Journal homepage: https://itce.com.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 2. 2025**

Article's History: Received: 21.01.2025 Revised: 17.05.2025 Accepted: 28.08.2025

UDC 004.056.5 DOI: 10.31649/vitce/2.2025.63

Algorithms for searching and analysing information from open sources in the context of cyber threats

Oksana Onyshchuk*

PhD in Technical Sciences, Associate Professor
Lesya Ukrainka Volyn National University
43025, 13 Voli Ave., Lutsk, Ukraine
https://orcid.org/0000-0002-8342-3011

Lyudmila Hlynchuk

PhD in Physical-Mathematical Sciences, Associate Professor
Lesya Ukrainka Volyn National University
43025, 13 Voli Ave., Lutsk, Ukraine
https://orcid.org/0000-0002-8943-9604

Abstract. The article presented the development of an algorithm for searching and analysing information from open sources in the context of cyber threats. The proposed algorithm is an effective tool for detecting, monitoring, assessing and neutralising threats in the digital environment. The work described the main stages of the algorithm, which include: quick access to relevant information, assessment of data reliability, trend analysis, identification of connections between objects, and prediction of potential threats. The development of the algorithm involved searching for and analysing information from open sources; noise filtering; contextual analysis and cross-checking to improve the reliability of results; and constructing relationship graphs to identify dependencies between objects and determine their potential danger. These tasks were implemented by collecting data through API (Application Programming Interface), web scraping (BeautifulSoup), using search operators, processing data with NLP (Natural Language Processing) tools, and classification using machine learning models and regular expressions. The article analysed the information obtained using relationship graphs, identifies key objects and evaluates the reliability of sources. The developed algorithm reduced the time required for searching and analysing information, increased the relevance and accuracy of the data obtained, and provides effective support for cybersecurity decisions. As an example, the algorithm was applied to monitor suspicious job postings on the LinkedIn platform, where phishing ads containing invalid links or false information were detected. The use of the LinkedIn API and web scraping made it possible to automate the collection of job postings and compare them with a database of known phishing websites. The developed algorithm reduced the time spent searching for and analysing information compared to manual methods. The implementation of such solutions helps prevent cyber threats and ensure security in the digital environment. The algorithm significantly improves the efficiency of working with open sources, providing an automated process for collecting, processing, and analysing data for further threat assessment

Keywords: cybersecurity information search; data analysis; OSINT; relationship graphs; NLP; Google Search Operators; Maltego

Introduction

The relevance of information security research in the modern world is determined by the rapid development of digital technologies, the increase in the volume of data processed, and the growth of cyber threats. Information security is one of the key aspects of the stable functioning of organisations, government agencies, and private users. Malicious actors constantly refine their attack methods, developing new approaches to information protection and preservation becomes crucial. In this regard, scientific research in the field of information security becomes

Suggested Citation:
Onyshchuk, O., & Hlynchuk, L. (2025). Algorithms for searching and analysing information from open sources in the context of cyber threats. *Information Technologies and Computer Engineering*, 22(2), 63-73. doi: 10.31649/vitce/2.2025.63

*Corresponding author

particularly relevant. Various authors propose different data protection strategies, consider issues of cyber hygiene, and develop systems for detecting and preventing cyber threats. Thus, the issue of information security requires a thorough analysis of current challenges and the search for effective solutions to overcome them.

The rapid growth of digitalisation is leading to an increase in cyber threats such as hacker attacks, data leaks, phishing and malware. The work of R. Khan *et al*. (2019) provides a detailed study and analysis of the security and privacy threats faced by 5G networks, as well as considering existing and promising approaches to addressing them. It emphasises the importance of security as a key aspect for the widespread adoption of 5G. The use of open source intelligence (OSINT) has become an effective tool for collecting and analysing data to identify potential threats and neutralise them. Thanks to automated OSINT algorithms, it is possible to significantly reduce the time needed to detect suspicious activities, assess the level of risk and develop protective measures. M. Bazzell (2021) discusses methods of using open sources to collect and analyse information from the Internet, which can quickly detect cyber threats such as phishing or malware, and provides practical recommendations for cybersecurity professionals on the effective use of OSINT tools to protect information systems.

The issue of using open data to ensure cybersecurity is actively discussed in the global scientific community. And the developers of tools such as Google Search Operators Cheat Sheet (2021), FOCA (2022), and Recon-ng (2022) have provided an analysis of current global trends in cybersecurity. Growing cyber threats and the importance of using innovative technologies to ensure resilience to attacks, such as OSINT and machine learning, have highlighted the importance of cooperation between government agencies and private companies to combat cybercrime. Recent studies have shown significant progress in the development of open source analysis methods. It is important to use tools such as Recon-ng, Censys, and OSINT Framework, which greatly facilitate the process of collecting data from Internet resources for further analysis and establishing links between threats.

Innovative methods proposed in the works of A. Nagy *et al*. (2025) used machine learning to improve the accuracy of cyberattack prediction based on open source data. The article by T. Yang *et al.* (2024) examines the effectiveness of anomaly detection algorithms in the digital environment and the importance of automating threat monitoring processes. This study focused on the use of Bayesian deep learning to improve the accuracy of anomaly detection in cybersecurity, and this approach significantly reduced the level of uncertainty in threat monitoring and detection processes. Systematic reviews by T.S. AlSalem *et al*. (2023) and M.T. Islam *et al.* (2025) addressed cybersecurity issues in the Internet of Things (IoT) environment: risks due to vulnerabilities and the importance of using open sources as a means of detecting potential threats, threat typology, specific vulnerabilities, and protection strategies. These works prove that the processing of open data in the IoT field can provide predictive models for early warning of attacks.

