UNIVERSITY OF PIRAEUS

School of Information and Communication Technologies
Department of Digital Systems | Postgraduate Program
M.Sc. in "Digital Systems Security"

Advancements in Open Source Intelligence (OSINT) Techniques and the role of Artificial Intelligence in Cyber Threat Intelligence (CTI)

Angeliki Gioti[1]

[1]mte2205 – Department of Digital Systems, University of Piraeus

[*] *Correspondence to agioti@ssl-unipi.gr*

Supervisor: Prof. Stefanos Gritzalis

Abstract

This dissertation addresses the multilayered connection among Open Source Intelligence (OSINT), Artificial Intelligence (AI), and Cyber Threat Intelligence (CTI). Focused on the impact of AI on CTI and advancements in OSINT techniques, as well as how these two fields interact in order to reshape the realm of information gathering and security analysis. The study involves a thorough examination of OSINT techniques, including the use of social media, forums, and unconventional data sources, while also discussing the ethical issues that are inherent in this developing field. The incorporation of artificial intelligence into CTI is revolutionizing the field by improving the processing of data, recognition of patterns, and predictive analysis. The study examines the ways in which AI enhances human intelligence, addresses emerging threats, and enables collaboration across many domains. An essential component is a thorough comparative examination, tracing the development of OSINT techniques and CTI methodologies prior to and following the emergence of AI. Examining the profound impact of AI on the process of collecting information, resulting in increased expertise, efficiency, and adaptability.

Keywords: Open Source Intelligence (OSINT), Artificial Intelligence (AI), Cyber Threat Intelligence (CTI), Information Gathering, Security Analysis, Technology Ethics.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Dr. Stefanos Gritzalis for the assistance, advice, inspiration, and guidance he provided me throughout this process. I am extremely grateful for our collaboration and your personal support in my academic and business endeavours. Furthermore, I want to express my gratitude to my parents, my sister, and my partner Bill for their patience, understanding, and support during my postgraduate studies and the preparation of my dissertation.

1 Introduction

The advent of the digital age led us in a period of interconnectedness and information accessibility, altering our way of life, employment, and communication. These technological advances have unquestionably significant benefits, yet rise a variety of cyber risks, from data security breaches and intrusions by malicious software to nation-state-perpetrated cyber espionage. Notably, these risks pose substantial obstacles to both organisations and countries on a global scale as they grow in complexity and reach.

Cyber Threat Intelligence (CTI) is seen as a crucial tool in this rapidly evolving digital environment. CTI enables proactive *defence actions* by providing insights into new threats, weaknesses, and adversaries. **Open Source Intelligence** (OSINT) is a critical component of CTI; it is a discipline that collects and analyzes publicly accessible information from many sources, including as websites, social media, forums, and more, to identify possible threats and vulnerabilities and produce valuable intelligence.

The significance of OSINT in contemporary cybersecurity strategies cannot be overstated. When it comes to threat detection, incident response, and risk assessment, it is a valuable resource that frequently contains early warning signs of cyber dangers. An attorney can use OSINT to search for publicly available state records to draw information for a case and a Red Team operator could search for mentions of the target in publicly available online services during reconnaissance. However, efficient OSINT analysis has been severely hampered by the sheer volume and complexity of data in the modern digital environment. AI integration has emerged as a game-changing strategy to tackle these issues and draw insightful conclusions from the massive sea of data. Advanced analytics and machine learning techniques enable AI to scan large datasets quickly, spot subtle trends, and find abnormalities that may escape human evaluation.

The main objective of this paper is to explore how OSINT techniques are evolving and how AI has taken on a bigger part in CTI.

i Assess Current OSINT Methodologies and Tools: This entails a critical assessment of current OSINT methods and their efficiency in aiding CTI initiatives.

ii Examine the Impact of AI on OSINT: We will examine how AI, in particular machine learning and deep learning, improve the procedures for gathering, analysing, and disseminating OSINT data.

iii Recognise and Analyse AI-Powered CTI Tools: This study will pinpoint and examine several AI-powered CTI systems, examining their features, effectiveness indicators, and practical applications.

iv Address limitations and Ethical Issues: Including ethical issues, integrating AI into OSINT is not without challenges. We'll look into concerns including bias, explainability, openness, and data privacy.

v Propose Future Trends and fields of Research: In light of the quick development of technology, we'll offer some analysis of current trends in OSINT and AI-CTI as well as promising fields for this field's future study.

While this research aims to provide a thorough examination of AI-driven OSINT in CTI, it is important to be aware of several limitations:

The cybersecurity industry is dynamic, with new technology emerging all the time. As a result, advancements in AI and CTI technologies might take place throughout the research time, potentially affecting the results.

Data Accessibility: The majority of the information used in the present research will come from publicly accessible OSINT sources. It won't include confidential or exclusive data sources.

Notwithstanding AI's potential, human judgement and knowledge are still essential in cybersecurity. This study recognises the necessity of finding the right equilibrium between automated AI methods and human analysis.

Finally, this study aims to discover the intersections between OSINT and AI in the field of CTI. This dissertation seeks to improve cybersecurity tactics and strengthen relevant operations in the face of constantly changing cyber threats by examining current practises, correcting issues, and predicting future trends.

2 Open Source Intelligence and Threat Intelligence

2.1 Overview of Open Source Intelligence

Open Source Intelligence (OSINT) is described by NATO as "intelligence derived from publicly available information that is collected, exploited, and disseminated to an appropriate audience in a timely manner for the purpose of meeting a specific intelligence requirement"[1]. Particularly, OSINT includes "grey literature", or unclassified publications with restricted public circulation such as technical and economic reports, official and unofficial government papers, newsletters, subscription-based magazines, and digital resources that cross political, socioeconomic, military, and civilian boundaries.

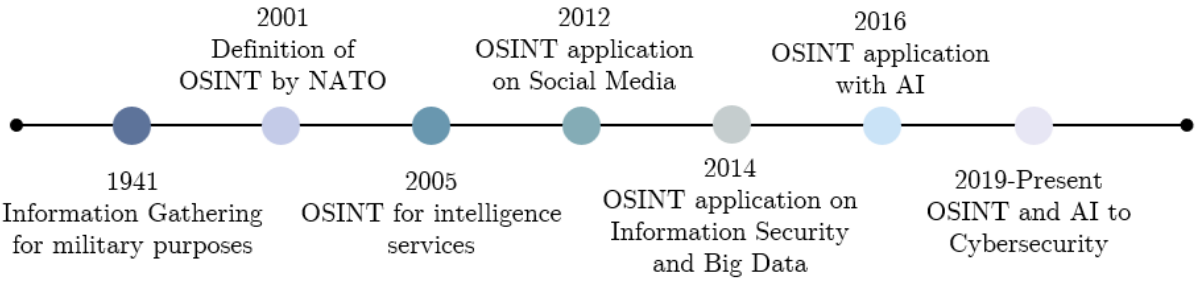

Figure 1: The OSINT Timeline.

ADVANTAGES	DISADVANTAGES
unclassified source	not a full-coverage solution
easy to share	information needs to be verified
does not compromise sensitive sources	big amount of noise
passive activity - low risk	valuable information may not be public
broad coverage	limited scope

Table 1: Advantages & Disadvantages of OSINT.

Table **4** presents a concise overview of the primary advantages and disadvantages relating to OSINT. An inherent benefit of OSINT is that it enables intelligence collection through passive means, without necessitating direct engagement with the subject. OSINT endeavours have minimal danger of notifying an opponent about the investigator's existence and intentions, resulting in a favourable outcome. One further benefit is that public information is more readily exchangeable among agencies compared to classified material. It can serve as an alternate source of intelligence without jeopardising a sensitive source that may disclose a technological or strategic edge. A major issue of OSINT is that the information accessible to the public domain often contains a significant quantity of irrelevant or misleading data, sometimes referred to as noise. An important issue within the data Community is the escalating challenge and duration involved in sifting through the irrelevant material and uncovering the valuable pieces of data from the constantly expanding reservoir of public information.

DEFINITION & IMPORTANCE

Open Source Intelligence, as previously discussed, revolves around information gathering from publicly accessible sources such as websites, social media platforms, news sources, academic publications, government reports, and public forums discussions and analysis of said data [1]. It can provide insights into various domains, from cybersecurity and national security to business intelligence and competitive analysis.

OSINT METHODOLOGIES & SOURCES

The field of OSINT includes a variety of methodologies and leverages diverse sources for information extraction. Traditional methods include web scraping, manual data collection, and basic keyword-based searches. However, as the digital landscape evolves, advanced techniques have taken center stage. Natural Language Processing (NLP) algorithms that enable the automated extraction of relevant information from unstructured textual data, is a great example. Furthermore, geospatial analysis tools have come to the fore, allowing analysts to correlate data with geographical locations, thereby providing context and depth to their intelligence. The deep web and dark web, although challenging to navigate, also hold valuable OSINT resources for those equipped with the necessary skills and tools.

OSINT is a subset of collection disciplines often referred to as "Intelligence Collection Disciplines", "INTs"[2].

1. **Human Intelligence (HUMINT)**: Information collection from human sources. The collection may be done openly, with witness or suspect interviews, or it may be done through covert means (espionage).

2. **Signals Intelligence (SIGINT)**: Electronic transmissions collected by various means, using ships, ground sites, planes and satellites. Communications Intelligence (COMINT), a type of SIGINT, is the intercepted communication between two parties.

3. **Imagery Intelligence (IMINT)**: Image Intelligence, also known as Photo Intelligence (PHOTINT), has changed over form due to the development of imaging technology. In the American Civil War hot air balloons with cameras were used for observation. In the first and second World Wars aircrafts were equipped with cameras in order to gather intelligence. Nowdays imagery satellites are used due to their high quality photographic capabilities. Geospatial Intelligence (GEOINT) refers to the analysis and visual rendering of security related activities on the earth produced by combing imagery, imagery intelligence and geospatial information.

4. **Measurement and Signatures Intelligence (MASINT)**: Advanced Information processing from IMINT and SIGINT systems. Weapon's telemetry can be intercepted - Telemetry Intelligence (TELINT), as well as sensor data from modern weapons and tracking systems - electronic intelligence (ELINT).

5. **Open-Source Intelligence (OSINT)**: Collection of information that is publicly available.

INTELLIGENCE DISCIPLINES	DESCRIPTION
COMINT	Communication Intelligence
CULTINT	Cultural Intelligence
DFINT	Digital Forensics Intelligence
ELINT	Electronic Intelligence
GEOINT	Geospatial Intelligence
HUMINT	Human Intelligence
IMINT	Image Intelligence
MARKINT	Market Intelligence
MASINT	Measurement & Signature Intelligence
SIGINT	Signal Intelligence
SOCMINT	Social Media Intelligence
TECHINT	Technical Intelligence
TELINT	Telemetry Intelligence

Table 2: Intelligence disciplines.

Open source intelligence is primarily associated with military intelligence and organizations yet, its scope of users is significantly broader. Open Source Intelligence, is becoming more popular among corporations, banks, and numerous industries that use it to collect valuable information and data for making informed decisions, gaining a competitive edge, and safeguarding [3]. The primary user groups can be enumerated as follows:

- Government entities, such as the military, security services, and law enforcement organizations.
- International organizations.
- Business corporations.
- Red Team.
- Criminal organizations, terrorist groups.
- Privacy-conscious people.

Government bodies are recognized as the primary consumers of open source intelligence. The government employs Open Source Intelligence (OSINT) for several objectives, including safeguarding national security, countering terrorism, preventing crime, conducting criminal profiling, and analyzing both domestic and foreign perspectives and occurrences on subjects of significance. Government entities distinguish themselves from other OSINT user groups by their ability to integrate OSINT intelligence with classified intelligence obtained through alternative methods. Moreover, they often possess greater capability and resources for data collection and processing compared to other user groups. It is anticipated that this pattern will persist in the future, with the government allocating additional attention and resources to OSINT. Government entities are highly regarded as excellent sources for OSINT due to their abundant resources and capacity to conduct thorough OSINT analysis [4].

International organizations use OSINT to ascertain impartial and transparent perspectives on subjects of significance, rather than relying on reports generated by influential governments or other sources that may present biased analyses in support of their own goals. An exemplary instance of an international institution employing Open Source Intelligence is the United Nations (UN), which employs OSINT to bolster peacekeeping endeavors on a global scale [4].

Business corporations have increasingly recognized the potential of open source intelligence to harness information effectively. The increasing popularity of OSINT within this user group can be attributed to the widespread usage of the Internet, which has resulted in a wealth of information becoming accessible. As a result, OSINT has also been accessible to small enterprises. Previously, this opportunity was exclusively available to businesses with substantial financial resources. Business corporations use OSINT for marketing purposes, conducting investigations into existing and emerging markets, monitoring competitors' actions, assessing their operational landscapes, and analyzing prevailing trends and shifts. Businesses utilize OSINT to safeguard against data leaks, monitor private data breaches, and observe network behaviors to protect against cyber threats. Numerous private firms have devised sophisticated algorithms and methodologies to get data from public sources for the purpose of generating commercial profits [4].

Penetration testers and **black hat hackers** employ OSINT with a greater level of precision. Their primary goal is frequently to collect information regarding specific targets on the internet in order to prepare for penetration testing or social engineering attacks.

Additionally, **criminal syndicates** and terrorist factions exploit open-source intelligence (OSINT) channels to strategize attacks, gather intelligence on potential targets, recruit fresh members through social media analysis, acquire military information disclosed by governments, and employ OSINT to devise optimal channels for disseminating their propaganda.

The concerns regarding the online exposure and security of private data among ordinary individuals are leading privacy-conscious **individuals** to make use of OSINT as well. They employ it to monitor their digital identities in order to safeguard their privacy and collect data from public sources for commercial purposes.

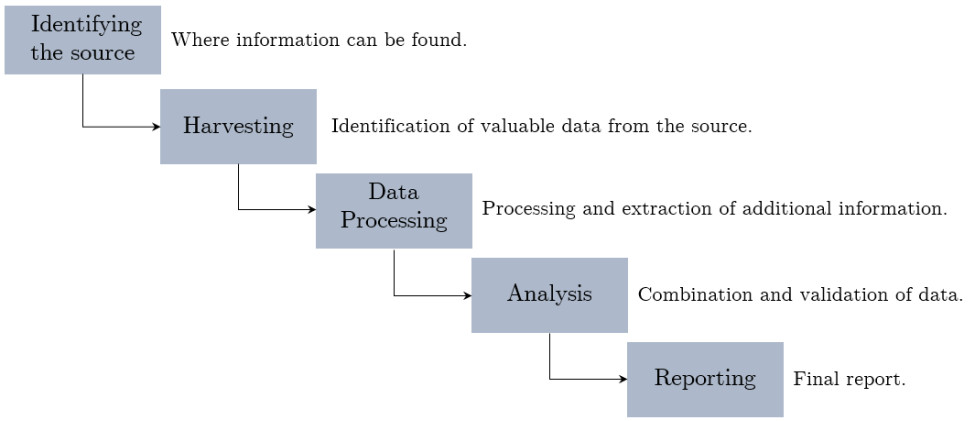

Figure 2: The OSINT Process.

OSINT, similar to other forms of intelligence, adheres to a precise and well-defined methodology. We will try to break it down in three phases [5].

Initially, the **collection phase** during which data that is accessible to the public is obtained from pertinent open sources in accordance with the intended purpose or aim. The Internet stands out as the preeminent resource owing to its extensive collection of content and convenient accessibility. The collection process holds significant importance as it initiates the entirety of the intelligence generation process.

Subsequently, during the **analysis phase**, insightful and practical information is generated through the processing of the gathered primary materials. The data is useless in and of itself; therefore, it must be interpreted in order to obtain the initial facts that result from an in-depth analysis.

In the final step of **knowledge extraction**, the previously cleansed data is utilized as input for inference algorithms that are more sophisticated. The current era's computational advancements enable the detection of patterns, profiling of behaviors, prediction of values, and correlation of events.

Notably, the second and third stages consist of technologies that are commonplace and well-known within the data mining domain. The OSINT acquisition method, on the other hand, is distinct from existing data-driven services. Presently, prevalent data analysis applications utilize pre-established data sources to collect substantial amounts of information while implementing transparent data collection processes. On the contrary, OSINT solutions ought to gather precise information from the vast expanse of accessible and feasible open resources.

A OSINT COLLECTION

Prior to conducting intelligence extraction and analysis, the investigator must augment the target-related dataset. In pursuit of this objective, we suggest several OSINT methodologies that symbolize various collection approaches. OSINT techniques pertaining to search engines, social networks, email addresses, usernames, actual names, locations, IP addresses, and domain names have been specifically examined. Each will contain an infinite number of OSINT services employing comparable methods of data collection. At this stage, it is presumed that a minimum of a single element of information pertaining to the target is accessible (e.g., their actual name, username, email address, and so forth). Based

on the characteristics of the initial sample, the investigator employs the most appropriate OSINT techniques in order to extract additional data. Thus, the outcomes acquired through a particular methodology constitute a data transfer that is to be utilized by an alternative methodology. The transactions that have been depicted serve as examples of potential methods for advancing the investigation, in which the results obtained from the method of origin are utilized as input to the method of destination.

