Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Dark Web Threat Detection System using Crawlers and NLP-based Threat Analysis

Mr. B. Ashok
Department of Information Technology
Keshav Memorial Institute of Technology
Hyderabad, India

Chaganti Sree Maanvith
Department of Information Technology
Keshav Memorial Institute of Technology
Hyderabad, India

Chinthakindi Keerthi
Department of Information Technology
Keshav Memorial Institute of Technology
Hyderabad, India

Ale Harini
Department of Information Technology
Keshav Memorial Institute of Technology
Hyderabad, India

Sunke Sruthi
Department of Information Technology Keshav Memorial Institute of Technology Hyderabad, India

***Abstract*—The rapid growth of illicit activities on the dark web including the exchange of stolen credentials and sensitive organizational data, has rendered traditional monitoring techniques insufficient. Existing dark web monitoring systems often generate high volumes of irrelevant alerts and lack contextual relevance, making it difficult for organizations to derive actionable threat intelligence. This paper proposes an organization-centric Dark Web Threat Intelligence System that integrates controlled web crawling, keyword-driven filtering, and Natural Language Processing (NLP)-based analysis to enable efficient and targeted threat detection. The proposed system employs a scheduled crawler to navigate dark web environments under predefined constraints, ensuring ethical and resource-efficient data collection. A keyword detection mechanism filters content specific to organizational assets, reducing noise and unnecessary processing. Relevant data is then analyzed using an NLP to understand contextual meaning and assess potential security risks. The system incorporates a threat scoring mechanism to classify detected incidents based on severity, enabling prioritized response and improved decision-making. Experimental observations indicate that the integration of NLP-based contextual analysis significantly reduces false positives compared to conventional keyword-based approaches while improving the accuracy of threat identification.**

Keywords—Dark Web, Cyber Threat Intelligence, Web Crawling, Keyword-Based Filtering, Natural Language Processing (NLP), Threat Detection, Data Leak Detection, Security Analytics

I. INTRODUCTION

The rapid advancement of digital technologies has significantly increased the volume of data generated and exchanged across online platforms. Alongside this growth, cybercrime has evolved in both scale and sophistication, with the dark web emerging as a prominent hub for illicit activities such as data breaches, identity theft, and unauthorized trading of confidential information [1], [9]. Sensitive organizational data, including employee credentials, internal documents, and financial records, is frequently exposed and circulated within these hidden networks [1]. This escalating threat landscape highlights the need for effective monitoring systems capable of identifying risks at an early stage. Traditional cybersecurity solutions primarily focus on securing surface web infrastructures and internal systems but lack the capability to monitor dark web environments due to their anonymous and decentralized nature [8]. Consequently, organizations often become aware of data breaches only after substantial damage has occurred [9]. This limitation underscores the importance of proactive dark web monitoring as a critical component of modern cybersecurity strategies.

To address these challenges, this paper proposes a Dark Web Threat Intelligence System that integrates controlled crawling, organization-specific keyword filtering, and NLP-based contextual analysis for improved threat detection. The system is designed to identify relevant data, assess its significance, and classify threats based on severity. By combining automated data collection with intelligent analysis, the proposed approach aims to provide a scalable and efficient solution for continuous monitoring of dark web environments and proactive cybersecurity defense.

II. LITERATURE REVIEW

The growing importance of dark web monitoring in cybersecurity has led to extensive research on data collection and threat intelligence extraction techniques. Several studies have focused on the development of automated crawlers designed to access and navigate dark web environments. These crawlers are capable of discovering hidden services and collecting large volumes of data from forums and marketplaces [3], [4]. However, existing crawling approaches primarily emphasize scalability and data acquisition, with limited focus on filtering relevant information or understanding the contextual significance of the collected content. To improve the relevance of collected data, keyword-based filtering techniques have been widely adopted for identifying sensitive information such as leaked credentials, organization names, and email addresses. While these methods are computationally efficient and easy to implement, they often suffer from high false positive rates due to the absence of contextual interpretation [6]. As a result, benign or unrelated mentions may be incorrectly

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

classified as potential threats, reducing the effectiveness of such approaches.

Recent advancements in machine learning and Natural Language Processing (NLP) have enabled more sophisticated analysis of dark web data. NLP-based techniques facilitate contextual understanding of textual content, allowing systems to distinguish between meaningful threats and irrelevant information [7]. These methods have demonstrated improved performance in identifying cyber threats, including data leaks and malicious discussions [2]. However, many existing solutions rely on large pre-collected datasets and apply NLP primarily as a post-processing step, rather than integrating it directly into the data collection pipeline.