The work of L.K. Vashishtha & K. Chatterjee (2025) focuses on the application of natural language processing (NLP) to analyse text data from social networks, which can detect phishing campaigns and disinformation. The research focused on new methods of threat detection using machine learning algorithms, in particular TestCloud IDS and SparkShield, which propose a dataset for training algorithms to detect anomalies and cyber threats in cloud environments. In this work, M. Alazab *et al*. (2024) propose an improved model for collecting and analysing threat data that can build more resilient cybersecurity systems and use modern OSINT tools to predict threats and mitigate risks in real time.

At the same time, an important aspect is the problem of the reliability of the collected data, since a large amount of information may be unreliable or unverified, which increases the risk of false positives. To overcome these challenges, it is proposed to use combined approaches that combine automated methods with manual data verification, which allows for more accurate results. At the same time, the use of machine learning and artificial intelligence, as shown in the works of A.K. Dey *et al*. (2023), can reduce the number of false positives and improve the effectiveness of real-time anomaly detection.

The aim of this study was to develop an algorithm for the automated collection, analysis and evaluation of open data for the detection of cyber threats. The algorithm combines machine learning, natural language processing (NLP) and graph analysis methods to improve the accuracy of threat detection. The main tasks to achieve this goal were to develop approaches for assessing data reliability, identifying trends, relationships between objects, and predicting potential threats. The use of automated tools will minimise the human factor, reduce time and resource costs, and increase the effectiveness of cyber incident prevention.

Materials and Methods

Data sources covering the LinkedIn social network were selected for the study. Data collection was carried out using official APIs, as well as web scraping methods using the BeautifulSoup library. The use of APIs provided official access to structured data, while web scraping obtained information from open web resources, including cybersecurity blogs and technical forums. To ensure the relevance of the information, data search and filtering mechanisms were implemented using: logical operators (AND, OR, NOT) to create complex search queries; regular expressions to identify specific data (phone numbers, IP addresses, domains, email addresses, file hashes); filtering by time, language, and location to improve search accuracy; data cleansing methods, including the removal of duplicates, advertisements, spam, and irrelevant information. Examples of search queries: “leaked database” AND “password” AND “2024”, “phishing site” OR “malware domain” AND “recent”, “ransomware” AND “new variant” AND “CVE-2024”, “credential stuffing attack” AND “affected services”.

The collected data was stored in JSON and XML formats, as well as in SQL/NoSQL databases (MongoDB, PostgreSQL). Natural language processing (NLP) methods were used to analyse unstructured text data, identify keywords, perform thematic clustering, and determine the tone of messages. The information search and analysis algorithm (Fig. 1) consisted of several key stages that ensured the efficiency of data collection, processing, and verification.

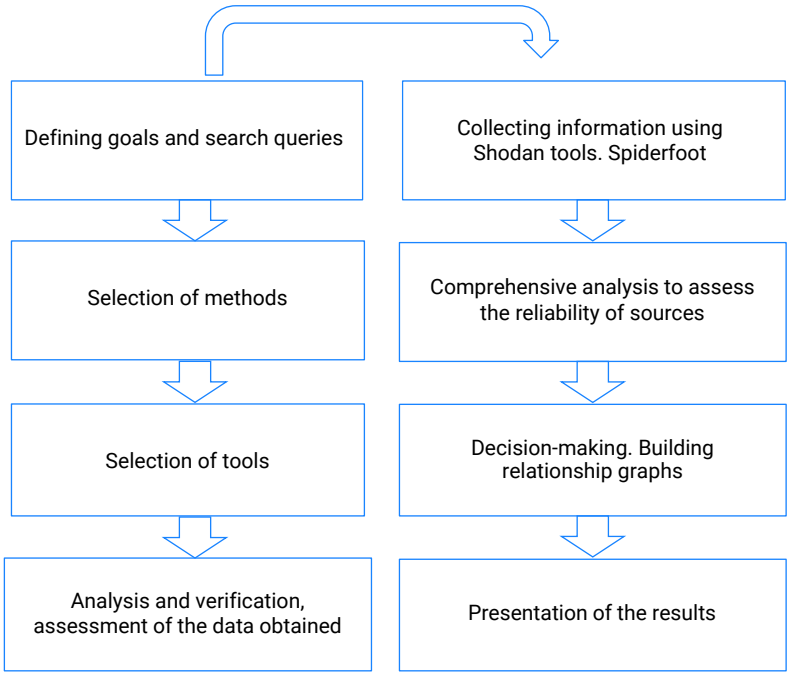

Figure 1. Algorithm for searching and analysing large volumes of data in the context of cyber threats
Source: developed by the authors based on Google Search Operators Cheat Sheet (2021), Spiderfoot (2022), Maltego Technologies (n.d.), Shodan (n.d.)