B OSINT ANALYSIS

Analyzing and comprehending the continuous iterations of the various OSINT techniques is necessary in order to produce valuable information. The literature presents a growing array of analysis techniques that can be employed to accomplish this task. Listed below are some particularly alluring procedures that are relevant to our particular scenario:

- ✧ **Lexical analysis**

  Entities and relationships within text should be extracted from raw data. It is critical to implement translation processes for the language used in the OSINT investigation and to eliminate non-value-adding sentences that contain noise.

- ✧ **Semantic analysis**

  It is useless to possess a bundle of words if their meanings are not extracted. Modern applications of natural language processing algorithms are designed to comprehend data. Furthermore, sentiment analysis methodologies enable the contextualization of subjective comments or viewpoints in order to categorize the author's emotional state as positive, negative, or neutral. In conclusion, truth discovery procedures tackle the formidable challenge of reconciling inconsistencies in multi-source data that present contrasting viewpoints on the same topic.

- ✧ **Geospatial analysis**

  Involves the examination of retrieved data from sources such as IP addresses, social networks, events, or sensors, with a focus on location. The utilization of maps or diagrams enhances the comprehension and representation of data, enabling the identification of significant relationships between incidents or individuals.

- ✧ **Social media analysis**

  Researchers are able to conduct comprehensive analyses of users by utilizing the functionalities introduced by contemporary social media platforms. Social data analysis enables the formation of a network encompassing the subject's contacts, interactions, locations, behaviors, and preferences.

The outcomes generated by implementing the aforementioned methods are referred to as output data and fall into three primary categories:

- ✧ PERSONAL INFORMATION

  consists primarily of details that establish an individual's identity, such as their actual name, email address, user name, social media profiles, and search engine usage.

✧ ORGANIZATIONAL INFORMATION

constituted by the distinct components of a team or business. Primarily, it is gathered through the utilization of search engines, social networks, location, domain name, and IP address techniques.

✧ NETWORK INFORMATION

consists of technical details regarding communication topologies and systems; this is typically accomplished via location, domain name, and IP address techniques. These three informational sections can be logically expanded to include additional elements. Furthermore, it is possible for a single investigation to yield various forms of complementary output data.

C OSINT KNOWLEDGE EXTRACTION

Without a doubt the information gathered thus far possesses significant value. Nevertheless, the intelligence derived from those discoveries ultimately culminates in a compelling identification of the target. In pursuit of this objective, knowledge elicitation is defined as the application of data mining and artificial intelligence techniques to the analysis results (output information). Following this, an example of such technologies is presented below:

i **Correlation**

The identification of associations among individuals, occurrences, or data points in a broader sense. Robust related features are particularly advantageous in exposing latent associations that may be present in the dataset.

ii **Classification**

Supervised learning enables the division of the data into groups based on predefined categories. This methodology enables the structuring of substantial volumes of data to facilitate more efficient knowledge retrieval.

iii **Outlier detection**

This process identifies anomalies in the dataset through analysis. An area of particular interest pertains to the observation of malevolent agents, whose conduct or behaviors deviate from those of the general populace.

iv **Clustering**

This process divides data points into clusters while taking into account a substantial number of conditions or heuristics. This may unveil, for instance, diverse forms of network behavior, various categories of online profiles, or methods of assaulting organizations, individuals, or infrastructures without the learner being aware of such diversity in advance (unsupervised learning).

v **Regression**

A method employed to anticipate or predict numerical values or facts. Illustratively, linear regression yields a value corresponding to a linear function, neural networks are structures that map intricate combinations of inputs to outputs, and deep learning consists of multiple layers that amalgamate and perform operations on the input.

vi **Pattern recognition**

As opposed to anomaly detection, is a procedure that identifies recurring patterns in data. The aforementioned methodologies may be incorporated into this expansive notion of knowledge discovery. In reality, any artificial intelligence technique is applicable to the extraction of knowledge from open data.

These sophisticated methodologies enable the derivation of abstract, intricate, and enticing matters pertaining to the subject matter that are not overtly disclosed on the internet. Nevertheless, this methodology presents a number of obstacles, the most significant of which are the investigation and development of this knowledge extraction process in order to detect, attribute cyberattacks, identify, profile, or monitor perpetrators, as well as to identify and investigate malicious organizations. Furthermore, the prospective ability to draw strong inferences gives rise to a number of privacy–related concerns. The information that is extracted regarding an individual, business, or organization may be particularly sensitive, and its manipulation can result in ethical and legal complications.

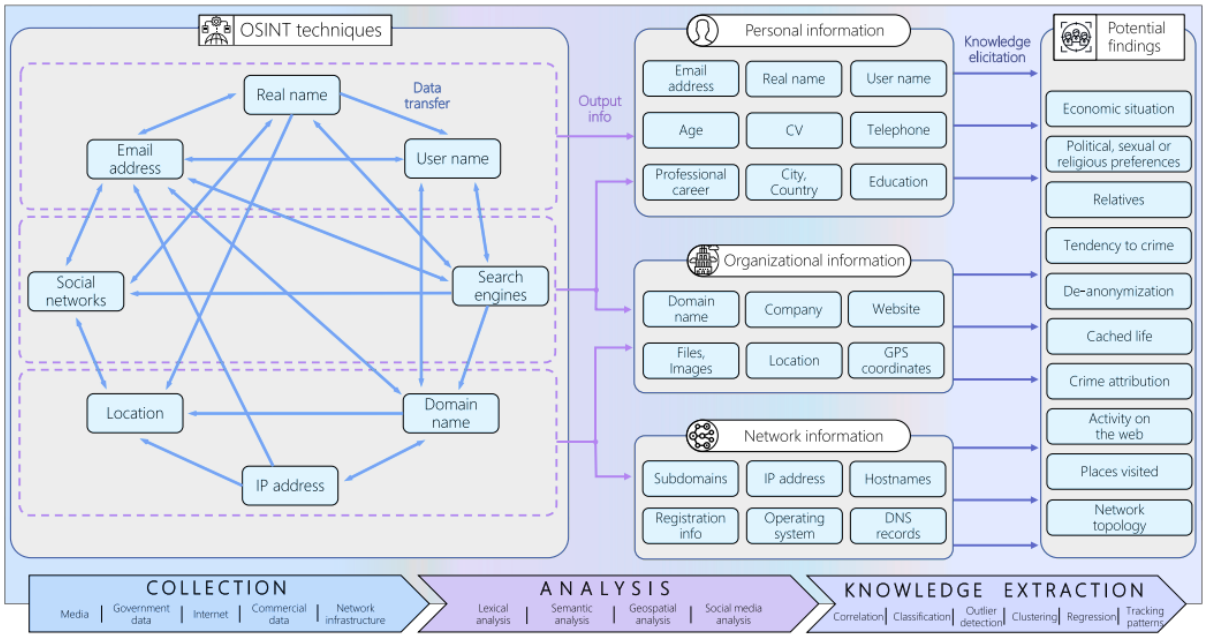

Figure 3: Principal OSINT workflows and derived intelligence.

A brief summary of OSINT's primary usage is the following:

i SOCIAL OPINION AND SENTIMENT ANALYSIS

With the proliferation of online social networks, it has become feasible to extract implicit knowledge from users' interactions, messages, interests, and preferences. The assemblage of evidence derived from social media is extensive and advantageous to a broad spectrum. For example, such data acquisition and analysis could be implemented in the fields of disaster management, political campaigns, or marketing.

ii ORGANIZED CRIME AND CYBERCRIME

OSINT processes match and continuously analyze public data to detect criminal intentions in their nascent stages. By considering the patterns exhibited by adversaries and the interconnections among transgressions, OSINT enables security forces to expeditiously

identify illicit activities. In this regard, it would be feasible to monitor the activities of terrorist organizations, which are progressively more active on the Internet, through the utilization of publicly available data.

iii CYBERSECURITY AND CYBERDEFENSE

Criminals persistently target ICT systems with the intention of causing service disruptions. Therefore, in order to protect these systems from cyberattacks, it is imperative to conduct research that addresses the unresolved issues in the field of cybersecurity. Consequently, data sciences are being implemented not only for the purpose of footprinting during penetration testing, but also for the proactive safeguarding of businesses and organizations. In practice, data mining techniques can be of assistance by conducting analysis of routine assaults, establishing correlations between them, and facilitating decision-making processes to ensure not only an effective defense but also a timely response. Similarly, OSINT can be regarded as a means of obtaining information for investigations and tracebacks in this particular context. OSINT can be utilized in forensic digital analysis to supplement digital evidence left by an incident.

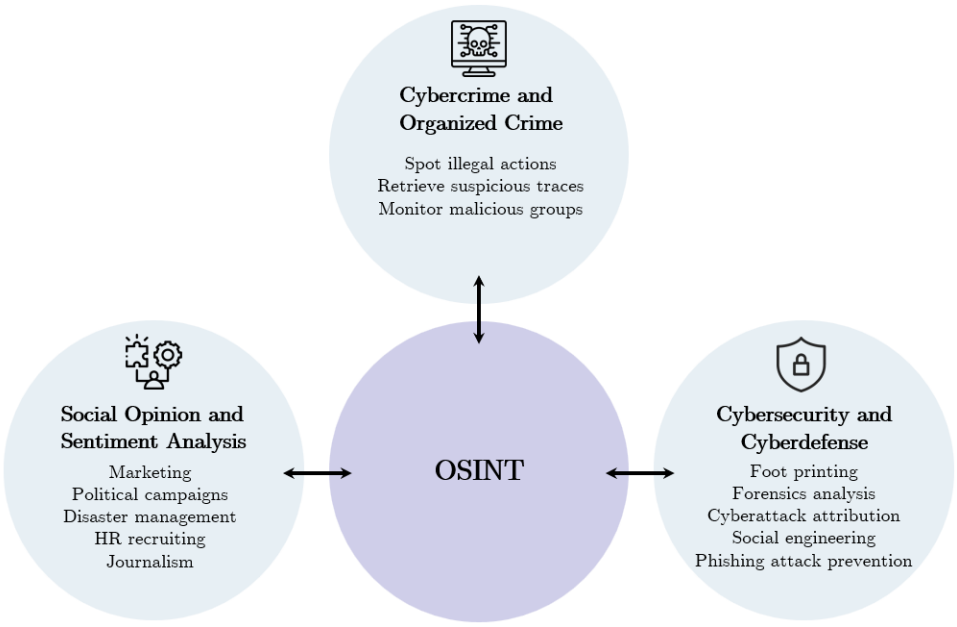

Figure 4: OSINT principal use cases.

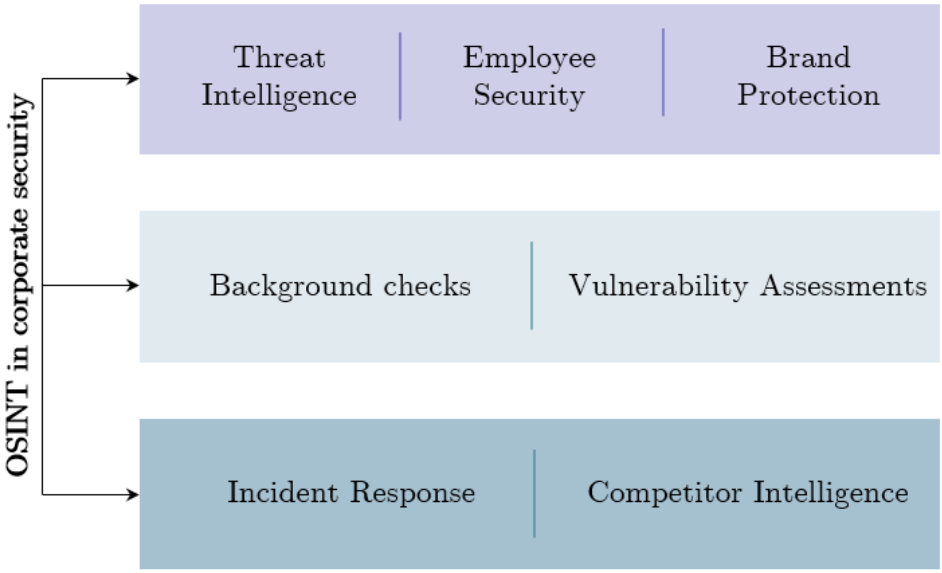

Figure 5: OSINT in corporate security.

2.2 The Intelligence Cycle

The Intelligence Cycle [6] consists of six main stages: planning & direction, collecting, processing & exploitation, analysis & production, dissemination & integration, and evaluation & feedback.

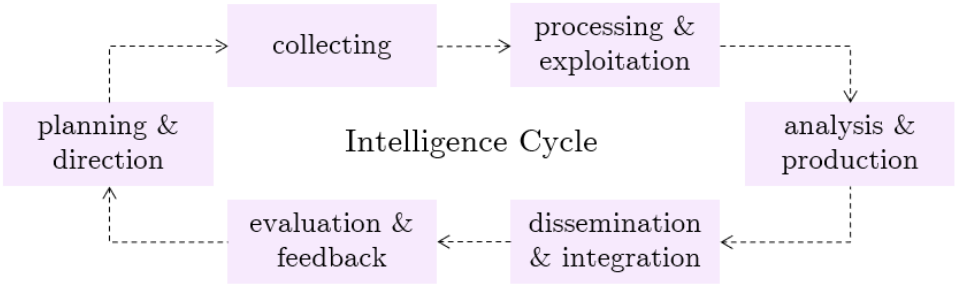

Figure 6: Six core stages of the Intelligence Cycle.

✧ PLANNING AND DIRECTION

This step is present at the beginning to establish goals and requirements for the following phases of the cycle, and at the end because finished intelligence requires action and unavoidably influences the next iteration. The management of the entire intelligence industry is mostly influenced by the end-user.

✧ COLLECTING

Collection is the process of acquiring and documenting unprocessed data, as the fundamental elements that will be refined into practical intelligence. The data can be gathered from a variety of sources, including open sources such as social media, online forums, public documents, publications, and essays, as well as closed sources like CCTV video and interviews. Private sector entities typically employ comprehensive and adaptable collec-

tion strategies, while public sector agencies tend to have a narrower and more specialized focus.

✧ PROCESSING AND EXPLOITATION

The processing stage involves transforming the raw data into a state that is appropriate for the analysts to begin their analysis. This process entails the organization, categorization, and arrangement of the unprocessed data, and may additionally require converting any analog information pertaining to a particular instance into a digital format. Processing is not a separate and isolated step in the cycle, but rather it overlaps smoothly with a significant portion of the gathering and analytic tasks.

✧ ANALYSIS AND PRODUCTION

This step represents the transformation of information from raw data into actionable intelligence. In essence, it is a procedure of merging and evaluating data to provide cohesive intelligence information. The analysts conducting this task are highly skilled experts who assess the data in terms of its reliability, validity, currency, and relevance. The data is consolidated, organized, and transformed into practical intelligence that includes detailed event breakdowns and analysis, along with the potential consequences for the end-user.

✧ DISSEMINATION AND INTEGRATION

After the intelligence has been obtained, it is necessary to convey it to the end-users, who are the ones who initially requested the service. Typically, this stage of the process initiates feedback, which often serves as the starting point for a new cycle. Effective communication is crucial at this point, and various clients will necessitate diverse formats such as written reports, briefings, and PowerPoint presentations.

✧ EVALUATION AND FEEDBACK

After providing the intelligence, we do not disengage from the matter; ongoing communication with the client is necessary. Analysts must finally address any overlooked deficiencies or, more broadly, assess the extent to which the intelligence fulfilled the client's requirements. Through implementing final adjustments or refinements, analysts can enhance client satisfaction, while reflecting on deficiencies can enhance internal processes for future endeavors.

The introduction of the World Wide Web revolutionized the concept of open data. In the midst of an unparalleled surge in digital information, OSINT, which was previously confined to analogue domains such as television broadcasts and newspaper publications, experienced a renaissance. The proliferation of open data presented analysts with an array of utterances, but presented tremendous obstacles and challenges.

Since we possessed data and in great quantity, we were also aware that this information could be utilized to accomplish a wide range of duties and resolve a multitude of inquiries. To accomplish this, however, needles in the haystacks or patterns from oceans of data had to be extracted. It is unsurprising that the tools and methods utilized prior to automation were completely ineffectual, while the volume of data continued to increase at an exponential rate.

