Routledge
Taylor & Francis Group

Journal of Applied Security Research

ISSN: 1936-1610 (Print) 1936-1629 (Online) Journal homepage: https://www.tandfonline.com/loi/wasr20

Systematic Literature Review to Investigate the Application of Open Source Intelligence (OSINT) with Artificial Intelligence

João Rafael Gonçalves Evangelista, Renato José Sassi, Márcio Romero & Domingos Napolitano

To cite this article: João Rafael Gonçalves Evangelista, Renato José Sassi, Márcio Romero & Domingos Napolitano (2020): Systematic Literature Review to Investigate the Application of Open Source Intelligence (OSINT) with Artificial Intelligence, Journal of Applied Security Research, DOI: 10.1080/19361610.2020.1761737

To link to this article: https://doi.org/10.1080/19361610.2020.1761737

JOURNAL OF APPLIED SECURITY RESEARCH
https://doi.org/10.1080/19361610.2020.1761737

Routledge
Taylor & Francis Group

Systematic Literature Review to Investigate the Application of Open Source Intelligence (OSINT) with Artificial Intelligence

João Rafael Gonçalves Evangelista, Renato José Sassi, Márcio Romero, and Domingos Napolitano

Programa de Pós-graduação em Informática e Gestão do Conhecimento, Universidade Nove de Julho, São Paulo, Brazil

ABSTRACT

Open Source Intelligence (OSINT) is a concept to describe the search, collection, analysis, and use of information from open sources, as well as the techniques and tools used. OSINT emerges out of a military need to collect relevant and publicly available information. Through the use of OSINT, it is possible to find specific information that has some knowledge or provides an advantage. Since its emergence, some studies have been done proposing and developing new ways of using OSINT in different areas. In addition to OSINT, another field of study that has also been a worldwide trend and is being used together with other areas is Artificial Intelligence (AI). AI is the area of computer science responsible for the development of intelligent systems. However, a systematic literature review that investigates the use of OSINT over the years and your application with AI was not found. So, this work has an objective to develop a systematic literature review on OSINT to investigate the application of OSINT with AI. This work was motivated to fill this research gap, for this, consolidate the publications on OSINT divided into the publication bases. As for its contribution, this work presents a systematic literature review composed of 9-step and also brings consolidated information to support the next OSINT studies. This research searched for publications between January 1990 and October 2019, finding a total of 244 publications. The 9-steps of the systematic literature review are Definition of Keywords, Query string definition, the definition of publication bases, the search on the publications bases, the base search results analysis, download of publications, importing the publications into Mendeley, Importing. Ris file into VOSviewer and Keyword Map Analysis. Analyzing the results, we find some relevant information about the publications that address OSINT and OSINT with AI or other areas. With this information, it was possible to understand where the largest concentration of publications, which countries and continents develop the most research and the characteristics of these publications. What are the trends for the next studies on OSINT with AI. Which AI subareas are used with OSINT. What are the most used keywords, and how do these keywords relate to others

KEYWORDS

Systematic literature review; OSINT; Open Source Intelligence; Artificial Intelligence

CONTACT João Rafael Gonçalves Evangelista jrafa1607@gmail.com Programa de Pós-graduação em Informática e Gestão do Conhecimento, Universidade Nove de Julho, São Paulo, Brazil.

over the years. Which publication bases have the highest concentration of publications and what are the types of these publications? Also, a timeline describing the application of OSINT. It also became evident how OSINT has been used with AI to solve problems in different areas with different objectives. Based on these results, it is concluded that the application of a systematic literature review can show the application of OSINT with AI.

1. Introduction

The internet and other media sources have been increasingly used for publishing and sharing information. The ease of access to the internet allows people to find and publish any type of information more easily (Edwards et al., 2017). Although the publication of information has significant advantages, such as the collection of information for researchers in different areas such as marketing, social psychology and, information security. There is also the disadvantage of information security problems, for example, in the preservation of privacy (Medkova, 2018).

This makes organizations dependent on information security to protect their information. This dependence coupled with a large amount of sensitive data and information available on the internet has given rise to standards, methods, services, tools, and technologies to assist the management and practice of information security (Haufe et al., 2016; Haqaf & Koyuncu, 2018).

