Artificial Intelligence Review
https://doi.org/10.1007/s10462-023-10454-y

Open-source intelligence: a comprehensive review of the current state, applications and future perspectives in cyber security

Ashok Yadav[1] · Atul Kumar[2] · Vrijendra Singh[1]

Abstract

The volume of data generated by today's digitally connected world is enormous, and a significant portion of it is publicly available. These data sources are web archives, public databases, and social networks such as Facebook, Twitter, LinkedIn, Emails, Telegrams, etc. Open-source intelligence (OSINT) extracts information from a collection of publicly available and accessible data. OSINT can provide a solution to the challenges in extracting and gathering intelligence from various publicly available information and social networks. OSINT is currently expanding at an incredible rate, bringing new artificial intelligence-based approaches to address issues of national security, political campaign, the cyber industry, criminal profiling, and society, as well as cyber threats and crimes. In this paper, we have described the current state of OSINT tools/techniques and the state of the art for various applications of OSINT in cyber security. In addition, we have discussed the challenges and future directions to develop autonomous models. These models can provide solutions for different social network-based security, digital forensics, and cyber crime-based problems using various machine learning (ML), deep learning (DL) and artificial intelligence (AI) with OSINT.

Keywords Artificial intelligence · Open-source intelligence · Cyber security · Public data · Social networks · Deep learning

✉ Ashok Yadav
ashok.smsit@gmail.com

Atul Kumar
atul.kumar@dsci.in

Vrijendra Singh
vrij@iiita.ac.in

[1] Department of Information Technology, Indian Institute of Information Technology Allahabad, Prayagraj, Uttar Pradesh 211015, India

[2] Data Security Council of India, New Delhi 110025, India

Published online: 15 March 2023

1 Introduction

Open sources have existed for many years, but the explosion of the Internet and the World Wide Web (WWW) motivates several cyber security professionals and researchers to publish journals and articles on cyber threats, cyber-criminal profiling, and information gathering (Amaro et al. 2018). The current state of Artificial Intelligence (AI)/Machine Learning (ML) analysis in the context of open source is challenging. To effectively utilize publicly available unstructured data in cyber crime investigation, the researchers are developing methods for identifying, gathering, and organizing it. The term "OS" of OSINT stands for "Open Source"; it refers to the publicly accessible source from which the user obtains information for his intelligence. The term "Information" is a crucial component of OSINT, which is freely available information. You don't need to be a hacker to use OSINT in your daily life. Maybe you've already used OSINT but didn't realize it. All internet users use OSINT tactics in some way, such as in the online search for a firm, school, university, or individual. It does not matter where the information is stored and whether it is obtained from social networks, photographs, videos, blogs, newspapers, or tweets, as long as it is public, accessible, and legitimate. With the help of proper knowledge gained using OSINT, we can achieve significant competitive advantages such as profiling criminals, tracking employees of companies, and tracing organized crime. The availability of information and the methods for gathering it are continuously changing with time.

Previously, open-source focused on social events, open addresses, and interviews. However, today's information is on the Web, and methods for retrieving it are becoming far more complicated, innovative, and open to everybody. Actionable intelligence may be gathered from public and unclassified sources due to the proliferation of social networks and the real-time information exchanges accessible today. The significance of OSINT has accelerated a dispute about how intelligence from data should be gathered from multiple sources between the military, the government, and the commercial sector. Some challenges include relevant data acquisition, exploitation, and dissemination to address specific intelligence requirements (Nouh et al. 2019).

In recent years, several journals have published work on cybercrime, cyber security vulnerabilities, and the potential threats posed to individuals, businesses, and countries. Open source data has brought challenges in acquiring and maintaining the necessary skills to use the heterogeneous information and the electronic media available. Several tools and techniques are publicly available that can help cyber investigators to investigate cyber threats using OSINT (Revell et al. 2016). The various characteristics of cyber criminals can be easily collected using OSINT tools and techniques. These characteristics of cyber criminals help to profile the type of criminals (Edwards et al. 2022). OSINT is becoming a more important discipline in current criminal investigations as a technique for resolving several types of cases. LEAs and investigation agencies rely primarily on open data for information verification, evidence gathering, and intel generation for various investigations, including money laundering, fraud detection, human and arms trafficking, etc.

To facilitate crucial decision-making and security measures like sentiment analysis and riot control, disaster response, counter-extremism, misinformation measures, threat intelligence, incident response, blockchain and dark web monitoring, data leak detection and social engineering countermeasures, etc., a variety of government bodies, including military organizations and intelligence agencies, relying on the huge and up-to-date nature of open data.

The remaining of the paper is structured as follows. The need for OSINT review is described in Sect. 1.1. Where we can used the OSINT is introduced in Sect. 1.2. The most recent OSINT applications are introduced in Sect. 2. We covered OSINT workflow in Sect. 3. A thorough examination of the various OSINT tools and methodologies, as well as suggestions on how to select the right option for a particular task, are covered in Sect. 4. The discussion of the paper in which we discussed the present practices is described in Sect. 5 of the paper. The work is concluded in Sect. 6, which covers the challenges and future research directions. The contribution of the paper is graphically summarised as shown in Fig. 1.

1.1 Need for OSINT review

The amount of information available online can be helpful to individuals and businesses. Lack of awareness and knowledge can lead to potential harm by developing false beliefs about unreliable information, which raise information security assurance issues. Due to internet access and security vulnerabilities, it is common for the number of published research to expand rapidly before they are assessed. Some studies may give unclear, confusing, or conflicting results or be available in different languages. So based on the above information, the following research gaps are found, lack of clarity to generalize the system security of other countries, lack of cyber security and cyber defense problem solving using OSINT, Lack of ways to use OSINT in robust and automated models, lack of knowledge about proper selection of tools, techniques, and processes, based on data availability and target. Automatic and self-propelled cyber crime and cyber threats investigation model according to the requirement. Therefore a systematic review is necessary for OSINT applications from a perspective of cyber security, which provides the right and unbiased information.

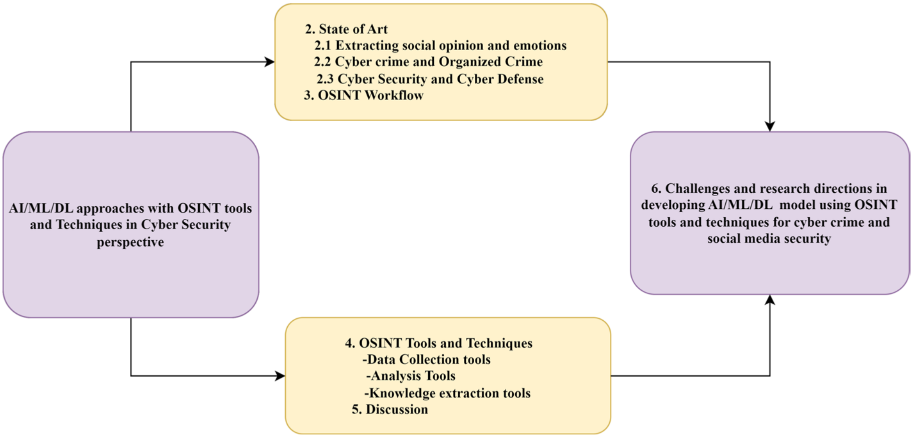
2. State of Art
2.1 Extracting social opinion and emotions
2.2 Cyber crime and Organized Crime
2.3 Cyber Security and Cyber Defense
3. OSINT Workflow
AI/ML/DL approaches with OSINT tools and Techniques in Cyber Security perspective
6. Challenges and research directions in developing AI/ML/DL model using OSINT tools and techniques for cyber crime and social media security
4. OSINT Tools and Techniques
-Data Collection tools
-Analysis Tools
-Knowledge extraction tools
5. Discussion

Fig. 1 Graphical overview of the paper

1.2 Where can open source intelligence be used?

Ethical hackers, penetration testers, and security experts identify potential flaws in the system using open-source tools and techniques. Defects commonly found include web-based life, open ports or unbound web-related gadgets, unpatched programming (e.g., websites running old customization of essential CMS elements), leaked resources, or unprotected unintended whole-code pasting containers for sensitive data such as resources. Some of the significant fields where OSINT is used are Law enforcement agencies (LEAs), Governmental bodies, and Corporate and cyber security.

1.2.1 LEAs

We have seen a reduction in HUMINT (Human Intelligence), for example, inspectors marching the streets and knocking on people's doors. OSINT use massively increased due to individuals spending so much of their lives online. This is especially helpful for criminal investigations such as cases of human trafficking to money laundering.

1.2.2 Governmental bodies

Military groups use open data to carry out efficient investigations and as counter-espionage. National security agencies use OSINT methods to detect threat groups like terror cells, maintain efficient incident response for riots and natural disasters, perform sentiment analysis, disprove rumors, and conduct many other civic responsibilities.

1.2.3 Corporate and cyber security

Companies are prioritizing their information security more due to significant financial losses caused by a single ransomware attack or compromise of corporate data. OSINT frameworks play a crucial role in threat intelligence, including penetration testing and incident response, but can also be used to counter persistent issues like social engineering-based attacks or even just irresponsible ways of data handling. Although OSINT strategies vary from case to case, there are fundamental methods for producing valuable intelligence throughout the data collection and processing phases. OSINT helps security experts to tackle cyber threats, such as identifying which vulnerabilities can be effectively exploited, blocking threats of imminent attacks, etc. Typically in cyber-investigating work, investigators need to distinguish and connect many focal points of information to identify cyber threats. For example, a single compromising tweet may not be a source of concern, but if it is linked to a dangerous cluster known to be dynamic in a particular industry, the associated tweet may be viewed with a suspicious perspective. One of the most important things to know about OSINT is its use in conjunction with other intelligence subtypes. Information extraction from closed sources, such as internal telemetry, dark Web networks, and shared external intelligence networks are typically used to verify and validate OSINT. The various applications of OSINT are shown in Fig. 2.

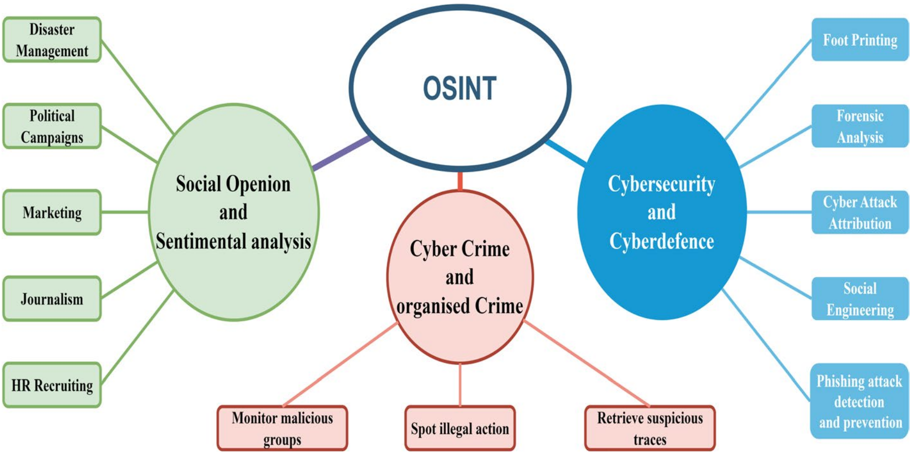

Fig. 2 Applications of OSINT (Khera 2021)

2 State of art

This paper has tried to incorporate the findings from previous research related to OSINT (open-source intelligence) tools and techniques. To further understand and enhance progress in OSINT research, we have formulated five Research Questions (RQs) based on a comprehensive review of OSINT tools, techniques, and their applications. The main objective of these questions is to enhance our understanding of OSINT and its utilization.

