

Dark Web Data Leak Monitoring System

S. Amaresan¹, S. Pranav², M. Sivasakthi³, G.K. Varshini⁴, S. Vignesh⁵

¹ Associate Professor, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India
Email: amaresansai2009@gmail.com

²UG Student, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India
Email: Pranavshankar313@gmail.com

³UG Student, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India
Email: sivasakthi8424@gmail.com

⁴UG Student, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India
Email: varshinisekaran0537@gmail.com

⁵UG Student, Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur, Tamil Nadu – 613006, India
Email: vigneshselvakumar.off@gmail.com

Abstract:

The dark web has emerged as a major platform for the illegal trade of stolen personal, financial, and corporate data, posing serious security risks to organizations. This project presents a Dark Web Data Monitoring system that continuously scans hidden forums, marketplaces, and leak sites to identify potential data exposure using predefined, organization-specific keywords. The proposed system employs automated crawlers operating over secure Tor networks to collect publicly accessible dark web content, while keyword-based detection monitors email domains, usernames, company names, and other sensitive terms associated with organizations and individuals. In addition, the system incorporates data filtering and pattern-matching techniques to improve detection accuracy and minimize false positives, and it analyzes collected data based on relevance, source credibility, and potential impact. The findings are stored in a structured database to enable historical tracking and trend analysis, while a user-friendly dashboard provides centralized visibility for security teams to review alerts and respond efficiently. Designed with scalability and modularity, the system can adapt to evolving cyber threats and integrate with existing security infrastructures, thereby strengthening an organization's overall cybersecurity posture while also providing practical exposure to cyber threat intelligence, OSINT methodologies, and ethical monitoring practices.

Keywords — Dark Web Monitoring, Tor Network, Cyber Threat Intelligence, Data Leak Detection, Keyword-Based Crawling, OSINT, Web Scraping, Intrusion Detection.

I. INTRODUCTION

The increasing dependence on digital platforms has led to a significant rise in cyber threats, including data breaches, identity theft, and unauthorized access to sensitive information. Many of these threats originate from the dark web, a hidden section of the internet that operates through anonymizing technologies such as Tor and is not accessible

through conventional search engines. The dark web is widely used for illegal activities such as trading stolen credentials, personal data, financial information, and confidential organizational records.

Due to its anonymous and decentralized nature, monitoring activities on the dark web is complex and resource-intensive. As a result, organizations often become aware of data

exposure only after serious security incidents have occurred. The domain of Dark Web Monitoring focuses on proactively identifying potential data leaks and cyber threats by continuously scanning hidden online sources.

These systems analyze collected data, assess risk levels, and generate alerts for high-risk findings, enabling early detection and timely response to security incidents. This project falls under the Cybersecurity domain, specifically targeting dark web threat intelligence and risk analysis. The proposed system enables users to monitor specific keywords, evaluate security risks, and visualize insights through an interactive dashboard. Although the application uses simulated data for ethical and legal reasons, it reflects real-world cybersecurity workflows and highlights the importance of proactive monitoring in protecting sensitive digital assets.

II. RELATED WORK

Sayuj Shah and Vijay K. Madiseti [1] introduced MAD-CTI (Multi-Agent Dark Web Cyber Threat Intelligence), a novel multi-agent Large Language Model framework designed to extract actionable Cyber Threat Intelligence (CTI) from dark web forums, blogs, and articles. Their research addresses the growing surge in cyber threats — with organizational data breaches rising 72% between 2021 and 2023 — by targeting the dark web, where over 60% of illicit activity occurs and where emerging threats are often discussed before surfacing in the real world. The proposed system, built on OpenAI's GPT-4o mini and Microsoft AutoGen, employs a sequential pipeline of specialized agents: a Translator Agent, Text Analysis Agent, Relevancy Classifier, and Category Classification Agent, each handling a distinct sub-task to classify dark web content as Hack, Malware, or Vulnerability. Evaluated against the CoDA dataset of 650 cybersecurity-related texts, MAD-CTI achieved a relevancy classification accuracy of 68.01% and threat categorization accuracy of 63.52%, outperforming comparable models such as DarkBERT and ThreatCrawl, while demonstrating an 11.55% improvement over single-agent approaches — enabling scalable threat detection.