First, the research objectives were defined and search queries were formulated. This stage was critical, as clearly defined objectives allowed for a focus on relevant information sources. Search queries were formulated based on the selected topics, which facilitated effective data collection. In the second stage, information was collected using specialised tools. Automated tools such as Shodan (n.d.) and Spiderfoot (2022) were used to: conduct open source intelligence; obtain technical information about servers, IP addresses, domains, etc.; and identify potential security threats. Automating data collection significantly sped up the process and reduced the human factor. After collecting the info, they picked the right analysis methods, which included: checking how reliable the data was (cross-checking from different sources); finding anomalies in the data; using machine learning algorithms to spot hidden dependencies. Maltego Technologies (n.d.) and Google Search Operators (2021) tools were used to process and analyse open data, taking into account their features and specific purposes. Thanks to a comprehensive approach to assessing the reliability of sources, the risks of obtaining irrelevant or false information were reduced. To visualise the relationships between cyber threat objects, graphs were created to help identify key nodes and their connections, analyse the structure of potential threats, and determine potential attack vectors. Verification of the data obtained and assessment of its accuracy included checking the data against real threats. The results of this study were obtained according to the selected algorithm, using combined methods to ensure a balance between accuracy and efficiency. The proposed algorithm was verified using the example of monitoring suspicious job postings on the LinkedIn platform, performed by manual and automated data collection.

Results and Discussion

The proposed algorithm is presented as a clear and structured sequence of actions that provides a comprehensive approach to analysing and verifying information in order to ensure the accuracy and reliability of results. It enables effective data collection from various sources; minimises noise and excludes irrelevant information; improves the quality of cyber threat analysis; and ensures quick access to the necessary information. The choice of information search methods plays a key role in the quality of the results obtained. There are several approaches to data collection, which differ in the level of automation, accuracy, and speed of information processing. Table 1 lists the main methods, their advantages and disadvantages. The choice of search method depends on the specific goals and requirements of the user. If accurate and verified results are required, manual search may be optimal, but for large amounts of data, automated systems provide significant time savings.

Table 1. Data search methods

Method	Description	Advantages	Disadvantages
Manual search	Searching for information directly by the user through search engines and databases	Accuracy; control over the process	Time consuming; dependent on user skills
Automated systems	Using programs and algorithms to collect and analyse information	Speed; processing large volumes of data	Requires technical knowledge; risk of obtaining irrelevant data
Combined methods	Combining manual search and automated solutions	Flexibility; efficiency	Requires coordination; possible integration difficulties

Source: developed by the authors based on Google Search Operators Cheat Sheet (2021)

To effectively collect, process, and analyse open data, you need the right tools, each with its own features and specific purpose. The choice of tool depends on the type of information being analysed, the required data processing speed, and the user's technical skills. Some of them specialise in visualising connections between objects, while others are needed to find vulnerabilities in connected devices or to do deep analysis of metadata. Table 2 shows the main tools for OSINT analysis, their functionality, and how they are used.

Table 2. Tools for analysis

Tool	Functionality	Features of use
Maltego	Visualisation of connections between objects, metadata analysis.	Highly effective in investigations, difficult for beginners to master.
Shodan	Search for information about devices connected to the Internet, open ports and vulnerabilities	Requires a paid subscription to access advanced features.
Google Search Operators	Creating complex search queries using operators for accurate search.	Easy to use, effective for quickly finding information.
Censys	Search and analysis of Internet assets such as servers or SSL certificates	Powerful tool for finding vulnerabilities, but requires registration.
OSINT Framework	Structured catalogue of tools for gathering information.	Convenient for beginners, needs to be combined with other platforms.
Spiderfoot	Automation of data collection from social networks, domain registries, and web resources.	Flexible settings, integration with other tools.
FOCA	Analysis of metadata in files (documents, images, etc.).	Convenient for detecting hidden data in documents.
Recon-ng	Cloudwork for collecting OSINT data from various sources	Flexibility thanks to modular architecture, requires programming skills.

Source: developed by the authors based on Google Search Operators Cheat Sheet (2021), FOCA (2022), Spiderfoot (2022), Censys (n.d.), Maltego Technologies (n.d.), Shodan (n.d.)

Each of the proposed tools for OSINT analysis has its own unique advantages and disadvantages. Maltego is a powerful tool for visualising connections between objects, making it indispensable in complex investigations, but at the same time, it requires a significant amount of time to master. Shodan and Censys allow for rapid acquisition of information about network vulnerabilities; however, access to their full capabilities is often restricted by a paid subscription. Google Search Operators are an effective method for quickly finding specific information but demand proficiency in query formulation. Spiderfoot and Recon-ng provide comprehensive data collection from various sources, but require configuration and software knowledge to use effectively. In terms of accuracy and speed, specialised automated tools such as Shodan and Spiderfoot are the most effective, while manual searches using Google Search Operators are highly accurate but take more time. The combined use of these tools allows for the most comprehensive results and enhances cybersecurity. Overall, effective OSINT analysis necessitates selecting tools that best match the specific goals and requirements of the investigation.

Combining different approaches and automated tools, such as Shodan and Spiderfoot, with more accurate but time-consuming manual searches using Google Search Operators, allows for more complete and comprehensive data collection. The choice of tools also depends on the analyst's level of technical training and the need for timely information. Using these tools in combination can significantly increase the effectiveness of investigations and ensure a high level of cybersecurity, minimising the likelihood of missed threats or vulnerabilities.

A specific example is the monitoring of suspicious job postings on LinkedIn to check for phishing campaigns. To solve the problem, the presented data analysis algorithm was used (Fig. 2). First, data was collected using the LinkedIn API to search for job postings by keywords: "remote job", "work from home", using manual search, analytical tools, Python, requests and BeautifulSoup libraries to collect information.