1. **RQ1** What are the various categories of OSINT and how we can utilize them? In previous studies, OSINT is used only for the purpose of translation, exploitation, analysis, and dissemination but due to the advent of technologies, it can be used in various applications. OSINT is categorized according to applications such as Geospatial intelligence, Signal Intelligence, Imagery Intelligence, Human intelligence, and Social media Intelligence. The details of these categories are discussed in Table 1.
2. **RQ2** What are the main strengths and weaknesses of different OSINT tools and techniques? This question's actual aim was to identify the strengths and weaknesses of various OSINT tools and techniques from several aspects such as publicly available, application, processing time, input type, and reliability score. The details of tools and techniques are described in tables of Sect. 4.
3. **RQ3** what are the benchmark research works available that integrate OSINT with different fields to maximize the utilization of OSINT? This question mainly focussed on the sources of research work published and how much it is relatable to OSINT and these works are being regularly cited and preferred in previous studies. The details are described in Sect. 2.
4. **RQ4** Is there any mechanism or workflow to utilize the OSINT to exploit publically available data? There are some research work, government guidelines, and reports published in which they explored the OSINT workflow. This workflow is quite helpful to exploit the OSINT tools and techniques according to requirements. Our aim is to explore the different relevant workflows on the basis of their characteristics and applications. The general workflow is shown in Fig. 5

5. **RQ5** How we can utilize OSINT tools and techniques in social network analysis, opinion extraction, cyber security, counter-terrorism, cyber defense, cybercrime investigation, criminal profiling, surveillance, etc. There are some very good research articles, letters, magazines, and blogs available that exhaustively exploited OSINT tools and techniques. It signifies the applications of OSINT in various domains are quite useful and suitable for real-life based scenarios (like tracing criminals, counter-terrorism, tracking of ships/ Flights, and surveillance). Therefore the main aim of this question is to identify the framework which integrates OSINT with Machine Learning/Deep Learning/Artificial Intelligence that can give better results. This question also focuses on the model result without OSINT so we can easily compare them by including them with OSINT. The all above points we have addressed in Sects. 2 and 4.

2.1 Research process

We have used various keywords for state of art such as OSINT, OSINT tools and techniques, cybercrime and organized crime detection using OSINT, counter cyber criminals using OSINT, Threat intelligence using OSINT, application of OSINT in cyber defense, security intelligence, disaster management, social opinion extraction, sentiment analysis, malware analysis, vulnerability assessment, national security surveillance, and counter misinformation. Using these keywords we have searched the following electronic databases like google scholar, ACM digital libraries, Web of Sciences, science direct, and IEEE Xplore. We also searched different tools and techniques of OSINT, different magazines related to OSINT, blogs of OSINT, and newsletters of OSINT. We have included those works which are related to OSINT, applications of OSINT, and possible domains to integrate OSINT. Other characteristics are citation score, reputed journals/publications, and real-life-based scenarios. Figure 3 shows the analysis of resources accessed and used to study the OSINT trends in cyber security using various AI/ML/DL techniques.

2.2 Extracting social opinion and emotions

Aye and Aung (2020) proposed a model that used adjectives, negations and intensifiers of text to determine the user's opinion on a given text. Their model is only for the Myanmar language. Kandias et al. (2017) the author has done an experiment on 450 users of Facebook to determine the level of stress. Yadu and Shukla (2020) use five different

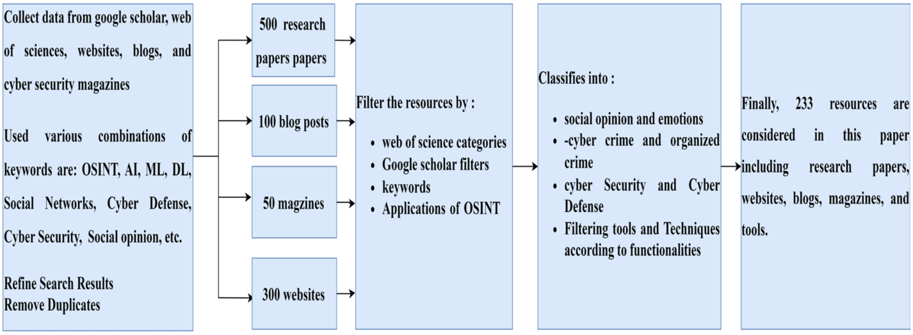
Collect data from google scholar, web of sciences, websites, blogs, and cyber security magazines
Used various combinations of keywords are: OSINT, AI, ML, DL, Social Networks, Cyber Defense, Cyber Security, Social opinion, etc.
Refine Search Results
Remove Duplicates
500 research papers papers
100 blog posts
50 magzines
300 websites
Filter the resources by :
• web of science categories
• Google scholar filters
• keywords
• Applications of OSINT
Classifies into :
• social opinion and emotions
• -cyber crime and organized crime
• cyber Security and Cyber Defense
• Filtering tools and Techniques according to functionalities
Finally, 233 resources are considered in this paper including research papers, websites, blogs, magazines, and tools.

Fig. 3 Overview of resources used for the analysis

classification techniques to analyze the emotions contained in tweets from the Indian Air Asia Service. Prabhakar et al. (2019) used AdaBoost (Ensemble), which also integrates other classifiers for building a robust classifier. The precision of the model is 84.5%. According to Wadawadagi and Pagi (2020) the use of the deep neural network can be elongate to other tasks of sentiment analysis. These analysis tasks include partnership categorization, computer translation, query response, and subject recognition. Other tasks such as polarity detection and opinion symbols are also included in sentiment analysis. Hashida et al. (2018) proposed a model based on distributed multichannel representation to enable a hybrid interpretation of text data.

Naseem et al. (2019) proposed a model based on BiDirectional short-term memory (BiLSTM) and hybrid word representation, which increases the accuracy of the model. They are considering terms out of vocabulary (OOV), grammar, polysemy, syntax, and word feelings in airline sentimental research tweets. Soomro et al. (2020) proposed a model in which they analysed more than 18 million tweets. All tweets are associated with the novel coronavirus. Tweets have been analyzed to see a relationship between the number of coronavirus infections and public mood. Due to this the number of cases is increasing or decreasing. The author of the paper (Abdul-Mageed and Diab 2014) proposed a model in which sentiment analysis was used mainly to understand arabic sentiments through YouTube comments and Twitter tweets. Garcia and Berton (2021) proposed a model using Twitter data to perform a sentiment analysis in Portuguese. Their aim was to analyze the consequences of the pandemic in two geographic locations, the United States and Brazil.

Mishra et al. (2019) proposed a model that analyzed the sentiments of various reviews. Using model results they created a hotel recommendation system. Using ML methods, Jain and Dandannavar (2016) studied different steps to accomplish sentiment analysis using Twitter data. They collected data from Twitter and used Natural Language Processing (NLP) for preprocessing. Then to obtain features related to the sentiment, they performed feature extraction. They used classifiers such as decision tree (DT), support vector machine (SVM), and naive bayes (NB) to train a model. Shuai et al. (2018) proposed a model with the help of Chinese hotel reviews. They also used Doc2Vec to organize the data into balanced negative and positive emotions. For training of the Doc2Vec model, data are integrated and assessed using various classifiers such as SVM, logistic regression (LR), and NB. According to the findings of his work, with an F measure score of 81.16%, the SVM gave the best result compared to the other classifiers. Based on the state of art of extracting social opinion and emotions we analysed that extracting opinions and emotions is important because they can cause damage to individual, community, and country. In most of the cases the data set is collected randomly and no standard data set is available. So if we use the OSINT tools and techniques with AI/ML/DL then it enhanced the results.

2.3 Cyber crime and organized crime

Digital forensics and open-source intelligence are the two broad types of cyber crime investigations. Cyber crime is divided into the following categories by the Council of europe convention: crime against the CIA (confidentiality, integrity, and availability), content crime, computer crime, and other types of cyber crime. We have considered terms like computer crime, technological crime, high-tech crime, digital crime, and electronic crime to describe cyber crime. Human trafficking, pornography, child pornography, assassination, drug sales, terrorist activity, cybercrime markets, and cryptocurrency exchange are among the eight primary cyber crimes highlighted by Nazah et al. (2020). According to the information, the choice of tools and techniques has a greater impact on cyber investigations. We found that no single tool or procedure can collect all the evidence that investigators require due to which they use different combinations of tools and techniques to conduct a cyber-crime investigation. Quick and Choo (2018) proposed an framework based on OSINT that increases the accuracy of arresting the culprit and applied OSINT in digital forensics to improve the analysis of criminal intelligence. Delavallade et al. (2017) proposed a model that is totally based on social network data. The model extracts the crime indicator from different social networks and predicts future crimes.

Shestak and Koscheeva (2021) accurately described crimes that belong to the category of cybercrimes, how the coronavirus pandemic led to an increase in cybercrime rates, and their possible mitigation procedures. A framework has been implemented for the analysis of incidents using related tweets, number of retweets, hashtags, and Uniform Resource Locator (URL) connections. The geo-location, expert direction text position, and tweet time were extracted from the phrase using the n-gram method, which has an accuracy rate of 80% (Roddy and Holt 2022; Martin et al. 2020; Sinha et al. 2022; Kakkar 2020). Malhotra et al. (2021) proposed a model to assess whether a tweet is related to Crime Hub events. They used feature depiction, tokenization, stop, trailing, and TF-IDF extraction for classification. A study was conducted to keep track of the frequent crime hub in the Indian States. Hub Watch can be used to collect information on potential risks, fraud, and the supply of public transportation. Three classifiers were used for decision planning namely Bayes Network (BN), KNearest Neighbors (kNN), and DT.

The authors proposed a fusion architecture to help investigators identify and recreate crime scenes. Each stage combines logical procedures with abstract principles. To ensure that LEAs can enhance the effectiveness and efficiency of an investigation, some open-source tools of each phase in the proposed fusion architecture are shown in Fig. 4. Crime

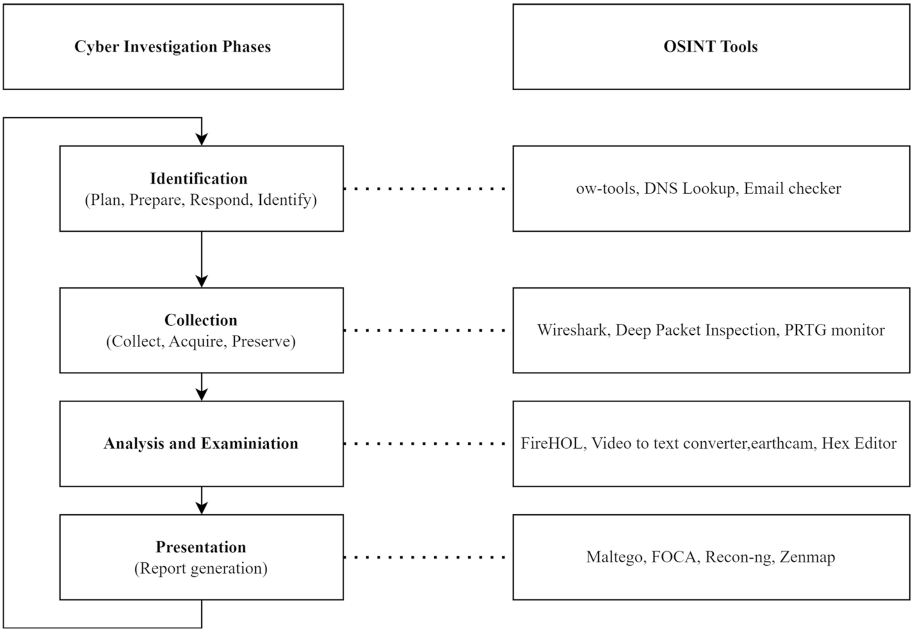
Cyber Investigation Phases
OSINT Tools
Identification
(Plan, Prepare, Respond, Identify)
ow-tools, DNS Lookup, Email checker
Collection
(Collect, Acquire, Preserve)
Wireshark, Deep Packet Inspection, PRTG monitor
Analysis and Examiniation
FireHOL, Video to text converter,earthcam, Hex Editor
Presentation
(Report generation)
Maltego, FOCA, Recon-ng, Zenmap

Fig. 4 cyber investigation with OSINT tools (Kao et al. 2018)

Hub data are classified into point division and link division. This approach properly categorised 467 tweets into link division and 3168 tweets from Crime Hub into point division with an accuracy of 94.66% and 76.85%, respectively (Valluripally et al. 2019). Kadoguchi et al. (2020) research led to the development of a device. This device can be used for real-time tracking of various crime hub locations across the Indian States. They used various ML classifier for classification Crime Hub fraud, SVM has the highest precision among all classification methods that is 97.28%. Statistics on the approaches to the attacker's target have been displayed as a strategy to assist hackers in avoiding cyber crime hub, incident, and hub maintenance (Roddy and Holt 2022).

The system was given the term (Intelligent Transportation System) by the researcher. Detection and tokenization methods based on ontologies have been used to preprocess Twitter data as textual context information to identify a real-time crime hub. Then, the value was quantified with IDF extraction and then classified. The SVM method offers the maximum precision, according to the results. It has a 91.1% accuracy rate when looking at the current state of crime hub delays and 86.3 % of the Crime Hub jam. Crawling the dark web can be useful for investigators as it can help to learn about new risks, new drug varieties, and new vendors (Celestini et al. 2017). Five methods for data mining are described by Edwards et al. (2015): natural language processing (NLP), information extraction, social network analysis (SNA), computer vision (CV), and ML. These methods use technologies to scrape data from an internet source about a criminal or terrorist group's links. Liao et al. (2016) proposed iACE, a technology that can be used to automatically gather intelligence from numerous sources and analyse data relationships. AI may be trained to search for patterns that signal criminal activity in forensic data, such as network traffic. Phishing is a social engineering attack as stated by Krombholz et al. (2015) and Ivaturi and Janczewski (2011). Both Pienta et al. (2018) and Gupta et al. (2017) suggested a phishing assault taxonomy, but did not provide sufficient analysis of email phishing attacks.