Daeun Kim et al. [2] explored the use of a keyword-based dark web collection framework for automated monitoring of Tor hidden services and illegal online activities. Their research focuses on gathering Tor hidden service addresses and associated Bitcoin wallet information using crime-related keywords, addressing challenges posed by the anonymous nature of the Tor network and the frequent expiration of .onion domains. The proposed framework analyzes the lifecycle of onion domains, keyword relationships, and cryptocurrency transaction patterns to support cyber threat intelligence. The study demonstrates that automating data collection and analysis through a structured database enables efficient dark web

monitoring and assists investigators in understanding emerging trends in illegal activities such as hacking, financial fraud, and cryptocurrency abuse.

Gun-Yoon Shin et al. [3] explored the use of a TextCNN-based classification framework integrated with topic modeling keyword weighting for automated dark web content analysis. Their research focuses on extracting and prioritizing class-specific keywords to reduce word vector dimensions and improve classification efficiency. The proposed approach filters out non-essential textual content and emphasizes influential keywords, outperforming conventional machine learning and deep learning models on real dark web datasets.

Yong Fang et al. [4] explored the use of an automated system for identifying data breach-related threads in underground forums across both the surface web and the dark web. Their research focuses on feature extraction using the Latent Dirichlet Allocation (LDA) topic model combined with statistical identifiers specific to breach-related content. The proposed approach employs a Random Forest classifier, successfully identifying over 92% of data breach threads. Experimental results demonstrate its effectiveness in proactive cyber threat intelligence and early incident response, outperforming traditional data leakage detection approaches.

Saiba Nazah et al. [5] explored a comprehensive review of existing dark web threat analysis and detection methodologies, examining key trends in criminal behavior and evaluating the effectiveness of current analytical techniques. Their research highlights the growing role of forum analysis, underground marketplace monitoring, cryptocurrency transaction tracking, and machine learning-driven approaches in cyber threat intelligence. The study identifies critical research gaps and emphasizes the necessity for scalable, adaptive, and legally compliant solutions to strengthen dark web threat detection and prevention efforts against cybercriminal activities.

Jesper Bergman and Oliver B. Popov [6] explored a systematic literature review of dark web crawling techniques by analyzing 34 peer-reviewed studies to identify common crawler architectures, programming languages, and scraping tools. Their research highlights that Python, Selenium, and Scrapy are the most commonly used technologies for Tor-based crawling, addressing challenges posed by strong encryption, limited traceability, and the unstable nature of onion services. The study additionally develops and evaluates a Tor-enabled crawler model, demonstrating effective data collection from both clear web and dark web sources to support digital forensics and cyber threat intelligence.

Seokhee Lee et al. [7] explored the use of social network-driven cyber threat intelligence by analyzing cybercrime-related data collected from X (formerly Twitter) to identify attack trends and tactics, techniques, and procedures (TTPs) targeting the Arab world. Their research employs keyword-based scraping, natural language processing, and statistical analysis to examine cyberattacks

across countries, industries, attack types, and threat groups. The findings reveal that sectors such as gas, government, and healthcare are frequently targeted, with social engineering, ransomware, and DDoS attacks being the most discussed threats, demonstrating the effectiveness of social network-driven CTI in supporting proactive threat detection and regional cybersecurity preparedness.

Sapna Sadhwani et al. [8] explored the use of an intelligent cyber threat detection framework called SmartSentry, leveraging machine learning and deep learning techniques for improved intrusion detection in Industrial IoT (IIoT) networks. Their research evaluates several algorithms including Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbor, and Deep Neural Network (DNN) using the Edge-IIoTset benchmark dataset, with data preprocessing methods such as PCA and SMOTE applied to enhance model accuracy. Experimental results indicate that the DNN-based approach consistently outperforms other models, demonstrating SmartSentry's scalability and effectiveness in securing industrial IoT infrastructures against sophisticated and rapidly evolving cyber threats.

