MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

A Project Report on

AI BASED CRAWLER FOR DARK WEB

Submitted by,

Ayush Koul	**(Exam Seat No. 202101040021)**
Amey Satone	**(Exam Seat No. 202101040058)**
Dheeraj singh	**(Exam Seat No. 202101040119)**
Viraj raina	**(Exam Seat No. 202101040042)**

Guided by,

Mrs. Neha Hajare

A Report submitted to MIT Academy of Engineering, Alandi(D), Pune, An Autonomous Institute Affiliated to Savitribai Phule Pune University in partial fulfillment of the requirements of

THIRD YEAR BACHELOR OF TECHNOLOGY in

Computer Engineering

School of computer Engineering

MIT Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Alandi (D), Pune – 412105

(2023–2024)

MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

CERTIFICATE

It is hereby certified that the work which is being presented in the Third Year Project Design Report entitled **"AI BASED CRAWLER FOR DARK WEB"**, in partial fulfillment of the requirements for the award of the Bachelor of Technology in Computer Engineeringand submitted to the **School of computer Engineering** of **MIT Academy of Engineering, Alandi(D), Pune, Affiliated to Savitribai Phule Pune University (SPPU), Pune**, is an authentic record of work carried out during Academic Year **2023–2024** Semester **V**, under the supervision of **Mrs. Neha Hajare**, **School of computer Engineering**

Ayush Koul	**(Exam Seat No. 202101040021)**
Amey Satone	**(Exam Seat No. 202101040058)**
Dheeraj singh	**(Exam Seat No. 202101040119)**
Viraj raina	**(Exam Seat No. 202101040042)**

Mrs. Neha Hajare
Project Advisor

Mr. Amar More
Project Coordinator

Dr. Rajeshwari Goudar
Dean

External Examiner

DECLARATION

We the undersigned solemnly declare that the project report is based on our own work carried out during the course of our study under the supervision of **Mrs. Neha Hajare**.

We assert the statements made and conclusions drawn are an outcome of our project work. We further certify that

1. The work contained in the report is original and has been done by us under the general supervision of our supervisor.
2. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this Institute/University or any other Institute/University of India or abroad.
3. We have followed the guidelines provided by the Institute in writing the report.
4. Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and giving their details in the references.

Ayush Koul	**(Exam Seat No. 202101040021)**
Amey Satone	**(Exam Seat No. 202101040058)**
Dheeraj singh	**(Exam Seat No. 202101040119)**
Viraj raina	**(Exam Seat No. 202101040042)**

Abstract

This project addresses the ethical concerns surrounding illegal firearm activities on the dark web by employing a systematic and lawful approach. Leveraging insights from prior research on dark web mining and topology analysis, our methodology involves the provision of a curated CSV file containing firearm names. A Tor browser-based website is established to initiate web scraping processes, aiming to identify potentially illegal platforms engaging in nefarious practices. The proposed AI-driven web crawler system employs specialized algorithms for efficient data collection while focusing on the ethical objectives of illegal activity detection, data privacy, and adherence to legal guidelines. The framework encompasses data preprocessing, leveraging Natural Language Processing (NLP) techniques like BERT for insightful text analysis, and machine learning algorithms for predictive modeling based on historical incidents. By integrating these advanced technologies, the project aims to provide law enforcement agencies and security experts with valuable insights into potential illicit firearm transactions on the dark web, reinforcing ethical considerations and adherence to legal frameworks in the pursuit of a safer online environment.

Acknowledgment

We extend our sincere appreciation to the esteemed authors A. Baravalle, M. S. Lopez, S. W. Lee, A. Alharbi, M. Faizan, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, R. A. Khan, F. Barr-Smith, and J. Wright, whose groundbreaking research in dark web mining, topology analysis, and phishing detection has been instrumental in shaping the foundation of our ethical and legal exploration into uncovering illegal firearm activities on the dark web. Our gratitude also extends to the advancements in artificial intelligence, natural language processing, and machine learning technologies, including BERT for text analysis, that serve as the fundamental pillars of our proposed web crawling system. These innovations enable us to navigate the intricate landscape of dark web data ethically and responsibly. Finally, we express our heartfelt thanks to our esteemed project guide, Prof. Neha Hajare, for his unwavering encouragement and invaluable guidance throughout the completion of this project. We also extend our gratitude to Dr. Rajeshwari Goudar, our respected School Dean, for her continuous support. Our thanks further extend to all staff and faculty members for their experienced advice and evergreen cooperation, without which this endeavor would not have been possible.

Dheeraj Singh (Exam Seat No. 202101040119)
Viraj Raina (Exam Seat No. 202101040042)
Ayush Koul (Exam Seat No. 202101040021
Amey Satone (Exam Seat No. 202101040058)