The experiments of Albladi and Weir (2017) and Halevi et al. (2013) predicted that personality trait has a direct influence on a user's ability to detect phishing email attacks but due to inconsistencies in their results they developed a hypotheses model that stated, that user detector ability is influenced by personality traits, trust, competence, and motivation. Tandale and Pawar (2020) has provided an overview of several forms of phishing attacks and detection methods. Additionally, they presented several different phishing mitigation techniques. They concluded that among all existing techniques of anti-phishing using ML 100% accuracy can be achieved in detecting phishing. Alabdan (2020) provided a review and complete evaluation of recently phishing attack strategies for awareness of phishing techniques and about the different types of attacks. Rastenis et al. (2020) presented an e-mail-based phishing attack classification model that addresses all the shortcomings of the previous phishing attack classification model. Kathrine et al. (2019) discussed many phishing attacks and the most recent preventive methods. Their study demonstrates how to detect and distinguish phishing attack using ML algorithms.

Kunju et al. (2019) surveyed various methods to detect phishing attacks. Their survey demonstrated various ideas and techniques to detect attacks. They also state that several of the methodologies offered are ineffective in providing effective attack solutions. Aleroud and Zhou (2017) presented the taxonomy of phishing strategies, as well as their vectors and countermeasures. The paper highlighted the vulnerabilities frequently exploited, and the taxonomy provides guidelines for developing various successful and efficient anti-phishing techniques. Cui et al. (2017) proposed a method to calculate the number of HTML tags utilized in DOM attacks. They used clustering for creating clusters of attacks that take place within a specified range of distance and stated that these clusters can be aggregated and used for detecting the phishing attacks. Their findings revealed that his strategy can detect a substantial number of new phishing attacks.

Wang et al. (2020) proposed a phishing avoidance approach. On an Android smartphone, they installed an optical character recognition system. To verify the effectiveness of the suggested preventive technique they tested hijacking attacks. They stated that their suggested OCR approach overcomes the drawbacks and constraints of existing solutions and is effective enough to detect phishing websites. Churi et al. (2017) proposed a prototype model for determining a website is a phishing site. They claimed in their research that current phishing protection frameworks are ineffective for the detection of phishing sites. They used combination of visual cryptography and code-generating techniques. Their suggested methodology creates an image, divides it into two parts using visual cryptography, and then combines these two shares to create an image captcha. In order to identify the genuine site from phishing sites, the user is requested to match the site with the image captcha. Stafford (2020) explored the effects and causes of phishing attacks and also how users are affected by attacks. Humans are susceptible to phishing because of their characteristics and behaviors such as narcissism, susceptibility, and frequent email use. The results show that spear phishing is the most targeted phishing technique.

2.4 Cyber security and cyber defense

Senekal and Kotzé (2019) used NLP to analyze WhatsApp chats for a remarkable investigation, which estimates the large-scale vandalism of South Africa. Some researchers observed the use of OSINT in criminal investigations that involve coordinated misbehavior and cybercrime (Kao et al. 2018). Currently, the availability of online services is more, which routes the growth of a large amount of digital information (Herrera-Cubides et al. 2020). AlKilani and Qusef (2021) for additional security, proposed integration between some OSINT techniques and the corresponding clauses of the ISO 27001 standard. In the context of integrating them with the global standard ISO 27001, they presented a collection of OSINT tools/techniques that address hiring screening, background checks, and vendor risk assessment.

Raj and Meel (2022) proposed a model of fake news detection. In their model, they used two real world data sets Medival202 (Pogorelov et al. 2020) and CovidHeRA (Dharawat et al. 2020). They classify the tweets into two classes that is real and fake. To identify fake and real news, they used the following features: gender, media usage, sentiment polarity, follower count, friends count, status counts, retweet counts and favorite counts. Edwards et al. (2017) proposed a classification model in which they classify the person of the organization as an employee and a non-employee, with the help of which we can find vulnerability and also prevent attacks on the organization by social engineering. The classifier used for classification is DT. They used the sub-classifiers such as 'Name subclassifier', 'Activity subclassifier', 'Writing fingerprint subclassifier', 'Link analysis subclassifier', 'Friend subclassifier', and 'Geographic subclassifier'. Yuan et al. (2021) proposed a model named Domain Adversial and Graph Attention-Neural Network for detection of fake news. Where two real-world datasets "2015 MediaEval twitter dataset (Boididou et al. 2015) and Weibo dataset" (Hu et al. 2020) were used for training and testing purposes. In the presented model, a domain discriminator, multi-modal data features extractor, and a graph-attention-base fake news classifier are integrated. To extract textual features, Bi-LSTM and image features, pre-trained VGG-19 model were used.

Cinelli et al. (2022) investigated the coordinated and non-coordinated Twitter account by considering the 2019 UK political election. For data collection they used twitter API by considering the election-related hastag, influential political accounts, parties offical accounts, and political leader accounts. Islam et al. (2022) proposed a validation tool using cyber threat intelligence to automates the validation of security alerts and incidents on the basis of security operation centres. They collected 'Indicators of compromise' using OSINT form public websites and MISP (MISP F 2021). Ch et al. (2020) proposed a model to analyse the rate of cyber crimes state-wise. They use various ML techniques to classify the cyber crime. NB and K-means is used for classification and clustering respectively. They used various attributes for classification like a victim, incident, offender, age of the offender, harm, year, location, and cybercrime.

Ganesan and Mayilvahanan (2017) proposed a methodology that helps to identify the unpredicted patterns. They collected the data from various web pages and databases that helps to classify the cyber crimes. The classes are of cyber crimes are cyberbullying, identity theft, scams, stalking, robbery, harassment and defamation. A multi-functional cybercrime intelligent system framework was proposed by Nouh et al. (2016). The main objective of this framework was to reduce the use of cognitive biases throughout the investigation process. This technique provides six main steps: identifying the issue, developing a hypothesis, gathering data, assessing the hypothesis, selecting a related hypothesis, and continuously monitoring incidents. Aslan et al. (2018), proposed a methodology to detect the social networks accounts (like twitter accounts) which are related to cyber security. They used various ML techniques such as Random Forest, DT and SVM etc to automatically detect the suspicious accounts. They also used some behavioural features that are extracted from the collected tweets to identify the suspicious accounts.

Abbass et al. (2020) proposed a methodology that helps to predict the various social networks based cyber crimes such as cyberbullying, cyber harassment, cyber stalking, cyber hacking and cyber scam etc. They used the data gathered from various social networks websites. They used Multinomial Naïve Bayes (MNB), KNN, and SVM for classification of cyber crimes into different classes. Kumar et al. (2020) proposed a methodology in which they used the type of crime data, time, and location for predicting crime in certain regions in India. They used KNN for prediction and crime which are predicted by this methods are robbery, accident, violence, gambling, murder, and kidnapping. The demographic and geographic information from past years' incidents have been utilized to predict the terrorist activities in India. They used the AI to predict the terrorist incidents (Verma et al. 2019). Carloni (2014) proposed a methodology based on various ML techniques to detect and predict the cyber attacks. They used previous cyber crime data to train and test the model.

3 OSINT workflow

The OSINT workflow involves the following phases such as collection, processing, analysis, knowledge extraction, dissemination, and planning as shown in Fig. 5.

The collection is a phase in which we collect data from open sources according to the objective (Herrera-Cubides et al. 2020). In this phase, we collect data from social media, websites, forums, reports (NGO, government, court, and law enforcement), articles (academics research, journalists), media (news, interviews, video, and audio recordings), booklets, books, etc. In the processing phase, the collected data is processed according to the

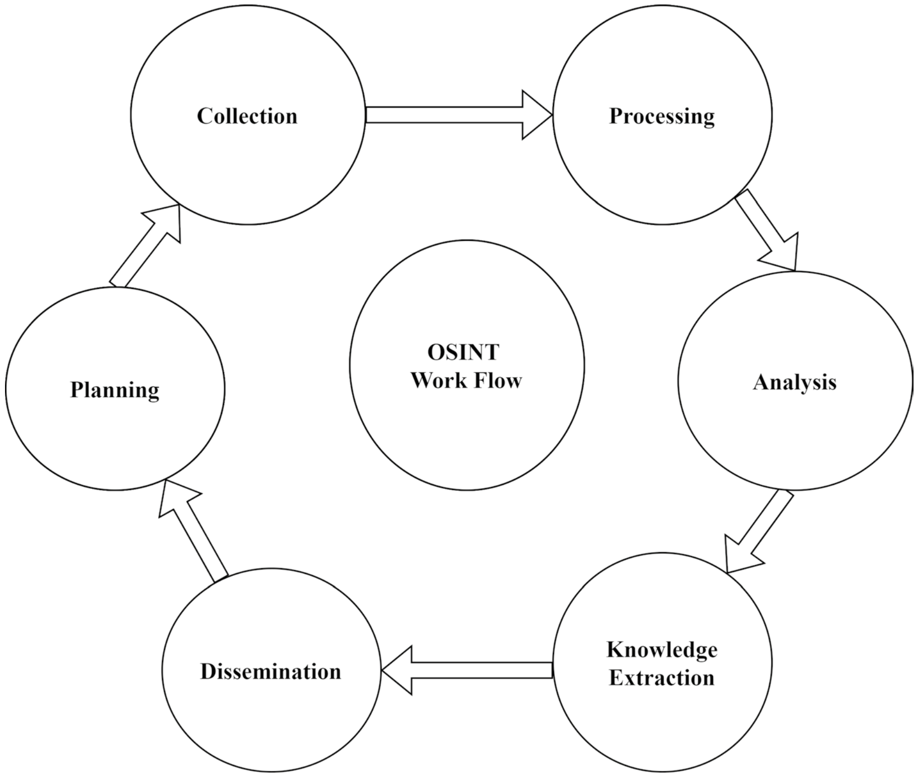

Fig. 5 OSINT workflow (Akhgar et al. 2017)

objective and synthesized in such a way that can be easily understandable. The data will be categorized into relevant and non-relevant. Additionally, it checks the reliability of the sources from where the data is collected. In this phase, we also do a translation if required such as if we get data in a different format/language then we translate it into the required format/language.

In the analysis phase, we evaluate the processed or relevant data. We do lexical analysis (Ghazi et al. 2018), SNA (Stieglitz et al. 2018) and Geo-spatial analysis (VoPham et al. 2018) etc. The lexical analysis involves aggregation and analysis of text data collected from open sources. We also analyze the most frequently used terms, social media accounts of people, and their demographic characteristics. In SNA, we examine the social media accounts, connections, interest areas, etc. This analysis also analyzes the connected network of particular users/communities and their objectives. In Geo-spatial analysis, we find the location of targets or groups of targets using different open-source tools and analyze the location coordinates. In the Knowledge extraction phase, we do the extraction of relevant information from the collected data to fulfill the objective of the task.

Dissemination is a process that helps policymakers, students, researchers, law enforcement, governments, organizations, etc. Dissemination is done from an educational and research point of view. For education purposes, we educate the person about their data privacy and security to handle their data from cybercrime or other security issues. In research cases, dissemination helps to improve the methods, tactics, and techniques to prevent the data for digital forensics. Additionally, it helps to utilize open-source data for investigation. The plan is prepared for a particular task by analyzing and processing the other stages of OSINT workflow in the planning phase.

4 OSINT tools and techniques

OSINT has some generic categories such as Geo-Spatial intelligence (GEOINT), Human Intelligence (HUMINT), Signal intelligence (SIGINT), Imagery Intelligence (IMINT), and Social Media intelligence (SOCMINT). The advantage, disadvantages, and applications of these categories are listed in Table 1 (Omand et al. 2012; Williams and Blum 2018).

The data can also be physically assembled, but this is time-consuming and easier to use at a later stage. The tool facilitates the selection phase by allowing you to collect information from multiple targets in minutes. It's your job to detect if a username is available and assume it's on all social media sites. One way is to log in to all internet-based life sites (perhaps you don't have the niftiest idea about them all!) And test your username on those sites. Another option is to use an open-source device that links to different websites you can easily remember and checks usernames on each website without delay and it takes minimal time. Operate multiple devices to gather all data related to the target then combine and analyze it. The general attack surface for utilizing the different OSINT-based tools and techniques is shown in Fig. 6.