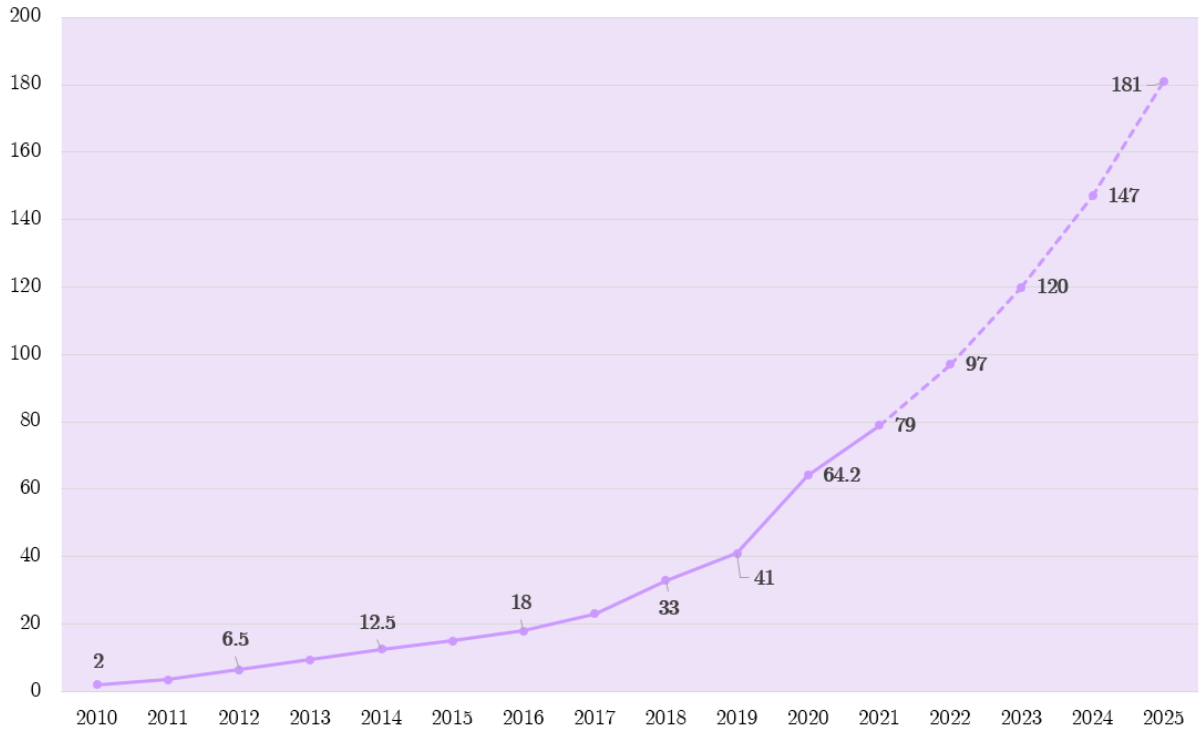

Figure 7: Volume of data created, stored, copied and consumed worldwide.

Online data has increased exponentially over the past decade, from around 7 zettabytes to an astounding 97 [7].

Since algorithms can now perform the majority of work, programmers started working on the development of AI models whose parameters could be modified based on the specific task such as the extraction or filtering of data leaving the human factor to deal with the decision-making process, which accelerated the achievement of any objective.

Machine Learning models has also evolved in parallel with data volumesan obstacle that AI can bypass, fundamentally transforming information analysis across multiple domains [8].

2.3 The Role of Artificial Intelligence in OSINT

INTEGRATION OF ARTIFICIAL INTELLIGENCE IN OPEN SOURCE INTELLIGENCE

Automated Data Collection and Aggregation: the process of gathering and combining data using automated methods.

Artificial intelligence algorithms are crucial in automating the process of gathering and combining data from many web sources. These techniques, which include web crawlers and machine learning-based systems, greatly improve the effectiveness and breadth of data retrieval. AI optimises the collection of extensive volumes of unorganised data by scanning, indexing, and categorising content received from social media platforms, websites, forums, and public databases. For example, AI may be coded to monitor modifications on websites, monitor discussions on social media, and provide thorough reports - all in real-time. This increased efficiency enables analysts to concentrate on more advanced activities that need cognitive reasoning and analysis.

Techniques for Processing and Analysing Data:

The fundamental aspect of Artificial Intelligence (AI) in Open Source Intelligence (OSINT) is its capacity to efficiently handle and examine unorganised material. Machine learning algorithms, Natural Language Processing (NLP), and computer vision techniques play a crucial role in analysing data and generating practical insights. AI allows analysts to comprehend extensive and intricate information found in OSINT sources, ranging from sentiment analysis of social media postings to entity recognition in textual data. Natural Language Processing is a specialised branch of Artificial Intelligence that enables machines to comprehend, analyse, and produce human language. Natural Language Processing (NLP) is crucial in Open Source Intelligence (OSINT) for tasks such as sentiment analysis, identifying important items in text, and condensing long documents. This offers a comprehensive perspective on public mood and opinions, which is useful for making well-informed judgements.

The identification and classification of patterns, as well as the detection of abnormalities:

The advanced capabilities of artificial intelligence in recognising patterns and detecting anomalies greatly assist in the identification of trends, abnormalities, and possible risks inside datasets. Machine learning algorithms effectively identify trends in network traffic, user behaviours, and content, aiding in the detection of new dangers that may bypass traditional detection approaches. Through the use of varied datasets, researchers may train AI models to create tools capable of discerning nuanced correlations among different bits of information. This facilitates a more profound comprehension of situations, possible dangers, and emerging patterns that might otherwise stay hidden.

Anticipatory Analysis and Prospective Projections:

AI-driven predictive analytics utilise past data to anticipate likely future occurrences or patterns. Through the examination of patterns, correlations, and historical settings, these models aid in proactive evaluation of potential dangers, allowing analysts to predict and make arrangements for upcoming cyber attacks.

ADVANTAGES OF AUTOMATED OSINT
Faster
Relatively cheaper
Improves data analysis
Minimizes the need of human factor
Increases the available sources

Table 3: Automated OSINT vs. Manual OSINT.

APPLICATION OF AI IN OSINT

Social Media Intelligence: the process of gathering and analysing information from social media platforms.

Artificial intelligence systems analyse the content of social media, collecting patterns, emotions, influential individuals, and possible risks. They enable the surveillance of public opinion, the recognition of growing concerns, and the identification of security hazards, allowing for preventative measures to be taken against prospective dangers.

Analysis of Images and Videos:

Computer vision systems powered by artificial intelligence (AI) analyse visual content, extracting important information from photographs and videos that are shared on web platforms. These systems offer improved insights from multimedia sources, ranging from geographic analysis to object identification and content verification.

NLP (Natural Language Processing):

Natural Language Processing (NLP) approaches enhance the capabilities of artificial intelligence (AI) systems to analyse and comprehend written information from various sources such as news articles, forums, and blogs [9]. Sentiment analysis, entity recognition, subject modelling, and language translation are prominent applications employed in OSINT.

Platforms for Threat Intelligence:

AI-powered threat intelligence solutions utilise machine learning to gather, correlate, and scrutinise data from diverse sources. These technologies detect possible dangers, analyse signs of compromise (IOCs), and offer practical information to cybersecurity experts.

Monitoring of the Dark Web:

Artificial intelligence (AI) techniques are utilised to monitor and analyse activity occurring on the dark web. Their role involves revealing unlawful actions, detecting cybercriminal conduct, and monitoring the trade of compromised data or illicit merchandise, so improving the understanding of clandestine networks.

Named Entity Recognition (NER)

Named entity recognition, also referred to as entity chunking, is an essential activity in the extraction and processing of textual data [10]. Text analysis refers to the identification and categorization of a given text split on words based on their semantic content. For example, a Named Entity Recognition (NER) model may identify the phrase “Unipi” and classify it as a “University of Piraeus”.

It is built upon natural language processing, which focuses on the analysis and manipulation of natural language including the one used in everyday communication (slang or formal).

The majority of open data is textual, such as social media comments, blog posts, site pages, and similar sources. In addition to this, it is worth noting that automated transcription systems have the capability to rapidly translate various data forms, such as video and audio, into textual form. This further emphasizes the crucial significance of NER.

Speech To Face (STF)

Speech-to-Face AI is an impressive machine learning system that can generate a highly realistic digital image of a person based on their voice recording. This image includes

important characteristics like age, gender, and ethnicity. Prominent systems in this field comprise Nvidia's Audio2Face [11] and MIT's Speech2Face [12].

Undoubtedly, these exceptional technologies are highly intriguing to digital forensic analysts. They enable investigators to formulate hypotheses about the behaviors and character traits of a specific speaker, as well as provide visual depictions in the absence of any eyewitness testimonies to rely on. Although these visuals may not be as dependable as sketches provided by firsthand witnesses, they can provide valuable clues that help advance an inquiry.

THE ROLE OF AI IN OSINT
Used to automate processes and tasks
Improves accuracy and impartiality of data analysis
Boosts the process of data gathering
Enables data driven decision-making

Table 4: The Role of AI in OSINT.

2.3.1 AI in the Intelligence Cycle

✧ PLANNING AND DIRECTION

Challenges: Planning and direction are often burdensome and time-consuming processes, particularly in large institutions like public sector enterprises. Departments frequently establish their own protocols and intelligence strategies, although a unified process is uncommon. Furthermore, analysts often need to recommence the intelligence process from the early phase as a result of inadequate elaboration during the planning stages.

Artificial Intelligence Solutions: Concise, automated, multi-step intelligence processes. Through the process of dividing large intelligence cycles into smaller ones, project managers are able to efficiently adjust the workflow prior to the ultimate distribution of the intelligence insights. However, to prevent overwhelming the entire process, it is necessary to fully automate the lower-tier cycles.

✧ COLLECTING

Challenges: The vast volume of user-generated data available on the internet, in its diverse forms such as text, audio, video, and images, has rendered the manual collection of all essential information very unfeasible. In addition, worldwide investigations necessitate the utilization of several languages that cannot be effectively translated using conventional internet translation tools.

Artificial Intelligence Solutions: Certain OSINT systems include machine learning-based capabilities for acquiring information. These programs have several parameter configuration choices that enable them to automatically search for any information that is associated with the initial input. These systems may also incorporate AI-powered translation plugins that enable users to accurately interpret the subtle meaning of text in different languages.

✧ PROCESSING AND EXPLOITATION

Challenges: The processing of data is equally complex as its acquisition, and this is due to identical factors. Although numerous systems demonstrate exceptional efficiency in textual analysis, some challenges arise regarding to processing additional input types such as audio, image, and video. With the introduction of deepfake technology, distinguishing pertinent information from irrelevant noise is becoming rather challenging.

Artificial Intelligence Solutions: In an early stage of development, there are a few systems for analyzing non-textual data such as voice recognition, image reconstruction, and deepfake detection.

✧ ANALYSIS AND PRODUCTION

Challenges: Even after undergoing processing, data may remain fragmented, disorganized, and contradictory. In addition to this, considering the possibility of a massive influx of data, potentially reaching terabytes in size, one can grasp the immense difficulty faced by the analyst in their task. With the exponential growth of online data collection, analysts are facing increasing challenges in identifying connections among vast amounts of multi-source information and extracting important insights. Analysts frequently have to disregard substantial quantities of information because it is just impossible to integrate all of it into a cohesive and comprehensive report.

Artificial Intelligence Solutions: Graph analysis technologies powered by artificial intelligence are currently available in the market and have become essential solutions in contemporary investigations.

✧ DISSEMINATION AND INTEGRATION

Challenges:An inherent challenge arises from the fact that the very same attributes that contribute to being an exceptional analyst can also hinder effective communication skills. In addition, analysts commonly perceive this step of the cycle as arduous and monotonous, mostly because of its repetitive nature, which adversely affects the quality of intelligence presentation. Analysts often face reprimands from their supervisors due to the submission of subpar reports characterized by bad linguistic style, excessive use of jargon, inconsistency, and a lack of logical coherence.

Artificial Intelligence Solutions: Utilizing AI technologies to assist analysts in organizing, condensing, and refining initial versions of intelligence reports is an effective approach to tackle this problem. By assisting analysts in bypassing the tedious process of generating reports and offering dependable automated verification, reports can be generated with superior quality and in a shorter amount of time.

✧ EVALUATION AND FEEDBACK

Challenges: Feedback is frequently disregarded and typically ends up being nothing more than superficial comments on the report. Optimally, feedback should possess the same level of depth and understanding as the report itself, in order to facilitate continuous improvement within the intelligence cycle.

Artificial Intelligence Solutions: Automated textual organization can greatly aid in this task, just like it does in the reporting stage.

The advantages and its impact are significant and can be mapped into three distinct aspects:

SCALABILITY

Refers to the ability of a system or process to handle an increasing amount of work or data without compromising its performance or efficiency.The volume of data processed during an open-source study is typically constrained by the number of analysts engaged in the activity, as well as the combined cognitive abilities and stamina of these individuals. Nevertheless, the extent of AI-powered data processing is solely constrained by the capacity of computer power, meaning that it is virtually boundless.

ACCURACY

Although the expertise of particularly qualified OSINT analysts is crucial for conducting successful investigations using open data, the utilization of AI techniques can significantly enhance the efficiency and efficacy of their work. By emphasizing the most pertinent data and its organizational patterns, analysts can bypass much of the ambiguous background noise that typically accompanies the human factor in investigations.

LATENCY

Open data has a limited duration of usefulness and can rapidly become obsolete or disappear completely. Effective OSINT work is primarily the outcome of timely analysis. AI offers a significant advantage over labor-intensive human procedures by efficiently extracting and categorizing open data with minimal delay. This enables precise and ongoing monitoring of data, rapid discovery of patterns, immediate detection of potential issues, and even automated reassessments of entire investigation processes.

2.4 Introduction to Cyber Threat Intelligence

Depending on the parties involved, the needs established, and the overarching objectives, the threat intelligence lifecycle generates several types of intelligence [13]. Threat intelligence falls into three categories:

Tactical Threat Intelligence employed by security operations centers (SOC) to identify and deal with ongoing threats. It focuses on IoCs, such as email subject lines linked to phishing attacks, file hashes linked to known malware and ransomware assaults, or IP addresses linked to command and control servers. Tactical threat information is also used by threat-hunting teams to find advanced persistent threats (APTs) and other active but hidden attackers; false positives are filtered out, allowing the interception of actual attacks.

Operational Threat Intelligence helps organizations in predicting and stopping upcoming attacks. It describes the TTPs and behaviors of recognized threat actors, including the attack vectors they employ, the vulnerabilities they exploit, and the assets they target, and because of that it is frequently referred to as “technical threat intelligence”. Information security decision-makers like CISOs and CIOs use operational threat intelligence to pinpoint threat actors who are most likely to target their enterprises. They then implement security measures

and take additional steps to counter these attacks.

Strategic Threat Intelligence is high-level intelligence on the global threat landscape and how an organization fits within it. Strategic threat information provides CEOs and other executives, who make decisions outside of IT, with knowledge of the cyberthreats that their companies have to contend with. Strategic threat intelligence focuses on geopolitical events, cyber threat patterns in a specific sector, or how or why specific strategic assets of the business might be attacked and is used by stakeholders to integrate investments in and larger organizational risk management initiatives with the cyber threat landscape.

RELEVANCE & APPLICATIONS

Cyber Threat Intelligence (CTI) is an indispensable field dedicated to comprehending, analyzing, and mitigating cyber threats whilst comprehending the objectives, targets, and methods of a threat actor. It equips organizations, government entities, and security professionals with actionable insights that enable them to not only identify but also respond to and defend against cyberattacks. The applications of CTI span across a multitude of sectors, including critical infrastructure protection, national security, financial services, and healthcare. Its impact is felt profoundly in the world of proactive cybersecurity measures, where it empowers stakeholders to anticipate emerging threats and fortify their defenses accordingly.

CTI PROCESS & LIFECYCLE

The CTI process is a well-defined cycle that encompasses multiple stages. Typically, these stages include data collection, threat analysis, intelligence dissemination, and feedback loops for continuous refinement. This cyclic approach ensures that organizations remain adaptable in the face of evolving cyber threats. It is this dynamic nature of CTI that contributes to its effectiveness in enhancing overall cybersecurity posture.

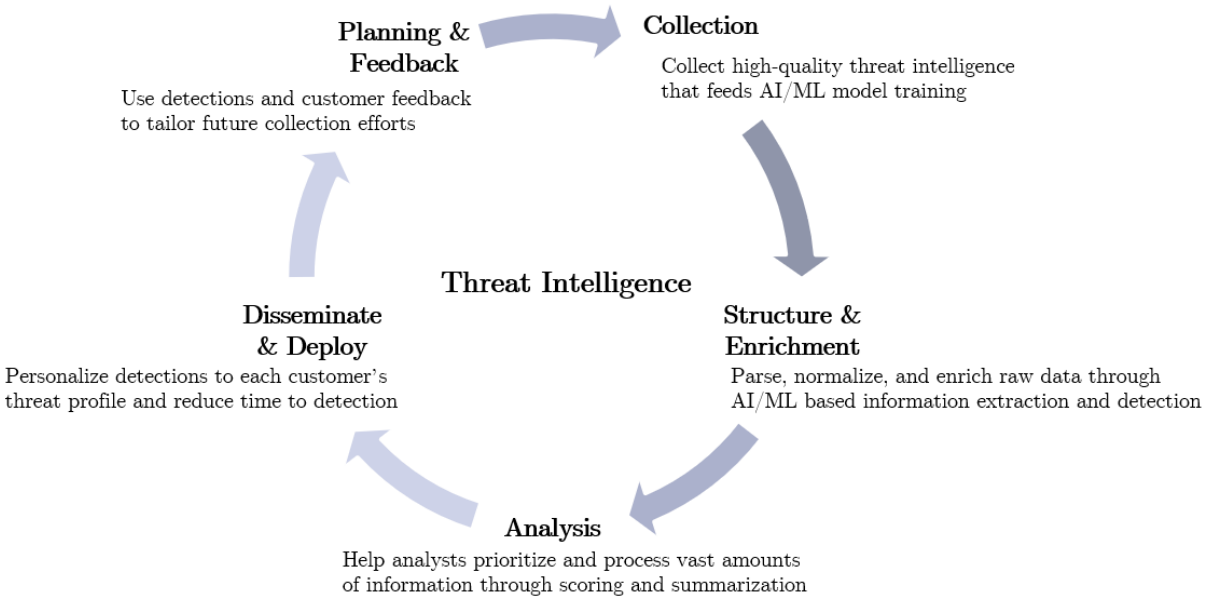

Figure 8: The Threat Intelligence Lifecycle.

