user-friendliness, rich feature set, and versatile applicability. Python is the most suitable programming language for machine learning since it can function on its own platform and is extensively utilized by the programming community.

Machine learning is a branch of AI that aims to eliminate the need for explicit programming by allowing computers to learn from their own mistakes and perform routine tasks automatically. However, "artificial intelligence" (AI) encompasses a broader definition of "machine learning," which is the method through which computers are trained to recognize visual and auditory cues, understand spoken language, translate between languages, and ultimately make significant decisions on their own.

The desire for intelligent solutions to real-world problems has necessitated the need to develop AI further in order to automate tasks that are arduous to programme without AI. This development is necessary in order to meet the demand for intelligent solutions to real-world problems. Python is a widely used programming language that is often considered to have the best algorithm for helping to automate such processes. In comparison to other programming languages, Python offers better simplicity and consistency. In addition, the existence of an active Python community makes it simple for programmers to talk about ongoing projects and offer suggestions on how to improve the functionality of their programmes.

ANACONDA

Anaconda is an open-source package manager for Python and R. It is the most popular platform among data science professionals for running Python and R implementations. There are over 300 libraries in data science, so having a robust distribution system for them is a must for any professional in this field. Anaconda simplifies package deployment and management.

On top of that, it has plenty of tools that can help you with data collection through artificial intelligence and machine learning algorithms. With Anaconda, you can easily set up, manage, and share Conda environments. Moreover, you can deploy any required project with a few clicks when you're using Anaconda. There are many advantages to using Anaconda and the following are the most prominent ones among them: Anaconda is free and open-source.

This means you can use it without spending any money. In the data science sector, Anaconda is an industry staple. It is open-source too, which has made it widely popular. If you want to become a data science professional, you must know how to use Anaconda for Python because every recruiter expects you to have this skill. It is a must-have for data science.

It has more than 1500 Python and R data science packages, so you don't face any compatibility issues while collaborating with others. For example, suppose your colleague sends you a project which requires packages called A and B but you only have package A.

You can easily collaborate on projects without worrying about any compatibility issues. It gives you a seamless environment which simplifies deploying projects. You can deploy any project with just a few clicks and commands while managing the rest. Anaconda has a thriving community of data scientists and machine learning professionals who use it regularly. If you encounter an issue, chances are, the community has already answered the same.

On the other hand, you can also ask people in the community about the issues you face there, it's a very helpful community ready to help new learners. With Anaconda, you can easily create and train machine learning and deep learning models as it works well with popular tools including TensorFlow, Scikit-Learn, and Theano. You can create visualizations by using Bokeh, Holoviews, Matplotlib, and Datashader while using Anaconda.

How to Use Anaconda for Python

Now that we have discussed all the basics in our Python Anaconda tutorial, let's discuss some fundamental commands you can use to start using this package manager.

Listing All Environments

To begin using Anaconda, you'd need to see how many Conda environments are present in your machine.

```
conda env list
```

It will list all the available Conda environments in your machine.

Creating a New Environment

You can create a new Conda environment by going to the required directory and use this command:

```
conda create -n <your_environment_name>
```

You can replace `<your_environment_name>` with the name of your environment. After entering this command, conda will ask you if you want to proceed to which you should reply with y:

```
proceed ([y])/n)?
```

On the other hand, if you want to create an environment with a particular version of Python, you should use the following command:

```
conda create -n <your_environment_name> python=3.6
```

Similarly, if you want to create an environment with a particular package, you can use the following command:

```
conda create -n <your_environment_name> pack_name
```

Here, you can replace `pack_name` with the name of the package you want to use.

If you have a `.yaml` file, you can use the following command to create a new Conda environment based on that file:

```
conda env create -n <your_environment_name> -f <file_name>.yaml
```

We have also discussed how you can export an existing Conda environment to a `.yaml` file later in this article.

Activating an Environment

You can activate a Conda environment by using the following command:

```
conda activate <environment_name>
```


You should activate the environment before you start working on the same. Also, replace the term `<environment_name>` with the environment name you want to activate. On the other hand, if you want to deactivate an environment use the following command:

```
conda deactivate
```

Installing Packages in an Environment

Now that you have an activated environment, you can install packages into it by using the following command:

```
conda install <pack_name>
```

Replace the term `<pack_name>` with the name of the package you want to install in your Conda environment while using this command.

Exporting an Environment Configuration

Suppose you want to share your project with someone else (colleague, friend, etc.). While you can share the directory on Github, it would have many Python packages, making the transfer process very challenging. Instead of that, you can create an environment configuration `.yaml` file and share it with that person. Now, they can create an environment like your one by using the `.yaml` file.

For exporting the environment to the `.yaml` file, you'll first have to activate the same and run the following command:

```
conda env export ><file_name>.yaml
```

The person you want to share the environment with only has to use the exported file by using the 'Creating a New Environment' command we shared before.

Removing a Package from an Environment

If you want to uninstall a package from a specific Conda environment, use the following command:

```
conda remove -n <env_name><package_name>
```

On the other hand, if you want to uninstall a package from an activated environment, you'd have to use the following command:

```
conda remove <package_name>
```

Deleting an Environment

Sometimes, you don't need to add a new environment but remove one. In such cases, you must know how to delete a Conda environment, which you can do so by using the following command:

```
conda env remove --name <env_name>
```

The above command would delete the Conda environment right away.

4.5 PROJECT MANAGEMENT PLAN

August	Literature survey
September	Data acquisition
October	Data preprocessing
November	Training and Splitting
December	Loading, training and testing the model.
January	Predicting the output and generating the final report

4.5 PROJECT MANAGEMENT PLAN

CHAPTER 5

IMPLEMENTATION DETAILS

This section allows you to showcase the effectiveness of your approach and provide insights into the implications of your research. Here is an outline of what you may include in this section:

- **Authentication Method Evaluation:**

- Provide the results of your evaluation of different authentication methods, including traditional passwords, biometrics, multi-factor authentication, and others.
- Include quantitative data on factors such as security, usability, and user satisfaction.
- Discuss the strengths and weaknesses of each method based on your findings.

- **User Profile Management Effectiveness:**

- Present the results of your analysis of user profile management systems, including user data protection, privacy controls, and user experience.
- Include metrics related to data security, user control, and compliance with privacy regulations.
- Discuss how different user profile management strategies impact user trust and satisfaction.

- **Security Metrics and Performance:**

- Share the security metrics and performance indicators you measured during your research, such as authentication success rates, false positives, false negatives, and response times.
- Discuss the trade-offs between security measures and system performance.

- **User-Centric Evaluation:**

- Describe the outcomes of user-centric evaluations, including user surveys,

interviews, or usability tests.

- Highlight user preferences, concerns, and feedback related to authentication method.
- **Privacy and Compliance Metrics:**
 - Present metrics related to data privacy and regulatory compliance, including how well the systems adhere to data protection laws (e.g., GDPR, CCPA).
 - Discuss any privacy-preserving technologies or mechanisms you implemented and their impact on user privacy.
- **Comparative Analysis:**
 - Compare the performance, security, and usability of different authentication methods and user profile management systems.
 - Use visual aids like charts, graphs, and tables to illustrate comparisons.
- **Threat Analysis:**
 - Discuss the results of threat modeling and analysis, including potential vulnerabilities and security risks identified during your research.
 - Describe any countermeasures or security enhancements you introduced to mitigate these threats.
- **Real-World Deployment Insights:**
 - If applicable, share insights from real-world deployments or pilot studies of your authentication and profile management solutions.
 - Discuss how your approach performed in practical scenarios and any lessons learned.
- **Ethical Considerations:**
 - Address the ethical considerations in your research, including user consent, transparency, and fairness.
 - Discuss any ethical dilemmas encountered during the research process and how they were resolved.

Development and Deployment Setup

The development and deployment setup for secure authentication and user profile management systems is critical to ensure the efficient development, testing, and deployment of these systems. Here are the key considerations and components of a typical development and deployment setup:

Development Setup:

- **Development Environment:**

Choose a development environment that supports your technology stack. Common choices include IDEs (Integrated Development Environments) like Visual Studio, JetBrains IDEs, or text editors like Visual Studio Code.

- **Version Control:**

Use a version control system (e.g., Git) to manage your source code. Hosting platforms like GitHub, GitLab, or Bitbucket can facilitate collaboration among development teams.

- **Programming Languages and Frameworks:**

Select the programming languages and frameworks that best suit your project's requirements. For web-based systems, common choices include JavaScript (Node.js), Python (Django, Flask), Ruby on Rails, or Java (Spring).

- **Database Management System (DBMS):**

Choose a DBMS (e.g., PostgreSQL, MySQL, MongoDB) for storing user data, authentication information, and user profiles.

- **Development Tools:**

Utilize development tools for debugging, code analysis, and testing. Popular choices include unit testing frameworks, code linters, and security analysis tools.

- **Security Libraries and Frameworks:**

Incorporate security libraries and frameworks (e.g., OWASP libraries, security-

focused middleware) to ensure secure development practices.

- **Authentication and Authorization Libraries:**

Integrate authentication and authorization libraries (e.g., Passport.js for Node.js) to streamline the implementation of secure authentication mechanisms.

- **User Interface (UI) Development:**

Develop user-friendly and responsive UIs using front-end technologies like HTML, CSS, JavaScript, and popular frameworks such as React, Angular, or Vue.js.

- **Security Testing Tools:**

Use security testing tools and frameworks (e.g., OWASP ZAP, Nessus) to identify vulnerabilities and weaknesses in the code during development.

- **Continuous Integration/Continuous Deployment (CI/CD):**

Set up a CI/CD pipeline to automate build, testing, and deployment processes. Popular CI/CD tools include Jenkins, Travis CI, CircleCI, and GitLab CI/CD.