Historical data is generally used to verify or investigate a particular incident. Several archives are available of historical data, such as the collection of newspapers, maps, books, manuscripts, science information, birth records, marriage license and death records, etc. Some historical data websites are listed in Table 2.

With the help of historical data, investigation and verification become easier. OSINT tools and techniques can automate historical data extraction and information verification.

Official data leak repositories provide us the leaked data such as spying and correction data, company information, and restricted official materials. These data can be collected using the official leak repositories. Some of them are listed in Table 3.

These data help in the verification and investigation of crime as well as in training the DL-based models. To access various online services, people use a unique username. Using the username, we can easily collect the data related to the particular user. Table 4 explores some username tools and techniques. These tools help to find the account on social networks., like lullar is an OSINT tool that automatically generates the URLs of user-profiles of different social networks.

The real name is also used to collect data about the target. Some tools and techniques based on real names are listed in Table 5. Using these tools, based on the real name, we can collect other information such as email address, location, social media accounts, images, phone numbers, and age.

Email investigation examines an email's header and body for information about the sender and recipient. In header analysis, we can obtain information such as Received by, X-Received, Return path, Received-SPF and authentication, etc., received by contains IP address of the server, SMTP-id and date and time at which the email was received. X-Received is a parameter not defined in the official protocol standard of the internet. They are created by the mail transfer agent and contain the same as the Received-by parameter. The return path includes the IP address of the server, the receiver email

Table 1 Advantages, disadvantages and applications of generic categories of OSINT

Categories of OSINT	Advantages	Disadvantages	Applications
GEOINT	Situational awareness, Timely and accurate information, Multi-source integration, Supporting a wide range of activities, Identifying patterns, Supporting location-based services	Limited accessibility, Cost, Data quality variation, Lack of standardization, Limited scalability, Limited accuracy, Limited resolution, Limited data visualization	Military and defense, Homeland security, Disaster response, Environmental monitoring, Law enforcement, maritime security, and safety
HUMINT	Identify and assess potential threats, including terrorist groups, criminal organizations, and foreign governments, used to develop and implement effective counterintelligence and counter-terrorism strategies, used to identify and track the movement of weapons, money	Unreliable and prone to bias, non verifiable, time-consuming and resource-intensive, prone to cultural, language and social barriers, vulnerable to counterintelligence, the chance of human error, affected by legal and ethical considerations	Counterintelligence operations, counter Terrorism, Military Operations, gather intelligence on cyber threats, Law Enforcement, used to gather intelligence on potential public safety threats, helping to identify patterns, trends, and potential threats
SIGINT	Real-time intelligence, Long-range surveillance, Non-intrusive, ersatility, Cyber warfare, Identifying vulnerabilities, Supporting military operations, Cost-effective	Limited accuracy, Privacy concerns, Dependence on technology, Complexity, Limited geographical coverage, Ethical considerations, Cyber security risks	Military operations, Counter-terrorism, Cyber defense, Law enforcement
IMINT	Real-time information, Cost-effective, support other types of intelligence (such as SIGINT, HUMINT), Non-intrusive, ability to provide historical data and long-term analysis, provide high-resolution images easily	Weather-dependent, Limited penetration (dense forests or underground structures), Expensive, Privacy and ethical concerns, Reliance on technology, Limited time-frame, vulnerable to electronic countermeasures (jamming or spoofing), vulnerable to hacking, malware, and other forms of cyber threats	Used to support military operations (such as reconnaissance, targeting, and battle damage assessment), can be used to detect and track potential threats such as smuggling, illegal immigration, and terrorism (Border and coastel security), Surveillance and reconnaissance
SOCMINT	Real-time information, Cost-effective, Large data sets, Unfiltered information, Targeted information, Wide reach, Public opinion, Identifying trends and patterns, Identifying potential crises, Identifying influencers, Supporting investigations	Privacy concerns, Data Quality, Information overload, Dependence on technology, Limited to online data, Legal and ethical considerations, Lack of standardization, Difficulty in determining the credibility of sources, Limited to specific languages	Gather intelligence and identify criminal activity on social media platforms, Political campaigns, potential fraud, such as phishing scams or identity theft, cyber threats, vulnerabilities and exploitations, Government intelligence, Criminal investigations, Fraud detection, Cyberbullying detection

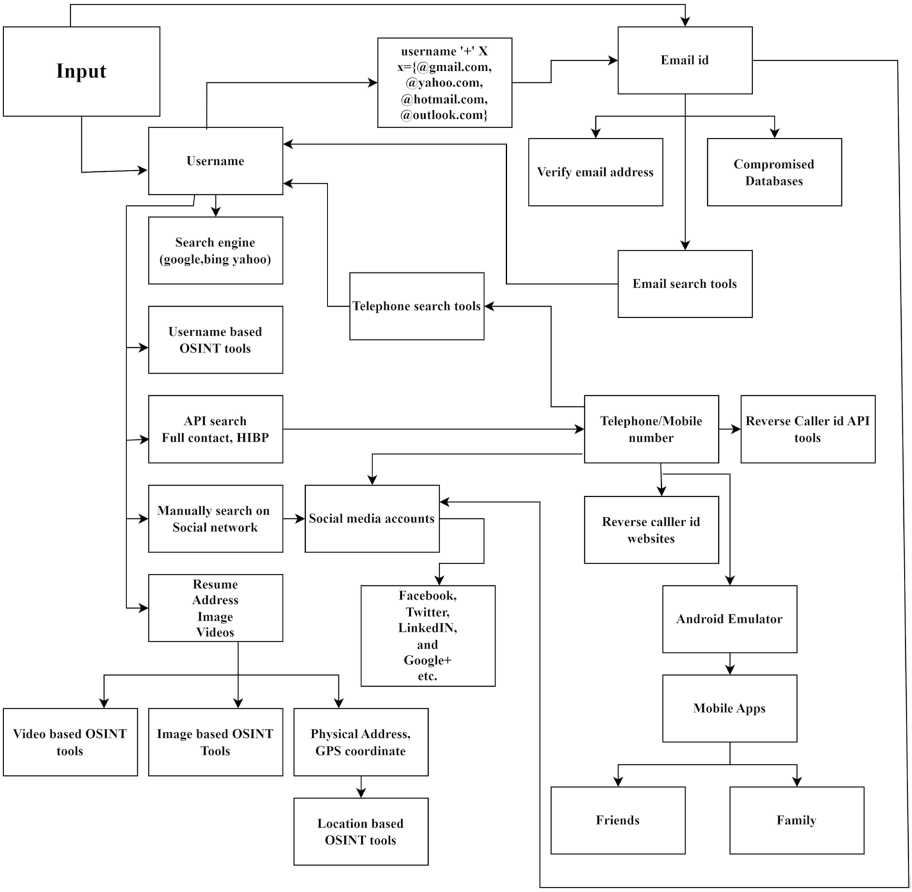

Fig. 6 General attack surface (Bazzell 2016)

Table 2 Historical data extraction websites

Name	Functions
Library of Congress (Hayden 2022)	Huge collection of newspaper archives, maps, photographs, books, and manuscripts
Vital record (Records 2022)	Birth record, marriage license, and death record of USA
Science.gov (Government US 2022)	Science information (more than 200 million pages)
Alexa (2022)	Detailed information on websites
Open access journals directory (Olijhoek 2022)	Provides access to peer-revied, high-quality, and open-access journals
Online book pages (Ockerbloom 2022)	More than 10 million books

Table 3 Official data leak repository

Name	Functions
WikiLeaks (Assange 2022)	Publication of large datasets of censored or otherwise restricted official materials involving war, spying and corruption
Cryptome (2022)	Publication of secret documents prohibited by governments worldwide
OffshoreLeaks (Desprat 2022)	Provides information about company owners, proxies and intermediaries in secrecy jurisdictions

Table 4 User name check tools

Name	Functions
Check user name (Wise 2008)	Helps to find user availability on social media
Namechk (2009)	Helps to find social media account of user
Namecheckr (2022)	Find social media account
Usersearch Team (2022)	Find someone by username or email on Social Networks, Dating Sites, Forums, Crypto Forums, Chat Sites and Blogs

Table 5 Real name search tools and techniques

Name	Functions
Truth finder (2015)	Public record search engines of America
Piple (Lead 2004)	Helps to find the real person
Spokeo (Tang 2006)	Search by name, phone, address, or email to confidentially lookup information about people
TruePeopleSearch (2022)	Helps to find real person
US search (Fast People Finder 2022)	Transparent and informative source for finding addresses, phone numbers, and email addresses
PeekYou Developers (2006)	Helps to find friends, relatives and colleagues across the Web
Whitepage (McMillan 1997)	Find people, contact info & Background checks
Beenverified (Virginia 2007)	Find People, Vehicle, Property and Contact Info
Address Search (2022)	Finding someone’s email or finding out who is emailing you
Lullar (2022)	Profile Search by Email First Last Name or Username
Yasni (2022)	People search
ProfileEngine (2022)	Provide complete website information
Fast People Finder (2022)	People search by name, address search, or reverse phone lookup
Thatthem (2014)	Find people by name, find who lives at an address, find people using phone number, email etc
Webmii (2022)	People search engine
Howmanyofme (LLC 2009)	People search engine
Genealogy (2022)	Search family history

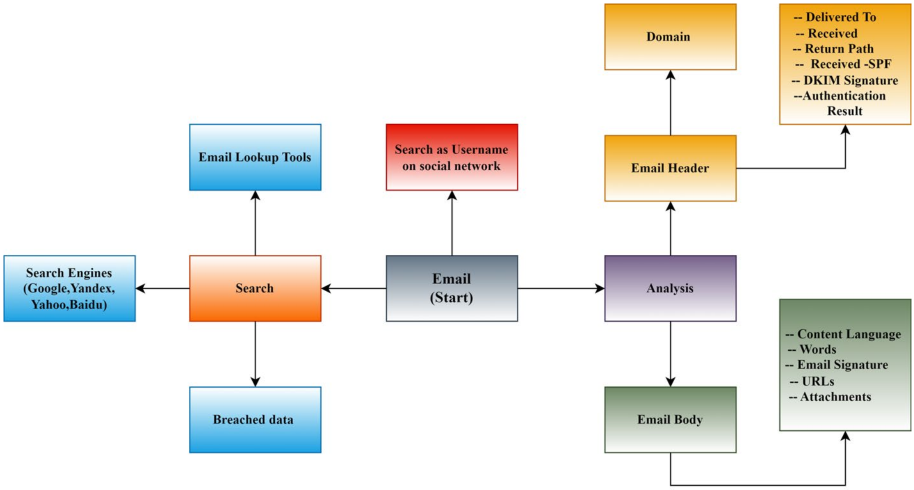

Fig. 7 Email attack surface for investigation

Table 6 Email investigation tools

Name	Functions
Email dossier (Technologies 1997)	Investigate email address
Emailhippo (Paul 2010)	Verify the email address
Hunter (Fink 2015)	Email verifier
Email Checker (2022)	Email checker
Byte plant email validator (GmbH 2003)	Email validator
Email format (2018)	Check email format
Scam dex (Mark 2004)	Scam emails archive
Email-header analysis (ipTRACKERonline 2010)	Email header analysis tool

address, encryption information etc. Received-SPF includes the IP address of the sender along with the hostname.

There are some parameters such as pass, which means the email source is valid; softfail, which means a fake source possible; neutral, which means source validity is difficult to understand as certain and unknown, representing SPF record is not found. In email body analysis, we check the language of the content, signature and attachments etc. And we also searched for email IDs in different search engines, email look-up tools, breached data related to that email, and on several social networks to collect information, as shown in Fig. 7. The open-source tools for email investigation are listed in Table 6.

The phone number is also the key entity that collects information about the target. Some phone number-based tools and techniques are listed in Table 7. We can quickly get the phone number owner details and device information using the phone number.

Internet protocol (IP) address helps to collect the information such as Geographic location of device, Time zone, Area code and ISP etc. Some IP address-based geo-location gathering tools are listed in Table 8.