CTI indicates an attack's delivery mechanism, aggression indicators, potential actors, and motivation. Reputation information, including malicious file feeds, IP reputation feeds, and phishing and botnet C&C URL feeds, is utilized to refine these. Such information can be obtained from large-scale security monitoring systems. The data mentioned above has the potential to enhance security systems such as firewalls, IPS/IDS, and SIEM. The majority of the gathered information consists of IP addresses, URLs, files, and mobile applications; it is preprocessed, categorized, and composed using various frameworks.

Generally, two types of data collection methods are employed to compose CTIs effectively: first, the **active collection** method, which includes web browsing and IP port scanning capabilities. The **passive collection** also includes log analysis of security solution logs, honeypots, and malware-infected user logs. By reducing noise and classifying the data, machine learning techniques are able to determine the association between particular cyber threats and the data. Moreover, they are associated with profile data of cyber threats in order to generate CTI. Due to the substantial data volume necessary to generate threat expressions with acceptable accuracy, security data feed collection infrastructures or third-party data feeds are used for CTI collection. Data consists of IP addresses, file hashes, URLs, and domains that are generated during network monitoring and web crawling, and additional observable information such as email addresses, domains, Whois details, or information that data collectors can derive by observing a specific IP address during a certain time period for surpicious activity. These risk indicators function as a point of reference for cyber-attacks and are identified so when the threat matrix surpassed a predetermined threshold.

2.5 The Intersection between OSINT and CTI

The confluence of OSINT and CTI is where the true potential of both fields is realized. OSINT serves as an invaluable source of information that can provide early warnings of cyber threats. It supplements other intelligence sources by offering a vast repository of publicly accessible data. When integrated into CTI operations, OSINT enriches the intelligence landscape, enabling more informed decision-making and proactive threat mitigation [14] [5] [13].

Data Collection and Analysis

INTERSECTION: OSINT and CTI both involve the collection and analysis of a wide range of data originating from both publicly accessible and confidential sources. OSINT provides fundamental data, whereas CTI is dedicated to the analysis of said data in order to identify cyber threats.

APPLICATION: CTI platforms establish a correlation between what is uncovered by OSINT tools such as Shodan and the identities of known threat actors who are targeting comparable vulnerabilities.

Threat Actor Analysis

INTERSECTION: OSINT profiles and identifies online organizations or individuals. This is expanded upon by CTI through the connection of profiles to cyber threats and attacks.

APPLICATION: OSINT detects the online presence of a cyber group, while CTI assesses their threat level by analyzing their past activities.

Incident Response

INTERSECTION: The contributions of both OSINT and CTI to incident response overlap. OSINT assists in the identification of potential compromise indicators, while CTI provides the necessary context to transform them into actionable intelligence.

APPLICATION: OSINT detects malicious IP addresses, while CTI determines whether or not they are associated with a larger campaign.

Phishing Detection

INTERSECTION: OSINT detects fraudulent emails and domains and these findings are linked by CTI to established threat campaigns.

APPLICATION: OSINT identifies a phishing domain, and CTI discloses that it is associated with a state-sponsored campaign that targets particular industries.

Vulnerability Management

INTERSECTION: OSINT identifies vulnerabilities and exposed systems whilst CTI conducts an intersection analysis of these vulnerabilities in relation to persistent cyber threats.

APPLICATION: OSINT identifies unpatched systems, while CTI ascertains whether or not they are being actively exploited by threat actors.

Supply Chain Security

INTERSECTION: OSINT conducts monitoring of suppliers and associates whose cybersecurity status is evaluated by CTI.

APPLICATION: OSINT detects the online presence of a supplier, while CTI discloses whether or not the supplier has been the target of cyber threats.

Dark Web Monitoring

INTERSECTION: OSINT includes the surveillance of the dark web and CTI establishes a connection between dark web activities and possible threats.

APPLICATION: OSINT identifies discussions concerning a potential attack, while CTI evaluates the threat's credibility and severity.

Predictive Threat Intelligence

INTERSECTION: Due to their overlap, both OSINT and CTI aid in the prediction of cyber threats. OSINT detects patterns, while CTI provides context for those patterns in order to forecast forthcoming threats.

APPLICATION: OSINT discerns an increase in discourse surrounding a particular susceptibility, while CTI forecasts the arrival of an imminent exploit campaign.

Attribution and TTPs

INTERSECTION: OSINT identifies the tactics, techniques, and procedures (TTPs) of malicious actors and CTI attributes TTPs to identified threat actors.

APPLICATION: OSINT detects a novel TTP, which is subsequently classified by CTI as a member of a recognized advanced persistent threat (APT) group.

Geopolitical Analysis

INTERSECTION: OSINT includes the surveillance of geopolitical incidents. CTI evaluates the potential cybersecurity implications of geopolitical events.

APPLICATION: The utilization of OSINT to monitor a political event and CTI to evaluate the likelihood of surveillance or hacktivist endeavors associated with it.

For instance, Shodan (OSINT) detects exposed industrial control systems. CTI establishes a connection between these systems and a known threat actor that specifically targets critical infrastructures.

2.6 The Role of Artificial Intelligence in CTI

The integration of Artificial Intelligence (AI) with Open Source Intelligence (OSINT) has emerged as a potent force in fortifying Cyber Threat Intelligence (CTI). We will examine the relationship between AI and OSINT, spotlighting their collective potential to support cybersecurity efforts [8].

AI enhances, rather than displaces, human analysts in threat intelligence. It collaborates with human expertise, aiding with the decision-making processes. AI solutions offer analysts data-driven insights, assisting them in making well-informed judgments regarding potential dangers. This collaboration exploits the capabilities of both artificial intelligence and human intelligence, thereby boosting the thoroughness and precision of threat assessments. AI systems utilized in threat intelligence exhibit exceptional scalability and adaptability, possessing the capability to manage escalating quantities of data and consistently acquire knowledge from emerging risks. The system's versatility enables it to enhance its threat detection capabilities over time, effectively keeping up with the dynamic and evolving threat landscape.

Organizations prioritize proactive threat hunting, using AI's predictive ability in order to predict and avoid attacks instead of just responding to them. The integration of artificial intelligence (AI) in threat intelligence has fundamentally transformed the process by augmenting data analysis, enhancing threat detection, enabling predictive capabilities, automating operations, facilitating collaboration between AI and human analysts, and shaping cybersecurity strategies towards proactive defense. AI technologies are progressing, and they are playing a crucial role in determining the future of threat intelligence.

Automating Threat Detection

AI has revolutionized threat detection by infusing automation into the process. Machine learning algorithms, encompassing supervised and unsupervised learning, are instrumental in examining massive datasets in real-time. Through this, AI can autonomously pinpoint anomalies and potential threats, thereby streamlining the initial phases of threat detection and diminishing the reliance on manual analysis.

Advanced Pattern Recognition

One of AI's unparalleled strengths lies in its ability to discern intricate patterns. Deep learning models, such as neural networks, exhibit exceptional prowess in recognizing subtle and mul-

tifaceted patterns within network traffic, user behaviors, and the characteristics of malware. This inherent capability enables AI to identify emerging threats that may elude traditional signature-based detection systems.

Natural Language Processing (NLP) for OSINT

AI's Natural Language Processing (NLP) underpins its efficacy in OSINT capacity is pivotal for extracting actionable insights from the voluminous corpus of text-based OSINT data.

2.6.1 AI-Driven Threat Analysis

Malware Detection and Analysis

AI-powered malware detection systems exhibit remarkable accuracy in identifying and categorizing malicious software. These systems dissect code, behavior, and known malware signatures to proactively detect and mitigate threats. By serving as a bulwark against malware infections, AI safeguards critical systems and networks from compromise.

Predictive Threat Intelligence

AI-driven predictive analytics harness historical and real-time data to anticipate potential cyber threats. By discerning trends, vulnerabilities, and attack vectors, predictive analytics empower organizations to implement pre-emptive security measures. This augments cybersecurity by reducing the likelihood of successful cyberattacks.

Threat Attribution and Actor Profiling

AI systems aid in threat attribution by scrutinizing diverse data points, including attack tactics, techniques, and procedures (TTPs), and correlating them with known threat actor profiles. This capability facilitates the identification of the source of an attack, be it a nation-state actor, a criminal syndicate, or a hacktivist collective. Threat attribution is pivotal for formulating precise and tailored responses.

2.6.2 Ethical and Privacy Considerations

Data Privacy and Ethical AI Use

As AI plays an increasingly central role in CTI, ethical considerations come to the fore. Ensuring responsible data handling practices and adherence to ethical principles is paramount. Safeguarding the privacy of individuals and organizations during data collection, analysis, and storage is a foundational tenet. Robust data anonymization techniques must be embraced to shield sensitive information from undue exposure.

Bias and Fairness in AI

Mitigating bias in AI algorithms is a non-negotiable imperative. Biased AI systems can inadvertently yield unfair or discriminatory outcomes. Ensuring fairness and transparency in AI-driven CTI is not solely an ethical obligation; it is also pivotal for engendering trust in the technology. Rigorous testing and continuous monitoring are essential to identify and rectify biases.

Compliance with Regulations

AI-driven CTI initiatives must align with data protection regulations and cybersecurity standards. GDPR [15], CCPA [16], and other privacy laws impose stringent requirements concerning the handling of personal data. Compliance with these regulations is not optional; it is a legal obligation. Stringent adherence is vital to sidestep potential legal ramifications.

AI IN CYBERSECURITY

Artificial Intelligence came to transform cybersecurity as we know it. The umbrella of AI encompasses a range of technologies, including machine learning, natural language processing, and deep learning, all of which exhibit remarkable capabilities in automating threat detection, analyzing vast datasets, and predicting cyberattacks. AI-driven cybersecurity solutions have the capacity to enhance and evolve human capabilities by processing information at a greater speed and scale.

AI TECHNIQUES APPLIED TO CTI

AI techniques are increasingly being integrated into CTI processes to augment human efforts. Machine learning models, can identify patterns in historical threat data, allowing the prediction of future cyberattacks. Natural language processing is instrumental in the automated analysis of textual threat intelligence reports, enabling quicker insights into potential threats. Deep learning, particularly in the realm of image and video analysis, assists in identifying visual indicators of compromise (IOCs) [17], adding a valuable layer of detection to the CTI toolkit. Additionally, AI can be harnessed in social media analytics, where it can monitor and assess online conversations and discussions related to cyber threats.

BENEFITS OF AI-DRIVEN CTI

The adoption of AI within CTI introduces a multitude of advantages. These include accelerated threat detection, a reduction in false positives, improved scalability, and enhanced decision support. AI-driven CTI tools are capable of processing and analyzing vast volumes of data swiftly, enabling human analysts to focus on higher-level tasks, such as strategy development and threat response. This symbiotic relationship between AI and human expertise is the hallmark of the future of CTI.

In summary, this chapter's review has illuminated the intertwined nature of OSINT, CTI, and AI in the domain of cybersecurity. It underscores the significance of OSINT as an invaluable information source, CTI as a proactive cybersecurity strategy, and AI as a catalyst for more efficient and effective threat detection. This chapter lays the foundation for subsequent chapters, which will delve deeper into the advancements, challenges, and future trajectories in these domains.

3 Advancements in Open Source Intelligence Techniques

In this chapter, we will be presenting a comprehensive analysis of the latest developments in OSINT techniques and their crucial role in Cyber Threat Intelligence. OSINT is the cornerstone of CTI, as it provides the initial data points required to detect, assess, and respond to cyber threats effectively. Recent advancements in OSINT methodologies have significantly improved the ability of cybersecurity professionals to gather intelligence from the open and accessible corners of the internet. This chapter delves into these advancements, highlighting their importance and their relevance in addressing contemporary cyber threats.

3.1 Traditional OSINT Methods and Challenges

Before we delve into recent innovations, it is essential to appreciate the foundations of OSINT. Traditional OSINT methods have long been utilized in cybersecurity. These include manual web searches, keyword analysis, and basic data collection techniques. However, these methods are grappling with several challenges:

1. **Information Overload**: The internet is an overwhelming repository of information.
2. **Skilled Analysts**: Effective OSINT analysis often requires skilled analysts who can discern meaningful patterns from the data.
3. **Unstructured Data**: The majority of OSINT data is unstructured, making it challenging to process and analyze.

3.1.1 Red Team Reconnaissance

As per the National Institute of Standards and Technology (NIST) [18], a red team is a formally authorised and organised group of individuals that simulate the attack or exploitation capabilities of a prospective adversary against the security measures of an organisation. The primary goal of the Red Team is to enhance corporate cybersecurity by showcasing the consequences of successful cyber attacks and by illustrating effective strategies for the defenders (known as the Blue Team) in a real-world setting. Commonly referred to as the Cyber Red Team.

Reconnaissance Techniques

To effectively simulate genuine adversary attacks, the red team must engage in reconnaissance activities. The initial phase of reconnaissance will focus on gathering Open Source information, as the level of engagement with the target may be limited to what is specified in the Standard Operating Procedure.

The MITRE ATT&CK® framework [19] is an invaluable resource for both red and blue teams. The ATT&CK framework is a publicly accessible repository of information about the methods and strategies used by adversaries. The reconnaissance category comprises the following techniques in the specified sequence:

> ACTIVE SCANNING: Engaging in proactive detection of open ports, vulnerabilities, or concealed directories on the target. Engaging in mass scanning that directly interacts with the target might generate significant noise and perhaps raise suspicion.

COLLECTING DATA ABOUT THE TARGETED HOST: Collecting data from the visible outside region of the subject. Exposing the operating system and endpoint management solution of workers in a snapshot shared on social media might assist attackers in customising their approach to exploit vulnerabilities. Furthermore, establishing public ties with firms specialising in the development of countermeasure solutions might provide as an indication of the target infrastructure. Ultimately, the disclosure or unauthorised release of defence solution configurations is highly significant, since it enables an attacker to enhance the precision and effectiveness of their assault. Discovering that an Endpoint Detection and Response program is set up to disregard a customised file path in Windows might aid the attacker in surreptitiously implanting their malware onto the victims' disc.

COLLECTING VICTIM IDENTITY INFORMATION: Acquiring information pertaining to the victim's identity, including personal particulars such as names, emails, phone numbers, as well as sensitive details like credentials. If Multi Factor Authentication (MFA) is not configured, leaked credentials provide a convenient entry point for an attacker to access the organization's infrastructure as an authenticated user.

COLLECT VICTIM NETWORK INFORMATION: Obtaining details on the target's networking architecture. The data in question pertains to internal subnets, topology, and trusts. An attacker with knowledge of the internal architecture might selectively target particular networks and evade detection by Network Based Intrusion Detection Systems (NIDS).

COLLECT TARGET ORGANISATION INFORMATION: Obtaining non-infrastructure related details about the victim organisation, such as branch locations, department names, internal job positions, and business processes.

PHISHING FOR INFORMATION: Sending deceptive emails or texts with the intention of obtaining sensitive information from the recipient. The target is deceived into providing confidential information, such as domain credentials and online account credentials, which will be utilised at a later time by the adversary.

CONDUCTING CLOSED SOURCE RESEARCH: Gathering information on the target from exclusive sources, such as private databases that store data about the organisation, procured technical data, and Thread Intelligence feeds. Reputable and illegal firms alike sell these categories of data in exchange for a monetary sum. Advanced Persistent Threats have the capability to purchase leaked databases or exfiltrated data from the organisation in order to obtain valuable information about their target. As a result of the increasing need for restricted data sources, several marketplaces have arisen on both the visible and concealed sections of the internet.

CONDUCT SEARCHES ON PUBLICLY ACCESSIBLE TECHNICAL DATABASES: Collecting data from sources such as DNS records, website certificates, WHOIS data, and databases collecting scans. Several Open Source search tools prioritise this aspect of reconnaissance as it involves limited to no involvement with the target and serves as a first phase for acquiring information by an attacker.