Government companies, organizations, and the public, in general, contribute to this growing volume of the information displayed on the Internet (Ghazi et al., 2018). To search and explore that information, several tools, techniques, and approaches are available on the Internet that performs searches on open sources. These searches happen on the most varied sources, such as code blocks, social media, academic pages, and social networking sites (Chen & Décary, 2018; McKeown et al., 2016).

Open Source Intelligence (OSINT) is the concept used to describe the search and acquisition of information from publicly available sources, as well as the techniques and tools used (Glassman & Kang, 2012). OSINT can be employed in a wide variety of open sources, such as social media, government reports, geolocations, code sites, social networks, satellite images, academic publications, vulnerabilities database, as well as a host of other information available through the internet and others open media features (Quick & Choo, 2018; Settanni et al., 2017).

So, the extent to which the information can be found and collected is large and open sources where this information meet is diverse (Watters & Layton, 2016). OSINT is still considered an emerging area, also to

information, the developed intelligence from the collection and analysis of publicly available information and open-source (Magalhães & Magalhães, 2018).

Another area that is also a global trend, wide and is being applied together with other areas, is Artificial Intelligence (AI). According to Dwivedi et al. (2019), literature offers some AI definitions, in general, the authors describe AI as an area of computing science responsible for the development of systems and/or machines capable of performing functions such as Learning and solving problems, functions that are used only by human beings.

The AI area is divided into subareas. Can cite as an example of these subareas the Natural Language Processing, Pattern Recognition, Robotics, and Machine Learning. All these subareas of AI can be applied in several other areas such as Health, agriculture, industry, information security, marketing, human resources, and education (Vijayakumar & Sheshadri, 2019). So, being OSINT as wide-reaching and diverse types of information in several open sources and, the area of AI being a worldwide trend, it is necessary a study that consolidates publications on OSINT that can investigate your application with AI.

Thus, this work has an objective to develop a systematic literature review on OSINT to investigate your application with AI. Regarding contributions to organizations and society, this work shows some trends for the application of OSINT with AI. These trends may offer business opportunities and the development of new studies on OSINT, making this work to be carried out based on the next studies on OSINT, such as the application of OSINT in areas where it has not yet been used.

2. Open Source Intelligence

OSINT is a concept that addresses the search, collection, processing, analysis, and use of information from open sources that can be legally accessed by any individual or organization. OSINT locates, selects, and extracts that information from open sources such as Twitter and LinkedIn, as well as analyzing this information to produce intelligence by how it can be used (Koops et al., 2013; Howells & Ertugan, 2017). Two important terms are related to OSINT by the US Department of the Army & US Department of the Army (2012). They are "Open sources" and "Publicly information available."

- Open sources: Any person, group or system that provides information without expectation of privacy. The information contained in these sources is not protected against public disclosure. Although this

information is publicly available, it is not necessarily information that should be open.

- Publicly information available: Data, facts, instructions or other material published, shared or transmitted for general public consumption; legally observed by any individual.

OSINT application may involve other concepts that address information gathering for other purposes. These concepts are denominated "Intelligence Disciplines." Each of these intelligence disciplines described in Table 1 can act in conjunction with OSINT or even be part of OSINT. Table 1 presents the intelligence disciplines and their descriptions (Evangelista et al., 2019).

Each of these intelligence disciplines is specific tasks for intelligence services. In each of these tasks, OSINT can act together, or even, be an essential element for this task to be completed. For example, for the GEOINT discipline, where geospatial information is searched, OSINT can be a fundamental tool to search for this type of information, using resources such as Google Maps.

The study of OSINT emerged as a trend that involved the business, military, and political intelligence. Conferences on the subject are held, large organizations are using OSINT strategically and academics are researching techniques to work in conjunction with OSINT (Yates & Zvegintzovi, 1999). The practice of data and information collection has been discussed since 1941 when German and Japanese radio broadcasts were launched with the creation of the Foreign Broadcast Monitoring Service, an organization that later became the Open Source Center.

From the creation of the Open Source Center to the present, numerous tools and techniques for collecting information from open sources have emerged that self-tune the search and analysis. These tools can provide various capabilities to find some sensitive information on websites, for example, Google Maps, Maltego, theHarvester, and, Carrot2 (Lee & Shon, 2016).