Table 7 Phone number search

Name	Functions
Zlookup (2022)	Check about owner of mobile number
ReversePhoneLookup (2022)	Info about owner of mobile number
Inter800 (Dianco 1995)	Locate products or services via a toll free number
Twilio (Lawson 2008)	Instantly delivers you a caller ID name and person type
Fonefinder (Brubaker 1997)	Info about owner of mobile number
Truecaller (Zarringhalam 2009)	Info about owner of mobile number
FreeCarrierLookup (2022)	The carrier name and whether the number is wireless or landline
PhoneLookup (2022)	Info about owner of mobile number

Table 8 IP geolocation information tools

Name	Functions
IPverse (2022)	Provide list of IPs
IP2Location (Hexasoft 2001)	Provide info about IP geolocation
IPfingerprints (MaxMind 2002)	Geographic location of an IP address along with some other useful information including ISP, Time Zone, Area Code, State etc
DB-IP (2010)	IP Geo-location API and database
IP Location (Media 2006)	Provide info about IP geolocation
Utrace (2022)	Provide info about IP geolocation

Table 9 Blacklisted IP address

Name	Functions
Block List (Martin 2007)	Report attacks on your server
FireHOL (Costa 2022)	Analyses all available security IP Feeds, mainly related to on-line attacks, online service abuse, malwares, botnets, command and control servers and other cybercrime activities
Directory of malicious IPs (Prince 2004)	Exact moment when the address was harvested and the IP address that gathered it

We can easily collect the list of blacklisted IPs using tools. These blacklisted IP are continuously updated, which may help train the DL model to detect cyber attacks. Some of the directories of blacklisted IPs are listed in Table 9.

Image is also one of the key entities with the help of which we collect information about the particular image. Image search tools help to collect images related to various incidents such as crime, education, breaking news, historical image, politics and elections. These tools help to investigate cyber incidents such as Google image search (2022), Bing image search (2022), Yahoo Images (2022), Yandex Image search (2022) and Baidu (2022). Some other Image search tools and techniques are listed in Table 10.

Nowadays, reverse image search is a trending technique that helps gather information about a particular image. Reverse image search tools help to collect information like the

Table 10 Image search tools

Name	Functions
Imgur (I2022)	Image Sharing and hosting site
Photobucket (2022)	Image hosting and video hosting website
SmugMug (2022)	Photo sharing site
Flickrmap (Norby 2022)	Photo and video sharing site
GettyImages (2022)	Capturing, creating and preserving content to elevate visual communications
Instant logo search (MaxCDN 2016)	Logo search
Reuters pictures (1993)	Provide archive of images, video, audio, graphics and news
News Press (Obits 2022)	Provides trending news of different types such crime, education and politics and elections
Associated press images portal (Meir 2013)	Collections of historical and contemporary photographs
PA images (Marshall 2018)	Image search in uk
European Pressphoto Agency (VisualRightsGroup 2003)	Photo and video coverage of breaking news
Canadian press images archive (2019)	Press images archieve of canada

Table 11 Reverse image search

Name	Functions
Google reverse search (RGoogle 2022)	All related images related to the name across the google
Karmadecay (Reddit 2022)	Source of image searched majorly across Reddit
TinyEye (Tineye 2008)	To know the interests or different accounts a person is connected through just from image
Reverse image search (2011)	Easy to access the sites containing the image was uploaded in it as a new tab opens and shows all links in it
Cam finds App (2019)	All the places across the web where the image is present. Limited as it's a mobile application
Image Identification project (Wolfram 1987)	The content of the image based on the result of AI. AI which determines the content inside the image, useful when bulk images to be analyzed

device used to capture, the location and source of the image etc. Some reverse image search tools and techniques are listed in Table 11.

Image manipulation check tools help to analyze the images. These tools also help to collect information like metadata of the image, location of the captured image, hidden pixels etc. Some tools and techniques for image manipulation check are listed in Table 12.

Video search tools help to find information such as the source of videos, type of content in video, and meta-data of video. These tools also help to collect intelligence on crimes and information about victims and help to investigate the crimes such as Google video search (Google 2022c), Yahoo video search (Yahoo 2022) and Bing Video (Microsoft 2022b). Some other video search tools are listed in Table 13.

Geospatial search tools help to collect location information and analyze incidents at any particular location, street view, and other information related to target locations. Google

Table 12 Image manipulation check

Name	Functions
Forensically (Fridrich 2015)	Magnifiers, String analyzers, Clone Detections
fotoforensics (Krawetz 2012)	ELA, hidden pixels, Metadata, Strings, etc
Ghiro (Ghirensics 2017)	Metadata extraction, GPS Localization, Error level analysis, etc. fully automated tool designed to run forensics analysis over a massive amount of images, just using a user-friendly and fancy web application
ExifTool (Harvey 2022)	Metadata editing tool
GeoSetter (Smith 2019)	Change the geo data linked with an image.exe file to be downloaded, but can be useful in modifying the info
Lets Enhance (2022)	Higher-resolution pic, Unblurring any blurred image of usefulness

Table 13 Video search

Name	Function
AOL (Lanzone 2022)	Videos are based on different categorizations. Very Less as pinpointing the information of the victim is difficult
Start Page video search (2006)	General web search BUT helps leave no footprints behind
Facebook video search (2022)	Browse through videos on Facebook
Internet archive open source movie (Kahle 1996)	All the related media files on the web. Gather huge amount of intel just from the name
Meta tube (2007)	Results spread across Facebook. Required intel from Facebook
Earthcam (1996)	Analyze the victim's surroundings
Insecam (2022)	Live camera feed of nearest CCTV camera
EzGif (Madars 2022)	Editand check hidden data in a video
Video to text converter (360Converter 2021)	Analyze the metadata of the video

maps (2022b), Bing Maps (Microsoft 2022a), and Yandex map (2022) are common geospatial tools. Some Geospatial tools and techniques are listed in Table 14.

Air movement tracking OSINT tools and techniques helps to gather a flight's path and current location and track the target's live location on the move. It also helps monitor air traffic at any location at any given time. Some air movement tracking OSINT tools are listed in Table 15

The main step of conducting maritime research is to detect a vessel's AIS (Automatic Identification System). You can see the vessel's name, position, destination, vessel type, and other information with the help of AIS. OSINT tools and technology can be used to track vessels that are breaking the law, smuggling drugs, smuggling humans, fishing illegally, and practicing naval maneuvers (Smithrae 2021). Some of OSINT tools and techniques of ships tracking are listed in Table 16.

Using the Siamese Network with the Region Proposal Network, Shan et al. (2020) proposed a marine ship tracker. They made changes to the CNN in the Siamese subnetwork to increase the effectiveness of feature abstraction and presented an adaptive search area extraction technique to lessen the impacts of shakings. However the

Table 14 Geospatial search tools

Name	Functions
Digital globe (ESRI 2022)	Map of the location, Explore the surroundings of a victim if the location is known
Daum (2022)	Map of location, Limited as everything in Korean
N2yo (2022)	Live feed from satellite to know relevant data
Wiggle (2001)	Network maps of the location
BB Bike (Schneider 2022)	Different maps placed side by side for easy comparisons
Newspaper map (OpenStreetMap 2022)	Intel over general interests in the area, News Papers brands used majorly in the area
USGS (2022)	The US provided a map finally with meeting the set properties, Useful in narrowing down on a target
Google map street view (2022a)	Street view of the location or around it
KartaView (Chris 2020)	Publicly available street views datasets
ZoomEarth (2022)	Live satellite feed with weather and all
Mapillary (2013)	Street view if available
Address lookup (Agarwal 2004)	Address of the clicker i.e. a process to make the victim click the button and get his address
Viamichelin (2022)	The path between the locations
Corona project (Wilson 1960)	Map made by a US secret satellite in the 1970s used to know the old name of a place
TripGeo (2022)	The animated path between the locations provided
Mapquest (2022)	Map with a nearest public place like restaurants etc. nearby, Easy to explore a wide area
Maphub (2022)	Collaborative Maps
Perry-Castaneda library map collection (2022)	Map of the selected period, Get some really old maps and new maps too WayBack Machine of Maps
Round shot (Seitz 2022)	Live Cam active nearby, Explore nature, and understand a country
Live earth map (GlobalIncidentMap 2006)	Marks on locations where a recent natural disaster occurred, Understand the different problems different countries face with respect to natural disasters

Table 15 Air movement tracking

Name	Functions
Flightaware (Sulak 2005)	Path and current location of a flight, track the live location of the target on the move
Flight radar 24 (2006)	Monitor Air Traffic at any location at any given time
Air cargo tracker (Track Trace 1998)	Track the cargo loaded on the flight, Limited flights allow this feature so functionality is limited too
Radar box 24 (Brandao 2001)	Monitor Air Traffic at any location at any given time
World Aircraft's database (PlaneMapper 2022)	Data related to flight, Dictionary of all the flights available

proposed tracker works for limited vessels and weather conditions. Yang et al. (2022) proposed a coastal ship tracking network that integrates segmentation and visual object tracking in general, which can track and segment ships simultaneously. They used ERM

Table 16 Maritime movements tracking tools

Name	Functions
Marine traffic (2007)	Map of all the marine traffic around the world, Know the marine traffic around the required location
Vessel finder (2011)	Map of all the marine traffic around the world, Know the marine traffic around the required location
Cruise Mapper (2022)	Know the marine traffic around the required location
Ship Finder (pinkfroot 2012)	All the ships (named) at the location, Know where the major famous ships are located live
Container prefix-list (Lorenz 1990)	All about different containers. Active, dead, in the dock, etc
Identification code of container owner (BIC 1970)	The exact location and details connected with code, Know in detail about the owner of the container
Volza (Volza 2022)	Know the port codes of all ports on the globe to make searches easy

with a more advanced feature pyramid fusion approach. The approach enhances the network's feature map during feature extraction and fusion which will help to increase tracking accuracy. They used the Large Maritime Dataset, which contains speed boats, cargo ships, passenger ships, fishing ships, and unmanned ships. Wang et al. (2021) have proposed a model for ship detection using various ML classifiers such as Random Forest, SVM, LR, KNN and LDA etc. They used satellite images to train the model. In preprocessing, they used an SDT filter to remove the noise from images. They used google earth data for the validation. Spadon et al. (2022) proposed a model which addresses the behavior of AIS messages using the various ML as well as DL algorithms under the irregular and noisy data. Zhang et al. (2021) proposed a model that is RoDAN, which fuses three dimensions such as scale, motion and region for tracking information. They used an ASPP module to scale dimensions that helps to get more accurate information to track the ships. They used marine public datasets named the SMD (Prasad et al. 2017) and HSD dataset suitable for ship tracking i.e MarDCT (Gundogdu et al. 2016). Package tracking tools helps to track the package which helps to get the update about the package. These tools helps to track the package of smuggling drugs, weapons, and illegal packages. Some OSINT tools and techniques of package tracking are listed in Table 17.

Table 17 Package tracking tools

Name	Functions
After ship (Chan 2022)	Live track after the drop
Trackingex (2022)	Live tracking the package from the company perspective
17 Track (Rustin 2010)	Live track after the drop
Package tracker (USP 2008)	Live tracking the package from the company perspective
Canada post (Ettinger 2019)	Locations and info about packages, Limited as restricted to only Canada
Royal mail (Thompson 2022)	Locations and info about packages, Limited as restricted to the only UK

5 Discussion

The challenges of OSINT investigation are: automating the collection process, analysis, and knowledge extraction, integration of several open data sources, filtering out misinformation and irrelevant data, globalization of OSINT tools and techniques, awareness of privacy, ethical and legal consideration, and prevent misuse of OSINT tools and techniques.

In ethical consideration, OSINT tools and techniques must be used sincerely and legally for legitimate works. The publicly available Data does not mean that they are not sensitive data. Because if someone's political views, medical data, religious beliefs, and family and friends' information are publically available these data might be used for threatening or blackmailing. For example, religious beliefs may lead to the government and public conviction towards terrorism. Due to this, we have to ensure that we must follow the ethics, data privacy rules and regulations such as General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and other applicable laws respective to their locations.

GDPR (EU 2016) was introduced by European Union to regulate the processing of the personal data of European citizens. OSINT also involves the processing of personal data. So some of the aspects of GDPR related to OSINT are accountability (everything used and how you are collecting data must be documented), legal basis (you need to have consent (article 6.1.a)), legal obligation and legitimate interest, principles(article 5) and data subjects rights (article 14). The principles involve lawfulness which states that the collection method which is going to be used for collection must be legally verified (do not break the authentication protocol to collect any type of data), Fairness states that collect only data that has related to your investigations, Transparency states that the data processor used, type of data processed, and purpose of processing collected data must be transparent, Data minimalization (you should try to process the minimal personal data as much as possible), Storage Limited (do not store the data longer than it needed), Integrity and confidentiality (the data must be accessed by only you and your client, you should store data in encrypted form). The data subject rights involve the right of notification, right of access, right of rectification, right to the erasure of data, and right to restriction of data processing.