III. PROPOSED METHODOLOGY

The proposed system is a Dark Web Data Leak Monitoring System designed to detect sensitive information leaks in real time. It continuously scans dark web sources such as forums, marketplaces, and hidden services using automated crawlers. The system uses the Tor network to securely access and collect publicly available dark web data. It applies keyword-based detection to identify leaked data related to organizations, such as emails, passwords, and confidential records. A central processing engine analyzes the collected data and filters relevant threats. The system stores identified leaks in a structured database for further investigation. It provides alerts and notifications to administrators when potential data exposure is detected. The admin dashboard allows monitoring, analysis, and management of detected leaks. Users can also view summarized reports and risk levels through a client interface. Overall, the system improves security by enabling early detection and response to data breaches.

a. System Overview

The Dark Web Data Leak Monitoring System consists of five interconnected modules that form a complete monitoring pipeline. The Tor Proxy Module routes all requests anonymously through the Tor network via SOCKS5, while the Spidering Module crawls .onion URLs, extracts hyperlinks, and manages a queue to avoid duplicate visits. The Web Scraper Module downloads and parses HTML through the Tor proxy using BeautifulSoup to extract meaningful text for analysis, and the CAPTCHA Handling Module detects verification challenges, solves them via APIs, and uses Selenium to simulate human interaction. The Database Manager Module stores all scan results, bcrypt-

hashed credentials, and keywords in SQLite with role-based access control for users and admins. Together, these modules create a modular, scalable cybersecurity tool for proactively detecting data leaks across hidden dark web platforms.

b. Authentication (RBAC)

The application handles secure user login and registration using bcrypt and jsonify. During registration, the user submits their credentials, and the plain-text password is hashed using bcrypt.hashpw() with an auto-generated salt, then stored in the SQLite database along with the assigned role (user or admin). During login, the entered password is verified against the stored hash using bcrypt.checkpw() — if it matches, the server responds with a jsonify object containing the authentication status, user role, and session token. If credentials are invalid, a jsonify error response with an appropriate HTTP status code (e.g., 401 Unauthorized) is returned immediately. The module also enforces role-based route protection — on every subsequent request, the user's role is validated before granting access, ensuring regular users only reach their own scan data while admins access the full control panel. This design keeps authentication stateless, secure, and consistent across all API endpoints.

c. Tor Connection

The proposed system focuses on establishing a secure and anonymous connection to the dark web by routing all requests through the Tor network. The system establishes a SOCKS5 proxy connection so that it can access .onion websites securely and anonymously. It also verifies Tor connectivity and rotates Tor circuits when connections fail.

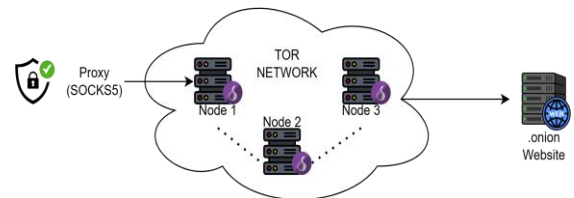


Fig 1. Tor Traffic Routing

During this stage, several important operations are performed, including:

- **Tor Network Connection:** The system initiates the scanning process by sending HTTP requests to .onion websites. The requests are routed through the Tor SOCKS5 proxy, which acts as an intermediary between the crawler and the Tor client.
- **Proxy Routing:** The proxy forwards requests into the Tor network instead of the system's direct internet connection, ensuring anonymous communication with .onion websites.
- **Circuit Building:** The Tor daemon receives proxy requests and builds encrypted circuits through the

Tor network. Traffic is routed through Entry, Middle, and Exit nodes, providing multi-hop anonymity before reaching the .onion website.

- **Circuit Rotation:** If blocking or connection issues occur, the system sends a NEWNYM signal via the Tor ControlPort. This generates a new Tor circuit and relay path, allowing the system to continue scanning anonymously.

d. Spidering Process

The system discovers and crawls dark web pages to find links that should be scanned. It navigates onion sites, extracts internal hyperlinks, and builds a queue of pages for further analysis while avoiding duplicate crawling.