CONDUCTING OPEN SOURCE INTELLIGENCE (OSINT) ON PUBLIC WEBSITES/DOMAINS: Extracting information voluntarily shared by the target on external websites, such as

social media platforms, search engine indexes, and public code repositories.

ANALYSING TARGET-OWNED WEBSITES: Extracting valuable data from websites owned by the target organisation to gain valuable insights. For instance, this data comprises of geographical positions, crucial details on operational protocols, contact information, and commercial affiliations with other entities.

As seen from the outlined approaches of ATT&CK, open source information is the foundation of reconnaissance. In a scenario where a red team is up against an organisation equipped with up-to-date security measures, every piece of information gathered is crucial for achieving the assessment's goals.

For instance, the social networking platform LinkedIn could potentially be utilised to gather information on the specific organisation of interest. By initially locating the organisation on Linkedin and subsequently cataloguing the personnel, the investigation for compromised login information and pilfered data gets progressively more accurate and focused. Individuals lacking expertise in cybersecurity and data protection can be identified and specifically targeted using a method known as Spear Phishing [20]. If the target of the phishing effort is an executive member or another high-value individual, the approach is referred to as Whaling. The impact of these attacks is magnified when the assailant has previously obtained entry to an email account owned by the organisation, a vendor, a client, or a partner, as the recipient considers the sender's address to be reliable.

3.1.2 Google Dorking and Advanced Search Techniques

Google Dorking [21], also known as Google hacking, is a method of using advanced search operators and specific search queries to extract information from the Google search engine that may not be readily accessible through standard search methods. This technique, when used responsibly and ethically, has various applications in research, information retrieval, and digital forensics. The concept of Google Dorking emerged as users began to realize the potential of refining their searches beyond simple keyword queries. It gained popularity among security professionals and researchers in the early 2000s when they sought to uncover vulnerabilities and sensitive information inadvertently exposed on the internet.

COMMON GOOGLE DORKING OPERATORS & TECHNIQUES:

SITE OPERATOR: By using the "`site:`" operator followed by a specific domain, researchers can narrow down their search to a particular website, limiting the results and making targeted searches. For example, "`site:wikipedia.org open source intelligence`" restricts search results to Wikipedia articles about open source intelligence.

FILETYPE OPERATOR: Researchers can utilize the "`filetype:`" operator to locate specific types of files, such as PDFs, Word documents, or spreadsheets. For instance, if we use "`filetype:pdf keyword`" we can limit our search results and retrieve retrieve PDF files containing a certain keyword, by a certain author etc.

INURL OPERATOR: The "`inurl:`" operator allows researchers to search for specific terms within the URL of a webpage. A query like "`inurl:password reset`" retrieves webpages associated with password reset links on various websites.

INTEXT OPERATOR: To search for specific terms within the body of a webpage, researchers can use the "`intext:`" operator. For example, "`intext:privacy policy`" yields webpages that contain mentions of "*privacy policy*".

RELATED OPERATOR: The "`related:`" operator is employed to discover websites related to a specific domain. A query like "`related:unipi.gr`" returns sites connected to University of Piraeus.

It is imperative to emphasize that Google Dorking should always be conducted within the boundaries of ethical and legal guidelines. Using these techniques for malicious purposes, such as extracting sensitive information or exploiting security vulnerabilities, is illegal and unethical. Respect for privacy and adherence to the law are fundamental principles that researchers and security professionals must uphold. Search engines like Google frequently update their algorithms and mechanisms to safeguard user privacy and security. As a result, some dorking techniques may become less effective over time. Researchers should stay informed about these changes and adapt their methods accordingly.

3.1.3 Digital Footprint Analysis

Recent advancements have revolutionized digital footprint analysis, providing cybersecurity experts with powerful tools to monitor and assess online activities.

WEB CRAWLING & SCRAPING: Automated web crawling and scraping tools can collect data from websites, forums, and social media platforms at scale. This process enables the extraction of valuable information for analysis. It helps in monitoring trends, tracking adversaries, and identifying potential threats in real-time.

SOCIAL MEDIA INTELLIGENCE: Social media platforms have become rich sources of information. Advanced OSINT techniques now allow for in-depth analysis of social media content. This includes sentiment analysis, which can gauge public sentiment around specific topics or entities. Such insights can be invaluable in assessing the public's reaction to security incidents or emerging threats.

3.1.4 Geospatial OSINT

Geospatial OSINT is a specialized field that leverages geographic data to enhance threat intelligence.

GEOLOCATION DATA: By analyzing geolocation data embedded in digital content, such as photos, posts, or metadata, analysts can determine the geographical origin of potential threats. This is particularly valuable for identifying the physical locations of cybercriminals or tracing the source of cyberattacks.

MAPPING & VISUALIZATION: Tools for geospatial mapping and visualization have become increasingly sophisticated, allowing cybersecurity professionals to create interactive maps that display threat data geographically. This visual representation aids in understanding the global distribution of threats and vulnerabilities.

3.1.5 Deep Web and Dark Web Analysis

Open Source Intelligence plays an essential part in the study of the deep web and the dark web, despite the fact that its implementation in these hidden sections of the internet that are are encrypted, inaccessible, and illicit, presents a several challenges.

Surface Web vs Deep Web vs Dark Web:

Websites and information that may be accessed through search engines such as Google are included in the phenomenon known as the **Surface Web**. Users are able to quickly and easily access the content that has been indexed.

Deep Web refers to content that is not indexed by search engines and requires certain access permissions. Examples of content that falls into this category include private databases, password-protected pages, and academic materials [22].

The **Dark Web** is a subset of the deep web that can only be accessed with the use of specialized software, such as Tor. It provides a platform for users to engage in anonymous and frequently unlawful activities, such as black markets, forums, and illegal services [22].

So how can OSINT aid the analysis of Dark and Deep Web?

— Access to Restricted Information: OSINT approaches have the ability to access data from sources that are not indexed by regular search engines, which increases the amount of information that is available for analysis.

— Private Databases and Forums: OSINT has the ability to access closed forums, websites with restricted access, and other hidden resources in order to collect material that is pertinent to particular investigations, research, or intelligence collecting.

— Legal and Ethical Restrictions: Professionals in the field of OSINT are required to handle legal and ethical considerations while accessing and exploiting material from the deep web, while also protecting individuals' privacy and avoiding engaging in criminal activity.

The Use of Open Source Intelligence in Dark Web Analysis:

— Monitoring Illicit Activities: The use of OSINT technologies and methodologies allows for the monitoring and analysis of activities that take place on the dark web. These activities include but are not limited to illegal markets, forums, and communication channels.

— Threat Intelligence: It assists in the collection of threat intelligence by identifying potential cyber threats, vulnerabilities, malware, and evolving attack patterns that could have an effect on cybersecurity.

— Law enforcement and investigations: OSINT on the dark web provides assistance to law enforcement authorities in their efforts to track illicit activities such as the trafficking of drugs, cybercrime, and human beings.

Obstacles and Consideration Points:

— Anonymity and Encryption: The dark web is famous for its anonymity and encryption, both of which provide difficulties for open source intelligence practitioners when it comes to identifying individuals and tracking activity.

— Considerations of an Ethical Nature: Participating in illegal activity gives rise to ethical conundrums. In order to extract information, practitioners of open source intelligence need to negotiate these concerns.

— Technical Expertise: proficiency in specialized tools, protocols, and security measures is required for effective operations in the dark web.

Tools and Procedures:

SPECIALIZED SEARCH ENGINES: OSINT specialists now use specialized search engines that can access deep web databases and retrieve information that traditional search engines cannot reach. These tools are crucial for locating hidden web resources and uncovering data that may be relevant to CTI. For example, TOR BROWSER a widely utilized tool that enables users to access the dark web while maintaining anonymity.

ANONYMIZATION TECHNIQUES: To operate safely in the dark web, OSINT practitioners employ anonymization techniques to protect their identity while gathering information. These techniques ensure their safety while they navigate these uncharted territories.

ADVANCED DATA COLLECTION TOOLS: Tools designed for dark web data collection have evolved, enabling the retrieval and analysis of data from underground forums, marketplaces, and encrypted communication channels. These tools provide valuable insights into cybercriminal activities, hacking tools, and the sale of compromised data. Crawlers and scrapers, for instance, are customized software designed to extract and analyze data from the dark web, with the purpose of gathering intelligence.

HUMAN INTELLIGENCE: OSINT practitioners utilize human sources, forums, and discussion groups to acquire information from the dark web.

3.2 OSINT Tools

OSINT relies heavily on a variety of tools for collecting, processing, and analyzing data from publicly available sources. Below are some common categories of OSINT tools along with notable examples:

Web Scraping Tools

BEAUTIFUL SOUP: A Python library for web scraping that simplifies parsing HTML and XML documents. It's particularly useful for extracting data from websites [23].

SCRAPY: A powerful web crawling framework for Python. It's highly customizable and capable of scraping large amounts of data efficiently [24].

OCTOPARSE: A user-friendly, cloud-based web scraping tool that allows non-technical users to extract data from websites through a visual interface [25].

Social Media Intelligence Tools

HOOTSUITE: A social media management platform that allows monitoring of multiple social media channels, tracking keywords, and scheduling posts [26].

BRANDWATCH: A social listening and analytics tool that helps organizations monitor brand mentions, trends, and sentiment across social media [27].

TALKWALKER: Offers real-time social media monitoring, sentiment analysis, and competitive intelligence, helping organizations track and analyze their online presence [28].

Geospatial Analysis Tools

QGIS: An open-source Geographic Information System (GIS) software that enables users to create, edit, and analyze geospatial data and maps [29].

ARCGIS: A comprehensive GIS platform developed by Esri, known for its advanced geospatial analysis capabilities and extensive mapping tools [30].

GOOGLE EARTH PRO: A versatile tool for exploring geospatial data and creating customized maps and visualizations [31].

Dark Web and Deep Web Tools

TOR BROWSER: The Tor network provides anonymity by routing internet traffic through a series of volunteer-operated servers, enabling access to the dark web [32].

FREENET: A decentralized network designed for secure, anonymous communication, and file sharing [33].

I2P (INVISIBLE INTERNET PROJECT): An anonymizing network that facilitates secure and anonymous communication, file sharing, and web browsing [34].

Data Analysis and Visualization Tools

TABLEAU: A powerful data visualization tool that allows users to create interactive and shareable dashboards [35].

POWER BI: Microsoft's business analytics service for creating interactive reports, dashboards, and visualizations [36].

PYTHON LIBRARIES: Python offers numerous libraries, including Pandas [37] for data manipulation, Matplotlib [38] for plotting, Seaborn [39] for statistical data visualization, and Jupyter [40] for interactive data analysis.

Domain Analysis Tools

WHOIS LOOKUP SERVICES: Domain name registration information can be obtained using WHOIS lookup services like ICANN's WHOIS or domain registrar-specific WHOIS tools [41].

DNS ENUMERATION TOOLS: Tools like DNSdumpster [42] and DNSlytics [43] can provide insights into domain names, IP addresses, and associated domains.

Email and Metadata Analysis Tools

MALTEGO: A powerful OSINT tool for visualizing and analyzing information about people, organizations, and relationships by aggregating data from various sources [44].

METADATA ANALYSIS TOOLS: Tools like ExifTool [45] and Metagoofil [46] allow the extraction and analysis of metadata from files, including geolocation data, authorship details, and timestamps. Special tools have been created that automate repeated processes and tasks.

FOCA: Fingerprinting Organizations with Collected Archives; a tool primarily used to uncover metadata and concealed information within scanned documents is called a scanner. It is capable of examining a diverse range of files, such as Microsoft Office, Open Office, and PDF documents. It also analyzes Adobe InDesign or SVG files. The archives are scanned using Google, Bing, and DuckDuckGo. FOCA gives the possibility to include local files to eliminate the EXIF data from image files, and prior to downloading and analyzing the files, an additional URL analysis is also performed [47].

Red Team Tools

AMASS: a Golang based tool created by OWASP which utilizes passive and active reconnaissance to map the attack surface and discover public facing assets of an organization. The techniques followed range from interacting with external APIs such as Shodan's, to looking up TLS certificate and IP info. The tool is widely trusted and used by red teams due to the automation it enables [48].

GITLEAKS: a Golang based tool used to search git directories for hardcoded secrets such as credentials and API keys. Leaked secrets have resulted in unauthorized access in company property such as the case of Toyota where attackers got access to customer data through a leaked access key. Moreover, notable is that fact that gitleaks offers a Github Action where each code commit of a repository would be vetted for leaked secrets [49].

SPIDERFOOT: an Open Source automation tool. It offers several great features with its 200 integrated modules, from domain enumeration to Threat Intelligence. The power of Spiderfoot lies in its simplicity, where only a single input such as a username, an email address or a domain can extract a lot of information about a person or company. The tool presents the pieces of information discovered in a connection oriented graph fashion in order to demonstrate the links between data [50].

LINKEDINT: an Open Source scraper for the social network Linkedin. It indexes the employees of a company based on their Linkedin profiles and discovers company email addresses based on naming conventions. Discovering the employees of a company is very useful for a red team since several pieces of information, such as departments and personnel roles, can be discovered. Moreover, employees can be selected for targeted phishing in order to infiltrate the organization [51].

RECON-NG: a framework that is akin to Metasploit, designed to facilitate web-based open source reconnaissance. This tool comprises multiple autonomous modules that execute

distinct functionalities. Recon-NG is consistently consolidating all the acquired information in a local database. The user exercises control over the research process by choosing the specified module, and the tool then automatically generates information based on that selection. The system has exceptional scalability while handling intricate investigations [52].

THE HARVESTER: Intended for utilization in the initial phase of a red team assessment or penetration test. The tool conducts open source intelligence (OSINT) collecting to analyze the external threat environment of a domain. It enables the retrieval of public information associated with a domain or corporate name from search engines. Specifically, it has the ability to compile a comprehensive inventory of the company's email addresses and host names, along with any subdomains, IP addresses, and URLs associated with the domain. Additionally, it allows for the generation of user-friendly HTML or XML representations of the results [53].

Internet Device Search Engines

Due to the number of internet connected devices, special search engines have been created in order to catalog exposed services. The entire public IP range is scanned for open ports and the hosted services are enumerated. An attacker could utilize the data provided by the following search engines in order to discover exposed services without interacting with the targets. An organization can opt out of scans by blocking the IP addresses associated with each the scanning service.

SHODAN: a search engine for internet connected devices. Internet wide port and service scans are performed and the results are uploaded to their website. When performing a query the familiar layout of a search engine is presented, with filtering options such as country, service or IP owner. A user can quickly go through massive amounts of indexed open services with the help of filters. An advantage of Shodan is that information about services are saved. For example, in case the port 3389 is found open (Remote Desktop Protocol) a screenshot of the login screen is saved. Therefore, an attacker could get a high level overview of the services exposed in an IP range or an organization with communicating with the hosts. Many modules about Shodan's API have been written in a variety of programming languages, enabling automation and mass enumeration [54].

CENSYS: a platform comparable to Shodan. Internet connected devices are routinely scanned and their services are indexed. With Censys' search engine a user has access to the scan data, as well as to filters to aid investigators narrow down their results [55].

FOFA: a Chinese internet connected device search engine that indexes the publicly services and devices. FOFA seems to more aggressively enumerate the open services compared to the other scanners [56].

LEAKIX: it is comparable to the other platforms, though the the search and indexing is optimized for finding bad configurations and leaked info through website [57].

INTELTECHNIQUES is an online tool developed by Michael Bazzel that provides a wide range of search utilities categorized by technique. When utilizing this tool, the investigator

chooses the desired services, and the program automatically generates the corresponding inquiry URLs. Subsequently, the user can input them into the browser in order to initiate the query. Nevertheless, the process of seeing and gathering the information is still done manually [58].

A brief comparison of some of the aforementioned tools is described in the table below:

OSINT Tool	Input				Output	Extensibility	Interface	Platform	Other
	Identity data	Network data	File data	Selectable data source					
FOCA	x	domain	File name, folder	Google, Bing, DuckDuck Go	Identity info, network info, file info	x	Standalone program	Windows	Server discovery module
Maltego	Personal information, company, community	domain	File URL	x	Identity info, network info, file info	Custom transforms	Standalone program	Linux, Windows, MAC	Location, auto input/output refeed, results in oriented graph
Metagoofil	x	domain	File type	x	Network info, file info	x	Command line	Linux, Windows	Option to narrow results
Recon-NG	Personal information	domain	x	several	Identity info, network info, file info	x	Command line	Linux	Location, modules for discovery and exploitation
Shodan	Country, city, keyword	Operating system, IP address, port, host name	x	x	Network info	x	Web interface	online	Location, webcam captures
Spiderfoot	Email, real name, phone number	Domain, IP address, subnet, host name	x	several	Network info	Custom modules	Web interface	Linux, Windows, MAC	Different types of scans, results in oriented graph
The Harvester	Company	Domain, DNS server	x	several	Identity info, network info	x	Command line	Linux, Windows, MAC	Results in reports, options to narrow files and results
IntelTechniques	Personal information, company, community	Domain, IP address	File name, file type, file URL	several	Identity info, network info	x	Web interface	Online	Location, public records, OSINT VM

Based on the user's requirements, certain tools may be more appropriate than others for a specific task. Therefore, if our objective is to retrieve concealed data from files, FOCA and Metagoofil are specialized tools created specifically for this task. Specifically, the first product appears to be more comprehensive, refined, and robust compared to the second one. FOCA offers further features, in addition to its ability to analyze file metadata, to enhance the concealed information.