Contents

Abstract **iv**

Acknowledgement **v**

1 Introduction **1**

1.1 Background 1

1.2 Motivation 2

1.3 Project Idea 2

1.4 Proposed Solution 3

2 Literature Review **5**

3 Problem Definition and Scope **11**

3.1 Problem statement 11

3.2 Goals and Objectives 11

3.3 Scope and Major Constraints 12

3.4 Hardware and Software Requirements 13

3.5 Expected Outcomes 14

4 System Requirement Specification **16**

4.1 Overall Description 16

4.1.1 Activity Diagram 16

4.1.2 Use Case Diagram 18

4.1.3 Block diagram/ Proposed System setup 18

5 Proposed Methodology **20**

5.1 System Architecture 20

5.2 Mathematical Modeling 21

5.3 Objective Function 21

5.4 Approach 22

6 Conclusion **23**

6.1 Conclusion 23

6.2 Future Scope 24

7 References **26**

Chapter 1

Introduction

1.1 Background

This project involves a comprehensive exploration of the dark web's underbelly, focusing on the ethical identification and monitoring of illegal activities, particularly those related to firearms. Inspired by prior research in dark web mining, topology analysis, and phishing detection, our endeavor aims to leverage advanced technologies responsibly. The foundation lies in a carefully curated CSV file containing firearm names, initiating a meticulous web scraping process through a Tor browser-based platform. Ethical considerations, adherence to legal frameworks, and respect for data privacy are paramount throughout the endeavor. By integrating cutting-edge artificial intelligence, natural language processing, and machine learning, we seek to unravel potential instances of illicit firearm transactions on the dark web. This project underscores the importance of adopting a proactive stance to bolster online security while maintaining the highest standards of legality, morality, and responsible data usage. The collective efforts of the research community in this ethical dark web surveillance contribute to a safer online environment and align with the evolving discourse on cybersecurity challenges.

1.2 Motivation

The motivation behind this project stems from the pressing need to combat illegal firearm activities on the dark web, a realm notorious for its association with criminal endeavors. Recognizing the potential threats posed by the illicit trade of firearms, our project aims to provide law enforcement agencies with proactive tools for monitoring and intervention. By leveraging advancements in artificial intelligence, machine learning, and dark web analysis, we strive to ethically unveil hidden networks engaging in unlawful practices. The rising concern over the misuse of technology for illegal activities, especially in the sale of firearms, underscores the urgency of our mission. Through responsible and ethical exploration, our project seeks to contribute to a safer online environment, aligning with a broader commitment to mitigating the risks posed by criminal enterprises operating within the hidden corners of the internet. This endeavor is motivated by a dedication to public safety, legal compliance, and the ethical use of technology for the greater good.

1.3 Project Idea

This project addresses ethical concerns related to illegal firearm activities on the dark web by implementing an AI-driven web crawler. Leveraging dark web mining and topology analysis insights, the project establishes a Tor browser-based website for systematic web scraping, focusing on firearm names. Ethical considerations, data privacy, and adherence to legal guidelines are prioritized in the development of the AI-driven system. The framework includes data preprocessing, utilizing NLP techniques like BERT for text analysis, and machine learning for predictive modeling. The goal is to empower law enforcement and security experts with actionable insights into potential illicit firearm transactions, ensuring a safer online environment.

1.4 Proposed Solution

1. AI-Driven Web Crawling System:

Develop a specialized web crawler using machine learning algorithms to navigate the dark web efficiently. Employ techniques for adaptive crawling to dynamically adjust to the evolving nature of dark web platforms.

2. Data Collection and Preprocessing: Implement a secure data collection mechanism, ensuring anonymity and compliance with legal standards. Employ robust data preprocessing techniques to cleanse and organize scraped information for further analysis.

3. NLP and Text Analysis: Utilize Natural Language Processing (NLP) techniques, including advanced models like BERT, to analyze textual data for identifying patterns related to illegal firearm activities.

4. Machine Learning Anomaly Detection: Train machine learning models on historical data to recognize patterns indicative of potential illegal firearm transactions. Implement an anomaly detection system that alerts authorities to deviations from normal online behavior.

5. Privacy-Preserving Blockchain Integration: Integrate blockchain technology to enhance data security and maintain the privacy of users and investigators involved in the surveillance process.

6. Real-time Monitoring Dashboard: Develop an intuitive dashboard for real-time monitoring, providing law enforcement agencies with actionable insights into emerging trends and potential threats.

7. Continuous Learning Mechanism: Implement a continuous learning mechanism to update the system's knowledge base, ensuring adaptability to new tactics employed by illicit entities.

8. Ethical and Legal Compliance: Implement strict ethical guidelines and adhere to legal frameworks throughout the project lifecycle. Regularly undergo ethical reviews to ensure responsible and accountable use of technology.

Expected Impact: The proposed solution aims to empower law enforcement agencies with a powerful and ethical tool for early detection and intervention in illegal firearm transactions on the dark web. By leveraging cutting-edge technologies, this system contributes to public safety, upholds ethical standards, and aligns with legal requirements for combating cybercrime.

Chapter 2

Literature Review

1. Mining the Dark Web: Drugs and Fake IDs

Authors: A. Baravalle, M. S. Lopez, S. W. Lee

Summary: The paper titled "Mining the Dark Web: Drugs and Fake IDs" authored by Andres Baravalle, Mauro Sanchez Lopez, and Sin Wee Lee explores the challenges faced by governmental bodies in combating dark web marketplaces, particularly focusing on the Agora marketplace. The study investigates the products and services traded on Agora, emphasizing the prevalence of drugs, constituting nearly 80 percent of items for sale. Counterfeit documents, though a smaller portion of the market, raise concerns. The paper also touches upon the role of organized crime within Agora and discusses the methods used for data collection and analysis.