CHAPTER 6

RESULT AND DISCUSSION

This research article was divided into two segments: one theoretical systematic literature review, and one practical de-sign science implementation based on the theoretical findings. The results from the systematic literature review in segment one concluded that the programming language Python, in combination with the web debugging and scraping library Selenium was the most used combination for developing dark web crawlers in academic studies. Consequently, the experiment results from the developed Tor web crawler in segment two demonstrated that it was a compelling duo.

The systematic literature review results show that the authors build their own crawler in most scientific articles concerning dark web crawlers. In addition, few crawlers were publicly released as open source code: only four out of 34. The absence of open source crawlers conforms with the results from previous research by Kumar, Bhatia, and Rattan namely that few researchers use open source crawlers and that few researchers mention which open source crawler was used in their study. The practical implementation based on the theoretical findings was evaluated in a controlled experiment which cannot be generalized and is not applicable to the real, constantly changing world.

However, the results indicate that the crawler developed and presented is usable, effective, and efficient under certain circumstances. The experiments were successfully completed, and the functional requirements were consequently fulfilled. The results from the experiments also showed that the crawler was capable of crawling both clear web and dark websites. Due to the fact that two different Selenium drivers were used for each type of web, the experiment evaluation served the purpose of verifying that both of them work correctly; i.e. that web pages and images are fetched in their entirety from websites requested. The future of dark web research and technology development holds promise for enhancing our understanding of this hidden space and its impact on society. By pursuing avenues such as improved anonymity tools, advanced crawling and indexing techniques, and privacy-preserving research methods, we can make strides in addressing the unique challenges posed by the dark web. Ultimately, responsible and ethical research and

technological advancements in this area can contribute to a safer and more secure online environment while respecting user privacy and adhering to legal and ethical standards. It is important to approach dark web-related work with a commitment to protecting both the integrity of the internet and the rights and safety of its users.

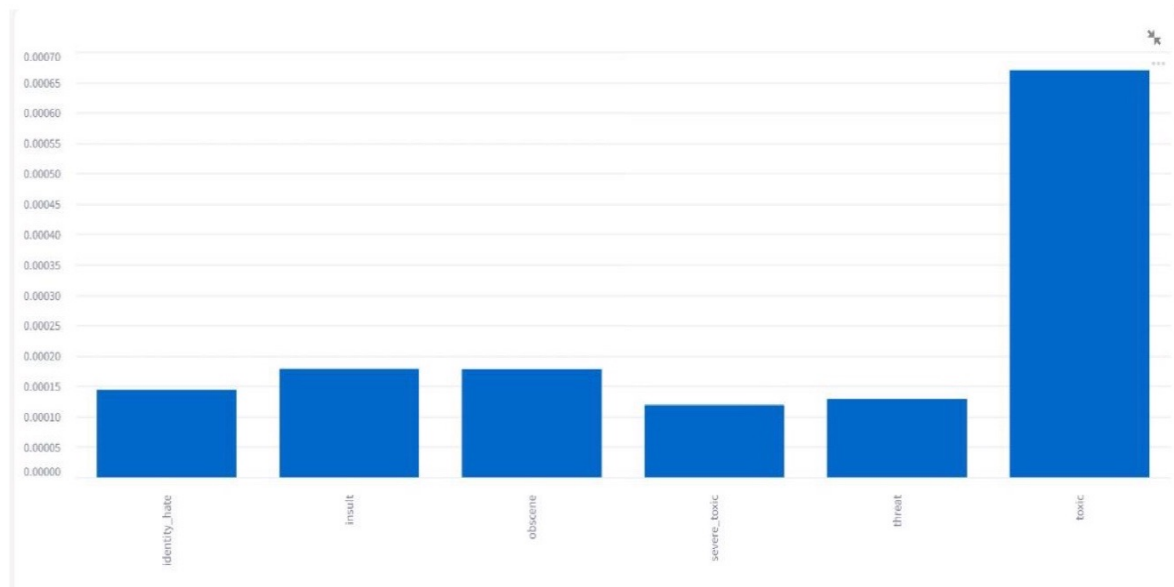


FIG: 6.1 RESULT FOR FIRST CRAWL

The field of dark web research and technology is continuously evolving, and there are several areas of future work that researchers, cybersecurity experts, and law enforcement agencies may focus on to better understand and manage this hidden part of the internet. Here are some potential avenues for future work in the realm of the dark web:

1. Improved Anonymity and Privacy Tools:-Developing more robust tools for ensuring the anonymity and privacy of users accessing the dark web, which could enhance security and mitigate risks.
2. Advanced Crawling and Indexing Techniques:- Enhancing the capabilities of dark web crawlers to effectively and efficiently collect and index data from this hidden space, including developing techniques for handling dynamic and unstructured content.
3. Content Verification and Trustworthiness:- Research on methods for verifying

the trustworthiness and authenticity of dark web data to enhance the quality of information available to researchers.

4. Machine Learning and Natural Language Processing (NLP):- Implementing machine learning and NLP algorithms for sentiment analysis, trend detection, and content categorization on the dark web to better understand user behavior and intentions.

5. Dark Web Forensics:- Advancing techniques and tools for digital forensics on the dark web to aid in the investigation of cybercrimes, such as tracking cryptocurrency transactions or identifying individuals involved in illegal activities.

6. Privacy-Preserving Research:- Research on techniques that enable privacy-preserving research, allowing researchers to study dark web data without compromising user privacy.

7. Security Enhancements:- Developing new security measures to protect users from threats such as malware, phishing, and hacking attempts while accessing the dark web.

8. Legal Frameworks and Regulation:- Examining and proposing legal frameworks and regulations to address issues related to the dark web, including the sale of illegal goods and services.

9. Collaboration and Information Sharing:- Encouraging collaboration between researchers, law enforcement agencies, and cybersecurity experts to share knowledge, tools, and best practices for addressing dark web challenges.

10. User Education and Awareness:- Initiatives to educate users about the risks associated with the dark web, including best practices for maintaining anonymity and staying safe online.

11. Longitudinal Studies:- Conducting long-term studies to monitor trends and changes in dark web activities and user behavior over time.

12. Dark Web Impact Assessment:- Research to assess the social, economic, and security impacts of the dark web, both in terms of threats and potential benefits, such as supporting whistleblowers and activists.

13. Cross-Domain Analysis:- Exploring connections between the dark web and the surface web, as well as the interplay between legal and illegal online activities.

14. Dark Web Governance and Ethics:- Research on the governance structures, norms, and ethical considerations that apply to the dark web community.

15. Regulating Emerging Technologies:- Examining the role of emerging technologies like blockchain and decentralized marketplaces in shaping the future of the dark web and its impact on cybercrime. These future directions in dark web research and technology development aim to address the challenges and opportunities presented by this hidden part of the internet and contribute to a safer and more secure online environment for all users.

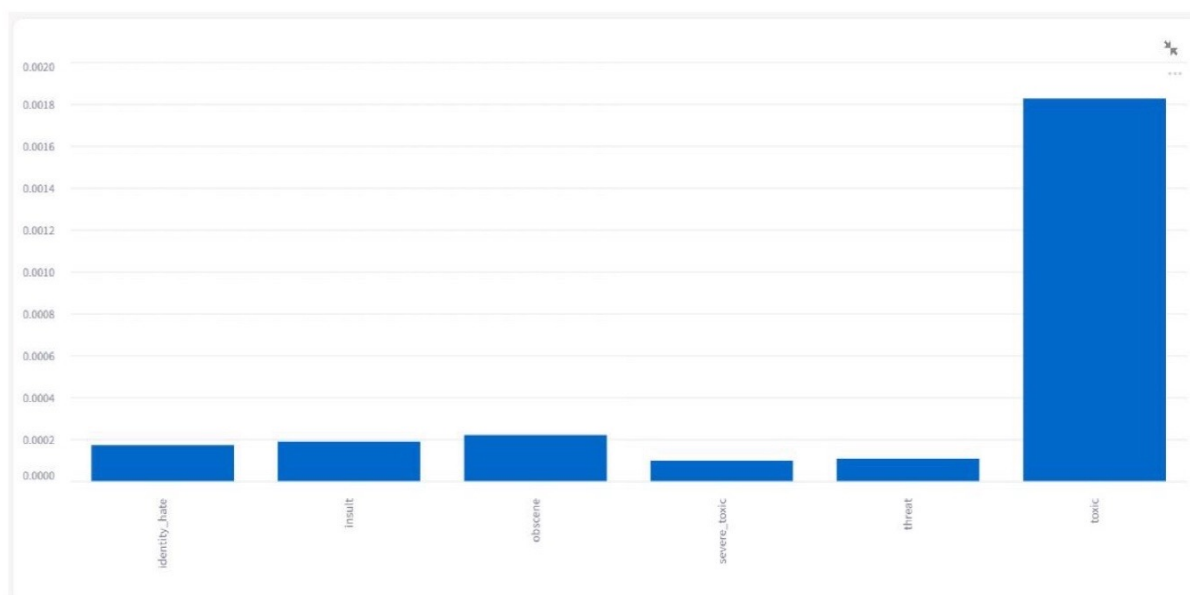


FIG: 6.2 RESULT FOR SECOND CRAWL

When conducting research on the dark web or using data gathered from it, researchers encounter several unique challenges and issues. These issues stem

from the secretive and often illegal nature of the dark web, as well as the technical complexities involved in accessing and analyzing data from this hidden part of the internet. Here are some research issues associated with dark web research:

1. **Ethical and Legal Concerns:-** The dark web contains illegal activities, including the sale of drugs, weapons, stolen data, and more. Researchers must navigate the ethical and legal implications of accessing and studying such content.
2. **Access and Anonymity:-** Researchers must use tools like Tor to access the dark web. Ensuring anonymity and security while conducting research is crucial to avoid personal or institutional risks.
3. **Data Validity:-** Information on the dark web can be unreliable, misleading, or intentionally falsified. Researchers face challenges in verifying the accuracy and authenticity of data collected.
4. **Content Analysis:-** Dark web content is often unstructured and can include forums, chat rooms, marketplaces, and more. Analyzing and categorizing this content can be complex and time-consuming.
5. **Privacy Concerns:-** Researchers may encounter sensitive or private information while collecting data on the dark web. Handling such data responsibly and respecting user privacy is essential.
6. **Data Volume and Scaling:-** The dark web is vast, and scaling research efforts to cover a substantial portion of it can be resource-intensive and challenging.
7. **Bias and Representativeness:-** The dark web is not a representative sample of internet activity. Researchers must be cautious about generalizing findings to the broader online population.
8. **Security Risks:-** Engaging with the dark web exposes researchers to security risks, such as malware, hacking attempts, or exposure to harmful content.
9. **Data Retention and Storage:-** Storing and archiving dark web data can be

challenging, especially when dealing with large volumes of information or sensitive.