Consequently, it has the capacity to deduce more information about the target. However, when seeking network information, it is advisable to consider using Shodan, Spiderfoot, and The Harvester as recommended tools for this specific activity. Firstly, we recommend utilizing Spiderfoot to examine the structure of the target and gather internal (yet publicly accessible) data regarding the target organization. Alternatively, we would enhance the findings by using data from Shodan, which provides detailed information about Internet of Things (IoT) devices, surveillance cameras, webcams, Voice over Internet Protocol (VoIP) systems, and other smart services.

Lastly, if the objective of the search is to get the maximum amount of information for a specific input, the tools Recon-NG and Maltego are the most comprehensive options available. These tools will provide a wide range of data and linkages.

The initial module is equipped with numerous components and interfaces with a database that expands in size as the inquiry progresses. This framework is highly suitable for conducting penetration tests, preventing phishing and social engineering assaults, and even creating a profile of an individual. Conversely, if we wish to evade the use of the command line and instead choose a more user-friendly interface, Maltego serves as a suitable substitute for OSINT activities. The system utilizes automated inference methods with transforms that expand the scope of the initial

search. Furthermore, it can be expanded using personalized discovery methods. Although the aforementioned comparison has been conducted based on the desired outcome, in reality, the user's options will be limited by the available input and the data format supported by the selected OSINT tools. It is important to understand that these technologies are complementary and not mutually exclusive. This means that a comprehensive OSINT study can benefit from using multiple tools simultaneously. While certain tools may yield comparable outcomes for a given search, each tool may uncover unique facts that others do not capture.

3.2.1 Utilizing Social Media as OSINT tools

Due to the widespread use of social media a great deal of information exists on personal or business social media pages. Online social networks provide a wealth of OSINT (Open Source Intelligence) including social connections, activities, and personal information about individuals. However, it is important to note that not all information on the internet is readily available. Investigators encounter many challenges, such as limitations on privacy and platform usage, as well as issues related to data accessibility and durability [59][60].

LIMITATIONS ON PERSONAL PRIVACY

In response to increasing concerns over privacy, several social networking sites have included additional privacy control features to limit access to personal information. Despite the rising concerns about privacy and the public need for more laws to safeguard individual privacy, one of the most effective approaches to get information from consumers is through straightforward requests. A prevalent strategy employed by third-party application developers is creating programs that request unnecessary permissions from users, with the intention of obtaining extra information. However, this may provide a misleading perception of privacy while interacting with online social networks.

LIMITATIONS OF THE PLATFORM

Data influx into online social networks occurs on a vast magnitude, while its outflow is subject to stringent monitoring by the social media platform. Social networking systems often regulate the dissemination of information using mechanisms such as social connections, user-defined privacy settings, rate limiting, activity tracking, and IP address limitations. While access control measures are commonly used to safeguard user privacy, certain mechanisms may be purposely modified or limited to safeguard the economic viability of the service platform.

AVAILABILITY OF DATA

As previously said, OSINT, by its very definition, lacks the capability to uncover material that is not already available in the public domain. In the event of an extreme circumstance, the required data could not have been collected by the platform, for instance, if a user decides not to disclose information on their social networking profile. If the necessary information was not previously documented and is not publicly accessible, OSINT is unable to uncover the needed information. In a conventional scenario, the needed information is present, but privacy and platform limitations create an obstacle that obscures

significant chunks of the information. Consequently, investigators are compelled to collect the needed information from the digital traces left by user actions, which are located outside of the obscured area.

LEGAL MATTERS

Although social networking networks like Facebook prohibit the use of screen scrapers and other data mining tools in their terms of service agreements, the legal validity of these prohibitions is still uncertain. Courts have acknowledged that robot web spiders and screen scrapers can be legally responsible for engaging in digital trespassing in certain situations. Issues of concern include:

1 Issues with the protection of personal information:
   — **Data Collection**: It is crucial to adhere to the terms of service and privacy regulations of social media platforms while collecting information for OSINT (Open Source Intelligence) objectives. Noncompliance with these terms may result in legal consequences.
   — **Personal Information**: The act of gathering, retaining, or disseminating personal data acquired via social media platforms without obtaining permission might violate privacy regulations. It is imperative to safely handle and utilise such data in accordance with relevant privacy rules, such as the General Data Protection Regulation (GDPR) in the European Union or the California Consumer Privacy Act (CCPA) in the United States.

2 Patents, copyrights, trademarks, and other legal protections for intellectual creations:
   — **Copyright Issues**: Content disseminated on social media platforms is frequently safeguarded by copyright legislation. Engaging in the use, replication, or dissemination of copyrighted content without authorization may result in the violation of copyright laws.
   — **Trademark Infringement**: Improper use of trademarks or logos on social media platforms may be a violation of trademark legislation. Any use that may cause ambiguity regarding the origin or approval of products/services might result in legal complications.

3 Ethical considerations:
   — **Deception and Misrepresentation**: Participating in deceitful activities, such as fabricating false accounts or mischaracterizing one's identity, may infringe either both platform regulations and legal norms. Engaging in impersonation of another individual with the intention of gathering data might lead to legal repercussions.
   — **Stalking and Harassment**: Engaging in aggressive or improper data collecting from social media, such as stalking or harassing individuals, not only violates the regulations of the site but may also infringe against laws pertaining to harassment and stalking.

4 Data Protection Regulations:

When gathering data from social media sites that involves the movement of information across national borders, it is essential to adhere to international data protection rules. Guarantee that data is sent and handled in compliance with applicable legislation.

5 Defamation and libel: refers to the act of making false statements about someone that harm their reputation.

Publication of False Information: Sharing false or defamatory content discovered on social media may result in potential legal responsibility for defamation or libel. It is imperative to authenticate information before to using or disseminating it in order to prevent any legal ramifications.

SNAPCHAT

Snapchat is a messaging app developed by Snap Inc. which allows users to exchange photos, videos, and messages. It also includes features such as Stories and Discover, which allows users to access content from various media outlets. One of the most useful tools for OSINT researchers is Snapchat Map, a feature of Snapchat that allows anyone to view other people's locations on a map where a video or a picture has been uploaded. The map updates in real-time and provides location-based filters, which can be used to enhance snaps taken in certain locations. The map features a heatmap where major world events can be easily discovered since the more pictures or videos are uploaded from a certain location the more noticeable the point becomes on the map. The researcher can then click on the map and get an idea of what is happening from the uploaded pictures and videos.

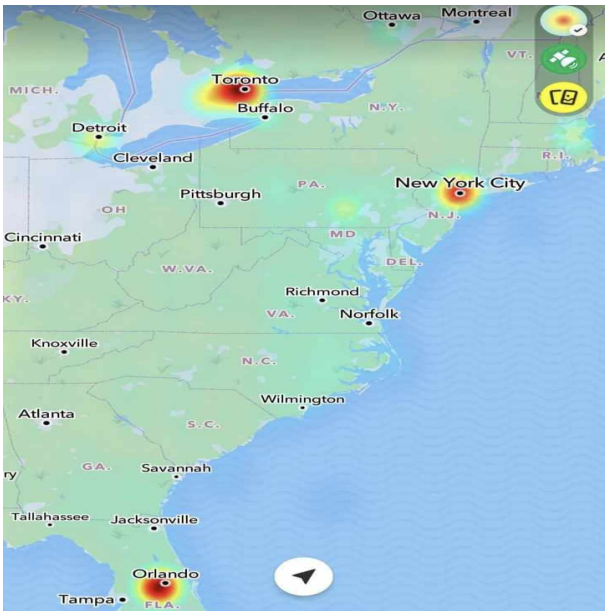

Figure 9: Snapchat map demonstrating areas of high user activity through shared audio-visual posts.

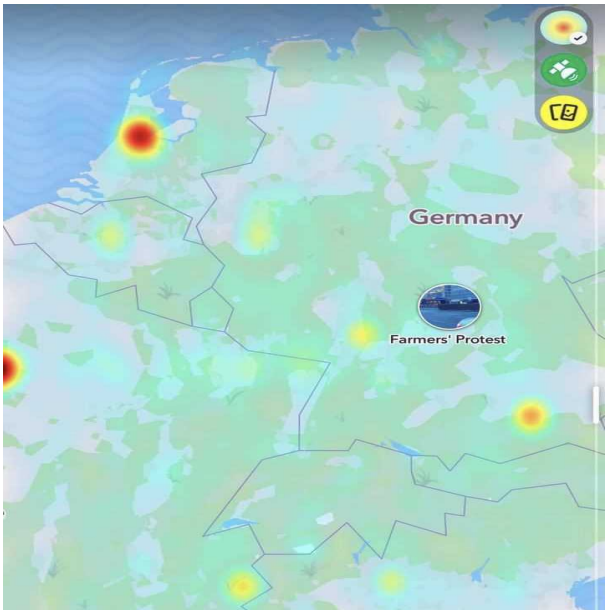

Figure 10: Demonstration of pressing events such as the Farmers' protest.

INSTAGRAM

Instagram can be used for OSINT in various ways, such as gathering information about trends, people, locations, and more. OSINT on Instagram can be used to identify patterns and trends in user behavior, which is used for marketing purposes such as understanding customer habits. Additionally, OSINT on Instagram can be used to uncover relationships between people, in the form of follows, and can provide valuable insight about a particular person or location. The Story feature on Instagram is disappearing pictures or short videos often accompanied by the location. Public profiles that upload stories or posts tagged with a specific location are indexed and can be discovered by anyone searching for that specific location.

TWITTER

Twitter can be a useful tool for monitoring world events and gathering open source intelligence. Because Twitter is a widely used platform, many people, including journalists, government officials, and other sources of information, use it to share news and information about current events. This makes it a valuable source of information for anyone interested in gathering open source intelligence. The instant sharing of information achievable through Twitter can be utilized for OSINT purposes with the following methods:

Following relevant accounts: Accounts and Hashtags related to the topic of choice that share news and updates. For example during the invasion of Ukraine journalists created accounts that gather and fact check information and footage from the war. Such intel is then used to analyze enemy movements or to spread awareness.

Using Alerts: Twitter has an advanced search feature that allows the set up of alerts. The user will then be notified when new content is posted related to the search parameter of choice.

Analytics: Twitter analytics are tools provided by Twitter to businesses to mainly gauge the engagement with followers. Open Source tools such as birdwatcher use the Twitter

API to download and then analyze the tweets locally.

3.2.2 Case Studies

Illustrating the utilization of social media in Open Source Intelligence (OSINT) in various scenarios:

`Ukrainian Conflict` (2014-Present) [61]: Social media platforms played a crucial role in providing information during the ongoing Ukrainian conflict. OSINT specialists conducted surveillance and examined social media posts to verify and enhance conventional intelligence. This involved validating the movements of military personnel, discerning the kind of weaponry utilized, and exposing possible violations of international humanitarian law.

Throughout the Russia-Ukraine conflict, social media has been essential in spreading information, recording incidents, and influencing perspectives at both local and international levels. Twitter and messaging systems like Telegram were inundated with photographs captured by soldiers and videos depicting battles or moments of respite. The intelligence agencies conducted vigilant surveillance on established communication channels to gather information on enemy activities, including their strategies and the whereabouts of their squads. A significant number of these videos and photographs were disseminated openly and reached a substantial audience of hundreds of thousands of individuals via news outlets or social media platforms, providing a perspective of the combat via the soldiers' vantage point.

— **Dissemination of Information and Propaganda**: OSINT analysts surveil social media platforms such as Twitter, Facebook, and VKontakte for geotagged posts, images, and videos in order to authenticate the locations of military activities, attacks, and troop movements.

— **Propaganda Analysis**: Multiple factions engaged in the conflict utilize social media platforms as a means to disseminate propaganda. OSINT specialists monitor and examine both authorized and unauthorized sources to detect deceptive storylines, misinformation, and strategies of propaganda designed to influence public sentiment.

— **Civilian Documentation and Human Rights Monitoring:**
User-Generated Content: Social media platforms serve as repositories of user-generated content that vividly depict the consequences of the conflict on non-combatant individuals. OSINT specialists collect and examine these messages, photographs, and videos to record instances of human rights violations, devastation, and harm to civilians.

— **Military Movements and Equipment Identification:**
Monitoring Military Activities: Users frequently provide visual content depicting the movement and deployment of military forces and equipment. OSINT professionals scrutinize this data to monitor the mobility of military forces, discern the categorization of armaments, and determine possible breaches of international treaties or ceasefires.

— **Verification and Fact-Checking of Information**:

Claims Verification: Social media posts frequently include assertions and data regarding the dispute. OSINT analysts authenticate the veracity of such assertions by cross-referencing diverse sources, such as videos, photographs, and eyewitness accounts.

— **Aid and Humanitarian Efforts**:

Aid Coordination: Social media has been utilized to facilitate the organization and allocation of humanitarian assistance. OSINT analysts surveil these platforms to detect locations requiring immediate assistance and to monitor the distribution of humanitarian relief.

Platforms like Reddit have become crucial centers for the immediate sharing of information in modern war situations, providing uninterrupted livestreams and up-to-date news continuously. These forums function as dynamic archives, offering a wide range of news, combat-related resources, and reports, along with other vital information, that demonstrate the significant influence of OSINT in areas of war.

The screenshots provided here illustrate the significant significance played by these platforms. Videos depicting assaults and conveying vital up-to-the-minute data graphically illustrate the seriousness of the situation. These examples highlight the powerful impact of OSINT and the quick spread of information through social media platforms, stressing their importance in today's information economy.

The promptness and ease of use of these platforms highlight their crucial function in communicating the actual conditions of areas affected by war. The existence of live videos capturing events as they happen illustrates the influence of social media in rapidly and comprehensively spreading important information.

The forums have far-reaching repercussions that go beyond simple information sharing. They have a substantial impact on the perception and comprehension of crisis situations worldwide. This demonstrates the vast capacity of social media to offer unedited, firsthand viewpoints, emphasizing the significant influence of OSINT.

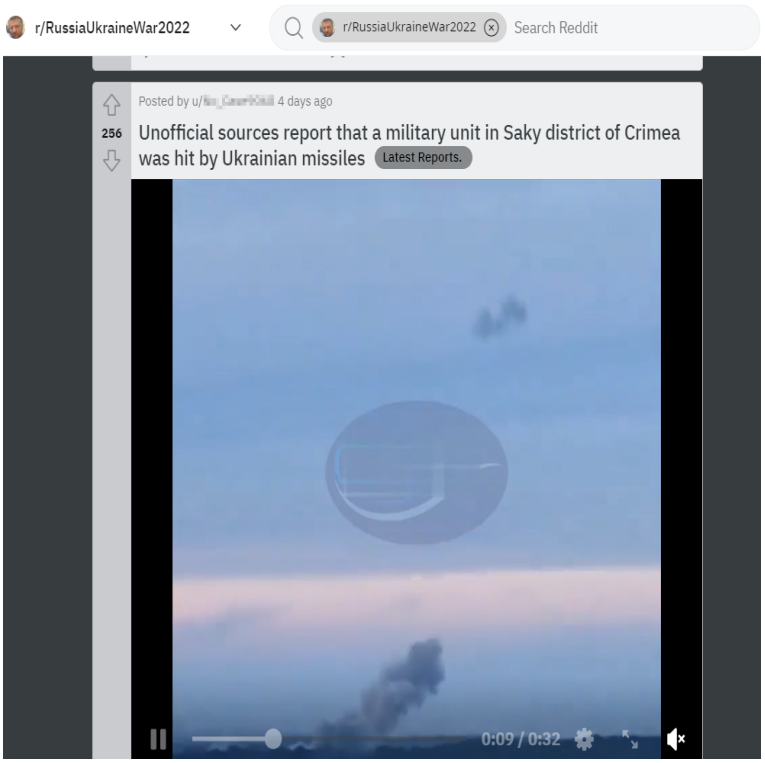

Figure 11: Military Unit hit by missiles.

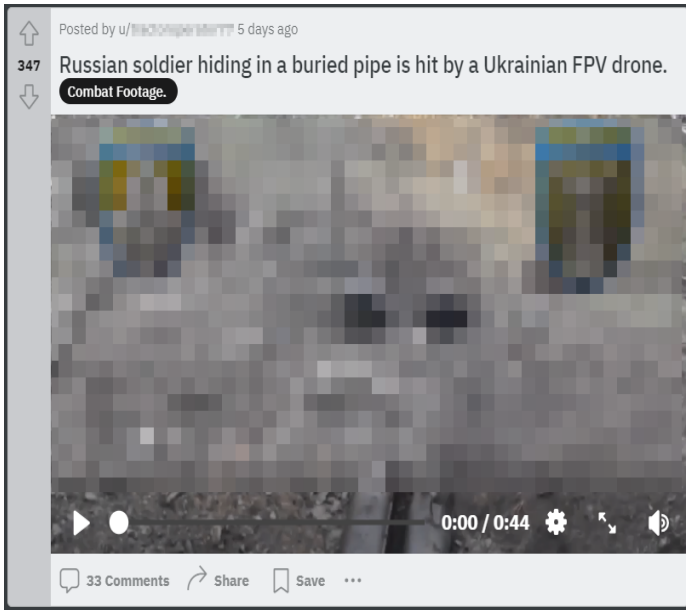

Figure 12: Soldier hiding after an attack.