In the case of surveillance (Miller et al. 2022), we gathered a huge amount of information from multiple sources that helps to track and monitor the individual, group, and organization. This information can raise the issue of civil liberties and freedom of expression and that's why we have to ensure that we do not violate any rights of individuals. There are various issues such as data privacy, data protection, issue of civil liberties, freedom of expression, etc. According to GDPR, you can process and store the data until it is needed. Otherwise, you must permanently delete all information regarding a particular investigation. The collected data must be stored in encrypted form and accessed by only you and your client. OSINT is helpful for government, LEAs, and cyber security professionals but it must be used ethically for legitimate work. You should ensure that data collection must be done with the rules and regulations and must not break the authentication protocols. Furthermore, organizations/individuals should have a clear understanding of the laws and regulations that govern the collection, storage, and use of OSINT. This includes understanding the rights of individuals and organizations concerning their personal information and being aware of the penalties for non-compliance.

OSINT must be used according to law and by considering the data privacy and protection policies (Rajamaki and Simola 2019). OSINT is legal by definition but it is necessary to ensure that data collected by researchers, investigators, government agencies, and LEAs did not publish even if data is publicly available. We should not break the authentication protocol for collecting data. In other words, we can say that OSINT must be used in a restricted manner for legitimate and non-harmful activities. We need to make a balance between what we collect and what we store or mention in the report after investigation. In any inquiry, we collect every possible data but we need to ensure that in the final report, the information is under the scope of GDPR and data must be in encrypted form and accessed only by you and your client.

In misinformation consideration, During the early stages of the COVID-19 pandemic, states were desperate for any means to combat the economic, health, and human consequences of the disease. The severity of the crisis prompted states to utilize or consider using any available tools, even those that had been previously limited to national security purposes. The utilization of intelligence and surveillance tools traditionally used for security, intelligence, and law enforcement for pandemic surveillance highlights the drastic measures that many states have taken to curb the spread and minimize the impact of COVID-19.

The technologies which are used for national security surveillance (like location monitoring, facial recognition, and social media intelligence) are deployed to combat the covid-19 pandemic. For example, The United States, developed technology using social media intelligence (like blogs, news, social networks, and government sources) to counter the spread of the Covid-19 pandemic (Berman et al. 2020). Some other countries (like Israel, Pakistan, South Korea, Hong Kong, and China, etc.) also used their national security intelligence (such as location monitoring, CCTV surveillance, Face recognition, tracking individuals/groups, GPS and cell phone tracking, etc.) to control the spread and counter Covid. Due to covid millions of people are died and it affected the economic conditions worldwide. Due to this using the national security intelligence system and OSINT tools and techniques to counter the pandemic is justified. But the OSINT tools and techniques must be utilized by considering the cyber laws and regulations, data privacy, data protection, freedom of expression and civil liberties, etc.

For the collection process web crawlers, web scrapers, and API-based tools are used. In analysis and knowledge extraction, we generally do semantic analysis, analysis of threat patterns, correlation with other events, and also occurrences of data to find the relation between different separated pieces of information to get the relevant information. And also used data mining techniques, NLP techniques, and SNA to extract relevant information from collected data. In the integration of multiple open-source data, we integrate the aggregated data from various social networks, the internet, and also from the dark web. In the case of the detection of fake information, filtering of irrelevant data, and misinformation generally ML and DL models are proposed. When we are using OSINT tools and techniques we need to focus on user privacy, user family privacy, user-friend privacy, and the privacy of their co-workers. And Follow GDPR protection rules and regulations. OSINT is legal because the data sources are publicly available. We also need to remember that publicly available data are prone to abuse (Castelle 2018), cyber aggression (Kumar et al. 2018), cyberbullying (Hosseinmardi et al. 2015; Wachs et al. 2019), cyber gossip (García-Fernández et al. 2022), and cyber harassment (Abarna et al. 2022).

Misuse of the OSINT tools and techniques causes, loneliness, depression, distress, and victims may commit suicide in the worst case. We can positively use the OSINT for HR recruitment, counter cyber criminals, prevent the spread of misinformation, detection of fake information, profiling of cybercriminals, and doing digital forensics, etc. Profiling cybercriminals in the darknet market is one of the use cases of OSINT tools and techniques. The challenge, however, is that profiling a person who in reality does not pose a threat leads to discriminatory and unfair attitudes then may affect victims. OSINT helps to profile the cyber attacks and improves sophisticated cyber attacks (Akinrolabu et al. 2018). It is benificial in private sector and also work as resource of public interest in governments (Lande and Shnurko-Tabakova 2019). This also helps in classified investigation and secret operations (Larsen et al. 2017). It helps in investigation and strategic planning to counter crimes (Akhgar 2016).

6 Conclusion and future work

This paper discussed the current state of open source intelligence (OSINT). It revealed that current techniques are inadequate due to their lack of effectiveness in real-world scenarios. This paper focuses on highlighting the importance of integrating OSINT into cyber defenses, social networks, digital forensics, and other possible domains to identify the profile of culprits, cyber crimes, counter-terrorism, and cyber incidents. The paper also outlined basic OSINT search techniques and describes advanced OSINT tools. Proper tool selection based on available data and goals is crucial, but using a combination of tools is the best way to achieve accurate results.

In this paper, we also described open source data, social networks, various types of cybercrime, and OSINT tools and techniques with NLP/AI/ML/DL and this can be helpful for investigators to enhance the OSINT research and applications. Generally, IOCs are extracted from traditional blacklists like cleanMX and Phishtank covering only URLs, domains, IPs, and MD5 signatures but it does not include the criminal groups and other context information. So using OSINT we can add psychological and behavioral features. We conclude that OSINT can help to improve cyber security issues in various fields like phishing detection, threat intelligence, hate speech detection, fake news detection, human trafficking, child trafficking, criminal profiling, etc. OSINT also helps to monitor APT groups, fake images/video analysis, monitor malicious activities, and trace violent acts. In the case of phishing detection, we use the various available data sets to train the model. But if we use the automatic updating features, attack patterns, and IOCs, then the model might become more robust and efficient.

The challenges that need to be addressed are: extracting indicators of compromise and their relations from unstructured threat intel reports and using that extracted knowledge for threat hunting, making a consortium of criminals to track them, developing a framework of every possible OSINT tool and technique, automation of information gathering and extracting the intelligence from open source data using AI, establish real-time model using the OSINT and DL techniques to monitor the Advanced Persistent Threat(APTs) groups automatically, make a multi-platform and multi-lingual social networks based benchmark dataset for cybercrime detection with the help of OSINT tools and techniques. Automatic detection of human trafficking and child trafficking with OSINT and DL techniques also needs to be explored.

References

360Converter (2021) 360converter provides different kinds of converters, primarily focusing on video, audio, speech and voice to text. http://www.360converter.com/

Abarna S, Sheeba J, Jayasrilakshmi S et al (2022) Identification of cyber harassment and intention of target users on social media platforms. Eng Appl Artif Intell 115(105):283

Abbass Z, Ali Z, Ali M et al (2020) A framework to predict social crime through twitter tweets by using machine learning. In: 2020 IEEE 14th international conference on semantic computing (ICSC), IEEE. pp 363–368

Abdul-Mageed M, Diab M (2014) Sana: a large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In: Proceedings of the ninth international conference on language resources and evaluation (LREC'14)

Address Search (2022) Free email address lookup and mailing address search. https://www.addresssearch.com/. Accessed July 2022

Agarwal A (2004) Where am I—know your location and postal address on google maps. https://ctrlq.org/maps/address

Akhgar B (2016) Osint as an integral part of the national security apparatus. In: Akhgar B (ed) Open source intelligence investigation. Springer, New York, pp 3–9

Akhgar B, Bayerl PS, Sampson F (2017) Open source intelligence investigation: from strategy to implementation. Springer, New York

Akinrolabu O, Agrafiotis I, Erola A (2018) The challenge of detecting sophisticated attacks: insights from SOC analysts. In: Proceedings of the 13th international conference on availability, reliability and security. pp 1–9

Alabdan R (2020) Phishing attacks survey: types, vectors, and technical approaches. Future Internet 12(10):168

Albladi SM, Weir GR (2017) Personality traits and cyber-attack victimisation: multiple mediation analysis. In: 2017 internet of things business models, users, and networks, IEEE. pp 1–6

Aleroud A, Zhou L (2017) Phishing environments, techniques, and countermeasures: a survey. Comput Secur 68:160–196

Alexa (2022) End of service notice. https://www.alexa.com/. Accessed July 2022

AlKilani H, Qusef A (2021) Osint techniques integration with risk assessment iso/iec 27001. In: International conference on data science, E-learning and information systems 2021. pp 82–86

Amaro L, De Los Santos T, Joseph N et al (2018) An application of the process of responsive identification on social media (prism) model in the online mommy platform context

Aslan ÇB, Sağlam RB, Li S (2018) Automatic detection of cyber security related accounts on online social networks: Twitter as an example. In: Proceedings of the 9th international conference on social media and society. pp 236–240

Assange J (2022) Wikileaks is a giant library. https://wikileaks.org/. Accessed July 2022

Aye YM, Aung SS (2020) Contextual lexicon based sentiment analysis in myanmar text reviews. In: 2020 23rd conference of the oriental COCOSDA International Committee for the Co-ordination and standardisation of speech databases and assessment techniques (O-COCOSDA), IEEE. pp 160–165

Baidu (2022) Baidu image. http://image.baidu.com/. Accessed July 2022

Bazzell M (2016) Open source intelligence techniques: resources for searching and analyzing online information. CreateSpace Independent Publishing Platform, Scotts Valley

Berman E, Fowler LR, Roberts JL (2020) Covid-19 surveillance

BIC (1970) The international register of container owner codes. https://www.bic-code.org/bic-codes/

Bing Image (2022). http://www.bing.com/images. Accessed July 2022

Boididou C, Andreadou K, Papadopoulos S et al (2015) Verifying multimedia use at mediaeval 2015. MediaEval 3(3):7

Brandao A (2001) Global flight tracking intelligence: live flight tracker and airport status. https://www.radarbox.com/

Brubaker W (1997) finds the geographic location of any phone number. http://www.fonefinder.net/

CamFind (2019) Mobile visual search engine. http://camfindapp.com/

Canadian Press (2019) Canada's leading collection of photos and video Canadian Press images. http://www.cpimages.com/fotoweb/index.fwx

Carloni G (2014) Content and language integrated learning: a blended model in higher education. Int J Technol Knowl Soc Annu Rev 9:4

Castelle M (2018) The linguistic ideologies of deep abusive language classification. In: Proceedings of the 2nd workshop on abusive language online (ALW2). pp 160–170

Celestini A, Me G, Mignone M (2017) Tor marketplaces exploratory data analysis: the drugs case. In: International conference on global security, safety, and sustainability. Springer, pp 218–229

Ch R, Gadekallu TR, Abidi MH et al (2020) Computational system to classify cyber crime offenses using machine learning. Sustainability 12(10):4087

Chan T (2022) Shipment tracking. https://www.aftership.com/couriers. Accessed July 2022

Chris (2020) Karta view. https://kartaview.org/landing

Churi T, Sawardekar P, Pardeshi A et al (2017) A secured methodology for anti-phishing. In: 2017 international conference on innovations in information, embedded and communication systems (ICIIECS), IEEE. pp 1–4

Cinelli M, Cresci S, Quattrociocchi W et al (2022) Coordinated inauthentic behavior and information spreading on twitter. Decision support systems. p 113819

Costa T (2022) Ip blacklists: Ip reputation feeds. http://iplists.firehol.org/. Accessed July 2022

CruiseMapper (2022) Cruisemapper. http://www.cruisemapper.com/. Accessed July 2022

Cryptome (2022) Unauthorized disclosures of official secrets. https://cryptome.org/. Accessed July 2022

Cui J et al (2017) Tracking phishing attacks over time. pp 667–676

Daum (2022) Geospatial search tool. http://map.daum.net/. Accessed July 2022

DBIP (2010) Ip geolocation api and database. https://db-ip.com/

Delavallade T, Bertrand P, Thouvenot V (2017) Extracting future crime indicators from social media. In: Larsen HL et al (eds) Using open data to detect organized crime threats. Springer, New York, pp 167–198

Desprat C (2022) Offshore leaks database. https://offshoreleaks.icij.org/. Accessed July 2022

Dharawat A, Lourentzou I, Morales A et al (2020) Drink bleach or do what now? covid-hera: a dataset for risk-informed health decision making in the presence of covid19 misinformation. Preprint at http://arxiv.org/abs/2010.08743