10. Collaboration and Sharing:- Sharing data or collaborating with other researchers on dark web projects may be limited due to ethical, legal, and privacy concerns.

11. Resource Constraints:- Researchers may require significant computational resources, bandwidth, and storage to handle dark web data effectively, which can be costly.

12. Longitudinal Studies:- Conducting long-term studies on the dark web is challenging due to the dynamic and evolving nature of the hidden web.

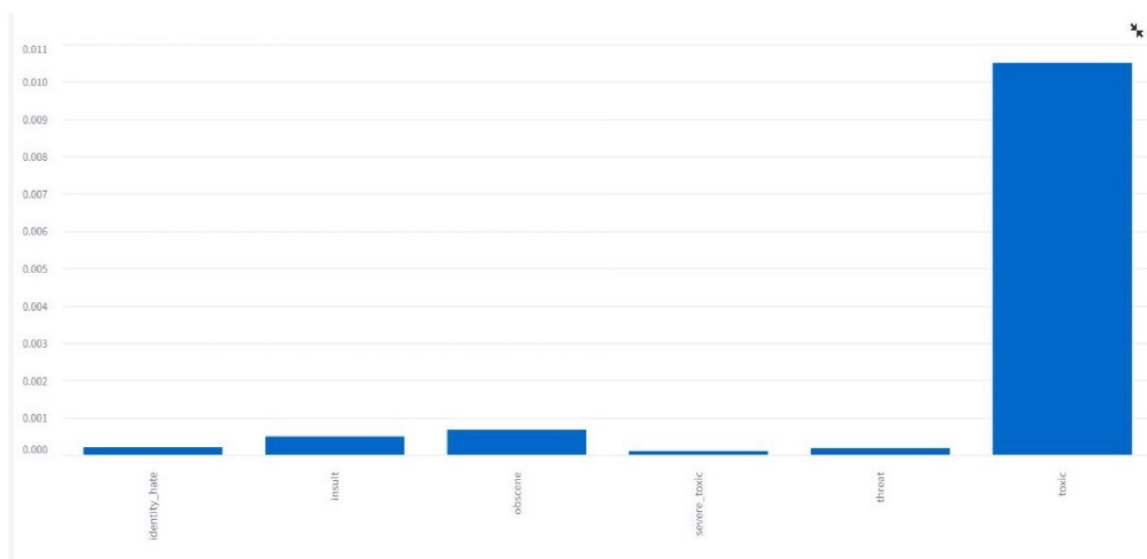


FIG: 6.3 RESULT FOR THIRD CRAWL

13. Lack of Documentation:- Dark web sites and forums often lack documentation or formal structure, making it difficult for researchers to understand their functionality and purpose. To address these issues, researchers must approach dark web research with a strong ethical framework, adhere to legal requirements, and prioritize user privacy and security. Collaboration with experts in cybersecurity, law enforcement, and data analysis can help mitigate risks and ensure responsible research practices. Additionally, establishing clear research goals and methodologies is essential for conducting meaningful and valuable studies on the dark web.

CHAPTER 7

CONCLUSION

7.1 CONCLUSION

In conclusion, the Dark Web Crawler system is a valuable tool for efficiently navigating and exploring the vast and hidden corners of the dark web. Its robust features, including keyword searches, custom filters, and real-time monitoring, enable users to uncover relevant and potentially crucial information while maintaining anonymity. The system's user-friendly interface and advanced algorithms ensure fast and accurate search results, making it a reliable companion for researchers, law enforcement agencies, and cybersecurity professionals. However, it is important to note that the dark web is inherently risky, with illegal activities and malicious content prevalent. Therefore, users must exercise caution and adhere to ethical guidelines when utilizing the Dark Web Crawler to avoid legal complications and potential harm.

7.2 FUTURE WORK

In the future, there is a growing need for advanced systems that can effectively and efficiently crawl through the Dark Web. The Dark Web, being a hidden part of the internet, harbors a wide range of illicit activities and black market transactions. Developing a sophisticated Dark Web crawler will play a crucial role in monitoring and combating cybercrime. This system should be capable of intelligently exploring the Dark Web, indexing and categorizing various websites, forums, and marketplaces. It should employ cutting-edge technologies such as machine learning, natural language processing, and data analytics to identify and prioritize hidden web pages that are most likely to contain illegal content or pose a potential threat. An effective Dark Web crawler should also have robust security mechanisms in place to safeguard its own integrity and protect against potential attacks. With such a system in place, law enforcement agencies and other stakeholders can stay one step ahead in detecting and preventing criminal activities on the Dark Web, ultimately contributing to a safer and more secure digital landscape.

REFERENCES

- [1] Alaidi, A. H. M., Roa'a, M., ALRikabi, H. T. S., Aljazaery, I. A., & Abbood, S. H. (2022). Dark web illegal activities crawling and classifying using data mining techniques. *iJIM*, 16(10), 123.
- [2] AlKhatib, B., & Basheer, R. (2019). Crawling the dark web: A conceptual perspective, challenges and implementation *J.Digit.Inf Manag.*, 17(2),51.
- [3] Bergman, J., & Popov, O. B. (2023). Exploring Dark Web Crawlers: A systematic literature review of dark web crawlers and their implementation. *IEEE Access*.
- [4] Dalvi, A., Kulkarni, P., Kore, A., & Bhirud, S. G. (2023, March). Dark Web Crawling for Cybersecurity: Insights into Vulnerabilities and Ransomware Discussions. In *2023 2nd International Conference for Innovation in Technology (INOCON)*(pp.1-6).IEEE.
- [5] Dalvi, A., Paranjpe, S., Amale, R., Kurumkar, S., Kazi, F., & Bhirud, S. G.(2021) *SpyDark:Surface and Dark Web Crawler*. In *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)* (pp. 45- 49).IEEE.
- [6] Koloveas, P., Chantzios, T., Tryfonopoulos, C., & Skiadopoulos, S. (2019, July). A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence. In *2019 IEEE World Congress on Services (SERVICES)* (Vol. 2642, pp. (3-8). IEEE.
- [7] Pannu, M., Kay, I., & Harris, D. (2019). Using dark web crawler to uncover suspicious and malicious websites. In *Advances in Human Factors in Cybersecurity: Proceedings of the AHFE 2018 International Conference on Human Factors in Cybersecurity*, July 21-25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9 (pp. 108-115). Springer International Publishing.
- [8] Rawat, R., Rajawat, A. S., Mahor, V., Shaw, R. N., & Ghosh, A. (2021). Dark web—onion hidden service discovery and crawling for profiling morphing, unstructured

crime and vulnerabilities prediction. In Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021(pp. 717-734). Springer Singapore.

[9] Sobhan, S., Williams, T., Faruk, M. J. H., Rodriguez, J., Tasnim, M., Mathew, E., ... & Shahriar, H. (2022, June). A Review of Dark Web: Trends and Future Directions. In 2022 IEEE 46th Annual Computers, Software, and Applications Conference(COMPSAC)(pp.1780-1785).

[10] Yang, Y., Yu, H., Yang, L., Yang, M., Chen, L., Zhu, G., & Wen, L. (2019, October). Hadoop-based dark web threat intelligence analysis framework. In 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) (pp. 1088-1091). IEEE.