The following operations are performed during the Spidering process:

- **Initial URL Discovery:** The system starts with initial .onion URLs and automatically finds new links from them. It helps in discovering more pages without manual input, which increases the number of pages available for scanning.
- **HTML Parsing and Link Extraction:** The system reads the HTML content of each webpage and analyzes its structure. It looks for anchor tags <a> to extract hyperlinks (href), and only valid .onion links are selected for further processing.
- **Queue Management:** All extracted links are stored in a queue for organized crawling. The system maintains a visited list to avoid visiting the same page again, ensuring efficient and non-repetitive crawling.
- **Crawl Control:** The system follows rules such as depth limits and domain restrictions to control exploration. This prevents infinite loops and unnecessary crawling, ensuring smooth and structured traversal of dark web pages.

e. Web Scraper

The system retrieves and extracts readable text content from dark web pages by downloading HTML pages through the Tor proxy, removing unnecessary tags, and preparing clean text for keyword detection. It is a core module of the system, responsible for transforming raw web data into usable information.

The following steps are performed during this process:

1. Input Data:

The system accepts initial .onion URLs along with user-defined keywords and scan depth settings as input. These parameters guide the crawling and scanning process throughout the system. The keywords define what content to search for, while the scan depth controls how deep the system navigates into dark web pages.

2. Spidering:

The system starts with the provided .onion URLs and crawls each page to discover and extract internal

hyperlinks. All found links are stored in a queue and organized based on the defined scan depth limit. The system maintains a visited list to avoid duplicate crawling and ensures structured traversal of dark web pages.



Fig 2. Data Extraction Workflow

3. HTML Download Page:

The system retrieves each queued .onion page by sending HTTP requests through the Tor SOCKS5 proxy. The raw HTML content of the page is downloaded and prepared for structured analysis and data extraction. This ensures that all page content is securely fetched through the Tor network before further processing.

4. Content Extraction:

The downloaded HTML page is parsed using the BeautifulSoup library to convert it into a structured DOM tree. The system traverses the DOM elements such as tags, text blocks, links, and data fields to extract meaningful textual content. Unnecessary tags and irrelevant elements are removed to retain only clean and readable text data.

5. Keyword Matching:

The extracted text content is analyzed and compared against the user-provided keywords. The system scans through the cleaned text to detect any occurrences or matches of the defined keywords within the page content. Matched results are flagged and recorded for further reporting and analysis within the monitoring system.

f. CAPTCHA Handling

The system includes a CAPTCHA handling component that detects and manages CAPTCHA challenges encountered during the web crawling and data collection process. When a CAPTCHA is identified, the system analyzes the webpage structure and captures the CAPTCHA image for further processing. The captured image is preprocessed using the Pillow library by adjusting properties such as cropping, filtering, and resizing to improve readability. The preprocessed CAPTCHA is then sent to CAPTCHA solving APIs, which use automated or human-assisted services to decode and return a valid response. Finally, browser automation tools such as ChromeDriver and GeckoDriver with Selenium simulate human-like interactions to submit the solved response and continue the crawling process.

During this stage, several important operations are performed, including:

- **CAPTCHA Detection:** The system detects CAPTCHA challenges encountered during automated web crawling by analyzing webpage

structures and identifying verification elements.

- **CAPTCHA Solving:** The processed CAPTCHA is sent to CAPTCHA solving APIs, which use automated or human-assisted services to decode the challenge. The solved CAPTCHA response is then submitted to the webpage to continue the crawling process.
- **Image Preprocessing:** Captured CAPTCHA images are processed using the Pillow library for image preprocessing. This helps improve readability by adjusting image properties such as cropping, filtering, or resizing.
- **Browser Automation:** The system uses browser automation tools such as ChromeDriver and GeckoDriver with Selenium to load webpages and interact with CAPTCHA-protected pages. These drivers control real browsers to simulate human-like interactions.

g. Database Manager

The system includes a database manager component that stores and manages scan results, users, and dark web sources. It records scan outputs, maintains keyword data, and provides historical scan information for the frontend dashboard and admin panel.

This process includes several key features such as,

- **Structured Data Storage:** The system stores all important data in an organized way, including user details, .onion URLs, scan results, and keywords. All data is saved in structured tables or collections for easy access.
- **Password Security:** User passwords are securely stored using the bcrypt library, which converts plain text passwords into irreversible hash values with added salt for security. During login, the entered password is hashed again and compared with the stored hash to ensure secure authentication.
- **Scan Result Recording:** The system saves scan results along with important details such as URL, content, and matched keywords. Each record includes a timestamp for tracking, which helps in analyzing what data was found and when.
- **Scan History and Access Control:** The system maintains a history of all scans for future reference. Users can view their results, while admins can manage data and users, ensuring proper access control and efficient monitoring.