Parsing the search results showed the collection of vacancy descriptions, contact details, and links to companies. As an alternative, an automated search was performed using specific queries, for example: site:linkedin.com/jobs “remote job” “apply now”, which resulted in a large amount of data about vacancies, including links, descriptions, companies, and contact details. The example of searching for job information on social networks showed that automating data collection using APIs and search operators significantly speeds up the process of obtaining relevant information, while the use of specific search platforms and indexed web resources provides effective access to large data sets.

```
import networkx as nx
import matplotlib.pyplot as plt
from googlesearch import search

# Task 1: Gathering information from open sources
query = "cybersecurity job postings site:linkedin.com"
results = []
for result in search(query, num_results=10): # Collecting the first 10 results
 results.append(result)

# Task 2: Noise filtering (selecting only relevant URLs)
filtered_results = [url for url in results if "linkedin.com" in url]

# Task 3: Create link graphs
G = nx.Graph()

# Adding nodes and connections
for idx, url in enumerate(filtered_results):
 G.add_node(url, label=f"Job {idx+1}")
 G.add_edge("Cyber Security Jobs", url)

# Graph visualisation
plt.figure(figsize=(10, 8))
pos = nx.spring_layout(G)
nx.draw(G, pos, with_labels=True, node_color='lightblue', font_size=10, node_size=3000,
font_weight='bold')
plt.title("Graph of Job Postings and Relationships")
plt.show()
```

Figure 2. Data analysis (Python)

Source: developed by the authors

Data analysis using Python showed the specified search queries, filtering, and construction of relationship graphs. Using Google Search to automatically collect URLs for a given query, only URLs belonging to linkedin.com were filtered to leave only relevant results (Google Search Operators Cheat Sheet, 2021). Next, a graph was created where nodes represent job vacancies and connections illustrate interrelationships. Algorithmic data processing for noise filtering, contextual analysis, and cross-checking, shown in Figure 3, includes filtering aimed at removing irrelevant or redundant information from the data obtained. To do this, preliminary data cleaning was performed, i.e., removal of duplicates, filtering by keywords and phrases, and detection and removal of spam, advertising, or malicious content. To apply regular expressions to search for specific patterns in texts (e.g., IP addresses, email addresses) and use text processing libraries such as NLTK and Spacy, the content was classified using machine learning models to determine relevance.

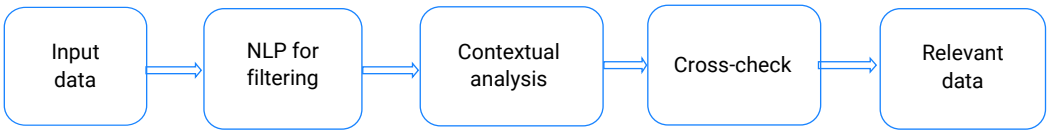

Figure 3. Noise filtering algorithm

Source: developed by the authors

The next stage involved contextual analysis, which allowed for an evaluation of the content of the acquired data and its relevance to the set objectives. Using NLP methods, we performed text tokenisation and lemmatisation to standardise word forms, sentiment analysis to understand the emotional tone of the information, and employed relationship graphs to establish interconnections between data objects, identifying key entities and concepts within the texts. In addition, it was important to evaluate the reputation and reliability of information sources using rating systems or historical data. Subsequently, cross-checking of data was carried out, which ensured greater accuracy of results by comparing data from different sources using cross-checking via various platforms (e.g., searching for

mentions on social networks, forums, official registers) and searching for confirmation through independent sources.

By using APIs to collect information from multiple platforms and automatically comparing the results using data analysis libraries such as Pandas or NumPy, discrepancies between data were identified using clustering and anomaly models. These methods reduced the amount of information processed and increased the relevance of the data. Similarly, key topics and their significance were identified, and text interpretation for strategic purposes was improved. In fact, this ensured effective information processing for decision-making in the context of cyber threats.

In the example presented, noise was filtered using NLP to remove irrelevant vacancies, i.e., records that did not contain contact details or contained suspicious text patterns ("we will provide your training free of charge, but you must pay") were removed. In addition, a contextual analysis was performed, including an analysis of vacancy descriptions for key phrases such as "no experience? no problem" or "only through this site", and suspicious contact details were highlighted: emails from publicly available domains (gmail.com, yahoo.com). Next, a cross-check was performed by comparing the collected information with the databases of well-known companies to identify fake profiles. As a result, relevant information was identified that indicated possible fraudulent activity.

During the process, relevant information was extracted using filtering algorithms that eliminate irrelevant data such as duplicates, outdated posts, or advertising materials. Contextual analysis and cross-checking, such as matching email addresses and IP addresses, added accuracy and reliability to the results. Relationship graphs are an effective method for analysing complex relationships between different objects in the context of cyber threats. They visualise and identify structures, dependencies, and behavioural patterns of potential attacks.

During the information gathering phase, various data sources were used to help form a complete picture of potential threats. In particular, publicly available data on network connections and vulnerabilities, including through tools such as Shodan or Censys, allowed us to identify open ports and potential vulnerabilities in networks. Domains, i.e. the use of WHOIS data and other registration databases to identify domain owners, as well as searching for historical data that indicated links to potentially dangerous objects. Email addresses were collected from open sources such as social networks, forums, or specialised services for checking email addresses against databases containing data exposed to the network. Analysis of profiles on social networks and forums, as well as checking accounts for possible threats or abnormal activity, and collecting file metadata, analysing malware or other compromising information through open sources such as file repositories, threat databases, or specialised tools such as VirusTotal, have enabled the use of publicly available data to collect, analyse, and draw conclusions about potential threats, as well as to identify links between threat objects (Table 3).