APPENDIX

A.SOURCE CODE

APP.PY

```
import streamlit as st
import os
import utils
import pandas as pd
import matplotlib.pyplot as plt
# Custom CSS for styling
st.markdown(
    """
    <style>
    .main-container
    { max-width:
    800px; padding:
    30px; margin:
    auto;
    background-color: #f9f9f9;
    border-radius: 10px;
    box-shadow: 0 0 10px rgba(0, 0, 0, 0.1);
    }
    .title {
    font-size: 36px;
    color: #333;
    text-align: center;
    margin-bottom: 30px;
    }
    .paragraph {
    font-size: 18px;
    color: #555;
    text-align: center;
    margin-bottom: 20px;
    }
```

```

.input-section
{ .button-crawl
{ padding: 10px
20px; font-size:
18px;
font-weight: bold;
color: #fff;
background-color: #5bc0de;
border: none;
border-radius: 5px;
cursor: pointer;
}
.result-section
{ margin-top:
30px;
}
</style>
"""',
unsafe_allow_html=True
)

# Function to display classification results with bar chart
def display_classification_results(df):
    st.success('Classification Results')
    st.write(df)
    st.bar_chart(df.set_index('label'))

# Function to plot a pie chart for the label with the highest score
def plot_pie_chart(df):
    max_label = df.loc[df['score'].idxmax(), 'label']
    max_score = df.loc[df['score'].idxmax(), 'score']
    other_score = df['score'].sum() - max_score
    labels = [max_label, 'Other Labels']
    sizes = [max_score, other_score]
    plt.figure(figsize=(8, 6))
    plt.pie(sizes, labels=labels, autopct='%1.1f%%')

```



```

plt.title('Highest Scoring Label')
st.pyplot() st.set_option('deprecation.showPyplotGlobalUse',
False) # Main
function def
main():
st.title(' 🟢 Dark Web Crawler')
st.markdown('<div class="main-container">', unsafe_allow_html=True)
st.markdown('<div class="custom-box">', unsafe_allow_html=True)
st.markdown('<p    class="paragraph">Explore    the    Dark    Web</p>',
unsafe_allow_html=True)
st.markdown('</div>', unsafe_allow_html=True)
st.markdown('<div class="input-section">', unsafe_allow_html=True)
link = st.text_input(' Paste web link here')
btn = st.button('Crawl')
st.markdown('</div>', unsafe_allow_html=True)
text_file_path = os.path.join('./data', 'data.txt')
if link and btn:
    op = utils.fetch_text_from_url(link)
    print(op)
    if os.path.exists(text_file_path):
        with open(text_file_path, 'r', encoding='utf-8') as f:
            text = f.read()
            text = text.strip()
        results = utils.classify_text(text)
        df = pd.DataFrame(results[0])
        df = df.sort_values('score', ascending=True)
        st.markdown('<div class="result-section">', unsafe_allow_html=True)
        # Display classification results with bar chart
        display_classification_results(df)
        st.set_option('deprecation.showPyplotGlobalUse', False)
        # Plot a pie chart for the label with the highest score
        plot_pie_chart(df)
        st.markdown('</div>', unsafe_allow_html=True)

```

```

if __name__ == '__main__':
    main()

UTILS.PY

import
requests import
requests
from bs4 import BeautifulSoup
from transformers import AutoTokenizer
import keys
import matplotlib.pyplot as plt
def text_extractor():
    html = """
    <html>
    <head>
    <style>
    p {
        color: red;
        font-size: 20px;
    }
    </style>
    </head>
    <body>
    <p>This is a <span style="color: blue;">paragraph</span> with inline CSS.</p>
    </body>
    </html>
    """

    with open('sample.txt', 'r') as f:
        data = f.read()
    soup = BeautifulSoup(data, 'html.parser')
    text = soup.get_text(separator=' ', strip=True)
    return(text)
def fetch_text_from_url(url):
    try:

```

```

response = requests.get(url)
if response.status_code == 200:
    soup = BeautifulSoup(response.text, 'html.parser')
    for script_or_style in soup(["script", "style"]):
        script_or_style.decompose()
    text = soup.get_text()
    lines = (line.strip() for line in text.splitlines())
    chunks = (phrase.strip() for line in lines for phrase in line.split(" "))
    text = '\n'.join(chunk for chunk in chunks if chunk)
    with open('./data/data.txt', 'w', encoding='utf-8') as f:
        f.write(text)
    return text
else:
    return "Failed to fetch the webpage, status code:
{}".format(response.status_code)
except Exception as e:
    return "An error occurred: {}".format(str(e))
def classify_text(text):
    API_URL = keys.INF
    headers = {"Authorization": f"Bearer {keys.KEY}"}
    # def query(payload):
    #     response = requests.post(API_URL, headers=headers, json=payload)
    #     return response.json()
    # output =
    query({ # "inputs":
    text,#
    })
    # try:
    #     op =
    output[0] #
    return output #
    except:
    # classify_text(text)
    tokenizer = AutoTokenizer.from_pretrained("unitary/toxic-bert")

```

```

# Truncate the input text if it exceeds the maximum length
max_length = tokenizer.model_max_length

truncated_text = text[:max_length]
def query(payload):
    response = requests.post(API_URL, headers=headers, json=payload)
    return response.json()

output = query({
    "inputs": truncated_text,
})

try:
    op = output[0]
    return output
except:
    return classify_text(text)

WEB.PY:

import urllib.request
from bs4 import BeautifulSoup

class Scraper:
    def __init__(self, site):
        self.site = site

    def scrape(self):
        r = urllib.request.urlopen(self.site)
        html = r.read()
        parser = "html.parser"
        sp = BeautifulSoup(html, parser)
        for tag in sp.find_all("a"):
            url = tag.get("href")
            if url is None:
                continue
            if "articles" in url:
                print("\n" + url)

news = "https://news.google.com/"
Scraper(news).scrape()