IV. SYSTEM ARCHITECTURE

The system architecture is designed as a modular framework that includes Admin, User, and Core Crawler Engine components, each performing a specific role. The

Admin module is responsible for configuring, controlling, and monitoring the entire system. The User module provides a simple interface for users to access scan results and reports. The Core Crawler Engine handles data collection, processing, and detection of sensitive information from the dark web. All components work together in an integrated manner to ensure efficient, secure, and automated detection of data leaks.

a. Admin Module

The Admin Module acts as the central control unit of the system, allowing administrators to configure, monitor, and manage the entire dark web leak detection process. Through the Admin-side web application, the administrator can define scanning parameters such as target keywords, URLs, and frequency of crawling. This module also provides a dashboard interface that displays analytical insights, including detected leaks, scan results, and system performance metrics.

Additionally, the admin has the authority to initiate or terminate scanning operations, manage stored data, and review logs generated by the crawler engine. The integration components within this module ensure seamless communication with the backend services and the core crawler engine. Overall, the Admin Module ensures that the system operates efficiently, securely, and according to predefined configurations.

b. User Module

The User Module represents the client-side interface through which end users interact with the system. It is designed to provide a simple and accessible way for users to retrieve information about potential data leaks without requiring technical knowledge of the backend processes. Through the client-side web application, users can submit queries, view scan results, and access reports generated by the system.

This module may also include automated assistance features, such as bots, to help users navigate the system or interpret results. The output provided to users typically includes summarized reports, alerts, and structured data indicating possible exposures of sensitive information. Unlike the admin module, user access is limited to viewing and interacting with results rather than controlling system operations.

c. Core Crawler Engine

The Core Crawler Engine is the backbone of the system, responsible for performing all data collection, processing, and analysis tasks. It operates by connecting to the Tor network, which enables anonymous access to hidden .onion services on the dark web. Using multiple nodes, the system ensures secure and untraceable communication while crawling restricted environments.

The engine employs spidering techniques to navigate through web pages and a web scraping mechanism to extract relevant data. The collected data is then processed using