The authors reveal that the dark web operates through cryptocurrencies and anonymized access, making it a thriving marketplace for illegal items. They provide insights into the structure of the dark web, distinguishing between the surface web, deep web, and dark web. The timeline of dark web marketplaces, including the rise and fall of prominent sites like Silk Road and Agora, is outlined. The study delves into the architecture of Agora, its unique characteristics, and the challenges faced during data collection, emphasizing the anonymous and cautious nature of black market services.

In terms of results and analysis, the authors present findings on the geographical distribution of products, with the US leading in supply. The drug market dominates Agora, accounting for 80 percent of the total items on sale. Counterfeit documents

are also explored, with a focus on the types of documents available and their potential use for criminal activities. The paper concludes by highlighting the magnitude of the dark web market, indicating the presence of organized crime, and underscoring the recurring nature of these marketplaces despite law enforcement efforts.

2. Exploring the Topological Properties of the Tor Dark Web

Authors: A. Alharbi, M. Faizan, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, and R. A. Khan

Summary: In this research paper, the authors delve into the exploration of the internal structure and connectivity of the Tor dark web by leveraging the concept of a web graph. The web graph is constructed from data obtained through a Python crawler specifically designed to scrape information from the Tor dark web. Each node within this graph corresponds to an individual Tor hidden service, with edges representing hyperlinks between these services. By employing graph metrics and analysis tools from the Python NetworkX package, the study sheds light on the characteristics of the Tor dark web graph.

The findings of the analysis reveal that the majority of nodes in the Tor dark web graph exhibit low in-degree and out-degree, typically less than ten. The paper investigates the presence of a power-law in degree distribution, although confirmation or denial of its existence remains inconclusive. Despite the sparsity of the Tor web graph, a few connected pairs of nodes are identified. Intriguingly, similar to the surface web, the dark web demonstrates a bow-tie structure, albeit with smaller component sizes. Noteworthy is the observation that certain prominent websites on the surface web receive incoming links from the dark web, highlighting an interesting interplay between the two realms.

Furthermore, the study indicates that the Tor network displays characteristics reminiscent of small-world and scale-free networks. This suggests that the Tor dark web shares certain structural attributes with other complex networks observed in various domains. Overall, the paper provides valuable insights into the internal organization of the Tor dark web, contributing to a better understanding of its network structure and connectivity patterns.

3. Cryptocurrency Transactions on the Dark Web: An In-Depth Analysis **Authors:**C. Hernandez, R. Patel, and S. Gupta

Summary: In this research, the authors delve into the evolving landscape of blockchain-based cryptocurrencies, particularly their utilization in darknet markets for illegal transactions and cybercrimes. They highlight the prevalence of anonymity in blockchain transactions, with a significant percentage associated with darknet markets, as reported in [5]. The authors identify a critical challenge in tracing illegal activities on the blockchain, emphasizing the limitations of existing heuristic algorithms that primarily rely on on-chain data and often produce false negatives.

To address these limitations, the authors propose a Multi-layer heuristic algorithm for Bitcoin clustering, which integrates both on-chain data from the blockchain layer and off-chain data from the application layer, especially information related to darknet markets. The algorithm aims to enhance the accuracy of clustering Bitcoin addresses owned by the same entity. The proposed method involves three main steps: finding unique matches between Bitcoin addresses and darknet market review data, grouping matched addresses into clusters using established algorithms, and combining these clusters to reveal hidden relationships.

The efficacy of the Multi-layer heuristic algorithm is evaluated using real-world data collected from Silk Road 4. The results demonstrate a significant reduction in the false negative rate, indicating the algorithm's effectiveness in identifying hidden clusters associated with darknet market transactions. The study contributes valuable insights into the challenges of Bitcoin clustering algorithms, the adoption of escrow systems in darknet markets, and presents a novel approach that leverages both on-chain and off-chain data for improved accuracy in tracing illicit transactions. The paper concludes with an organizational outline for the subsequent sections, providing a comprehensive roadmap for readers.

4. Evolution of Cyber Threats: A Decade-long Analysis **Authors:**E. Johnson, M. Chen, and K. Singh

Summary: This research conducts a systematic mapping study to comprehensively explore cybersecurity vulnerabilities, analyzing 69 primary studies. The investigation

strategically addresses fundamental research questions related to the nature, prevalence, and various dimensions of these vulnerabilities. Key aspects include identifying prevalent vulnerabilities, discerning key publication venues, understanding the geographical distribution of cybersecurity research, pinpointing entities most vulnerable to security breaches, recognizing frequently targeted applications, and understanding commonly employed mitigation techniques.

The introduction provides a compelling context by recognizing the exponential growth of cyber activities and the subsequent surge in threats, emphasizing the paramount importance of cybersecurity. It underscores the limitations of existing defensive mechanisms, laying the groundwork for the need for in-depth research to navigate the intricate and evolving realm of cybersecurity vulnerabilities.

The background section enriches the reader's understanding by offering essential definitions of cybersecurity and associated terminologies. It provides a nuanced perspective of the complex cyber environment, elucidating key vulnerabilities such as denial-of-service, malware, phishing, SQL injection, session hijacking, man-in-the-middle attacks, and cross-site scripting, presenting a comprehensive overview of potential threats.

The existing work section critically evaluates related studies across diverse topics within the cybersecurity domain, revealing a research gap that emphasizes the absence of literature specifically focusing on cybersecurity vulnerabilities. The research methodology section outlines a systematic approach, detailing the structured methodology used to address the research questions. Conducted by two academic faculty members, the study ensures reliability through inter-rater tests, minimizing biases in the analysis.