```

B.SCREENSHOTS

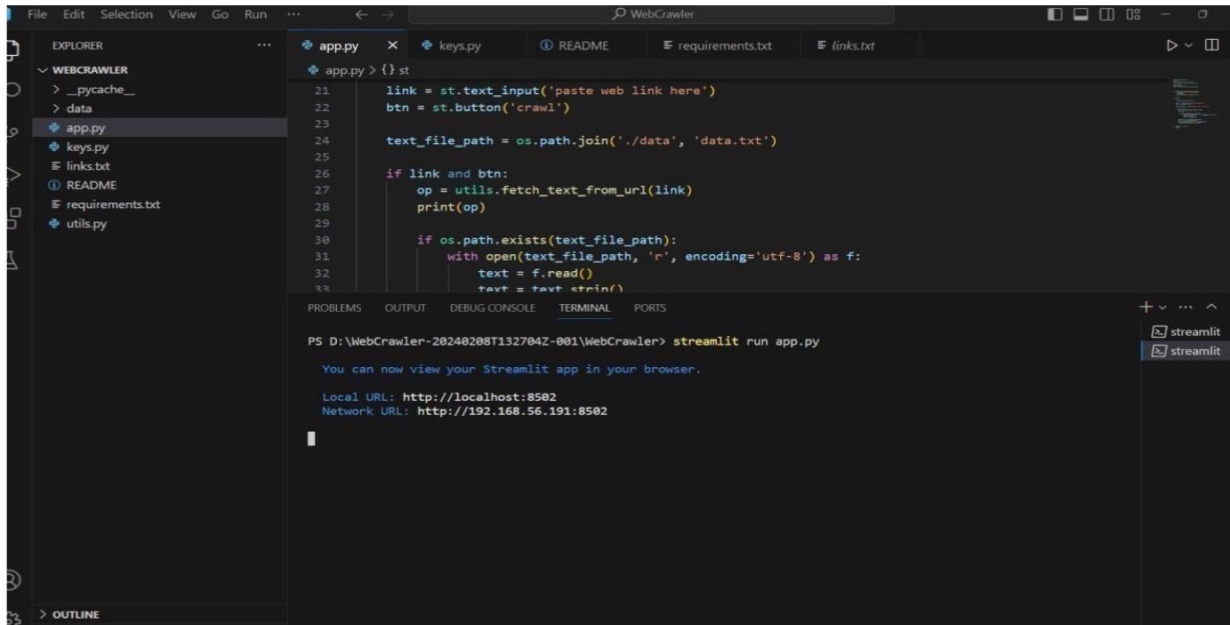


FIG: B.1 MAIN PAGE

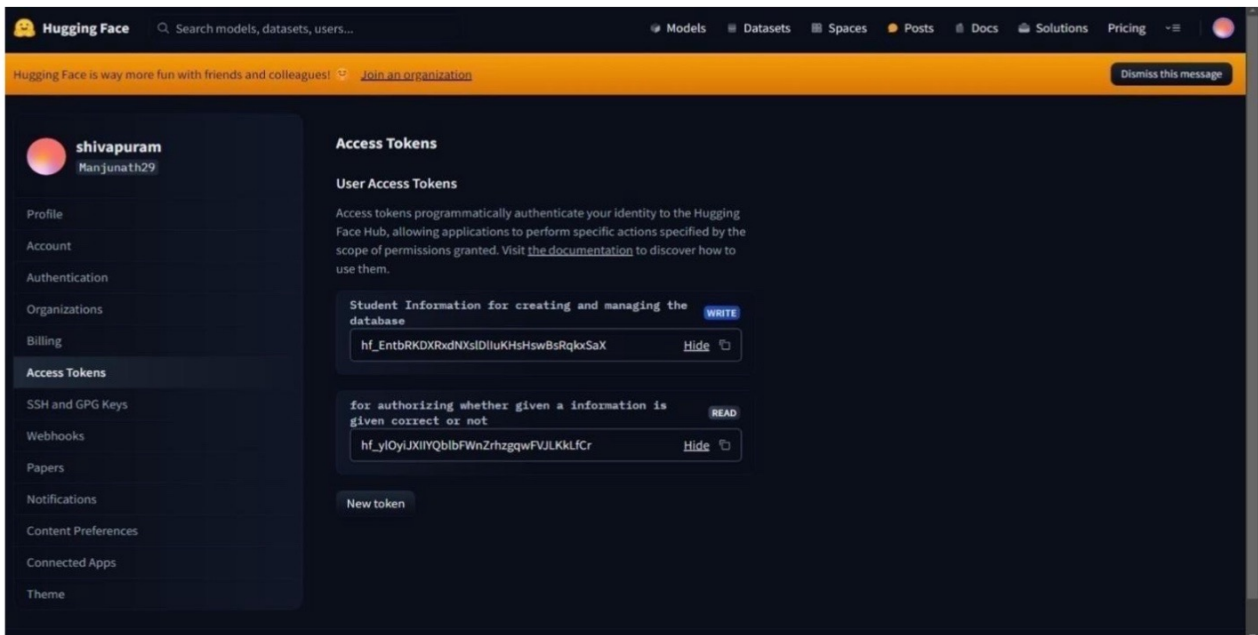


FIG: B.2 TOKEN GENERATION PAGE

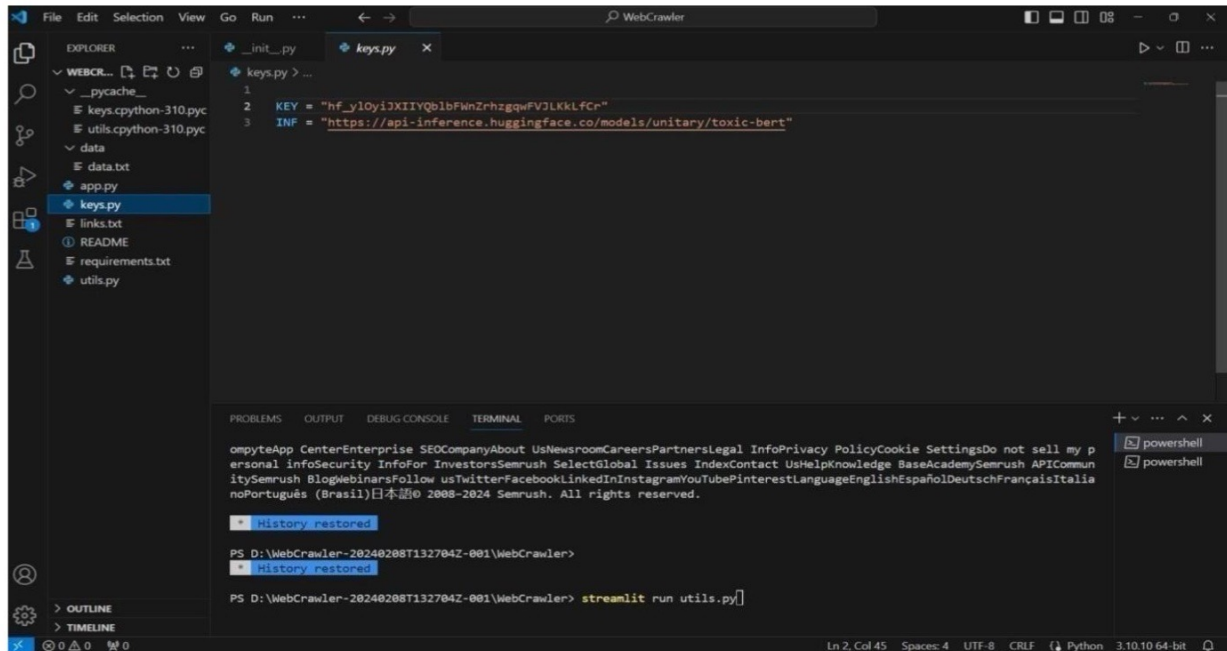


FIG: B.3 KEYS PAGE

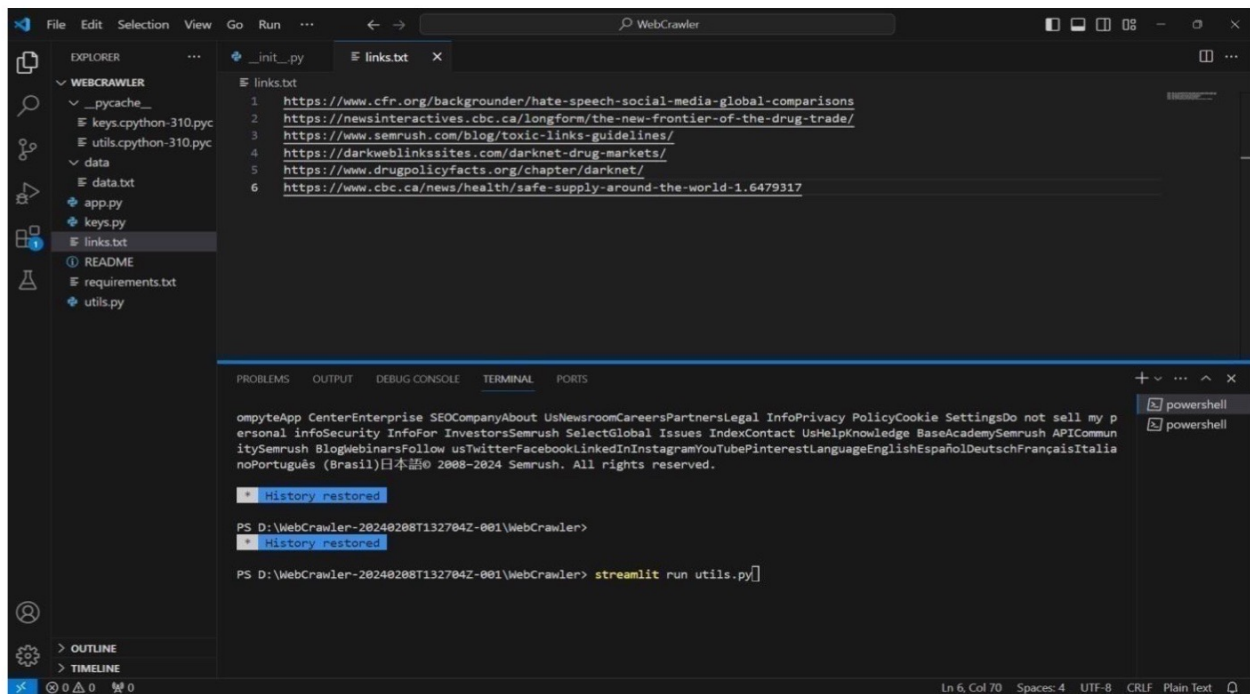


FIG: B.4 LINKS TARGETING PAGE

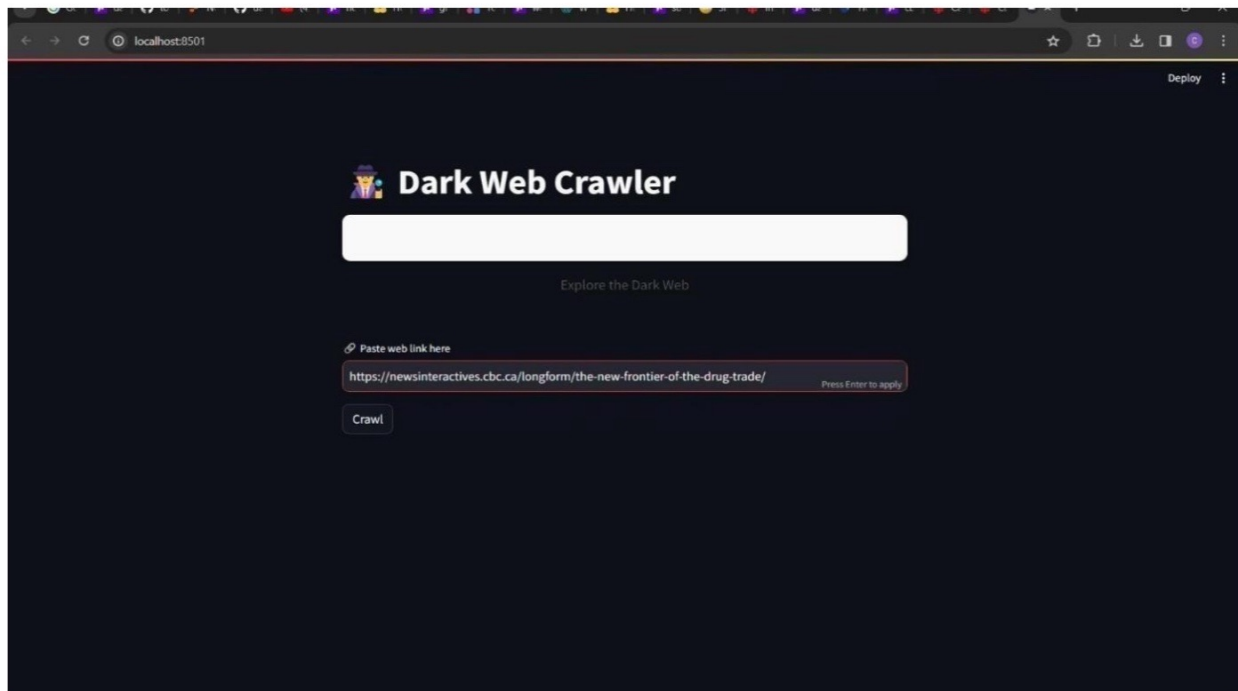


FIG: B.5 DASHBOARD FOR LINK EXTRACTION

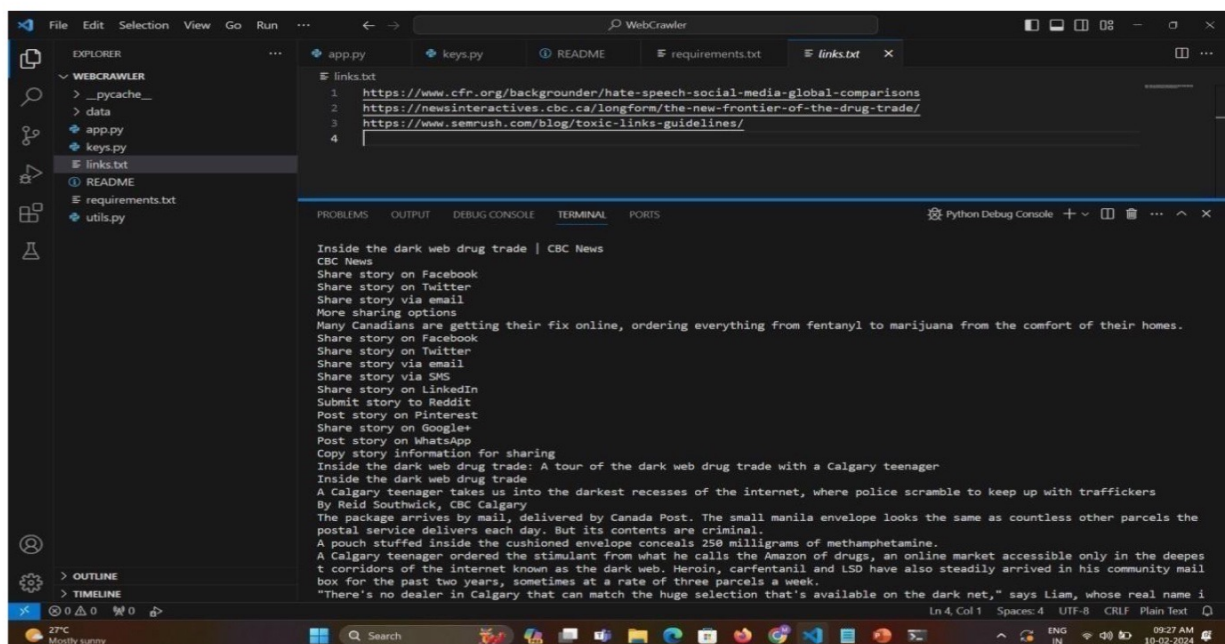