Table 3. Monitoring suspicious job postings on the LinkedIn platform

Data collection request and keywords	Noise and filtering of irrelevant data	Algorithmic data processing for noise filtering, contextual analysis and cross-checking	Potential threats
API LinkedIn: "remote job" ("remote job", "remote", "work from home")	Vacancies without mention of remote work, fake vacancies	Use algorithms to remove search noise (filtering by word frequency, NLP to identify irrelevant ads); check vacancies across multiple platforms (LinkedIn, Glassdoor, Indeed) to identify anomalies.	Potential phishing campaigns, payment fraud, providing false information to collect personal data or money.
API LinkedIn: "work from home" ("work from home", "home office", "telecommute")	Vacancies that only mention office work	Use text classification algorithms to identify incompatible job descriptions; check vacancies for terms confirming remote work, compare descriptions with other sources.	Data collection for phishing campaigns, using vacancies to collect email addresses and personal data.
Manual search: "remote job" ("remote job", "virtual", "home-based")	Advertisements for vacancies without specific information	Use text classifiers to automatically check vacancies that do not contain accurate information about remote work; check the context of the vacancy based on phrases that indicate the remote nature of the work.	Prepayment fraud, using invalid companies or empty vacancies to collect personal information.
Manual search: "work from home" ("work from home", "work at home")	Vacancies without the description "remote" or that require relocation	Algorithm for automatically filtering vacancies that do not contain the terms "work from home" in their descriptions, checking vacancies for the correctness of the description of working from home conditions.	Incorrect or fake job postings leading to personal information leaks or financial losses through fraudulent websites.
API LinkedIn: "remote job" + "security" ("remote job", "security", "cyber security")	Advertisements for vacancies that do not match the request (noise)	Identification of false job postings by analysing industry-specific terms (e.g., "security" in the context of vacancies) and checking job postings on specialised security websites to confirm the existence of vacancies in real companies.	Potential threats of phishing attacks, malicious use of job vacancies to collect user information or carry out cyber attacks.

Source: developed by the authors

Analysis of Table 3 showed that effective monitoring of job postings on LinkedIn requires a comprehensive approach that includes the use of algorithmic methods to filter out noise and verify data relevance, in particular by comparing job postings with other platforms. Algorithmic data processing, including text analysis, can significantly reduce the number of irrelevant job postings and improve the accuracy of identifying potential threats, such as phishing campaigns or fraudulent ads. Cross-checking job postings from multiple sources increases the reliability of results and allows for the detection of fake or unreliable job postings that could pose financial or information risks to users.

Next, a graph structure was formed, where the vertices are each object (threat, IP address, user) and the edges are the relationships between objects (shared IPs, similar domains, cross-activities), among which direct and indirect relationships are distinguished. During data processing, irrelevant objects were filtered out (using noise filtering algorithms) and classified by threat level (using risk ratings or trust indices). When visualising the graph, graph construction tools (Gephi, Cytoscape, Neo4j) were used and data layers (geolocation, activity time, attack types) were overlaid. To analyse the graphs, a centrality algorithm (Degree, Betweenness, Closeness) was defined to identify key objects and identify the most active or risky nodes in the network. The graph showed the connections between infected devices and command servers and made it possible to identify common operators by domain and email address. Building such a graph helped to understand the structure of threats, quickly identify critical relationships, and conduct an effective investigation.

For the example of job postings on social networks, a graph of connections was constructed using the Python library (networkx). The graph shows the vertices: job postings, contact details, IP addresses, companies, email addresses, and edges: connections between job postings (shared contact details, IP). The analysis revealed frequently recurring nodes (e.g., the same email address in different job postings) and grouped job postings by common attributes. The graph in Figure 3 shows the structure of relationships between suspicious job vacancies, which made it possible to identify the source of the cyber threat.

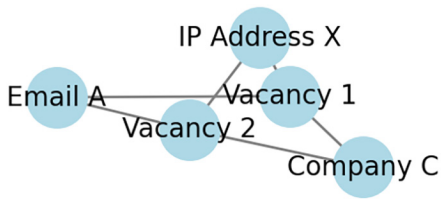

Figure 1. Graph of job vacancies on social media
Source: developed by the authors based on Google Search Operators Cheat Sheet (2021), Spiderfoot (2022), Censys (n.d.)

The relationship graph shows how interconnected objects (e.g., job postings, contact details, IP addresses) can help understand the structure of potential threats. This approach helped reveal hidden connections between different elements, allowing for more effective analysis of information and informed decision-making. As a result, we obtained an effective system for collecting data from open sources, reducing the time spent searching for and analysing information, and increasing the accuracy and relevance of the results obtained.

The issue of using open sources of information to ensure information security and combat cyber threats is extremely relevant, and recent studies have demonstrated various approaches to solving it. Research in this area is constantly evolving, but it is important to note that there are certain challenges and limitations that require further attention. The issues of information security and OSINT are extremely important in the modern digital age, when the number of cyber threats is constantly growing. As a result of research into methods of collecting and analysing data from open sources, it has been concluded that such approaches are highly effective in detecting and neutralising threats. However, the analysis revealed a number of issues that require further consideration and improvement.