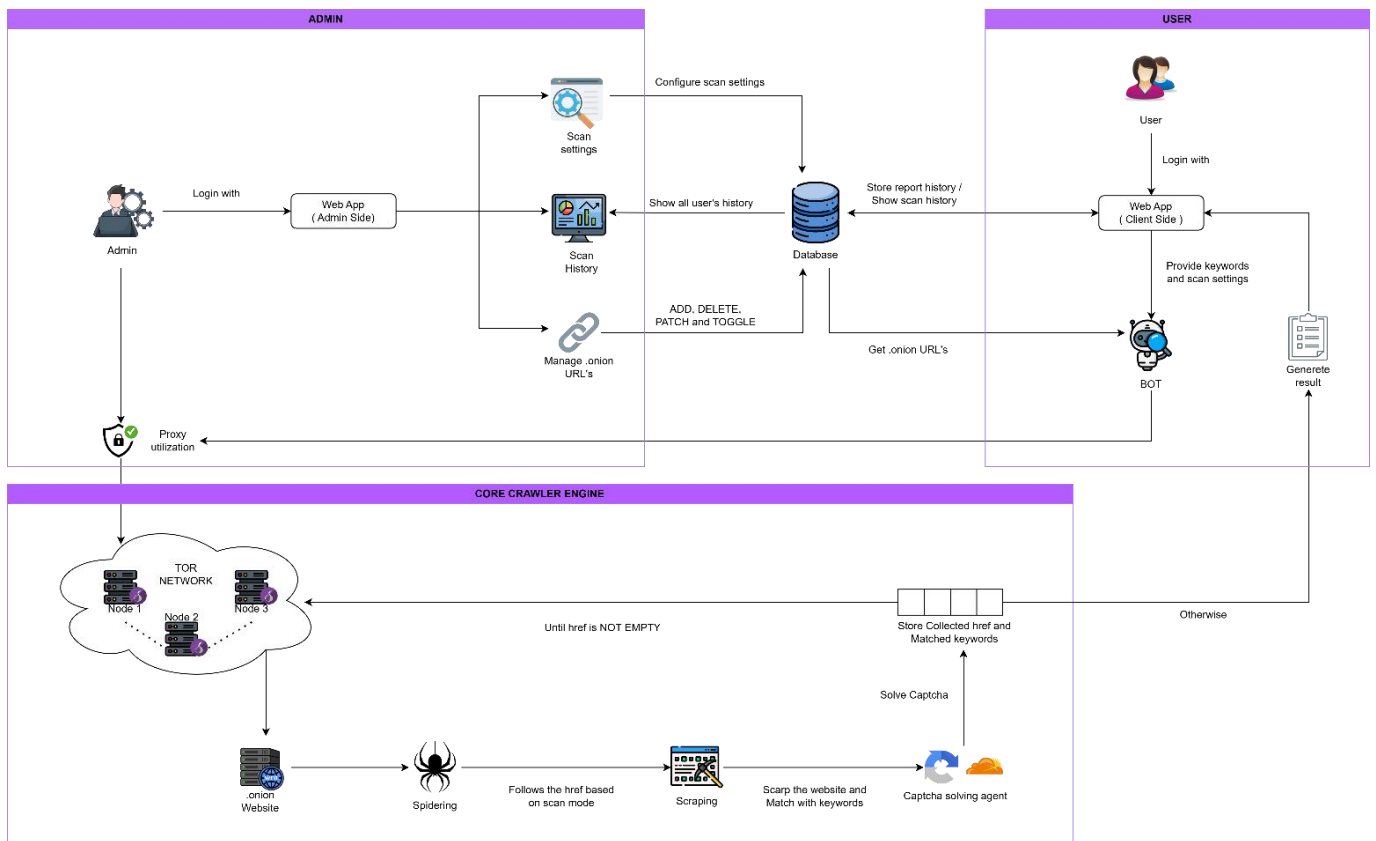


Fig 3. Architecture Diagram

keyword detection and pattern-matching algorithms to identify sensitive information such as credentials, personal data, or confidential records. CAPTCHA handling mechanisms may also be integrated to bypass access restrictions during crawling.

Once processed, the extracted information is stored in a database and made available to both the admin and user modules. The system supports continuous and automated scanning, ensuring real-time monitoring and timely detection of potential data leaks. This module essentially drives the entire functionality of the system by transforming raw dark web data into meaningful insights.

V. EXPERIMENTAL SETUP

The experimental setup is designed to evaluate the performance and efficiency of the proposed system in a controlled environment. It focuses on analyzing how effectively the system performs crawling, data extraction, and keyword detection. Different parameters are configured to observe system behavior under varying conditions. The setup also helps measure processing time, accuracy, and resource utilization during execution. Overall, it ensures reliable testing and validation of the system's functionality and performance.

a. Environment Setup

The proposed system is evaluated by conducting experiments in a secure and isolated environment using Oracle VirtualBox with Kali Linux as the guest operating system. This virtualized setup provides enhanced security by containing all dark web activities within a controlled space. Kali Linux is selected due to its compatibility with cybersecurity tools and network analysis utilities. The Tor Browser is installed and configured, and the Tor network is activated with a SOCKS5 proxy running on port 9050, enabling anonymous communication and secure access to .onion websites.

b. Dataset Preparation

For the experiment, a dataset of more than 50 .onion URLs is initially collected from publicly available sources. To ensure controlled testing and performance evaluation, 10 URLs are selected from this dataset. These URLs represent different types of content and active pages on the dark web, providing a diverse environment for testing. This selection helps in analyzing the system's behavior under realistic conditions while keeping the experiment manageable.

c. Spidering Configuration

The spidering module is configured with a queue length of 50 links per URL. This parameter defines the

maximum number of internal links that can be extracted and processed from each seed URL. The system maintains a queue to organize the crawling process and a visited list to avoid duplicate page visits. This configuration ensures efficient traversal of web pages while preventing infinite loops and unnecessary resource consumption.