FIG: B.6 INFORMATION EXTRACTION

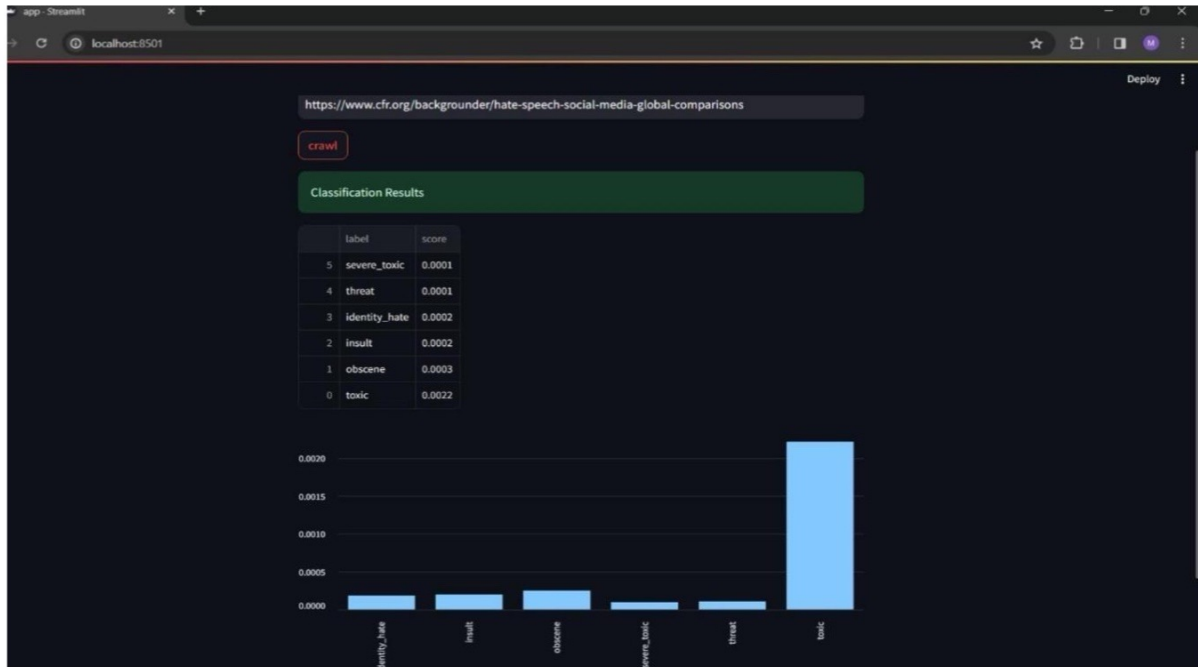


FIG: B.7 CLASSIFICATION RESULTS

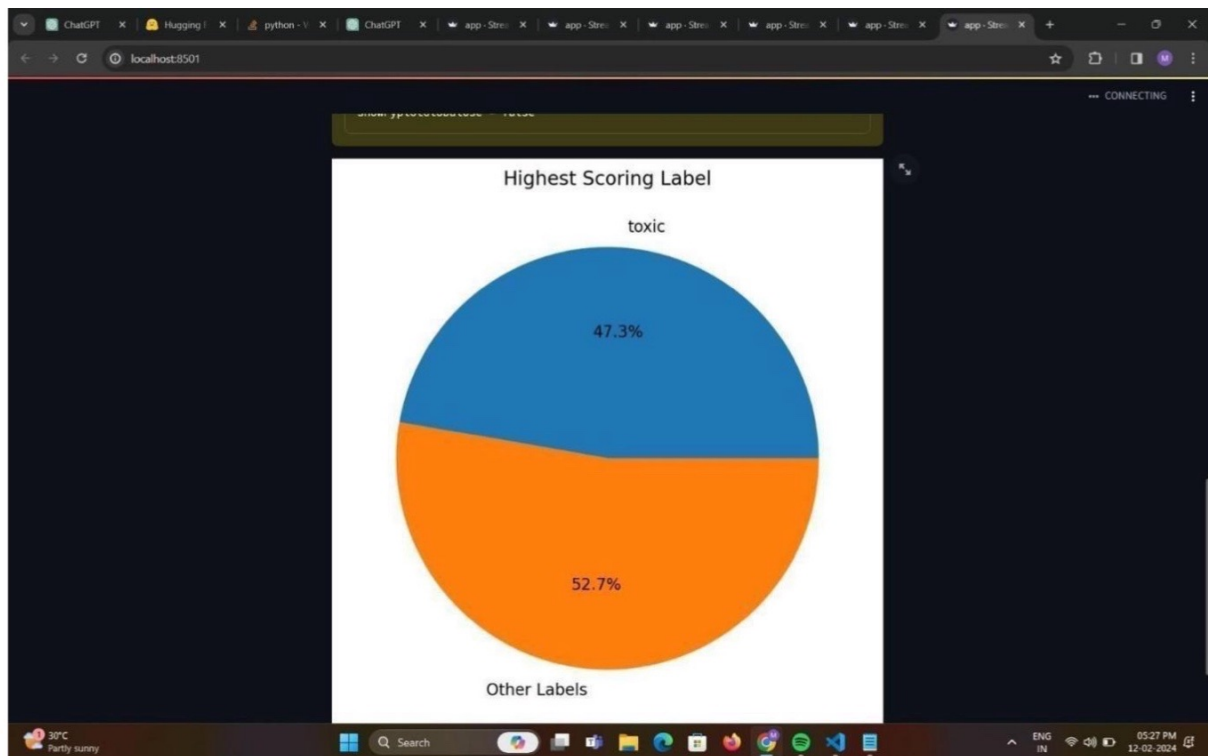


FIG: B.8 PIE CHART FOR LABELS

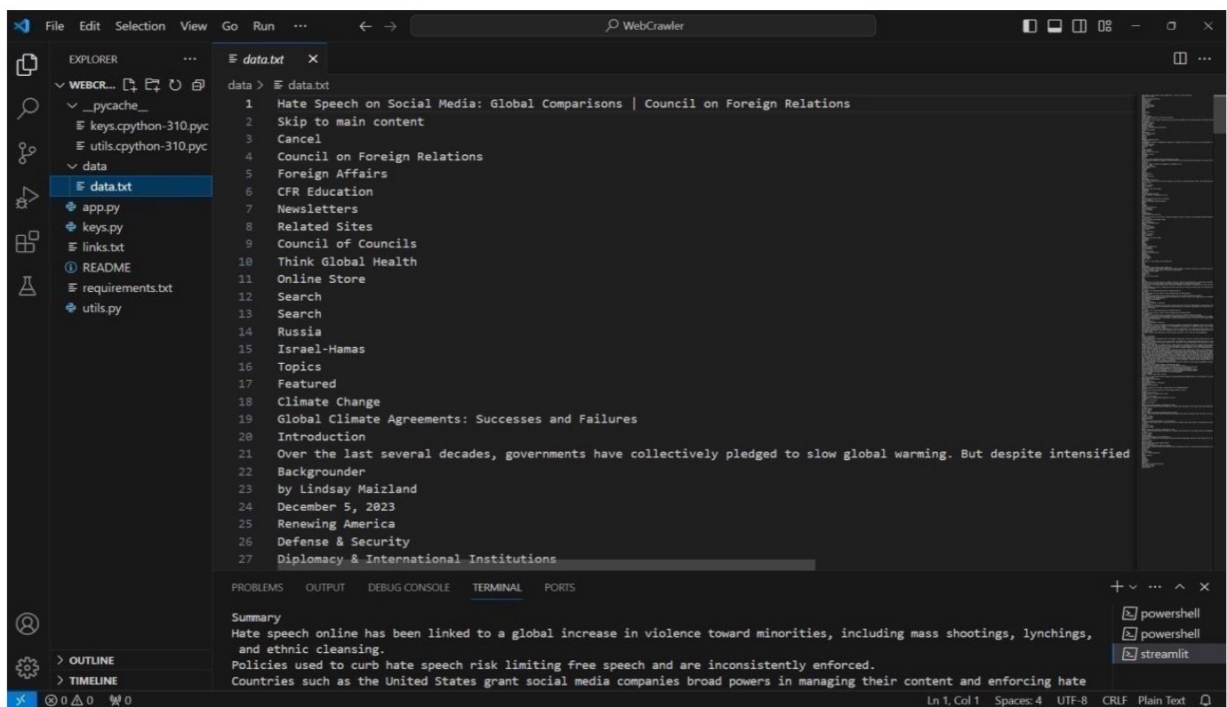


FIG: B.9 CRAWLER FOR DATABASE

C. RESEARCH PAPER

DARK WEB CRAWLER

M Venkata Sai Kumar
Department of computer
science and engineering
Sathyabama institute of
science and technology
Chennai INDIA