The proposed system introduces a hybrid and organization-centric approach that integrates controlled dark web crawling, targeted keyword-based filtering, and NLP-driven contextual analysis within a single pipeline. Unlike traditional methods, the system prioritizes relevant data at early stages and applies intelligent analysis only when necessary, thereby reducing noise, improving efficiency, and enabling more accurate and actionable threat intelligence.

III. METHODOLOGY

A. System Overview

The proposed system adopts a hybrid, pipeline-based architecture for efficient dark web monitoring and threat detection. Unlike traditional approaches that separate crawling and analysis, the system integrates keyword-aware intelligence directly into the crawling process. This hybrid design enables selective data acquisition, reducing unnecessary data collection and improving overall efficiency. The system consists of five primary components: (i) Hybrid Keyword-Guided Crawler, (ii) Organization-Specific Keyword Module, (iii) Data Filtering Layer, (iv) NLP-Based Analysis Module, and (v) Threat Scoring and Alert System.

B. Crawling Strategy

The proposed system introduces a hybrid crawling strategy that combines traditional web crawling with keyword-guided intelligence. Unlike conventional crawlers that indiscriminately collect data, the hybrid crawler incorporates a real-time keyword matching mechanism to guide navigation and data extraction decisions, inspired by focused crawling techniques [4].

During the crawling process, each visited page is evaluated based on the presence of organization-specific keywords. Pages with higher keyword relevance are prioritized for deeper exploration, while irrelevant pages are discarded early in the pipeline. This approach reduces bandwidth consumption, limits exposure to unnecessary content, and improves processing efficiency. The hybrid crawler operates under predefined constraints, including crawl depth, frequency, and ethical limitations, ensuring controlled and responsible access to dark web resources.

C. Workflow Description

The workflow of the proposed system is defined as follows:

- **Input Initialization:** Define organization-specific keywords and crawling parameters.
- **Hybrid Crawling:** Navigate dark web pages while performing real-time keyword checks.
- **Relevance Evaluation:** Prioritize or discard pages based on keyword presence.
- **Content Extraction:** Extract textual data from relevant pages only.
- **Data Filtering:** Remove duplicates and noise.
- **NLP Analysis:** Perform contextual evaluation of filtered data[7].
- **Threat Scoring:** Assign severity levels based on analysis results.
- **Alert Generation:** Notify users of significant threats.

D. Algorithm

Algorithm 1 Hybrid Keyword-Guided Dark Web Crawling

```
Input: Keyword set K, Seed URLs U, Crawl interval T, Threshold θ
Output: Threat alerts with severity levels
Initialize keyword set K and seed URLs U
Schedule crawler with interval T
While system is active do
    for each URL in U do
        Fetch page content C
        Extract textual data D
        RelevanceScore = KeywordMatch(D, K)
        if RelevanceScore > 0 then
            Prioritize page for deeper crawling
            D_filtered = RemoveNoise(D)
            ThreatContext = NLP_Analyze(D_filtered)
            SeverityScore = ComputeSeverity(ThreatContext)
            if SeverityScore ≥ θ then
                GenerateAlert(ThreatContext, SeverityScore)
            end if
        else
            Discard page
        end if
    end for
end while
```

The threat severity score S is computed as:

$$S = w_1 \cdot K_r + w_2 \cdot C_s + w_3 \cdot D_s$$

where:

K_r= Keyword relevance score

C_s = Context score derived from NLP analysis

D_s = Data sensitivity score

w1,w2,w3 = weighting factors

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

IV. System Architecture

A. Overview

System architecture is designed as a modular and pipeline-based framework that integrates hybrid crawling, keyword-based filtering, and NLP-driven threat analysis.. The architecture emphasizes early-stage relevance detection and controlled data processing to improve efficiency and reduce noise. Each module performs a specific function, and data flows sequentially from input to threat reporting. The system operates in a continuous monitoring mode, where dark web sources are periodically accessed, analyzed, and evaluated for potential threats related to the target organization.

B. Architecture Description

The architecture of the proposed system consists of the following key components:

- **Input Module:** This module allows organizations to define custom parameters such as keywords, domain names, email identifiers, and sensitivity levels. These inputs guide the behavior of the crawler and filtering mechanisms.
- **Scheduler and Control Unit:** The scheduler manages the frequency and timing of crawling operations. It ensures that the system performs periodic scans based on predefined intervals while maintaining constraints such as crawl depth and access limitations.
- **Hybrid Keyword-Guided Crawler:** The crawler is responsible for accessing dark web pages using controlled navigation strategies. Unlike conventional crawlers, it integrates keyword-based relevance checking during the crawling process. Pages containing relevant keywords are prioritized, while irrelevant pages are discarded early, reducing unnecessary data collection.
- **Data Extraction Module:** This module extracts textual content and metadata from the selected web pages. The extraction process is lightweight and avoids full-scale scraping, ensuring ethical and efficient data handling.
- **Data Filtering Layer:** The extracted data is refined by removing duplicate entries, noise, and irrelevant content. This step ensures that only meaningful data is forwarded for further analysis.
- **NLP-Based Analysis Module:** The filtered data is processed using Natural Language Processing techniques to understand context and identify potential threats. This module distinguishes between benign mentions and actual data leaks or security risks[2],[5].
- **Threat Scoring Engine:** This computes a severity score based on keyword relevance, contextual understanding, and data sensitivity. The scoring mechanism enables classification of threats into different risk levels.
- **Alert and Reporting Module:** The final module generates alerts for detected threats and provides actionable insights through dashboards or notifications. It enables organizations to respond proactively to potential risks.

As illustrated in Fig. 1, the proposed system architecture is presented as a high-level abstraction. The detailed internal components, including the hybrid crawler, NLP module, and threat scoring engine, are encapsulated within the backend layer.

C. Data Flow Description

As shown in Fig. 2, the system follows a sequential data flow beginning with user-defined inputs and progressing through controlled crawling, filtering, and analysis stages. Data is continuously refined at each stage, ensuring that only relevant and high-quality information reaches the threat analysis module. The final output consists of prioritized alerts, enabling efficient decision-making and threat mitigation.

D. Key Innovation

The novelty of the proposed methodology lies in the integration of keyword-based relevance evaluation directly within the crawling process, forming a hybrid crawling mechanism. Unlike traditional approaches where filtering and analysis occur after data collection, the proposed system performs early-stage decision-making to selectively process only relevant information. This significantly reduces computational overhead, minimizes false positives, and enhances the efficiency of NLP-based threat analysis.

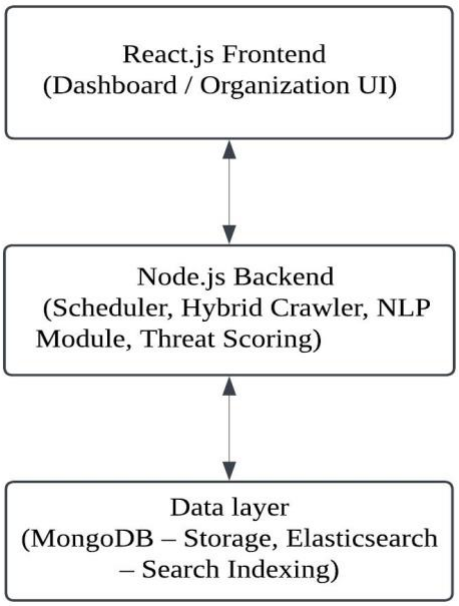

Fig. 1. Decoupled three-layer architecture of the proposed system.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

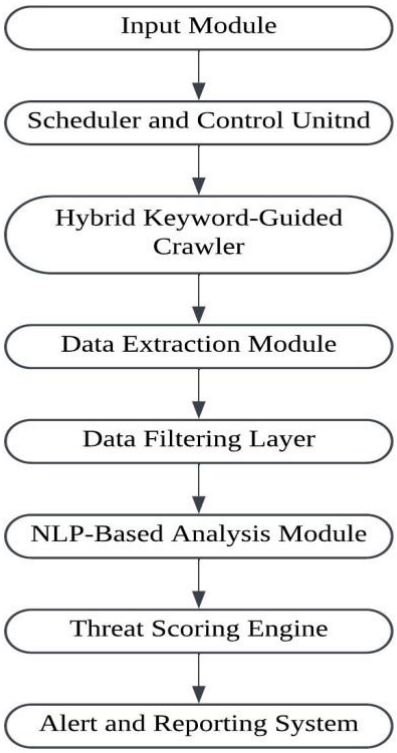

Fig. 2. Data flow of the proposed system

V. RESULTS AND DISCUSSION

A. Experimental Dataset

The proposed system was evaluated using a controlled dataset consisting of simulated dark web pages and publicly available text samples resembling real-world data leak scenarios. Organization-specific keywords such as email patterns, employee identifiers, and domain-related terms were used to guide the crawling and filtering process.

The performance of the system was compared against a baseline keyword-only detection model to highlight the effectiveness of the proposed hybrid approach integrating keyword filtering with NLP-based contextual analysis.

B. Dataset configuration

Due to the restricted and sensitive nature of dark web data, a fully real-world dataset could not be utilized. Therefore, a controlled experimental dataset was constructed to simulate realistic dark web data leak scenarios. The dataset consists of a combination of synthetic data and publicly available text sources that mimic the structure and content of dark web marketplaces and forums.