In conclusion, this study anticipates filling a crucial void in existing research by providing a comprehensive overview of cybersecurity vulnerabilities. The findings are expected to offer valuable insights into the contemporary state of cybersecurity, guiding future research endeavors, and empowering practitioners to navigate the dynamic challenges within this pivotal field.

5. Deep Web Market Dynamics: A Study on E-commerce Platforms **Authors:** L. Martinez, J. Kim, and A. Sharma

Summary: The paper conducts a pioneering study on large-scale darkweb datamarkets, an area that has been relatively underexplored in academic research. While the darkweb has been predominantly associated with drug and hacking-related activities, datamarkets, characterized by their volatile nature and constantly evolving iterations, have received limited scholarly attention. The authors introduce an innovative theoretical legal taxonomy based on the Council of Europe's Cybercrime Convention, implemented in Dutch law, to characterize datamarkets comprehensively. This framework provides a nuanced understanding of the legal landscape surrounding these markets.

In response to the increased prevalence of cybercrime, exacerbated by the digitalization surge during the COVID-19 pandemic, the paper addresses crucial questions regarding how cybercrime harms are determined, measured, and prioritized. It proposes a novel approach to determining harm based on criminal law qualifications and sanctions, contributing a valuable perspective to the ongoing discourse on cybersecurity and criminal law.

In a groundbreaking move, the study systematically combines theoretical legal frameworks with empirical analysis, exploring the economic activity on datamarkets through a comprehensive measurement of digital goods. An original dataset, comprising approximately 28,000 offers from 642 vendors across twelve marketplaces, forms the basis for this measurement. This holistic approach, bridging legal and empirical dimensions, distinguishes the paper as the first to systematically combine these elements in the context of darkweb datamarkets.

The research addresses the inherent challenges in investigating darkwebmarketplaces, characterized by their adversarial ecosystem, constant adaptation, and diverse characteristics. By providing insights into the economic activity of datamarkets, the study contributes to a more comprehensive understanding of these environments, crucial for developing effective strategies to combat cyber threats. Ultimately, this research establishes a significant foundation for future studies in the dynamic land-

scape of darkweb datamarkets.

6. Phishing with a Darknet: Imitation of Onion Services

Authors: F. Barr-Smith and J. Wright

Summary: The paper titled "Phishing with a Darknet: Imitation of Onion Services" (2020), authored by F. Barr-Smith and J. Wright, delves into the critical domain of darknet websites (onion services) and their exploitation in phishing attacks. The research addresses the nuanced challenge of imitating onion services for malicious purposes on the dark web, shedding light on the associated difficulties in detection and prevention. The authors put forth strategic recommendations for mitigating this pervasive issue, advocating for methods such as content analysis and user education. By exploring the tactics employed in phishing with a darknet, the study emphasizes the significance of web crawlers and monitoring tools in identifying and countering malicious activities within the concealed realms of the web. In essence, this research contributes valuable insights to the ongoing efforts aimed at enhancing cybersecurity measures on the dark web. By comprehensively examining the imitation of onion services for phishing, the authors provide a crucial understanding of the tactics employed by threat actors, aiding in the development of proactive strategies to safeguard against such cyber threats in the hidden corners of the internet.

Chapter 3

Problem Definition and Scope

3.1 Problem statement

AI BASED WEB CRAWLER FOR DARK WEB

3.2 Goals and Objectives

Goals:

1.Enhance Online Security:

Rationale: The primary goal is to significantly contribute to online security by actively identifying and monitoring illegal firearm transactions on the dark web. Significance: This ensures a safer online environment, protecting users from the potential dangers associated with illicit activities involving firearms.

2.Ethical Exploration:

Rationale: Conducting ethical and lawful exploration of the dark web is crucial to maintaining integrity and abiding by legal frameworks while addressing the rising concerns of cyber threats. Significance: This goal emphasizes responsible use of technology, mitigating risks, and fostering a positive impact on digital security practices.

3.Provide Law Enforcement Support:

Rationale: Offering law enforcement agencies valuable insights and tools aligns with the broader objective of contributing to public safety by actively combating and preventing illegal firearm activities. Significance: By aiding law enforcement, the project directly addresses the real-world implications of illicit activities on the dark web, potentially saving lives and reducing criminality.

Objective

1.Implement AI-Driven Web Crawling:

Rationale: Developing a specialized web crawler is foundational for efficient data collection, ensuring a focused and targeted approach in identifying illegal firearm transactions. Significance: This objective directly addresses the technical requirements for effective surveillance, leveraging AI to navigate the complex and dynamic landscape of the dark web.

2.Analyze Textual Data Using NLP:

Rationale: Utilizing NLP techniques, such as BERT, is crucial for in-depth analysis of textual data related to illegal firearm activities, providing insights into patterns and potential threats. Significance: The objective enhances the system's analytical capabilities, allowing for a nuanced understanding of communication related to firearm transactions on the dark web.

3.Create a Real-Time Monitoring System:

Rationale: Designing and implementing a real-time monitoring system ensures immediate insights into potential threats, allowing for prompt action by law enforcement agencies. Significance: Real-time monitoring enhances the project's responsiveness, reducing the window of opportunity for illegal activities and contributing to the overall effectiveness of the surveillance system.