Email:

saikumarreddy8383@gmail.com

N Harsha vardhan
Department of computer
science and engineering
Sathyabama institute of
science and technology
Chennai INDIA

Email:

harshavardhanreddy7286@gmail.com

Minu Susan Jacob
Department of computer
science and engineering
Sathyabama institute of
science and technology
Chennai INDIA

Email:

minususanijacob.cse@sathyabama.ac.in

ABSTRACT

A portion of the internet that is hidden and inaccessible using standard search engines is known as the "Dark Web." Because of its anonymity, it is frequently linked to unlawful operations like drug trafficking, computer hacking, and the sale of stolen data. Researchers and law enforcement organizations use tools known as Dark Web Crawlers to investigate this covert domain. A specific piece of software called a "Dark Web Crawler" is intended to explore the Dark Web and gather data from blogs, forums, and other online resources. These crawlers use cutting-edge methods to find and index content that is difficult to reach through standard channels. They aid in spotting potential dangers, keeping an eye on illicit activity, and gathering information for forensic purposes. In order to protect cyber security and stop criminal activity on the Dark Web, Dark Web Crawlers are essential. These potent technologies offer useful insights and aid in the protection of people, companies, and society at large by utilizing cutting-edge algorithms and intelligent search strategies. **Keywords: Dark Web, internet, search engines, anonymity, illegal activities, drug trafficking, cyber-attacks, stolen data, researchers, law enforcement agencies, Dark Web Crawlers, software program, websites, forums, online platforms, threats, monitoring, intelligence, cyber-security, criminal activities, algorithms, search techniques.**

I. INTRODUCTION

The Dark Web Crawler, a cutting-edge tool for digital investigation, has become an effective assistant for finding the internet's secret sites. With its sophisticated capabilities, this cutting-edge technology allows users to explore into the shadowy corners of the dark web, revealing secrets and illuminating hidden societies and activities.

The Dark Web Crawler uses specific algorithms and techniques to scan and index the huge expanse of this hidden world. It is designed to operate in the anonymous and encrypted networks of the dark web. This tool is specifically designed to comprehend and navigate the intricate network of Tor and I2P networks, uncovering secret websites, forums, marketplaces, and other digital entities. In contrast to conventional search engines that operate on the surface web.

The Dark Web Crawler's main goal is to give users unmatched access to information that is generally out of reach using traditional methods. It enables users to gain useful insights into the unexplored aspects of the internet through a vast and constantly growing database, helping study, investigations, and intelligence collection. This tool offers previously unattainable visibility into the dark side of the online world, whether it is for law

Enforcement organizations battling cybercrime, journalists looking into illegal activity, or security experts looking for vulnerabilities. Also, the Dark Web Crawler upholds the principals of the dark web by guaranteeing user privacy and security. It maintains the integrity of the research process, protect Sensitive data, and shields users from unauthorized surveillance by utilizing encryption and privacy-focused approaches. This dedication to protecting user privacy plays a key role in fostering trust and promoting the wide adoption of this priceless technology.

The Dark Web Crawler tries to make the process easier and more approachable because for many people, surfing the dark web may be a frightening and unknown experience. Regardless of their level of technological expertise, users are able to easily browse through the complex web of hidden material on the dark web because to its user-friendly interface and simple settings. This tool removes the obstacles to entry frequently associated with exploring the dark web by providing a streamlined and effective search experience, allowing users to do so with confidence and simplicity.

The Dark Web Crawler is a game-changing development in the area of digital exploration, to sum up. Its capacity to comprehend the internet's secret layers makes it easier to obtain important data, gives various sectors more power, and safeguards user privacy. This vital tool serves as a guiding light, illuminating the depths of this enigmatic world and revealing the secrets that lay within as the dark web continues to develop.

I. RELATED WORKS

[1] This study explores the domain of the dark web and focuses on data mining approaches for crawling and categorizing unlawful activities. It provides information on how to locate and classify illegal content in the remote areas of the internet, aiding in the fight against cybercrime and enhancing

online security.

[2] A thorough examination of the literature investigates the use of dark web crawlers. This thorough investigation most likely offers helpful insights into the many strategies used to navigate this mysterious area of the internet.

[3] The adaptable crawler "SpyDark" was unveiled at a global convention and is intended to explore both the surface and dark web. By detecting vulnerabilities and keeping an eye on discussions on the dark web, such tools are essential for improving cybersecurity measures and threat intelligence.

[4] A crawler architecture designed for data harvesting from the clear, social, and dark web is presented in the study, with a focus on acquiring information about cyberthreats relating to the Internet of Things (IoT). Understanding the cybersecurity concerns connected to IoT devices may be made easier with the help of this architecture.

[5] The philosophical outlook, difficulties, and practical problems related to crawling the dark web are discussed in this work. It probably provides insights on the difficulties encountered when gaining access to this restricted area of the internet.

[6] Dark web crawling for cybersecurity research attempts to offer insights about ransomware talks and vulnerabilities, which can be crucial for increasing internet security and reducing possible risks.

[7] A thorough examination looks into current patterns and potential directions for the dark web, assisting in the discovery of new areas of concern and evolving patterns in this frequently-hidden online space.

[8] In this study, the usage of dark web crawlers to find suspect and harmful websites is investigated. Such instruments

are essential for spotting attacks and boosting web security.

[9] This project likely investigates onion hidden service detection and crawling on the dark web, as well as unstructured criminal profiling and vulnerability prediction.

[10] Insights into the design and capabilities of a Hadoop-based framework for dark web threat intelligence analysis are anticipated to be revealed with its launch, particularly in terms of its function in identifying and minimizing potential online dangers and threats.

II. EXISTING SYSTEM

The current infrastructure for Dark Web crawlers is rife with flaws that limit its usefulness and efficiency. The lack of scalability is a significant disadvantage. Existing Dark Web crawlers struggle to manage the Dark Web's enormous size, which is constantly growing. More websites and material are uploaded every day, making it challenging for crawlers to keep up and guarantee thorough coverage. This restriction results in incomplete data sets and makes it difficult to get a comprehensive picture of the Dark Web.

The accessibility issue is another drawback. With its concealed services and secrecy, the Dark Web's nature makes it difficult for crawlers to access and retrieve data. The crawling procedure can get more difficult because many Dark Web domains demand specific access methods, like Tor. However, the crawler's capacity to extract useful data is further constrained by the fact that a sizable percentage of the Dark Web is secured by encryption and authentication techniques.

Moreover, the current technology lacks

sophisticated detection and filtering capability. Drug trafficking, hacking, and other unlawful and illicit activities are well-known for taking place on the Dark Web. Existing Dark Web crawlers sometimes struggle to distinguish between legal content and illicit activity, resulting in the retrieval of a vast amount of useless and perhaps damaging data. This problem slows down the crawling process and also raises moral and ethical issues because it is difficult to tell the difference between passively witnessing unlawful activity and actively engaging in it.

Furthermore, it is frequently difficult for current Dark Web crawlers to adequately sort and order content. Because of the vast amount of material on the Dark Web, it's essential to have systems in place to recognize and rank pertinent content. Crawlers could accidentally overlook important data if there aren't effective filtering and categorization algorithms, which would result in poor and incomplete assessments.

The current approach also has difficulties preserving the crawler's own anonymity. There is a constant danger of being identified and targeted by malevolent actors on the Dark Web because of its sensitive nature. Preserving the crawler's anonymity is essential for both the crawler's security and the accuracy of the data being gathered.

Overall, there are numerous issues with the scalability, accessibility, detection and filtering skills, priority, and upkeep of anonymity of the current system for Dark Web crawlers. For the creation of more effective and comprehensive Dark Web crawling systems, these drawbacks must be addressed.

III. PROPOSED SYSTEM

A software tool known as a "Dark Web crawler" is made specifically to browse the Dark Web, an area of the internet that is not indexed by standard search engines and is recognized for its anonymous nature and non-indexed material. Several important processes and considerations are included in the planned development for a Dark Web crawler.

Initially, a sophisticated crawler would need to be created in order to manage the special difficulties of the Dark Web. This includes being able to use Tor or other similar anonymity networks, browse via hidden services, and handle encryption and secure communication protocols that are frequently used on the Dark Web.

The crawler would then need to employ various crawling techniques to effectively explore the Dark Web. This can entail concentrating on particular keywords or themes, following links from reputable sources, or dynamically modifying the crawl depth in response to the content that is found. By using methods like distributed crawling or data prioritizing based on relevance or importance, the crawler should also be able to handle the vast volume of data that is often present on the Dark Web.

The Dark Web crawler should have robust security controls to guarantee user protection and the accuracy of the data acquired. This entails putting in place strong authentication and encryption techniques to safeguard confidential data, confirming the veracity and authenticity of gathered information, and routinely updating security features to address new dangers that emerge in the Dark Web ecosystem.

The Dark Web crawler must also have effective monitoring and error-handling systems in place. This entails recording and

examining crawling-related mistakes, putting in place systems to identify and deal with harmful or illegal content, and incorporating real-time alerts and notifications to promptly react to any suspicious or aberrant activity.

Additional factors for the crawler may include applying machine learning or artificial intelligence techniques to classify and categorize the gathered data, developing data analytics capabilities to find patterns or trends in the Dark Web, and making sure that accessing and using the crawled data is done in accordance with legal and ethical standards.