A similar situation is observed with the `Gaza-Israel Conflict` [62]. The screenshots provided offer a compelling demonstration of the pivotal role these platforms play. They showcase diverse content, including multimedia depicting the conflict's effects, personal testimonies, and the rapid transmission of critical real-time information. In the messaging app Telegram, activists and other groups are using OSINT to retrieve information about infrastructures and facilities; the example below presents how a group discovered the electrical facilities of Israel, collected data and managed to cause outage for a short duration

of time.

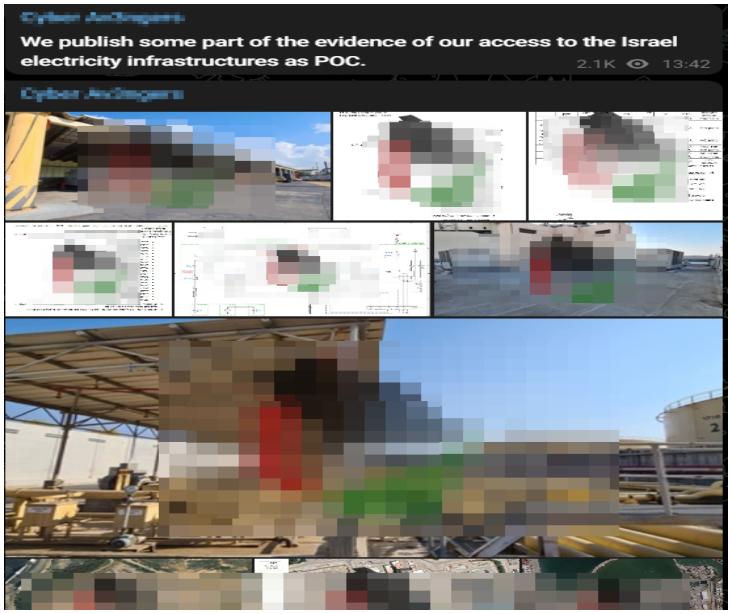

Figure 13: Information on electricity infrastructures.

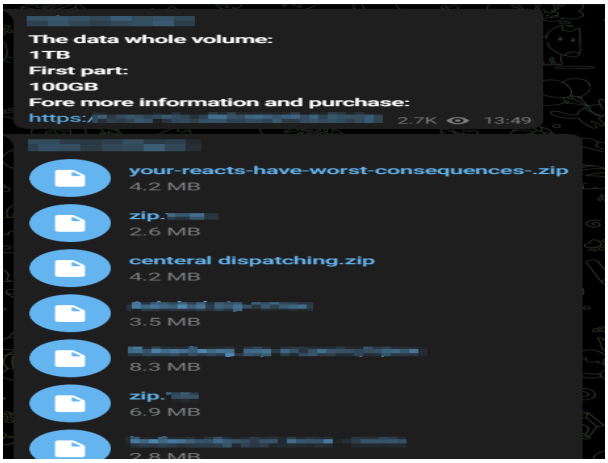

Figure 14: Example of data collected.

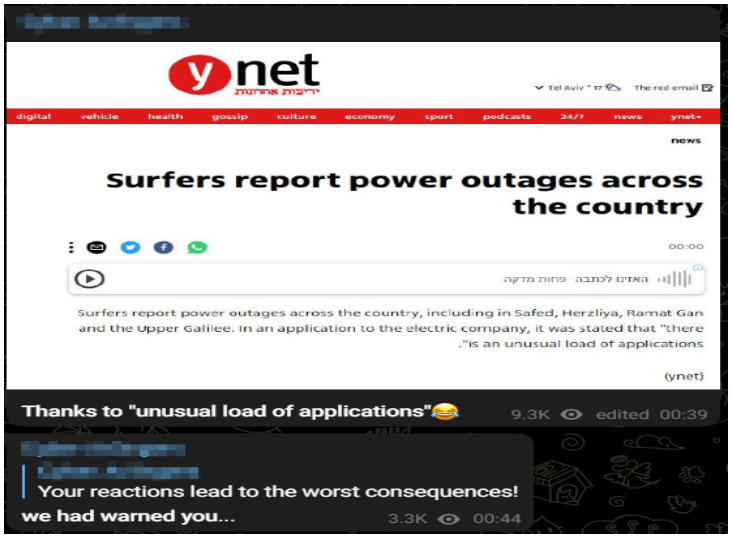

Figure 15: News report for the power outages.

`Boston Marathon Bombing` (2013) [63]: Following the Boston Marathon bombing, law enforcement authorities utilized social media platforms to aid in their investigations. OSINT analysts extensively searched networks such as Twitter to gather photographs, videos, and firsthand testimonies. This user-generated content aided authorities in identifying suspects and reconstructing the sequence of events preceding the attack.

`Legal Cases and Criminal Investigations`: The field of digital forensics. The utilization of social media data is progressively prevalent in legal proceedings. For example, in the context of criminal investigations or civil cases, OSINT practitioners gather and examine social media posts as digital evidence. Posts, geotags, timestamps, and conversations might offer crucial information into a suspect's behavior or an individual's psychological state.

`Arab Spring Uprisings` (2010-2012) [64]: Social media played a crucial role in organizing protests and spreading information throughout the Arab Spring. Protesters made full use of platforms such as Twitter and Facebook to coordinate demonstrations and exchange real-time information. OSINT specialists observed these platforms in order to comprehend feelings, monitor mobilization, and assess the political atmosphere in the afflicted regions.

`Ongoing Documentation of the Syrian Conflict`: OSINT researchers and journalists have widely utilized social media to record instances of human rights violations during the Syrian conflict. Platforms such as YouTube, Twitter, and Facebook have served as mediums for sharing films, images, and testimonials that reveal the extreme violence of the conflict, thereby assisting in the process of documenting and raising awareness through various campaigns.

The `#MeToo movement` [65] has facilitated the disclosure of sexual harassment and misconduct instances through social media, allowing victims to share their experiences. OSINT analysts collect and examine these testimonies, confirming accusations, discovering trends, and presenting evidence in court procedures.

`Influence Operations and Disinformation Campaigns`: Social media platforms have

played a crucial role in shaping public sentiment and disseminating false information. OSINT specialists monitor these campaigns during elections or geopolitical tensions, examining counterfeit accounts, bot behavior, and viral content to comprehend and counter these operations.

We will be presenting some additional real-life case studies to fully illustrate and describe the potential application of OSINT as well as its efficacy and potency in resolving intricate issues and revealing valuable information, allowing us to grasp the application in different sectors and situations.

— `Case Study 1: Missing Person`

In this instance, an individual inexplicably vanished under enigmatic conditions. Law enforcement agencies can use OSINT tools to collect information and track a person's location. This includes tracks, online presence, associations, and recent engagements of the missing individual. Since geolocation data can be acquired from check-ins posts, built-in maps the possibility of retrieving the probable areas visited by a person increases drastically. The solution to the problem where the person is found is a result of the combination of the aforementioned information with conventional investigation methods.

SUMMARY: Using SOCMINT and geolocation data can be highly effective in locating and identifying missing individuals.

— `Case Study 2: Business Intelligence for Competitive Analysis`

A corporation sought to get a competitive advantage by doing comprehensive Open Source Intelligence research on its industry competitors. Researchers can use web-based OSINT methods to track and analyse rival websites, social media accounts, consumer evaluations etc. Discovering information varies from marketing tactics to product range and customer satisfaction levels. This kind of knowledge enabled the organization to optimize their own strategy, pinpoint market deficiencies, and bolster their overall competitive standing.

SUMMARY: OSINT offers a variety of useful data that provide valuable insights regarding an organization's competition helping for better decision making.

— `Case Study 3: Investigative Journalism`

A diligent journalist employed open-source intelligence methods to expose malfeasance within a governmental organization. The journalist utilized a thorough examination of publicly accessible records, news stories, and social media postings to establish correlations between prominent individuals and questionable financial activities. The discoveries served as the foundation for a persuasive inquiry report, resulting in widespread recognition, inquiries, and eventual responsibility for the individuals implicated.

SUMMARY: OSINT methods enable journalists to conduct thorough investigations, unveil concealed information, and reveal illicit activities.

These case studies provide a limited view of the numerous real-world uses of OSINT in different fields. Through the proper utilization of open-source intelligence technologies and procedures, investigators, organizations, journalists, and people may discover significant insights, resolve intricate issues, and make well-informed judgments.

`Note: Please be aware that the case studies provided are fictional examples specifically designed for illustrative purposes.`

3.3 The Ethics of OSINT

Although OSINT is undeniably a valuable tool for gathering intelligence and improving decision-making, it raises important ethical concerns. The concept of privacy is being introduced first. Privacy is a crucial ethical consideration in OSINT, and it has been a highly contentious topic in recent times. During the collection and examination of publicly accessible data, individuals' sensitive and personal information may inadvertently get exposed. In order to tackle this issue, OSINT practitioners must adopt measures to safeguard individuals' privacy. These measures include censoring identifying information and ensuring that any acquired data is strictly used for its intended purpose.

3.3.1 Doxing and Swatting

Doxing, often known as "dropping documents", refers to the malicious act of collecting and publicly disclosing personal information about an individual [66]. The data may include the individual's name, address, phone number, email address, social media profiles, and other confidential information. These practices are employed by malevolent individuals or collectives with malicious intentions, such as cyberbullies, hackers, and trolls. Vigilante justice often serves as a common motive for doxing, which is frequently employed to intimidate, manipulate, or humiliate a target in order to force them to comply.

Hacktivism [67] is employed as a means to gain social or political benefits. Journalists reporting on far right controversies have been subjected to harassment due to the unauthorized disclosure of their personal information on the internet. Moreover, doxing might have adverse consequences on an individual. From both personal and professional spheres, individuals may experience blemishes in their lives caused by reputational injury, job loss, and other undesirable consequences. In extreme cases, doxing can even lead to fatalities.

The term "swatting" originates from the United States' Special Weapons and Tactics (SWAT) unit, which is sent to address high-risk and hazardous situations. The malevolent individual might contact emergency services and feign reporting a potentially perilous circumstance to the emergency operator, such as suicide prevention, an alleged family at risk from an invader, or the observation of an individual brandishing a weapon. The cops would then arrive equipped and prepared to neutralize a non-existent threat.

3.3.2 Biases

An important ethical concern that arises in the process of collecting information and drawing conclusions is the existence of bias. There are various mamay be unreliable due to human fallibility or intentional manipulation, with the aim of promoting specific ideologies or disseminating false information. The researcher must meticulously scrutinize every data point for any deficiencies or inaccuracies by cross-referencing sources or seeking corroborating evidence.

Personal bias can compromise manifestations of bias. Information obtained from internet or open sources the accuracy of a study, hence researchers must be cognizant of their own prejudices and strive to present facts in an impartial and objective manner. Additionally, it is vital for them to implement measures to guarantee the precision of the data they collect and remain vigilant regarding any potential biases in the sources they utilize.

3.3.3 Transparency

Transparency is another crucial ethical concern in OSINT. OSINT practitioners must openly disclose the procedures and sources they use to obtain and evaluate information. This entails delineating any potential conflicts of interest, biases, or limitations pertaining to the collected data. OSINT practitioners can establish the dependability and authenticity of their work, especially in investigations where precision and reliability are crucial, by openly disclosing the sources and techniques employed.

Furthermore, transparency enables others to verify the collected data. It is important for other researchers to be able to replicate the research and verify the accuracy of the results. Furthermore, when doing OSINT research and collecting information about individuals, it is imperative for researchers to be transparent and truthful regarding the sources they depend on, the objectives behind the data collection, and ensuring that no irrelevant material was analyzed during the study.

3.3.4 The Gray Area of Open Sources

The precise definition of "Open Source" is a subject that is open to debate. Typically, sources like publicly public records, personal social media accounts, and inadvertently disclosed or leaked documents are considered "Open". Occasionally, leaked documents may not include any sensitive information, although this may not be the case with documents labeled as "Internal Use Only" or "Private/Confidential". Downloading such files could potentially expose a researcher to sensitive information, including patient data, passwords, or financial records, which are not intended for public disclosure.

Furthermore, social media platforms divulge extensive information regarding an individual's identity, including potentially overlooked data stored within their personal profiles. Estimating the extent of personal information disclosed online can provide a challenge for individuals. OSINT experts can analyze the customary social media conduct of a user, including their posts, comments, likes, and shares, in order to gain insights into their behavior, preferences, interests, and relationships. In addition, they can utilize the user's social media profiles to track

events, observe trends, and ascertain their political ideologies. Hence, to maintain the privacy of individuals, it is imperative to adhere to ethical norms [68].

3.4 Best Practices for Optimizing OSINT Results

With these advanced OSINT techniques in play, it becomes crucial to assess their effectiveness. This evaluation involves several critical aspects:

DATA ACCURACY: Assessing the accuracy of data collected through these techniques is essential. The reliability of OSINT data directly impacts the quality of threat intelligence.

RELEVANCE TO THREAT PROFILES: Evaluating how well the gathered information aligns with specific threat profiles or indicators of compromise is vital. Relevant data is more likely to contribute to proactive threat detection.

SPEED & SCALABILITY: The speed at which these advanced OSINT techniques operate and their scalability in handling large volumes of data are crucial factors. In the fast-paced world of cybersecurity, timely intelligence is paramount.

In addition, we will explore key strategies and techniques that improve the quality of findings, promote ethical behavior, and maximize the effectiveness of research endeavors.

VERIFICATION OF THE SOURCE

Source verification is an essential stage in OSINT investigations to guarantee the precision and dependability of the information you collect. A brief presentation of optimal methods for verifying sources as follows:

— CROSS-REFERENCING

Verify the consistency and legitimacy of the material by cross-referencing it from several sources. Seek for supporting evidence or other viewpoints to bolster your conclusions.

— ASSESSING CREDIBILITY

Assess the credibility of the sources employed. Take into account variables such as the credibility of the website or platform, the proficiency of the author, and any possible prejudices that might impact the material.

— VERIFYING INFORMATION ACCURACY

Whenever feasible, corroborate the facts by using official or authoritative sources. This process guarantees the accuracy and currency of the data you gather.

DOCUMENTATION & RECORD-KEEPING

It is important to keep an adequate and detailed documentation during investigations to stay organized, manage references easily and ensuring compliance with legal requirements.

— COMPREHENSIVE DOCUMENTATION

Record comprehensive notes on your discoveries, encompassing the origin, date, and pertinent circumstances. This will aid in organizing and retrieving of the material gathered, ensuring efficient future reference.

— SCREENSHOTS & PRESERVATION

Take screenshots or keep copies of web sites, social media postings, or other online material that may undergo alterations or become inaccessible in the future. This action maintains the evidence and allows you to make future references to it, if necessary.

— SECURE STORAGE

Safely save gathered data and paperwork to ensure the confidentiality and accuracy of the information. It is advisable to utilize encrypted storage or password-protected systems in order to guarantee secrecy.

ETHICAL CONSIDERATIONS

Following ethical guidelines is of major importance for any OSINT investigation.

— PRIVACY & CONSENT Show regard for individuals' rights to privacy and acquire suitable consent while gathering or distributing personal information. Take care to adhere to the limits and legal obligations of the jurisdictions in which you conduct business.

— LAWFUL ACCESS Ensure that your Open Source Intelligence investigations adhere to relevant legal statutes and regulatory frameworks. Prevent unwanted intrusion or hacking endeavors, and gather information exclusively from publicly accessible sources.

— RESPONSIBLE USE Employ the knowledge you acquire in a responsible and ethical manner. Avoid disseminating inaccurate or deceptive information and utilize the data solely for valid investigation objectives.

`Note: Please ensure that all OSINT investigations are conducted in compliance with legal regulations and with due regard for the privacy rights of persons.`

4 Comparison

The incorporation of Artificial Intelligence into Open Source Intelligence and Cyber Threat Intelligence signifies a fundamental change in the way information is collected and security analysis is conducted. This chapter provides a thorough and analytical examination of the development of OSINT and CTI approaches, making important contrasts between the time before and after the introduction of AI.