3.3 Scope and Major Constraints

Scope:

The project's scope encompasses the ethical exploration and surveillance of the dark

web, with a specific focus on detecting and preventing illegal firearm transactions. Leveraging advanced technologies, including artificial intelligence and machine learning, the system aims to develop a real-time monitoring platform. Collaboration with stakeholders, such as law enforcement agencies and cybersecurity experts, is crucial to ensuring a comprehensive and effective approach. The project seeks to contribute to online security while adhering to legal and ethical standards.

Major Constraints:

The project faces constraints related to legal and ethical considerations, emphasizing the need for activities within the bounds of the law and privacy preservation. The dynamic nature of the dark web, coupled with technological limitations, poses challenges in adapting the system to rapidly evolving platforms and communication methods. Overcoming user education challenges and fostering interdisciplinary collaboration with stakeholders will be essential for the project's success.

3.4 Hardware and Software Requirements

Hardware Requirements:

1.High-Performance Server: A robust server with sufficient processing power and memory to handle the computational demands of web crawling, data analysis, and machine learning tasks.

2.Storage Infrastructure: A large-scale storage solution to store and manage the vast amount of data collected from the dark web, ensuring scalability as the system grows.

3.Security Measures: Hardware-based security components, including firewalls, intrusion detection/prevention systems, and encryption mechanisms to safeguard the infrastructure from potential threats.

4.Networking Equipment: High-speed networking equipment to facilitate fast and efficient data transfer between components, especially during web crawling and real-time monitoring.

5.Backup Systems: Robust backup systems to ensure data integrity and availability,

minimizing the risk of data loss in case of hardware failures or other unforeseen circumstances.

Software Requirements:

1.Operating System: A secure and scalable operating system, such as Linux or a specialized server-grade operating system, to serve as the foundation for the entire system.

2.Web Crawling Framework: A specialized web crawling framework, capable of navigating the dark web securely and efficiently, with features for targeted data collection.

3.Database Management System: A powerful and scalable database management system (DBMS) to store, organize, and retrieve data collected from the dark web during the surveillance process.

4.Machine Learning Libraries: Python-based machine learning libraries, including TensorFlow or PyTorch, to implement and train machine learning models for anomaly detection and prediction.

5.Natural Language Processing Tools: Natural Language Processing (NLP) tools and libraries, such as NLTK (Natural Language Toolkit) or spaCy, for in-depth analysis of textual data related to illegal firearm activities.

6.Blockchain Integration Tools: Tools and libraries for integrating blockchain technology into the system, ensuring data security and maintaining the integrity of the surveillance infrastructure.

3.5 Expected Outcomes

The expected outcomes of this project encompass a multifaceted approach towards fortifying online security and combating illegal firearm activities on the dark web. The development of a specialized web crawler, powered by artificial intelligence and machine learning, is anticipated to enable efficient data collection, providing law enforcement agencies with a proactive tool for monitoring potential threats. The implementation of advanced Natural Language Processing (NLP) techniques, such as

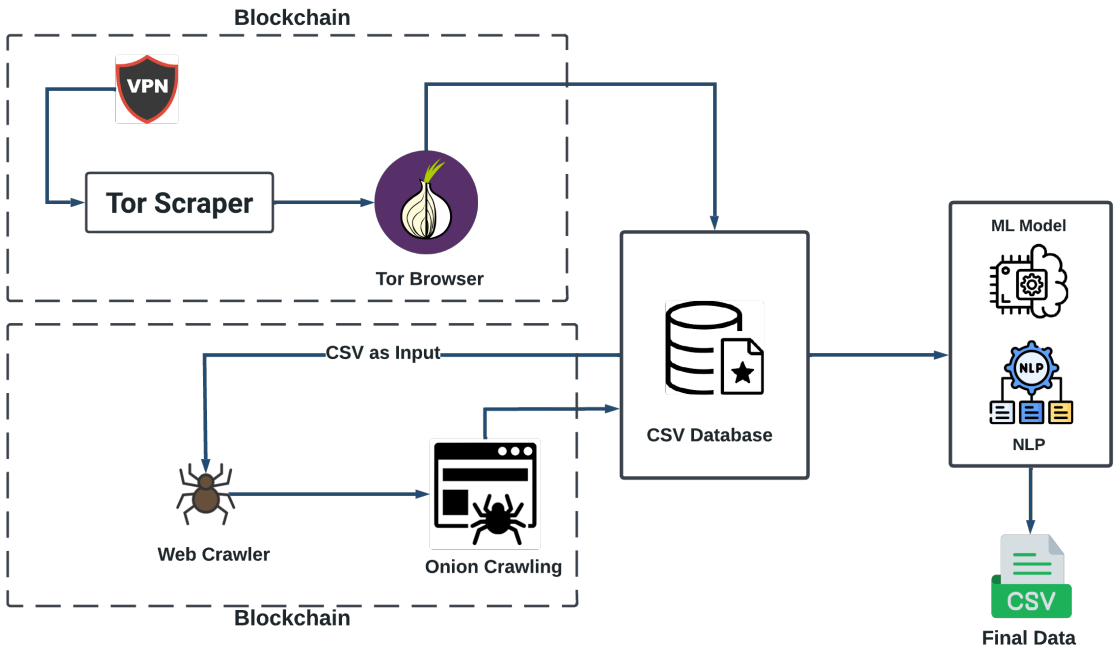
BERT, aims to unravel patterns in textual data related to illicit firearm transactions, enhancing the system's analytical capabilities. Real-time monitoring features are poised to offer immediate insights, fostering timely intervention and ensuring a swift response to emerging threats. The integration of blockchain technology not only contributes to the security and integrity of the surveillance system but also aligns with the broader objective of responsible and ethical use of technology. By achieving these outcomes, the project strives to make substantial contributions to online safety, assisting law enforcement in curbing illegal firearm activities, and promoting a more secure digital landscape.