One of the key sources in the field of OSINT is the work of M. Bazzell (2021), which examines the basics of collecting and analysing data from open sources. The author focused on the importance of understanding the specifics of information flows coming from open sources and the need for proper analysis to identify potential threats. The results confirmed the importance of this approach, as the correct interpretation of data can significantly reduce the risk of cyber threats. In addition, the integration of the methods shown was able to more effectively filter and verify information before its analysis, which is important for improving the accuracy of the results. In this study, these approaches were supplemented with modern methods of noise filtering, contextual analysis, and automated authenticity verification, which increased the accuracy of the results. The focus was not only on technical monitoring but also on the analysis of vacancy content, which indicates a broader coverage of OSINT aspects, including social engineering. Also, the work of J.M Clemen & J. Teleron (2023) was considered due to the development of encryption for communication protection, as encryption has an indirect impact on OSINT and ensures the integrity of analytical information transmission channels.

The article by A. Yadav *et al*. (2023) reviewed systematic modern approaches, sources, and challenges in the field of OSINT, with a particular focus on the use of NLP technologies, machine learning, and their integration with cyber defence systems. The authors emphasised that the effective use of OSINT significantly increases the ability to detect cyber threats and has significant potential for development through automation and the implementation of AI solutions. Unlike the review nature of the aforementioned study, the current work focuses on the practical implementation of a specific algorithm for detecting phishing threats. It automated the process of collecting and analysing data, particularly from professional platforms such as

LinkedIn. Thanks to applied testing of the algorithm, it was possible to improve the effectiveness of detecting phishing vacancies, in particular by significantly reducing the time required to process and classify suspicious information.

One of the areas of effective cybersecurity analysis is the use of graph methods to analyse the interaction of cybercriminal groups and build connections between threat objects. The Spiderfoot framework was used to automate the collection of OSINT data and was also used in this article to build graphs of relationships between suspicious accounts and domains. A study by G. Onoh (2018) examined the possibility of predicting cyber attacks based on publicly available data. This is consistent with the approach presented in the article to identifying suspicious job postings and predicting potential threats based on trends. However, this study proposed an automated algorithm that can not only analyse but also build relationship graphs for a deeper understanding of the threat structure.

The article by P. Goyal *et al*. (2018) discussed similar approaches to detecting cyber threats by analysing information flow in the network. The study also used similar signal indicators, in particular the frequency of keywords and similarities between advertisements. The importance of detecting cyber threats using signals from open web sources has been confirmed. In this article, this principle is implemented through monitoring the LinkedIn platform using API and web scraping, which made it possible to effectively detect phishing ads. In the study by G. Majumder *et al*. (2019), a method for comparing the semantic similarity of sentences was proposed. This directly formed the basis of a subsystem for detecting duplicate or modified job postings within phishing campaigns. Such a mechanism for detecting reformatted texts is critical for detecting variant attacks on different platforms.

A study by S. Kumar Birthriya *et al*. (2024) demonstrated phishing detection using deep learning. This publication also uses machine learning methods for data classification, but its distinctive feature is the combination of NLP analysis with regular expressions and cross-checking, which has significantly reduced the number of false positives. In this context, the classification of machine learning methods for threat detection conducted by K. Shaukat *et al*. (2020) in a review of general aspects of cybersecurity was useful in selecting approaches to NLP and classification of phishing texts in the current work. In particular, this article adapted the recommended ensemble approaches, which made it possible to improve the accuracy of classifying suspicious job postings.

The above approaches are complemented by the research of D. Kovalchuk (2025), which reveals the potential of large language models (LLMs) for automated analysis of cyber threats in a corporate environment, which is taken into account in this work. Of particular value to the current study is the work of S. Chen *et al*. (2024), which examines the application of machine learning algorithms to improve the effectiveness of threat detection using artificial intelligence. The authors emphasised the advantages of integrating AI solutions into cybersecurity systems, in particular due to the adaptability, scalability and high accuracy of such systems, which was important in building the architecture of the phishing vacancy detection system in the current study.

An additional contribution to the development of intelligent cyber defence systems is the study by N. Zaplatynskyi *et al*. (2024), which examined the possibilities of using artificial intelligence to improve the effectiveness of responding to cyber threats. The authors focused on combining analytical models with AI algorithms to form adaptive and self-learning security systems. The presented approach is consistent with the logic of the current work, which implements automated analysis of open sources with the subsequent use of machine learning methods to detect phishing vacancies. This emphasises the relevance of integrating intelligent systems into applied cybersecurity tasks and confirms the effectiveness of such solutions in a practical environment. A systematic review by C.S. Kruse *et al*. (2017) outlined current cyber threats in the healthcare sector. The parallel with this study is that both examples demonstrate the importance of early detection and filtering of threats. In addition, as in the case of phishing vacancies, attacks in the healthcare industry are often disguised as legitimate requests, which emphasises the importance of contextual analysis of detected objects.

Researchers point to the significant advantages of comprehensive approaches that combine automated methods with manual verification to reduce data processing time and improve the accuracy of cyber threat detection. The results confirmed these conclusions but revealed the need to improve the system for assessing the reliability of the data obtained. In fact, this system, which combines automated analysis methods with analytical assessment, showed better results in detecting potential threats, particularly in the context of verifying information sources. This work added new algorithms to improve the accuracy of detecting such threats, taking into account the ambiguity of messages in networks.

Conclusions

The article covered key aspects of developing algorithms for searching and analysing information from open sources (OSINT) in the context of cyber threats. The main goal was to create an effective toolkit for detecting, assessing, monitoring and neutralising potential threats in the digital environment. An algorithm was developed for collecting data from open sources, including APIs, web scraping, and search operators, as well as using noise filtering algorithms to extract relevant information. The article provided a detailed overview of data collection and analysis tools, including Google Search Operators, Maltego, Spiderfoot, Shodan and others, with a description of their advantages and limitations. Specific examples of use are provided, such as monitoring suspicious job postings on LinkedIn, which demonstrated the practical implementation of the proposed approaches. The use of cross-checking data from different sources increased the reliability of the

information, and the automation of the analysis significantly reduced the amount of irrelevant information and optimised the search process.