TABLE I
DATASET STATISTICS

Parameter	Value
Total Documents	1,000
Relevant Samples	400 (40%)
Irrelevant Samples	600 (60%)
Training Set	700 (70%)
Testing Set	300 (30%)

C. Evaluation Metrics

The performance of the system was evaluated using standard metrics:

- **Accuracy** – Overall correctness of classification
- **Precision** – Correctly identified threats out of detected threats
- **Recall** – Ability to detect actual threats
- **F1-Score** – Balance between precision and recall

D. Results

The performance comparison between the baseline model and the proposed system is presented in Table II.

TABLE II
PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-Score
Keyword-Based Detection	72%	68%	75%	71%
Proposed Hybrid System	89%	91%	87%	89%

The results demonstrate a significant improvement in performance when using the proposed hybrid approach. The keyword-based model achieves moderate accuracy but suffers from lower precision due to a high number of false positives. This occurs because keyword matching alone cannot distinguish between irrelevant mentions and actual data leaks. In contrast, the proposed system achieves higher precision (91%), indicating that most detected threats are indeed relevant. This improvement is attributed to the NLP-based contextual analysis, which enables the system to understand the meaning and intent of the extracted data.

E. Discussion

The integration of hybrid crawling and NLP-based analysis significantly enhances the efficiency and effectiveness of dark web monitoring systems. By filtering irrelevant data during the crawling phase, the system reduces unnecessary processing and improves scalability. The use of contextual analysis addresses one of the major limitations of existing systems, namely the inability to interpret the meaning of detected content. This

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

results in more accurate threat classification and actionable insights for organizations.

F. Security Considerations

Security is a critical aspect of the proposed system due to the nature of dark web interactions. The crawler operates in a controlled environment with restricted access policies to prevent exposure to malicious content.
Key security measures include:

- Isolation of the crawling environment
- Restricted network permissions
- Avoidance of direct interaction with suspicious or executable content
- Secure handling and storage of extracted data

G. Privacy Considerations

The system is designed to operate within ethical and legal boundaries by focusing only on organization-specific data provided as input. It does not perform indiscriminate data scraping or attempt to collect personal information beyond defined keywords.

The sensitive data processed by the system is handled securely and is not stored beyond its intended analytical purpose. The use of synthetic and publicly available datasets during evaluation ensures compliance with privacy regulations. This approach aligns the system with responsible cybersecurity practices while maintaining its effectiveness.

H. Limitations

The proposed system has certain limitations despite its advantages:

- The effectiveness of detection depends heavily on the quality and coverage of predefined keywords.
- Highly encrypted, or coded language may not be accurately detected.
- The simulated dataset may not fully represent the complexity and unpredictability of real-world dark web environments.
- NLP-based analysis, while effective, may still misinterpret ambiguous or context-poor text.

I. Future Enhancements

- Expansion of crawler capabilities to support multiple dark web networks and protocols.
- Incorporation of multilingual analysis to detect threats across different languages.

VI. Conclusion

This paper presents a Dark Web Threat Detection System that combines a hybrid keyword-guided crawler with NLP-based contextual analysis to detect potential data leaks related to organizations. Unlike traditional approaches, the proposed system performs early-stage filtering during crawling and applies context-aware analysis to reduce false positives and improve detection accuracy.

Experimental results show that the system outperforms conventional keyword-based methods in terms of accuracy and precision, while also providing meaningful threat severity assessment. The proposed approach offers an efficient and scalable solution for proactive cybersecurity monitoring, enabling organizations to identify and respond to potential threats more effectively.

References

[1] M. Chertoff and T. Simon, “The Impact of the Dark Web on Internet Governance and Cyber Security,” Center for International Governance Innovation, 2015.

[2] S. Dua and X. Du, “Data Mining and Machine Learning in Cybersecurity,” Auerbach Publications, 2011.

[3] M. K. Bergman, “The Deep Web: Surfacing Hidden Value,” Journal of Electronic Publishing, vol. 7, no. 1, 2001.

[4] S. Chakrabarti, M. van den Berg, and B. Dom, “Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery,” Computer Networks, vol. 31, no. 11–16, pp. 1623–1640, 1999.

[5] C. D. Manning, P. Raghavan, and H. Schütze, “Introduction to Information Retrieval,” Cambridge University Press, 2008.

[6] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries,” In Proceedings of the First Instructional Conference on Machine Learning, 2003.

[7] E. Cambria and B. White, “Jumping NLP Curves: A Review of Natural Language Processing Research,” IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48–57, 2014.

[8] M. Bishop, “Computer Security: Art and Science,” Addison-Wesley, 2003.

[9] Verizon, “Data Breach Investigations Report (DBIR),” 2024.