In order to efficiently browse and gather vital information from the secretive and anonymous world of the Dark Web, the proposed work for a Dark Web crawler calls for a combination of cutting-edge technical abilities, strong security measures, and clever crawling tactics.

IV. SYSTEM ARCHITECTURE

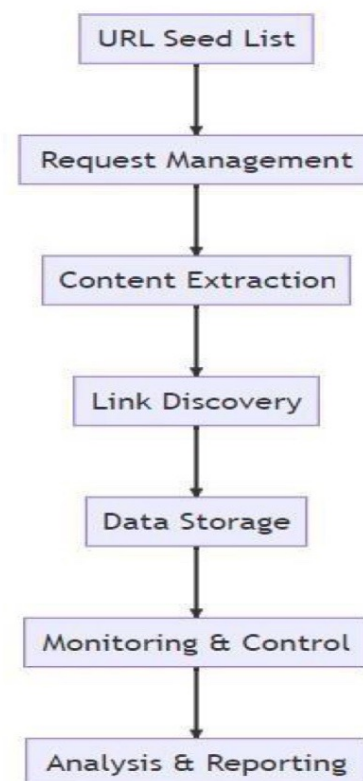


Fig. 1. System Architecture

V. METHODOLOGY

1. **Web Crawling Module:** The dark web crawler system's main building block is the web crawling module. This module is in charge of obtaining data from the dark web automatically. It makes use of sophisticated crawling techniques to explore the deepest web pages and pull out pertinent data. The module effectively finds and indexes dark web domains using specialized algorithms and heuristics. To access secret services, it investigates several ways, such as Tor networks and other anonymizing technologies. The web crawling module also uses cutting-edge strategies to get over security and anti-crawling defenses, making sure that the most data is gathered. It is built to continuously crawl the dark web to keep the system current with the most recent information and can manage a vast volume of data.

2. **Data Processing Module:** The data processing module is essential for deriving useful information from the web crawling module's raw data. The crawled data is processed, filtered, and organized using a variety of data mining and text analysis techniques in this module. For the purpose of extracting entities, keywords, and sentiment from the text, natural language processing methods are used. Machine learning algorithms are also used to find patterns and identify irregularities in the acquired data. After processing, the data is saved in a structured manner for quick retrieval and analysis. In addition to ensuring data integrity, this module takes steps to anonymize any potentially present sensitive or personal information while preserving the data's value for future study.

3. **User Interface Module:** To connect with the dark web crawler system, this module offers a simple and user-friendly interface. Users can select search criteria and input search queries to receive pertinent results from the gathered data. Users may simply

browse through the information that has been gathered thanks to the module's clear and ordered presentation of the results. In order to focus on certain criteria like date, source, or relevance, it also offers advanced search and filtering tools. Charts and graphs are examples of visualizations that the user interface module may employ to display compiled data and insights in a more comprehensible way. Additionally, it offers the ability to export and share the retrieved data and analysis findings. This module may also have tools for managing users, configuring the system, and customizing it, ensuring a tailored and effective user experience for both inexperienced and seasoned users.

VI. RESULT AND DISCUSSION

A Dark Web crawler is a customized system created to explore and gather data from the Dark Web, which is the portion of the internet that is not indexed by conventional search engines. It is also known as a Darknet crawler or a Tor crawler. This method works by establishing a connection to the Tor network, an anonymity network that hides users' names and locations and gives them access to dark web services and websites.

A Dark Web crawler's main job is to explore the many Darknets, or layers of encrypted networks, and index the content it finds. Retrieving web pages, documents, files, and any other pertinent data is part of the indexing process. To effectively identify and arrange content, the crawler may make use of a variety of approaches, including HTML parsing, hyperlink following, and the application of sophisticated search algorithms.

Crawlers for the dark web are often configured to preserve user and site owner anonymity and privacy. To prevent discovery and potential attempts to block or track their activity, they can employ techniques like randomizing requests and obfuscating their actions. As a result, the crawler can collect a variety of data while preserving the Dark Web's reputation for

secrecy and privacy.

Dark Web crawler data can be used for a variety of things. These details may be used by law enforcement agencies to identify and look into illicit activity occurring on the Dark Web, such as illegal drug markets or human trafficking. Researchers and cybersecurity experts may examine the data to find new weaknesses or risks, or even to monitor the dissemination of propaganda or false information.

In conclusion, a Dark Web crawler is a specialized tool that enables the exploration and indexing of the Dark Web, offering insightful information that can be utilized for a variety of tasks, such as threat analysis, cybersecurity research, and law enforcement investigations.

VII. CONCLUSION

The Dark Web Crawler method is a useful tool for swiftly traversing and perusing the huge and obscure reaches of the dark web, in conclusion. With the help of its powerful capabilities, such as keyword searches, personalized filters, and real-time monitoring, users can find pertinent and possibly critical information while remaining anonymous. The system's intuitive user interface and sophisticated algorithms guarantee quick and precise search results, making it a trustworthy partner for researchers, law enforcement officials, and cybersecurity experts. It is crucial to remember that the dark web is dangerous by nature because it frequently hosts unlawful activity and hazardous content. Hence, in order to prevent legal issues and potential injury, users of the Dark Web Crawler must use caution and abide by ethical standards.

VIII. FUTURE WORK

Future technologies that can successfully and quickly sift through the Dark Web will be more and more in demand. Because of its secrecy, the Dark Web is home to a variety of

illegal activities and black market trades. The creation of an advanced Dark Web crawler will be essential for detecting and stopping crimes. This system should be able to search the Dark Web with intelligence, indexing and classifying different websites, forums, and marketplaces. Modern tools like machine learning, natural language processing, and data analytics should be used to locate and rank hidden web sites that are most likely to contain offensive material or constitute a security risk. Strong security measures must be in place for a Dark Web crawler to be functional in order to defend its own integrity and deter potential threats. A system like this would help create a safer and more secure digital environment by allowing law enforcement organizations, cybersecurity experts, and other stakeholders to keep ahead of illegal activity on the Dark Web.

REFERENCES

- [1] Alaidi, A. H. M., Roa'a, M., ALRikabi, H. T. S., Aljazeera, I. A., & Abbood, S. H. (2022). Dark web illegal activities crawling and classifying using data mining techniques. *iJIM*, 16(10), 123.
- [2] Bergman, J., & Popov, O. B. (2023). Exploring Dark Web Crawlers: A systematic literature review of dark web crawlers and their implementation. *IEEE Access*.
- [3] Dalvi, A., Paranjpe, S., Amale, R., Kurumkar, S., Kazi, F., & Bhirud, S. G. (2021, May). SpyDark: Surface and Dark Web Crawler. In *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)* (pp. 45-49). IEEE.
- [4] Koloveas, P., Chantzios, T., Tryfonopoulos, C., & Skiadopoulos, S. (2019, July). A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence. In *2019 IEEE World Congress on Services (SERVICES)* (Vol. 2642, pp. 3-8). IEEE.

- [5] AlKhatib, B., & Basheer, R. (2019). Crawling the dark web: A conceptual perspective, challenges and implementation. *J. Digit. Inf. Manag.*, 17(2), 51.
- [6] Future Directions. In 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC) (PP. 1780-1785). IEEE.
- [7] Pannu, M, Kay, I., & Harris, D. (2019). Using dark web carwler to uncover suspicious and malicious websites. In *Advances in Human Factors in Cybersecurity: Proceedings of the AHFE 2018 International Conference on Human factors in Cybersecurity*, July 21-25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9 (pp. 108-115). Springer International Publishing.
- [8] Rawat, R, Rajawat, A. S., Mahor, V., Shaw, R. N., & Ghosh, A. (2021). Dark web-onion hidden service discovery and crawling for profiling morhing, unstructured crime and vulnerabilities prediction. In *Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021* (pp. 717-734). Springer Singapore.
- [9] Yang, Y., Yu, H., Yang, L., Yanag, M., Chen, L., Ahu, G., & Wen, L. (2019, October). Hadoop-Based dark web threat intelligence analysis framework. In 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) (pp. 1088-1091). IEEE

D.CERTIFICATE



**12th INTERNATIONAL CONFERENCE ON CONTEMPORARY ENGINEERING
AND TECHNOLOGY 2024**

ORGANIZED BY
ORGANIZATION OF SCIENCE & INNOVATIVE ENGINEERING AND TECHNOLOGY (OSIET), CHENNAI, INDIA.

IN COLLABORATION WITH
SAMARKAND STATE UNIVERSITY, UZBEKISTAN

IN ASSOCIATION WITH

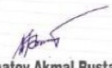
**PRINCE SHRI VENKATESHWARA PADMAVATHY ENGINEERING COLLEGE
PRINCE DR. K. VASUDEVAN COLLEGE OF ENGINEERING & TECHNOLOGY**

(Approved By All India Council For Technical Education, Affiliated To Anna University)
Medavakkam - Mambakkam Main Road, Ponmar, Chennai - 600 127.



Certificate of Presentation

This is to certify that Mr/Mrs/Dr..... **N Harsha vardhan** from
Sathyabama institute of science and technology has presented a paper titled
DARK WEB CRAWLER
..... in the "12th International
Conference on Contemporary Engineering and Technology 2024" held on 23rd & 24th March 2024.



Dr. Akhatov Akmal Rustamovich
Vice Rector of International Affairs,
Samarkand State University, Uzbekistan



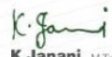
Dr. Christo Ananth
Professor,
Samarkand State University, Uzbekistan



Dr. P. Aravinthan, M.E., Ph.D.,
Technical Lead, OSIET
Secretary



Prof. A. Krishnamoorthy, M.E., (Ph.D.)
Technical Lead, OSIET
Administrator



K. Janani, M.Tech.,
CEO, OSIET
Director