The results of this study confirmed that automated algorithms based on machine learning can significantly reduce query processing time and increase the accuracy of threat detection. Improvements to the system reduced processing time and increased accuracy, demonstrating the high effectiveness of the proposed methods. In particular, it is important to combine automated technologies with manual evaluation, which allows the system to adapt to new challenges. The developed methods provide quick access to information for decision-making in the field of cybersecurity. Continuous updating of algorithms will make it possible to adapt them to changes in data structures and current challenges in the field of cyber threats. The results of the study form the basis for the creation of automated monitoring and analysis systems that can be used in the field of cybersecurity. The use of the proposed methods and algorithms has made it possible to increase the speed and accuracy of analysis, which is especially important in the context of modern threats. The use of OSINT to ensure information security is highly effective but requires a comprehensive approach that combines automated methods and human analytics, and the test results confirmed that the integration of various methods, in particular graph analysis and NLP, allows for high accuracy in threat detection.

Further development of the system should include improving algorithms to increase data reliability and minimise the impact of noise in information from open sources. Overall, the article proposed a systematic approach to solving cybersecurity problems using modern OSINT technologies and demonstrated the potential of automation in combating cyber threats. Prospects for further research lie in improving algorithms to take into account the rapidly changing conditions of the digital environment, which will increase their effectiveness and relevance in detecting threats.

Acknowledgements

None.

Funding

The study was not funded.

Conflict of Interest

None.

References

[1] Alazab, M., Abu Khurma, R., García-Arenas, M., Jatana, V., Baydoun, A., & Damaševičius, R. (2024). Enhanced threat intelligence framework for advanced cybersecurity resilience. *Egyptian Informatics Journal*, 27(3), article number 100521. doi: 10.1016/j.eij.2024.100521.

[2] AlSalem, T.S., Almaiah, M., & Lutfi, A. (2023). Cybersecurity risk analysis in the IoT: A systematic review. *Electronics,* 12(18), article number 3958. doi: 0.3390/electronics12183958.

[3] Bazzell, M. (2021). *Open source intelligence techniques: Resources for searching and analyzing online information*. Washington: IntelTechniques.

[4] Censys. (n.d.). Retrieved from https://censys.io.

[5] Chen, H., Shen, Z., Wang, Y., Hu, K., & Xu J. (2024). Threat detection driven by artificial intelligence: Enhancing cybersecurity with machine learning algorithms. *World Journal of Innovation and Modern Technology,* 7(6), 58-70. doi: 10.53469/wjimt.2024.07(06).09.

[6] Clemen, J.M., & Teleron, J. (2023). Advancements in encryption techniques for secure data communication. *International Journal of Advanced Research in Science Communication and Technology*, 3(2), 444-451. doi: 10.48175/IJARSCT-13875.

[7] Dey, A.K., Gupta, G.P., & Sahu, S.P. (2023). Hybrid meta-heuristic based feature selection mechanism for cyber-attack detection in IoT-enabled networks. *Procedia Computer Science,* 218, 318-327. doi: 10.1016/j.procs.2023.01.014.

[8] FOCA. (2022). *FOCA – metadata extraction tool*. Retrieved from https://www.elevenpaths.com.

[9] Google Search Operators Cheat Sheet. (2021). Retrieved from https://surl.li/bmjxzv.

[10] Goyal, P., Hossain, K.S.M.T., Deb, A., Tavabi, N., Bartley, N., Abeliuk, A., Ferrara, E., & Lerman, K. (2018). Discovering signals from web sources to predict cyber attacks. *ArXiv.* doi: 10.48550/arXiv.1806.03342.

[11] Islam, M.T., Niger, M., Kynatun, M., & Mission, M.R. (2025). Systematic review of cybersecurity threats in IoT devices focusing on risk vectors, vulnerabilities, and mitigation strategies. *American Journal of Scholarly Research and Innovation*, 1(1), 108-136. doi: 10.2139/ssrn.5190439.

[12] Khan, R., Kumar, P., Jayakody, D.N.K., & Liyanage, M. (2019). A survey on security and privacy of 5G technologies: Potential solutions, recent advancements and future directions. *IEEE Communications Surveys & Tutorials*, 22(1), 196-248. doi: 10.1109/COMST.2019.2933899.

[13] Kovalchuk, D. (2025). Utilising large language models for automated real-time cyber threat analysis. *Bulletin of Cherkasy State Technological University*, 30(1), 48-58. doi: 10.62660/bcstu/1.2025.48

[14] Kruse, C.S., Frederick, B., Jacobson, T., & Monticone, D.K. (2017). Cybersecurity in healthcare: A systematic review of modern threats and trends. *Technology and Health Care,* 25(1), 1-10. doi: 10.3233/THC-161263.