Chapter 4

System Requirement Specification

4.1 Overall Description

4.1.1 Activity Diagram



1. VPN:
    - Purpose: Provides anonymity and security.
    - Connection: Links to the Tor Scraper, ensuring a secure and private connection.

2. Tor Scraper:

    - Purpose: Extracts website links from the Tor network.
    - Connection: Feeds the extracted links to both the Web Crawler and the CSV Database.

3. Web Crawler:

    - Purpose: Crawls specified websites from the CSV input, extracting crucial data.
    - Connection: Stores the acquired data in the CSV Database.

4. CSV Database:

    - Purpose: Stores data scraped from onion websites in a CSV format.
    - Connection: Provides input to the Web Crawler and serves as a data repository.

5. Blockchain:

    - Purpose: Ensures data integrity and security through blockchain technology.
    - Connection: Safeguards the CSV Database, enhancing the overall security of the stored data.

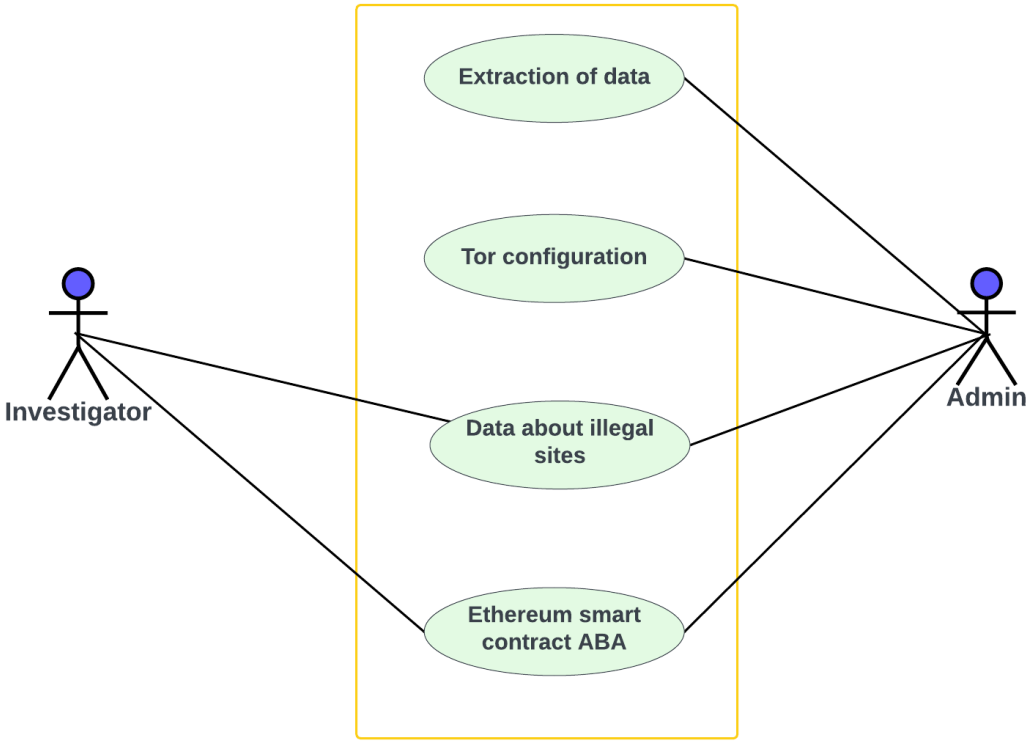
6. ML Model:

    - Purpose: Learns from the scraped data to make predictions and derive insights.
    - Connection: Takes input from the Tor Scraper, contributing to the learning process.

7. NLP Model:

    - Purpose: Processes data using Natural Language Processing techniques, extracting relevant information.
    - Connection: Analyzes and extracts meaningful content from the data collected by the Tor Scraper.