Dianco (1995) Internet 800 directory—directory of free toll free (tollfree), 800, 866, 877 and 888 listings of businesses by company, number and type of industry, regardless of long distance carrier. http://inter800.com/index.html

EarthCam (1996) The global network of owned and operated live streaming webcams. http://www.earthcam.com/

Edwards M, Rashid A, Rayson P (2015) A systematic survey of online data mining technology intended for law enforcement. ACM Comput Surv (CSUR) 48(1):1–54

Edwards M, Larson R, Green B et al (2017) Panning for gold: automatically analysing online social engineering attack surfaces. Comput Secur 69:18–34

Edwards M, Williams E, Peersman C et al (2022) Characterising cybercriminals: a review. Preprint at http://arxiv.org/abs/2202.07419

Email Checker (2022). https://email-checker.net/. Accessed July 2022

Email Format (2018) Email address format for millions of companies. https://email-format.com/

Enhance (2022) increase image resolution without losing quality. https://letsenhance.io/. Accessed July 2022

ESRI (2022) Digital globe maps. https://discover.digitalglobe.com/. Accessed July 2022

Ettinger D (2019) Online shipping and marketing tools. https://www.canadapost-postescanada.ca/cpc/en/tools.page

EU E (2016) Lex-02016r0679-20160504-en-eur-lex. https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04

Facebook (2022) Facebook videos. https://www.facebook.com/pg/facebook/videos. Accessed July 2022

Fast People Finder (2022) Fast people finder. https://fastpeoplefinder.com/. Accessed July 2022

Fink A (2015) Email verifier: verify email address with free email checker hunter. https://hunter.io/email-verifier

Flightradar24 (2006) Global flight tracking service that provides you with real-time information about thousands of aircraft around the world. https://www.flightradar24.com/

FreeCarrierLookup (2022) Phone number search tool. http://freecarrierlookup.com/. Accessed July 2022

Fridrich J (2015) Tools for clone detection, error level analysis, meta data extraction. https://29a.ch/photo-forensics/

Ganesan M, Mayilvahanan P (2017) Cyber crime analysis in social media using data mining technique. Int J Pure Appl Math 116(22):413–424

Garcia K, Berton L (2021) Topic detection and sentiment analysis in twitter content related to covid-19 from Brazil and the USA. Appl Soft Comput 101(107):057

García-Fernández CM, Moreno-Moya M, Ortega-Ruiz R et al (2022) Adolescent involvement in cybergossip: influence on social adjustment, bullying and cyberbullying. Span J Psychol 25:e6

Genealogy (2022) Research resource with over 14,000 online forums devoted to genealogy. https://www.genealogy.com/. Accessed July 2022

GettyImages (2022) Royalty free stock photos, illustrations, vector art, and video clips. http://www.gettyimages.com/. Accessed July 2022

Ghazi Y, Anwar Z, Mumtaz R et al (2018) A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources. In: 2018 international conference on frontiers of information technology (FIT), IEEE. pp 129–134

Ghirensics (2017) Automated digital image forensics tool. http://www.getghiro.org/

GlobalIncidentMap (2006) Find and map terrorism events, disease outbreaks, earthquakes, forest fires, aviation incidents, threats to churches, threats to schools, and threats to journalists etc. http://quakes.globalincidentmap.com/

GmbH B (2003) Verify email addresses with byteplant's email validator. https://www.email-validator.net/

Google (2022a) Discover street view and contribute your own imagery to google maps. https://www.google.com/streetview. Accessed July 2022

Google (2022b) Google maps. https://www.google.co.in/maps. Accessed July 2022

Google (2022c) Google videos. https://www.google.com/videohp. Accessed July 2022

GoogleImages (2022) Image tool. https://images.google.com/. Accessed July 2022

Government US (2022) Portal offers free access to research and development (r and d) results and scientific and technical information from Scientific Organizations across 13 Federal Agencies. https://www.science.gov/. Accessed July 2022

Gundogdu E, Solmaz B, Yücesoy V et al (2016) Marvel: a large-scale image dataset for maritime vessels. In: Asian conference on computer vision. Springer, pp 165–180

Gupta BB, Tewari A, Jain AK et al (2017) Fighting against phishing attacks: state of the art and future challenges. Neural Comput Appl 28(12):3629–3654

Halevi T, Lewis J, Memon N (2013) Phishing, personality traits and facebook. Preprint at http://arxiv.org/abs/1301.7643

Harvey (2022) The sudbury neutrino observatory. https://sno.phy.queensu.ca/?phil%2Fexiftool%2F. Accessed July 2022

Hashida S, Tamura K, Sakai T (2018) Classifying sightseeing tweets using convolutional neural networks with multi-channel distributed representation. In: 2018 IEEE international conference on systems, man, and cybernetics (SMC), IEEE. pp 178–183

Hayden C (2022) Home: library of congress. https://www.loc.gov/. Accessed 16 Sept 2022

Herrera-Cubides JF, Gaona-García PA, Sánchez-Alonso S (2020) Open-source intelligence educational resources: a visual perspective analysis. Appl Sci 10(21):7617

Hexasoft (2001) Ip address to ip location and proxy information. https://www.ip2location.com/

Hosseinmardi H, Mattson SA, Ibn Rafiq R et al (2015) Analyzing labeled cyberbullying incidents on the instagram social network. In: International conference on social informatics. Springer, pp 49–66

Hu Y, Huang H, Chen A et al (2020) Weibo-cov: a large-scale covid-19 social media dataset from weibo. Preprint at http://arxiv.org/abs/2005.09174

Imgur (2022) Database of trending images. https://imgur.com/. Accessed July 2022

Insecam (2022) Live cameras directory. http://www.insecam.org/. Accessed July 2022

IpTRACKERonline (2010) Complete email header analysis. Analyse, track ip here. https://www.iptrackeronline.com/email-header-analysis.php

Ipverse (2022) Ipv4/ipv6 address block lists by country code. https://xranks.com/ipverse.net. Accessed July 2022

Islam C, Babar MA, Croft R et al (2022) Smartvalidator: a framework for automatic identification and classification of cyber threat data. J Netw Comput Appl 202(103):370

Ivaturi K, Janczewski L (2011) A taxonomy for social engineering attacks

Jain AP, Dandannavar P (2016) Application of machine learning techniques to sentiment analysis. In: 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT), IEEE. pp 628–632

Kadoguchi M, Kobayashi H, Hayashi S et al (2020) Deep self-supervised clustering of the dark web for cyber threat intelligence. In: 2020 IEEE international conference on intelligence and security informatics (ISI), IEEE. pp 1–6

Kahle B (1996) A digital library of internet sites and other cultural artifacts. https://archive.org/details/

Kakkar A (2020) A survey on secure communication techniques for 5g wireless heterogeneous networks. Inf Fusion 62:89–109

Kandias M, Gritzalis D, Stavrou V et al (2017) Stress level detection via osn usage pattern and chronicity analysis: an osint threat intelligence module. Comput Secur 69:3–17

Kao DY, Chao YT, Tsai F et al (2018) Digital evidence analytics applied in cybercrime investigations. In: 2018 IEEE conference on application, information and network security (AINS), IEEE. pp 111–116

Kathrine GJW, Praise PM, Rose AA et al (2019) Variants of phishing attacks and their detection techniques. In: 2019 3rd international conference on trends in electronics and informatics (ICOEI), IEEE. pp 255–259

Khera DV (2021) An introduction to open source intelligence (osint). https://cyberprotection-magazine.com/an-introduction-to-open-source-intelligence-osint

Krawetz N (2012) Fotoforensics provides budding researchers and professional investigators access to cutting-edge tools for digital photo forensics. http://fotoforensics.com/

Krombholz K, Hobel H, Huber M et al (2015) Advanced social engineering attacks. J Inf Secur Appl 22:113–122

Kumar S, Hamilton WL, Leskovec J et al (2018) Community interaction and conflict on the web. In: Proceedings of the 2018 world wide web conference. pp 933–943

Kumar A, Verma A, Shinde G et al (2020) Crime prediction using k-nearest neighboring algorithm. In: 2020 International conference on emerging trends in information technology and engineering (IC-ETITE), IEEE. pp 1–4

Kunju MV, Dainel E, Anthony HC et al (2019) Evaluation of phishing techniques based on machine learning. In: 2019 international conference on intelligent computing and control systems (ICCS), IEEE. pp 963–968

Lande DV, Shnurko-Tabakova EV (2019) Osint as a part of cyber defense system. Igor Sikorsky Kyiv Polytechnic Institute, Kyiv

Lanzone J (2022) Aol video-serving the best video content from aol and around the web. https://www.aol.com/video. Accessed July 2022

Larsen HL, Blanco JM, Pastor RP et al (2017) Using open data to detect organized crime threats: factors driving future crime. Springer, New York

Lawson J (2008) Lookup api. https://www.twilio.com/lookup

Lead T (2004) Pipl collects, cross-references and connects online identity information. https://pipl.com/

Liao X, Yuan K, Wang X et al (2016) Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp 755–766

LLC AT (2009) Examines the commonness or uncommonness of names. http://howmanyofme.com/search/

Lorenz (1990) The world's largest container site. https://www.prefixlist.com/

Lullar (2022) Lullar—search people profile by email or username. http://com.lullar.com/. Accessed July 2022

Madars (2022) Ezgif is a free, simple to use toolset designed primarily for creating and editing. https://ezgif.com/reverse-video. Accessed July 2022

Malhotra P, Singh Y, Anand P et al (2021) Internet of things: evolution, concerns and security challenges. Sensors 21:1809

MapHub (2022) Maphub · create interactive maps. https://maphub.net/. Accessed July 2022

Mapillary (2013) Street-level imagery and map data available. https://www.mapillary.com/

Mapquest (2022) Mapquest-location-enabled geospatial solutions. https://www.mapquest.com/. Accessed July 2022

MarineTraffic (2007) Global ship tracking intelligence: Ais marine traffic. https://www.marinetraffic.com/

Mark S (2004) Scam reporting, email scams, internet fraud, id theft and phishing resource free repository. http://www.scamdex.com/

Marshall C (2018) UK's leading news, sport and entertainment image provider. https://www.paimages.co.uk/

Martin S (2007) Fail2ban-reporting service (we sent reports from attacks on postfix, ssh, apache-attacks, spambots, irc-bots, reg-bots, ddos and more) from fail2ban via x-arf. http://www.blocklist.de/en/index.html

Martin J, Munksgaard R, Coomber R et al (2020) Selling drugs on darkweb cryptomarkets: differentiated pathways, risks and rewards. Br J Criminol 60(3):559–578

MaxCDN (2016) Instant logo search. http://instantlogosearch.com/

MaxMind (2002) Ip address geographical location finder. http://www.ipfingerprints.com/

McMillan L (1997) Find people, contact info and background checks. https://www.whitepages.com/

Media B (2006) Ip-based geolocation is mapping of an ip address. https://www.iplocation.net/

Meir N (2013) Access to the world's largest collections of historical and contemporary photographs. http://www.apimages.com/

Metatube (2007) share and upload videos and photos. http://www.metatube.com/

Microsoft (2022a) Bing maps. https://www.bing.com/maps. Accessed July 2022

Microsoft (2022b) Bing videos. https://www.bing.com/videos. Accessed July 2022

Miller S, Regan M, Walsh PF (2022) National security intelligence and ethics. Taylor & Francis, Cambridge

Mishra RK, Urolagin S et al (2019) A sentiment analysis-based hotel recommendation using tf-idf approach. In: 2019 International conference on computational intelligence and knowledge economy (ICCIKE), IEEE. pp 811–815

MISP F (2021) Misp-open source threat intelligence platform & open standards for threat information sharing n2yo (2022) Real time satellite tracking. http://www.n2yo.com/. Accessed July 2022

NameCheckr (2022) Namecheckr: social and domain name availability search for brand professionals. https://www.namecheckr.com/. Accessed July 2022

NameChk (2009) Username and domain name checker—search all domain names and user names to see if they're available. https://namechk.com/

Naseem U, Khan SK, Razzak I et al (2019) Hybrid words representation for airlines sentiment analysis. In: Australasian joint conference on artificial intelligence. Springer, pp 381–392

Nazah S, Huda S, Abawajy J et al (2020) Evolution of dark web threat analysis and detection: a systematic approach. IEEE Access 8:171796–171819

Norby (2022) Online photo management and sharing application. https://www.flickr.com/map. Accessed July 2022

Nouh M, Nurse JR, Goldsmith M (2016) Towards designing a multipurpose cybercrime intelligence framework. In: 2016 European intelligence and security informatics conference (EISIC), IEEE. pp 60–67