[15] Kumar Birthriya, S., Ahlawat, P., & Kumar Jain, A. (2024). An efficient spam and phishing email filtering approach using deep learning and bio-inspired particle swarm optimization. *International Journal of Computing and Digital*

Systems, 15(1). doi: 10.12785/ijcds/150144.
[16] Majumder, G., Pakray, P., & Pinto, D. (2019). Measuring interpretable semantic similarity of sentences using a multi chunk aligner. *Journal of Intelligent & Fuzzy Systems,* 36(5), 4797-4808. doi: 10.3233/JIFS-179028.
[17] Maltego technologies. (n.d.). *Maltego evidence user manual*. Retrieved from https://support.maltego.com/en/support/solutions/folders/15000013724.
[18] Nagy, A., Du, X., Wang, X., Oates, M., Aronson, S., Plasek, J., Babb, L., Rehm, H., Zhou, L., & Lebo, M. (2025). P642: Facilitating machine learning and artificial intelligence in genetic databases: An open-source tool for data integration and summarization. *Genetics in Medicine Open,* 3, article number 103011. doi: 10.1016/j.gimo.2025.103011.
[19] Onoh, G. (2018). Predicting cyber-attacks using publicly available data. *Journal of the Colloquium for Information System Security Education (CISSE),* 6(1).
[20] OSINT Framework. (n.d.). Retrieved from https://osintframework.com.
[21] Recon-ng. (2022). *Recon-ng framework documentation. Recon-ng*. Retrieved from https://www.recon-ng.com.
[22] Shaukat, K., Luo, S., Varadharajan, V., Hameed, I.A., & Xu, M. (2020). A survey on machine learning techniques for cyber security in the last decade. *IEEE Access*, 8, 222310-222354. doi: 10.1109/ACCESS.2020.3041951.
[23] Shodan. (n.d.). *Search engine for the Internet of everythings*. Retrieved from https://www.shodan.io.
[24] Spiderfoot. (2022). *Spiderfoot OSINT Framework*. *GitHub*. Retrieved from https://github.com/smicallef/spiderfoot.
[25] Vashishtha, L.K., & Chatterjee, K. (2025). Strengthening cybersecurity: TestCloudIDS Dataset and SparkShield algorithm for robust threat detection. *Computers & Security,* 151, article number 104308. doi: 10.1016/j.cose.2024.104308.
[26] Yadav, A, Kumar, A. & Singh, V. (2023). Open-source intelligence: A comprehensive review of the current state, applications and future perspectives in cyber security. *Artifcial Intelligence Review,* 56, 12407-12438. doi: 10.1007/s10462-023-10454-y.
[27] Yang, T., Qiao, Y., & Lee, B. (2024). Towards trustworthy cybersecurity operations using Bayesian Deep Learning to improve uncertainty quantification of anomaly detection. *Computers & Security,* 144, article number 103909. doi: 10.1016/j.cose.2024.103909.
[28] Zaplatynskyi, N., Lub, P., & Zaporozhtsev, S. (2024). Improving cybersecurity with artificial intelligence. *Bulletin of Cherkasy State Technological University*, 29(4), 53-61. doi: 10.62660/bcstu/4.2024.53.

Алгоритми пошуку та аналізу інформації з відкритих джерел в умовах кіберзагроз

Оксана Онищук

Кандидат технічних наук, доцент
Волинський національний університет імені Лесі Українки
43025, просп. Волі, 13, м. Луцьк, Україна
https://orcid.org/0000-0002-8342-3011

Людмила Глинчук

Кандидат фізико-математичних наук, доцент
Волинський національний університет імені Лесі Українки
43025, просп. Волі, 13, м. Луцьк, Україна
https://orcid.org/0000-0002-8943-9604

Анотація. У статті представлено розробку алгоритму пошуку та аналізу інформації з відкритих джерел в умовах кіберзагроз. Запропонований алгоритм є ефективним інструментом для виявлення, моніторингу, оцінки та нейтралізації загроз у цифровому середовищі. У роботі описано основні етапи алгоритму, що включають: швидкий доступ до релевантної інформації, оцінку достовірності даних, аналіз трендів, визначення зв'язків між об'єктами, прогнозування потенційних загроз. Розробка алгоритму передбачала пошук та аналіз інформації з відкритих джерел; фільтрацію шуму; контекстний аналіз, перехресну перевірку для підвищення достовірності результатів; побудову графів зв'язків для виявлення залежності між об'єктами і визначення їх потенційної небезпеки. Ці задачі були реалізовані за допомогою збору даних через API (Application Programming Interface), веб скрейпинг (BeautifulSoup), використання пошукових операторів, обробки даних із застосуванням NLP-інструментів (Natural Language Processing), класифікації за допомогою моделей машинного навчання та регулярних виразів. У статті проаналізовано отриману інформацію за допомогою графів зв'язків, визначено ключові об'єкти та оцінено достовірність джерел. Розроблений алгоритм забезпечує скорочення часу пошуку та аналізу інформації, підвищення релевантності й точності отриманих даних, а також ефективну підтримку рішень у сфері кібербезпеки. Як приклад, алгоритм був застосований для моніторингу підозрілих вакансій на платформі LinkedIn, де були виявлені фішингові оголошення, що містили недійсні посилання або неправдиву інформацію. Використання API LinkedIn та веб скрейпинг дозволило автоматизувати збір вакансій та порівняти їх з базою даних відомих фішингових веб сайтів. Розроблений алгоритм дозволяє скоротити час пошуку та аналізу інформації порівняно з ручними методами. Впровадження таких рішень сприяє запобіганню кіберзагрозам і забезпеченню безпеки в цифровому середовищі. Алгоритм дозволяє значно підвищити ефективність роботи з відкритими джерелами, забезпечуючи автоматизований процес збору, обробки та аналізу даних для подальшої оцінки загроз

Ключові слова: кібербезпека пошуку інформації; аналіз даних; OSINT; графи зв'язків; NLP; Google Search Operators; Maltego