d. Processing Workflow

During execution, the system performs a complete pipeline that includes spidering, HTML page downloading, content extraction, and keyword matching. The spidering process begins by crawling each selected .onion URL and extracting internal links using HTML parsing. The system then downloads the HTML content of each page through the Tor proxy. The content is processed using parsing tools such as BeautifulSoup to extract clean and meaningful text by removing unnecessary tags and irrelevant elements.

e. Keyword Detection and Storage

The extracted text is analyzed using a keyword matching module, where it is compared against predefined keywords related to sensitive data such as credentials and personal information. When matches are found, the system records the details, including the source URL and matched content, into a structured database. This allows efficient storage, retrieval, and further analysis of detected information.

f. Performance Evaluation

The system's performance is evaluated based on metrics such as processing time and crawling efficiency. It is observed that scanning 10 .onion URLs with the configured settings takes approximately 1000 to 1500 seconds to complete the entire process. The variation in time depends on factors such as Tor network latency, page complexity, and the number of links discovered. The results demonstrate that the system is capable of effectively monitoring dark web content while maintaining security and scalability for future improvements.

VI. CONCLUSION

The dark web data leak monitoring system provides an efficient approach to identifying potential data breaches using automated crawling, scraping, and keyword detection techniques. By integrating Tor-based anonymous access, spidering, web scraping, CAPTCHA handling, and intelligent keyword matching, the system can securely analyze onion sites for sensitive information while maintaining anonymity. Its modular architecture, including components like the Tor Proxy, Spidering module, Scraper, Scan Manager, Detection module, and Database Manager, ensures efficient processing, maintainability, and flexibility for future improvements.

Additionally, the system supports structured data storage for better analysis, reporting, and tracking of detected information. It can be further enhanced by incorporating machine learning for improved threat detection and real-time alert mechanisms to notify administrators instantly. Overall, the project emphasizes the importance of automated cybersecurity solutions in proactively detecting threats and strengthening data protection in an increasingly complex digital environment.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to their project guide and faculty members of the Department of Computer Science and Engineering for their continuous support, valuable guidance, and encouragement throughout this research work.

The authors also extend their thanks to their external guide, Mr. T. Siva, from Cybersecurity Nerds Lab, for providing valuable insights, technical guidance and necessary resources that contributed to the successful completion of this work.

Additionally, the authors acknowledge the use of publicly available datasets and tools that supported the development and evaluation of the proposed system.

REFERENCES

- [1] S. Shah and V. K. Madiseti, "MAD-CTI: Cyber Threat Intelligence Analysis of the Dark Web Using a Multi-Agent Framework," *IEEE Access*, vol. 13, pp. 40158–40168, 2025.
- [2] D. Kim, Y. Park, and S. Kim, "A Measurement Study on Tor Hidden Services via Keyword-Based Dark Web Collection Framework," *IEEE*, 2024.
- [3] G.-Y. Shin et al., "Dark Side of the Web: Dark Web Classification Based on TextCNN and Topic Modeling Weight," *IEEE Access*, vol. 12, pp. 363 61–36375, 2024.
- [4] Y. Fang et al., "Analyzing and Identifying Data Breaches in Underground Forums," *IEEE Access*, vol. 7, pp. 48770–48780, 2019.
- [5] S. Nazah et al., "Evolution of Dark Web Threat Analysis and Detection: A Systematic Approach," *IEEE Access*, vol. 8, pp. 171796–171820, 2020.
- [6] J. Bergman and O. B. Popov, "Exploring Dark Web Crawlers: A Systematic Literature Review of Dark Web Crawlers and Their Implementation," *IEEE Access*, vol. 11, pp. 35914–35930, 2023.
- [7] S. Lee et al., "Leveraging Social Networks for Cyber Threat Intelligence: Analyzing Attack Trends and TTPs in the Arab World," *IEEE Access*, vol. 13, pp. 5679–5695, 2025.
- [8] S. Sadhwani et al., "SmartSentry: Cyber Threat Intelligence in Industrial IoT," *IEEE Access*, vol. 12, pp. 34720–34740, 2024.