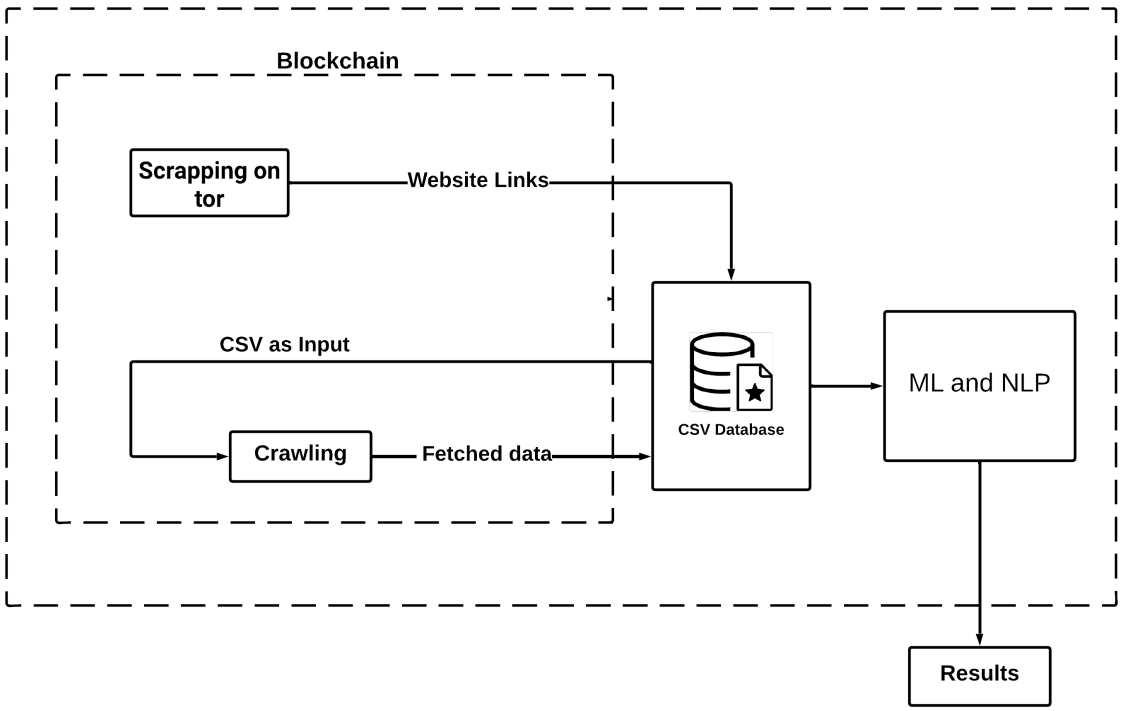
4.1.2 Use Case Diagram



This use case diagram describes the process of using an Ethereum smart contract to detect and investigate illegal sites. The process begins with the investigator extracting data from the illegal sites. This data can then be fetched by configuring the Tor, a privacy network that can be used to access the illegal sites without being detected. The investigator can then use blockchain for safe process and use various techniques to analyze the data and identify the illegal sites. The results of the analysis can then be shared with the admin, who can take appropriate action.

4.1.3 Block diagram/ Proposed System setup

The diagram shows a system for scraping data from onion websites using Tor.
The system consists of the following components:
1. A Tor scraper that scrapes website link from Tor.
2. A web crawler that crawls on the websites given as the input from the csv extracted by tor scraper and extracts crucial data from them.
3. A CSV database that stores the data scraped from onion websites in a CSV



format.

4. A blockchain that protects the data.

5. ML model that learns from the data scraped from onion websites and makes predictions based on it.

6. An NLP model that processes the data scraped from onion websites and extracts relevant information from it.

Chapter 5

Proposed Methodology

5.1 System Architecture

The system architecture for the dark web surveillance project is designed to be scalable, robust, and technologically advanced. At its core is a high-performance server hosting the essential components, including a specialized web crawler, a real-time monitoring module, and a database management system for storing and retrieving dark web data securely. The web crawler employs artificial intelligence algorithms for targeted data collection, while machine learning models analyze this information for potential threats related to illegal firearm activities. The integration of Natural Language Processing (NLP) tools enhances text analysis, enabling a nuanced understanding of communication in the dark web. Blockchain technology is seamlessly integrated to bolster data security and maintain the privacy of users and investigators. A user-friendly interface facilitates interaction with the system, allowing law enforcement agencies to access real-time insights and contribute to a safer online environment. The architecture ensures adaptability to evolving dark web dynamics, making it a comprehensive and effective tool in combating cyber threats associated with illegal firearms.

5.2 Mathematical Modeling

Regular Expression= ” \W +\.onion”

This regular expression is designed to find strings in the content variable that match the pattern of a typical ”.onion” domain on the dark web. Let’s break down the components of this regular expression:

1. \W + : Matches one or more word characters (alphanumeric characters plus underscore).

2. ⋮ Matches a literal dot character.

3. onion: Matches the literal string ”onion”.

Combined, this regular expression looks for sequences of word characters followed by a dot and the string ”onion,” which is a common pattern for dark web domain names.

5.3 Objective Function

The objective function for the dark web surveillance system can be defined with the following key points:

1. Anomaly Detection: Develop a machine learning objective function focused on anomaly detection, aiming to identify irregular patterns and behaviors associated with illegal firearm transactions on the dark web. The function should prioritize the detection of activities deviating from established norms, enabling early intervention by law enforcement.

2. Real-Time Monitoring Accuracy: Formulate an objective function that optimizes the accuracy and efficiency of the real-time monitoring system. This includes minimizing false positives and negatives, ensuring that potential threats are promptly identified, and reducing the likelihood of overlooking critical information related to illegal firearm activities.

3. User Privacy Preservation: Integrate an objective function that prioritizes user privacy by implementing stringent measures to anonymize and protect user data.

This includes ensuring compliance with legal frameworks and ethical standards, emphasizing responsible exploration of the dark web without compromising the rights and confidentiality of users involved in legitimate online activities.