Nouh M, Nurse JR, Webb H et al (2019) Cybercrime investigators are users too! understanding the socio-technical challenges faced by law enforcement. Preprint at http://arxiv.org/abs/1902.06961

Obits (2022) Santa barbara news. https://newspress.com/. Accessed July 2022

Ockerbloom JM (2022) The online books page. https://onlinebooks.library.upenn.edu/. Accessed July 2022

Olijhoek T (2022) Directory of open access journals. https://doaj.org/. Accessed July 2022

Omand D, Bartlett J, Miller C (2012) Introducing social media intelligence (socmint). Intell Natl Secur 27(6):801–823

OpenStreetMap (2022) Location of newspapers. https://newspapermap.com/. Accessed July 2022

Paul M (2010) Verify email address for free with email checker tool. https://www.verifyemailaddress.org/

PclMaps (2022) Pcl maps: the university of Texas at Austin. https://legacy.lib.utexas.edu/maps/. Accessed July 2022

PeekYou Developers (2006) Peekyou—people search made easy. https://www.peekyou.com/

PhoneLookup (2022) Phone lookup. https://www.phonelookup.com/. Accessed July 2022

Photobucket (2022) Photobucket. http://photobucket.com/. Accessed July 2022

Pienta D, Thatcher JB, Johnston AC (2018) A taxonomy of phishing: attack types spanning economic, temporal, breadth, and target boundaries

Pinkfroot (2012) Ship finder works by picking up ais ship feeds used by commercial vessels and pleasure craft to transmit their name, position, mmsi, status and lots more. https://shipfinder.co/

PlaneMapper (2022) World aircrafts database. http://www.planemapper.com/aircrafts. Accessed July 2022

Pogorelov K, Schroeder DT, Burchard L et al (2020) Fakenews: corona virus and 5g conspiracy task at mediaeval 2020. In: MediaEval

Prabhakar E, Santhosh M, Krishnan AH et al (2019) Sentiment analysis of us airline twitter data using new adaboost approach. Int J Eng Res Technol 7(1):1–6

Prasad DK, Rajan D, Rachmawati L et al (2017) Video processing from electro-optical sensors for object detection and tracking in a maritime environment: a survey. IEEE Trans Intell Transp Syst 18(8):1993–2016

Prince M (2004) The web's largest community tracking online fraud and abuse, project honey pot. https://www.projecthoneypot.org/

ProfileEngine (2022) Profileengine.com. http://profileengine.com/. Accessed July 2022

Quick D, Choo KKR (2018) Digital forensic intelligence: data subsets and open source intelligence (dfint+ osint): a timely and cohesive mix. Futur Gener Comput Syst 78:558–567

Raj C, Meel P (2022) People lie, actions don't! modeling infodemic proliferation predictors among social media users. Technol Soc 68(101):930

Rajamaki J, Simola J (2019) How to apply privacy by design in osint and big data analytics? In: ECCWS 2019 18th European Conference on cyber warfare and security, academic conferences and publishing limited. p 364

Rastenis J, Ramanauskaitė S, Janulevičius J et al (2020) E-mail-based phishing attack taxonomy. Appl Sci 10(7):2363

Records V (2022) Vital records—birth certificates, death records, marriage licenses and more. http://www.vitalrec.com/. Accessed July 2022

Reddit (2022) Reddit is home to thousands of communities, endless conversation, and authentic human connection. https://www.reddit.com/domain/karmadecay.com/. Accessed July 2022

Reuters T (1993) Reuters pictures. http://pictures.reuters.com/

Revell Q, Smith T, Stacey R (2016) Tools for osint-based investigations. In: Revell Q, Smith T, Stacey R (eds) Open source intelligence investigation. Springer, New York, pp 153–165

ReverseImage (2011) Find google images and lookup a photo online. http://www.reverse-image-search.com/

ReversePhoneLookup (2022) Reverse phone lookup: phone number lookup. https://www.reversephonelookup.com/. Accessed July 2022

RGoogle (2022) Reverse image search. https://www.google.com/imghp. Accessed July 2022

Roddy AL, Holt TJ (2022) An assessment of hitmen and contracted violence providers operating online. Deviant Behav 43(2):139–151

Rustin (2010) All in one package tracking. https://www.17track.net/en

Schneider W (2022) Extracts areas from planet.osm. https://extract.bbbike.org/. Accessed July 2022

Seitz P (2022) Livecam images are true digital marketing assets and can be shared not only on the web. https://www.roundshot.com/xml_1/internet/en/intro.cfm?userlg=en. Accessed July 2022

Senekal B, Kotzé E (2019) Open source intelligence (osint) for conflict monitoring in contemporary South Africa: challenges and opportunities in a big data context. Afr Secur Rev 28(1):19–37

Shan Y, Zhou X, Liu S et al (2020) Siamfpn: a deep learning method for accurate and real-time maritime ship tracking. IEEE Trans Circuits Syst Video Technol 31(1):315–325

Shestak VA, Koscheeva DA (2021) Cyber crime during the COVID-19 pandemic: key trends and ways to solve the problem Criminal policy of the Russian Federation in the field of ensuring economic security Round table (20 October 2021). Institute of Legislation and Comparative Law, Moscow

Shuai Q, Huang Y, Jin L et al (2018) Sentiment analysis on chinese hotel reviews with doc2vec and classifiers. In: 2018 IEEE 3rd advanced information technology, electronic and automation control conference (IAEAC), IEEE. pp 1171–1174

Sinha T, Chowdhury T, Shaw RN et al (2022) Analysis and prediction of covid-19 confirmed cases using deep learning models: a comparative study. In: Advanced computing and intelligent technologies. Springer, pp 207–218

Smith (2019) Geosetter is a freeware tool for windows for showing and changing geo data and other metadata. http://www.geosetter.de/en

Smithrae (2021) Investigating maritime vessels. https://wondersmithrae.medium.com/6-tips-for-investigating-maritime-vessels-77e9c8bf75

SmugMugInc (2022) Protect, share, store, and sell your photos. https://www.smugmug.com/. Accessed July 2022

Soomro ZT, Ilyas SHW, Yaqub U (2020) Sentiment, count and cases: analysis of twitter discussions during covid-19 pandemic. In: 2020 7th international conference on behavioural and social computing (BESC), IEEE. pp 1–4

Spadon G, Ferreira MD, Soares A et al (2022) Unfolding ais transmission behavior for vessel movement modeling on noisy data leveraging machine learning. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3197215

Stafford CD (2020) Weakest link: assessing factors that influence susceptibility to falling victim to phishing attacks and methods to mitigate. PhD thesis, Utica College

StartPage (2006) Startpage—private search engine. no tracking. No search history. https://www.startpage.com/eng/video.html

Stieglitz S, Mirbabaie M, Ross B et al (2018) Social media analytics-challenges in topic discovery, data collection, and data preparation. Int J Inf Manage 39:156–168

Sulak J (2005) World’s largest flight tracking and data platform. https://uk.flightaware.com/

Tandale KD, Pawar SN (2020) Different types of phishing attacks and detection techniques: A review. In: 2020 international conference on smart innovations in design, environment, management, planning and computing (ICSIDEMPC), IEEE. pp 295–299

Tang H (2006) People search: white pages: reverse phone lookup. https://www.spokeo.com/

Technologies H (1997) Tools for investigating, exploring, and troubleshooting internet resources such as domain names, ip addresses, email addresses, and urls. https://hexillion.com/

Thatsthem (2014) Find people by name: free people finder: Thatsthem. https://thatsthem.com/people-search

Thompson S (2022) Track and trace track your item royal mail group ltd. https://www.royalmail.com/track-your-item. Accessed July 2022

Tineye (2008) Tineye reverse image search. http://www.tineye.com/

Track Trace (1998) Air cargo tracking. http://www.track-trace.com/aircargo

Trackingex (2022) All-in-one package tracking. https://www.trackingex.com/. Accessed July 2022

TripGeo (2022) Directions map with animated street view. http://www.tripgeo.com/Directionsmap.aspx. Accessed July 2022

TruePeopleSearch (2022) True people search. https://truepeoplesearch.io/. Accessed July 2022

Truthfinder (2015) Popular people search tools. https://www.truthfinder.com/

Usersearch Team (2022) Find someone by username or email on social networks, dating sites, forums, crypto forums, chat sites and blogs. 600+ sites supported! largest reverse user search online! https://usersearch.org/lookups.php. Accessed July 2022

USGS (2022) Earthexplorer. https://earthexplorer.usgs.gov/. Accessed July 2022

USP (2008) Package trackr best site to track a package. https://www.packagetrackr.com/track/usps

Utrace (2022) Locate ip addresses and domainnames. http://en.utrace.de/. Accessed July 2022

Valluripally S, Sukheja D, Ohri K et al (2019) Iot based smart luggage monitor alarm system. In: International conference on internet of things and connected technologies. Springer, pp 294–302

Verma D, Yarlagadda R, Gartner S et al (2019) Understanding patterns of terrorism in India (2007–2017) using artificial intelligence machine learning. Int J Technol Knowl Soc 15(4):23–39

VesselFinder (2011) Free ais ship tracker. https://www.vesselfinder.com/

ViaMichelin (2022) Route planner, maps, traffic info, hotels. https://www.viamichelin.com/. Accessed July 2022

Virginia (2007) Mission is to help people discover, understand and use public data. https://www.beenverified.com/

VisualRightsGroup (2003) Home. https://www.epa.eu/

Volza (2022) Global export import trade data of 209 countries. https://www.volza.com/. Accessed July 2022

VoPham T, Hart JE, Laden F et al (2018) Emerging trends in geospatial artificial intelligence (geoai): potential applications for environmental epidemiology. Environ Health 17(1):1–6

Wachs S, Wright MF, Vazsonyi AT (2019) Understanding the overlap between cyberbullying and cyberhate perpetration: moderating effects of toxic online disinhibition. Crim Behav Ment Health 29(3):179–188

Wadawadagi R, Pagi V (2020) Sentiment analysis with deep neural networks: comparative study and performance assessment. Artif Intell Rev 53(8):6155–6195

Wang Y, Liu Y, Wu T et al (2020) A cost-effective ocr implementation to prevent phishing on mobile platforms. In: 2020 international conference on cyber security and protection of digital services (Cyber Security), IEEE. pp 1–8

Wang Y, Rajesh G, Mercilin Raajini X et al (2021) Machine learning-based ship detection and tracking using satellite images for maritime surveillance. J Amb Intell Smart Environ 13:1–11

Webmii (2022) Public information available on the web. http://webmii.com/. Accessed July 2022

WiGLE (2001) Wireless network mapping. https://wigle.net/

Williams HJ, Blum I (2018) Defining second generation open source intelligence (osint) for the defense enterprise. Rand Corporation, Santa Monica

Wilson J (1960) Corona atlas and referencing system. http://corona.cast.uark.edu/

Wise B (2008) Checkusernames. https://knowem.com/

Wolfram S (1987) The wolfram language image identification project. https://www.imageidentify.com/

Yadu R, Shukla R (2020) Sentimental classification method of twitter data for Indian air Asia services analysis. Tech. rep, EasyChair

Yahoo (2022) Yahoo video search. https://video.search.yahoo.com/. Accessed July 2022

YahooImage (2022) Yahoo image search. http://images.yahoo.com/. Accessed July 2022

Yandex (2022) Yandex maps. http://maps.yandex.com/. Accessed July 2022

YandexImages (2022) Search for images online or search by image. https://yandex.com/images. Accessed July 2022

Yang X, Wang Y, Wang N et al (2022) An enhanced siammask network for coastal ship tracking. IEEE Trans Geosci Remote Sens 60:1–11. https://doi.org/10.1109/TGRS.2021.3122330

Yasni G (2022) Yasni is a free search engine dedicated to finding people on the web. http://www.yasni.com/. Accessed July 2022

Yuan H, Zheng J, Ye Q et al (2021) Improving fake news detection with domain-adversarial and graph-attention neural network. Decis Support Syst 151(113):633

Zarringhalam N (2009) The world's best caller id and spam blocking app. https://www.truecaller.com/

Zhang W, He X, Li W et al (2021) A robust deep affinity network for multiple ship tracking. IEEE Trans Instrum Meas 70:1–20. https://doi.org/10.1109/TIM.2021.3077679

Zlookup (2022) Lookup the owner of any mobile or cell phone using zlookup. https://www.zlookup.com/. Accessed July 2022

ZoomEarth (2022) Live weather map, storm tracker, rain radar. https://zoom.earth/. Accessed July 2022

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.