4.1 Evolution of OSINT Techniques

Technology has advanced to overcome the two main obstacles—scale and speed—that prevent OSINT from being used to inform decisions that are vital to the mission. Before the advent of automated technologies, the volume of OSINT data grew exponentially, surpassing the capabilities of analysis tools and impeding the ability to generate timely insights that would enable one to proactively address threats. The capacity for real-time data analysis has been greatly enhanced as a result of artificial intelligence. By leveraging AI-enabled systems, intelligence analysts are now able to rely on OSINT to verify classified reports, identify potential threats, and determine which targets necessitate labor-intensive conventional intelligence collection. The progression of OSINT techniques demonstrates a consistent development, where conventional approaches relied on manual gathering and examination of data. Prior to the advent of AI, OSINT predominantly functioned within the boundaries of human capabilities, struggling with the overwhelming amount of information and the constraints of pattern identification. Nevertheless, the integration of AI has surpassed these limitations, allowing OSINT techniques to go beyond their previous boundaries.

Tools such as FOCA and Maltego, which formerly relied on human data correlation and enumeration, have now integrated artificial intelligence to make procedures more efficient. Artificial intelligence algorithms improve the efficiency and precision of FOCA in retrieving metadata, while Maltego's graphing capabilities now leverage predictive analytics, transforming the field of link research.

CRITERIA
DATA PROCESSING
PATTERN RECOGNITION
SPEED OF ANALYSIS
PREDICTIVE ANALYSIS

Table 5: Comparison criteria.

FOCA	MALTEGO	HARVESTER	METAGOOFIL
predictive	time consuming	manual	manual, limited
limited, predefined patterns	limited	limited	limited
moderate	varied response times	moderate	moderate
limited	not very efficient	limited	limited

Table 6: OSINT pre AI.

FOCA	MALTEGO	HARVESTER	METAGOOFIL
automated, efficient	automated, enhanced efficiency	AI-enhanced data gathering	predictive analytics enabled
advanced, adaptive	incorporates AI for dynamic	improved with AI	enhanced pattern recognition
real-time, rapid	swift real-time analysis	enhanced with AI	rapid analysis with AI support
proactive	AI-enabled	improved with AI	AI-driven

Table 7: OSINT post AI.

4.2 CTI Approaches

CTI is also a subject of continuous improvement and development as conventional approaches, which often react to incidents and depend on past information, encountered difficulties in effectively dealing with the ever-changing nature of cyber threats. This part explores the period before the advent of AI, emphasizing the constraints of rule-based systems and signature-based detection.

Metagoofil and Recon-ng, while efficient, were constrained by their responsive characteristics. The incorporation of AI has enabled these systems to actively analyze metadata, anticipate future dangers, and automate reconnaissance procedures, guaranteeing a more thorough and immediate collection of intelligence.

4.3 Integration of Artificial Intelligence in Open Source Intelligence and Cyber Threat Intelligence

The emergence of AI brought about a new era for both Open Source Intelligence (OSINT) and Cyber Threat Intelligence (CTI). This section outlines the ways in which machine learning algorithms and predictive analytics have become essential elements.

Reconnaissance tools such as The Harvester and IntelTechniques have integrated artificial intelligence (AI) to improve the gathering and analysis of data. The Harvester, currently enhanced with AI technology, carries out information retrieval with greater efficiency and precision. Meanwhile, IntelTechniques use AI to establish connections and analyze extensive datasets, resulting in a more sophisticated comprehension of threat landscapes.

CRITERIA
THREAT IDENTIFICATION
SIGNATURE-BASED DETECTION
INCIDENT RESPONSE TIME
TOOL PERFORMANCE

Table 8: Comparison criteria.

METAGOOFIL	RECON-NG	HARVESTER	INTELTECHNIQUES
reactive	reactive	reactive	reactive
predominant	predominant	predominant	predominant
variable	varied response time	varied response time	varied response time
focused on cataloging	focused on data gathering	focused on data gathering	diverse functionalities

Table 9: CTI pre AI.

METAGOOFIL	RECON-NG	HARVESTER	INTELTECHNIQUES
proactive, predictive	AI-driven, proactive	proactive	proactive
ML augmented	AI augmented	AI-enhanced	ML enhanced
reduced, real-time alerts	improved IR time	swift response, real-time alerts	AI enhanced
enabled	AI-optimized	AI-enhanced data gathering	AI comprehensive data analysis

Table 10: CTI post AI.

4.4 Ethical Considerations and Challenges

In addition to exploring the revolutionary potential of AI in OSINT and CTI, this chapter also examines the ethical considerations involved. The conversation encompasses issues related to privacy, biases in algorithms, and the necessity for transparent policies in the use of AI-powered intelligence tools.

PRIVACY CONCERNS

Data Collection and Surveillance

Issues related to the protection of personal information and confidentiality. The vast gathering of data inherent in AI-powered Open Source Intelligence and Cyber Threat Intelligence creates substantial privacy problems, especially when handling personal or sensitive information. The ethical concern of finding a balance between intelligence collecting and protecting individual privacy becomes crucial.

De-identification and Anonymization

Ensuring the ethical utilization of AI necessitates the implementation of strong de-identification and anonymization protocols. Insufficiently anonymizing data poses a threat to persons and can result in unforeseen repercussions.

BIAS AND FAIRNESS

Algorithmic Bias

The utilization of AI in Open Source Intelligence and Cyber Threat Intelligence carries the possibility for algorithmic bias, which can further reinforce and intensify biases that already exist in the training data. Addressing prejudice and guaranteeing fairness are crucial for the appropriate implementation of AI.

Fairness

It is crucial to prioritize the establishment of fairness in AI models to avoid unequal effects on different populations. Seeking fairness in intelligence results is not just a moral

obligation but also a necessary condition for dependable decision-making.

SECURITY RISKS

Adversarial Attacks

Due to the vulnerability of AI models to adversarial assaults, it is crucial to carefully assess security measures. Strong safeguards against alterations of input data are essential to preserve the integrity of intelligence results.

Due to the large amount of data handled in OSINT and CTI, it is necessary to implement enhanced data security measures. The significance of protecting against data breaches and illegal access is emphasized by ethical considerations.

RELIABILITY AND ACCOUNTABILITY

Transparency

Ensuring the openness of AI algorithms is essential for upholding accountability. For ethical implementation, it is important to clearly explain the methods used to make decisions, making sure that the reasons behind the intelligence outputs are understandable and can be justified.

Reliability

Reliability is a crucial need for the ethical implementation of AI. Ensuring the absence of incorrect positive or negative results is of utmost importance, considering the possible repercussions of inaccurate intelligence evaluations.

HUMANS FACTOR

Human supervision is essential for the ethical implementation of AI. Human knowledge is necessary for evaluating data, placing findings in perspective, and ensuring decisions are in line with ethical norms.

The ethical ramifications of AI-generated intelligence underscore the imperative of human engagement in the process of decision-making. The significance of a collaborative approach is emphasized by the need to ensure that decisions are in accordance with ethical ideals.

COMPLIANCE TO LAWS AND REGULATIONS

Complying with legal frameworks is a fundamental ethical consideration. Ensuring the proper incorporation of AI into OSINT and CTI necessitates thorough compliance with privacy laws, data protection rules, and other relevant legal frameworks.

DUAL-USE DILEMMA The dual-use dilemma refers to a situation where a certain technology or knowledge may be used for both beneficial and harmful purposes.

The dual-use problem highlights the potential danger of AI that is initially created for Open Source Intelligence and Cyber Threat Intelligence being redirected for harmful purposes. Responsible development and ongoing care are necessary due to ethical issues to prevent unwanted applications.

ENSURING THE ACCURACY AND CONSISTENCY OF DATA

The preservation of data integrity in the training of AI models is of utmost importance. Ethical issues emerge when data is manipulated, which can result in erroneous assessments and undermine the dependability of intelligence conclusions.

5 Conclusion

This dissertation explores the interconnectedness of Open Source Intelligence (OSINT), Threat Intelligence (TI), and Artificial Intelligence (AI) in the context of intelligence analysis. The session began by extensively examining the various aspects of OSINT in the Intelligence Cycle, recognizing the significant impact of AI. It concluded with a thoughtful examination of the ethical implications associated with these breakthroughs.

The investigation into OSINT techniques revealed not just the wide range of information sources but also the complexities of navigating this vast digital terrain. The practice of Red Team Reconnaissance has evolved into a sophisticated art form, involving a strategic dance that tests and improves intelligence procedures. Advanced Search Techniques, previously simple keystrokes, have developed into a complex and precise process, revealing the immense potential of semantic comprehension in the digital realm. The examination of the digital trail left by individuals and the unknown areas of the internet known as the Deep and Dark Web demonstrated the changing difficulties that intelligence experts encounter and the necessity of adjusting their approaches to the always changing environment of threats.

The resources available to intelligence professionals have extended beyond traditional limits. Social media, originally a medium for social interaction, has evolved into a valuable repository for information and knowledge. The inclusion of sites such as Twitter, Facebook, and LinkedIn, along with progress in sentiment analysis and natural language processing, has converted social media into a dynamic and immediate source of valuable information.

The comparative analysis explored the development of OSINT tactics, mapping their progression from conventional procedures to the advanced approaches required to address modern challenges. The convergence of OSINT, AI, and CTI addresses the demand for accuracy and efficiency in intelligence as the digital age progresses. The once-disparate pieces of the intelligence ecosystem now merge, providing a robust framework that not only satisfies the demands of the present but predicts the problems of the future.

However, in the midst of celebrating technological achievements, ethical concerns arise as the somber companions of advancement. The integration of AI with cognitive processes requires a purposeful and careful approach. The dissertation addressed the ethical challenges presented by the utilization of AI in Open Source Intelligence (OSINT) and Cyber Threat Intelligence (CTI). It emphasized the crucial requirement for transparent procedures, responsibility, and a nuanced comprehension of the intricate equilibrium between security priorities and individual liberties.

As we wrap up this exploration of the intersection of OSINT, AI, and CTI, the dissertation serves as both evidence of the existing state of intelligence approaches and a guiding light towards the future. To move forward successfully, it is necessary to possess not only advanced technological skills, but also a strong dedication to ethical behavior, an understanding of the changing nature of threats, and a willingness to work together with others in the modern intelligence field.

Essentially, the narrative presented here encompasses not just the evolution of intelligence, but also highlights the themes of adaptation, teamwork, and ethical awareness. As intelligence professionals venture into unknown realms of the future, may they be driven not just by technological prowess, but also by the sagacity that acknowledges the intricate balance between security imperatives and the principles that shape our societies.

References

[1] NATO. *OSINT Definition by NATO*. URL: https://nso.nato.int/natoterm/content/nato/pages/home.html.

[2] Richard Enbody Aditya K Sood. *Targeted Cyber Attacks, Multi-staged Attacks Driven by Exploits and Malware*. URL: https://doi.org/10.1016/B978-0-12-800604-7.00002-4.

[3] Rami Hijazi Nihad A. Hassan. *Open Source Intelligence Methods and Tools*. URL: https://doi.org/10.1007/978-1-4842-3213-2.

[4] Krishna Prasad Yogish Pai. *Open Source Intelligence and its Applications in Next Generation Cyber Security - A Literature Review*. URL: http://doi.org/10.5281/zenodo.5171580.

[5] Michael Bazzell. *Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information*. URL: ISBN1530508908.

[6] Central Intelligence Agency. *The Intelligence Cycle*. URL: https://www.cia.gov/spy-kids/parents-teachers/docs/Briefing-intelligence-cycle.pdf.

[7] Petroc Taylor. *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025*. URL: https://www.statista.com/statistics/871513/worldwide-data-created/.

[8] Sassi et. al Evangelista. *Systematic literature review to investigate the application of open source intelligence (osint) with artificial intelligence*. URL: https://doi.org/10.1080/19361610.2020.1761737.

[9] IBM. *Natural Language Processing*. URL: https://www.ibm.com/topics/natural-language-processing.

[10] IBM. *Named Entity Recognition*. URL: https://www.ibm.com/topics/named-entity-recognition.

[11] Nvidia. *Audio to Face*. URL: https://www.nvidia.com/en-us/omniverse/apps/audio2face/.

[12] MIT. *Speech to Face*. URL: https://speech2face.github.io.

[13] Aaron Roberts. *Cyber Threat Intelligence: The No-Nonsense Guide for CISOs and Security Managers*. URL: ISBN9781484272206.

[14] P. Nespoli Javier Pastor-Galindo. *The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends*. URL: https://ieeexplore.ieee.org/document/8954668.

[15] EU. *General Data Protection RegulationGDPR*. URL: https://gdpr-info.eu.

[16] US. *California Consumer Privacy Act*. URL: https://oag.ca.gov/privacy/ccpa.

[17] Fortinet. *Indicators of Compromise*. URL: https://www.fortinet.com/resources/cyberglossary/indicators-of-compromise.

[18] NIST. *NIST's Definition of Red Teaming*. URL: https://csrc.nist.gov/glossary/term/red_team.

[19] MITRE. *ATT&CK*. URL: https://attack.mitre.org/.

[20] Definition of. *Spear Phishing*. URL: https://www.imperva.com/learn/application-security/spear-phishing/.

[21] Google. *Dorking*. URL: https://www.exploit-db.com/google-hacking-database.

[22] Kaspersky. *Deep and Dark Web*. URL: https://www.kaspersky.com/resource-center/threats/deep-web.

[23] Python. *Beautiful Soup*. URL: https://pypi.org/project/beautifulsoup4/.

[24] *Scrapy*. URL: https://scrapy.org.

[25] *Octoparse*. URL: https://www.octoparse.com.

[26] *Hootsuite*. URL: https://www.hootsuite.com.

[27] *Brandwatch*. URL: https://www.brandwatch.com.

[28] *Talkwalker*. URL: https://www.talkwalker.com.

[29] *QGIS*. URL: https://qgis.org/en/site/.

[30] *ARCGIS*. URL: https://www.arcgis.com/index.html.

[31] *Google Earth*. URL: https://www.google.com/intl/el_ALL/earth/about/versions/#earth-pro.

[32] *Tor*. URL: https://www.torproject.org/.

[33] *Freenet*. URL: https://freenet.org.

[34] *Invisible Internet Project*. URL: https://geti2p.net/en/.

[35] *Tableau*. URL: https://www.tableau.com.

[36] *Power-BI*. URL: https://www.microsoft.com/en-us/power-platform/products/power-bi.

[37] *Pandas*. URL: https://pandas.pydata.org.

[38] *Matplotlib*. URL: https://matplotlib.org.

[39] *Seaborn*. URL: https://seaborn.pydata.org.

[40] *Jupyter*. URL: https://jupyter.org.

[41] *WHOIS*. URL: https://who.is.

[42] *DNSDumpster*. URL: https://dnsdumpster.com.

[43] *DNSlytics*. URL: https://dnslytics.com.

[44] *Maltego*. URL: https://www.maltego.com.

[45] *Exiftool*. URL: https://exiftool.org.

[46] *Metagoofil*. URL: https://www.kali.org/tools/metagoofil/.

[47] *Foca*. URL: https://github.com/ElevenPaths/FOCA.

[48] *Amass*. URL: https://github.com/owasp-amass/amass.

[49] *Gitleaks*. URL: https://github.com/gitleaks/gitleaks.

[50] *Spiderfoot*. URL: https://github.com/smicallef/spiderfoot.

[51] *LinkedInt*. URL: https://github.com/vysecurity/LinkedInt.

[52] *Recon-ng*. URL: https://github.com/lanmaster53/recon-ng.

[53] *harvester*. URL: https://www.kali.org/tools/theharvester/.

[54] *Shodan*. URL: https://www.shodan.io.

[55] *Censys*. URL: https://search.censys.io.

[56] *Fofa*. URL: https://en.fofa.info.

[57] *Leakix*. URL: https://leakix.net.

[58] *IntelTechniques*. URL: https://inteltechniques.com.

[59] Susnea Elena. *A Real-Time Social Media Monitoring System as an Open Source Intelligence (Osint) Platform for Early Warning in Crisis Situations*. URL: https://doi.org/10.1515/kbo-2018-0127.

[60] Pöhn Daniela Walkow Marcus. *Systematically Searching for Identity-Related Information in the Internet with OSINT Tools*. URL: https://doi.org/10.5220/0011644200003405.

[61] *Russo-Ukrainian War*. URL: https://en.wikipedia.org/wiki/Russo-Ukrainian_War.

[62] *Gaza–Israel conflict*. URL: https://en.wikipedia.org/wiki/Gaza%E2%80%93Israel_conflict.

[63] *Boston Marathon bombing*. URL: https://en.wikipedia.org/wiki/Boston_Marathon_bombing.

[64] *Arab Spring Uprisings*. URL: https://en.wikipedia.org/wiki/Arab_Spring.

[65] *MeToo movement*. URL: https://en.wikipedia.org/wiki/MeToo_movement.

[66] *Doxing*. URL: https://www.economist.com/the-economist-explains/2014/03/10/what-doxxing-is-and-why-it-matters.

[67] *Hacktivism*. URL: https://www.fortinet.com/resources/cyberglossary/what-is-hacktivism.

[68] Sweeney Eoghan. *Yes we can... but should we? OSINT Essentials*. URL: https://osintessentials.medium.com/osint-investigation-online-verification-some-thoughts-on-ethics-267a84418895.