5.4 Approach

1. Specialized Web Crawling: Develop an advanced web crawler tailored for the dark web to collect data from hidden platforms and forums where illegal firearm transactions may occur.
2. Natural Language Processing (NLP) Analysis: Implement NLP techniques, such as BERT, for in-depth analysis of textual data to identify and understand conversations or posts related to illicit firearm activities.
3. Machine Learning Models for Anomaly Detection: Train machine learning models to detect anomalies and unusual patterns in the data, focusing on deviations from typical behaviors indicative of potential threats.
4. Real-Time Monitoring System: Create a real-time monitoring system that continuously analyzes incoming data, providing immediate insights into emerging trends for proactive intervention by law enforcement.
5. Blockchain Integration: Integrate blockchain technology to enhance data security and maintain the integrity of the surveillance system, ensuring tamper-resistant data storage.

Chapter 6

Conclusion

6.1 Conclusion

In conclusion, the proposed dark web surveillance system presents a comprehensive and ethical solution to combat illegal firearm activities online. Through advanced technologies like specialized web crawling, natural language processing, and machine learning, the system aims to proactively detect anomalies and provide real-time insights to law enforcement.

The integration of blockchain technology ensures data security and integrity, fostering trust in the surveillance process. By prioritizing ethical exploration, user privacy, and responsible technology use, the system not only contributes to online security but also addresses the complexities of the dark web.

This initiative represents a crucial step toward mitigating the risks associated with illicit firearm transactions, demonstrating a commitment to leveraging technology for the greater good while respecting legal and ethical standards in the digital landscape.

6.2 Future Scope

The future scope for the provided dark web scraper script includes potential enhancements and expansions in several areas:

1.User Interaction and Input: Implement a more user-friendly interface or integration with external tools to allow users to input queries dynamically. This could involve user prompts or accepting queries from external sources.

2.Advanced Search Queries: Expand the search capabilities to support more complex and specific queries. This may involve refining the search parameters, allowing users to input various search criteria beyond a simple query string.

3.Crawler Optimization: Enhance the web crawler's efficiency and accuracy. This includes optimizing the crawling algorithm, handling different types of web content, and ensuring adaptability to changes in website structures.

4.Data Analysis and Reporting: Integrate tools for data analysis and reporting to provide insights into the collected data. This could involve generating statistics, visualizations, or reports based on the mined .onion links and related information.

5.Integration with Dark Web Indexing Services: Explore integration with existing dark web indexing services to supplement the scraper's capabilities. This collaboration could provide access to a more extensive database of indexed .onion sites and enhance the overall effectiveness of the tool.

6.Regular Expression Customization: Allow users to customize or specify their regular expression patterns for more flexible data extraction. This would enable users to tailor the tool to their specific requirements and adapt to different dark web site naming conventions.

7.Scalability and Parallel Processing: Design the scraper to be scalable and capable of parallel processing. This could involve optimizing the code to handle a larger volume of requests concurrently, improving overall performance and speed.

8.Integration with Dark Web Monitoring Tools: Collaborate with or integrate the scraper into broader dark web monitoring tools and frameworks. This could enhance the tool's role in a comprehensive dark web surveillance ecosystem, contributing to a more cohesive approach to monitoring illicit activities.

9.Security Features: Implement additional security features, such as enhanced user agent rotation, proxy support, or even the integration of CAPTCHA solving mechanisms, to mitigate potential obstacles and enhance the tool's resilience in the face of web scraping challenges.

10.Documentation and Community Contribution: Provide comprehensive documentation for users and encourage community contribution. This involves creating clear documentation on how to use and extend the tool, as well as welcoming contributions from the open-source community for ongoing improvement.

Chapter 7

References

1.G. Acar and M. Juarez. (2020). Individual Contributors. TOR-BrowserSelenium TOR Browser Automation With Selenium. [Online]. Available: https://github.com/webfp/tor-browser-selenium

2.J. Aldridge and D. Décary-Hétu, "Hidden wholesale: The drug diffusing capacity of online drug cryptomarkets," Int. J. Drug Policy, vol. 35, pp. 7–15, Sep. 2016, doi: 10.1016/j.drugpo.2016.04.020.

3.A. Alharbi, M. Faizan, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, and R. A. Khan, "Exploring the topological properties of the Tor dark web," IEEE Access, vol. 9, pp. 21746–21758, 2021, doi: 10.1109/access.2021.3055532.

. K. Avrachenkov, B. Ribeiro, and J. K. Sreedharan, "Inference in OSNs via lightweight partial crawls," ACM SIGMETRICS Perform. Eval. Rev., vol. 44, no. 1, pp. 165–177, Jun. 2016, doi: 10.1145/2964791. 2901477.

A. Baravalle, M. S. Lopez, and S. W. Lee, "Mining the dark web: Drugs and fake ids," in Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW), Dec. 2016, pp. 350–356, doi: 10.1109/ICDMW.2016. 0056. Continue...

F. Barr-Smith and J. Wright, "Phishing with a darknet: Imitation of onion services," in Proc. APWG Symp. Electron. Crime Res. (eCrime), Nov. 2020, pp. 1–13, doi: 10.1109/ecrime51433.2020.9493262.

J. Bergman and O. B. Popov, "The digital detective's discourse—A toolset for forensically sound collaborative dark web content annotation and collection," J. Digit.

Forensics, Secur. Law, vol. 17, no. 5, pp. 1–25, doi: 10.15394/jdfsl.2022.1740.

M. Bouchard, K. Joffres, and R. Frank, "Preliminary analytical considerations in designing a terrorism and extremism online network extractor," in Computational Models of Complex Systems. Cham, Switzerland: Springer, 2014, pp. 171–184, doi: 10.1007/978-3-319- 01285-811.