Table 1. Intelligence disciplines.

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

Although the Open Source Center came into being in 1941, OSINT was only defined in 2001 with the publication "The Open Source Intelligence Handbook" of the North Atlantic Treaty Organization (NATO). The publication defines OSINT as non-confidential information that has been deliberately discovered, broken down, distilled and disseminated to a target audience to address a specific issue (North Atlantic Treaty Organization, 2001).

Because OSINT is a fast and effective way of conducting in-depth security reviews, OSINT is increasingly being used by government agencies such as the Federal Bureau of Investigation (FBI), the Central Intelligence Agency (CIA), and Europol for criminal investigation purposes (Hayes & Cappa, 2018). If previously OSINT was associated only with government services, now OSINT is an area of interest for companies and individuals to provide specific information search and collection, analyze social media and find connections between entities, people, and organizations (Maciołek & Dobrowolski, 2013).

In the past, the task of those responsible for analyzing and executing OSINT was to find information that is hidden or difficult to find. Nowadays, in the face of increasing volumes of information, the other challenge is to find pertinent information, that is, information that brings knowledge. This large-scale sharing and dissemination of information are due to increased use of the Internet (Nicart et al., 2016).

So, the exploration of internet content is an indispensable asset to support the execution of OSINT. Its range allows for a rapid spread of information. Much of your content is searchable and available from open sources. The number and variety of tools on the Internet allow communications to be established not only through email and instant messaging, but also with social media coming from Web 2.0, such as Twitter and Facebook. This increased flexibility and, conversely, reduced face-to-face communication explains the growing need for security and availability of data and information over the internet. (Zunino et al., 2013; Krombholz et al., 2015).

Researchers from the OSINT, especially computer scientists, are following a different path from the traditional OSINT for military purposes. Academic researchers are concerned about the development of systems for processing, analysis, and auto-run to OSINT. As OSINT involves the production of intelligence from a large amount of data, the AI techniques used for data mining are easily employed with OSINT. Besides, natural language processing systems and algorithms have also been used for data structuring, automatic translation and, extraction of information, and additional analysis of results (Yang & Lee, 2012; Noubours et al., 2013).

3. Artificial Intelligence

Artificial Intelligence (AI) is a field of computer science oriented toward the development of computers and intelligent systems. To achieve this goal, an AI has maintained a useful area about research and development since its birth at the 1956 Dartmouth conference in the United States (Cantu-Ortiz, 2014; Mckinnel et al., 2019). According to Talwar and Koury (2017) and Mckinnel et al. (2019), AI has the potential to act in different areas. The inspiration for the development of intelligent systems and computers comes from aspects present in nature such as speech recognition, language translation, visual perception, learning, reasoning, planning, decision making, and intuition.

In many countries, specially developed, the production of studies involving the application of AI in several areas is encouraged. This makes other countries understand the importance that the area of AI plays for the development of Science, Research and Technology. With this, it can be said that the world has entered the era of AI, as its development is closely related to international modernization and directly affects the interests of countries and their peoples (Fu, 2019).

Naveen et al. (2019) present in their study the benefits of applying AI. Examples of these benefits can be cited: Systems more powerful than conventional systems and/or algorithms, ability to solve new problems, better engagement with information, greater ability to extract knowledge from databases, low error rate compared to human beings, this is clear if AI is programed correctly.

To understand the importance of the AI area, just view a variety of areas where it can work together. With the continuous development of technologies such as the Internet, Big Data, Cloud Computing, and the Internet of Things, computational resources have evolved along with the volume of data available. This variety of technologies allows AI to be applied in several areas such as industry, agribusiness, medicine, information security, human resources, transportation, public security, entertainment, education, among others (Vijayakumar & Sheshadri, 2019; Zhang et al., 2019). As it is wide, an area of AI has some sub-areas. Among these sub-areas, we can consider Natural Language Processing, Expert Systems, Computer Vision. Speech Recognition, Machine Learning, Robotics and Fuzzy Systems (Naveen et al., 2019).

4. Research Method

The research method adopted in this work was the systematic literature review. A systematic literature review consists of gathering evidence from previously published material, consisting mainly of books, journal articles

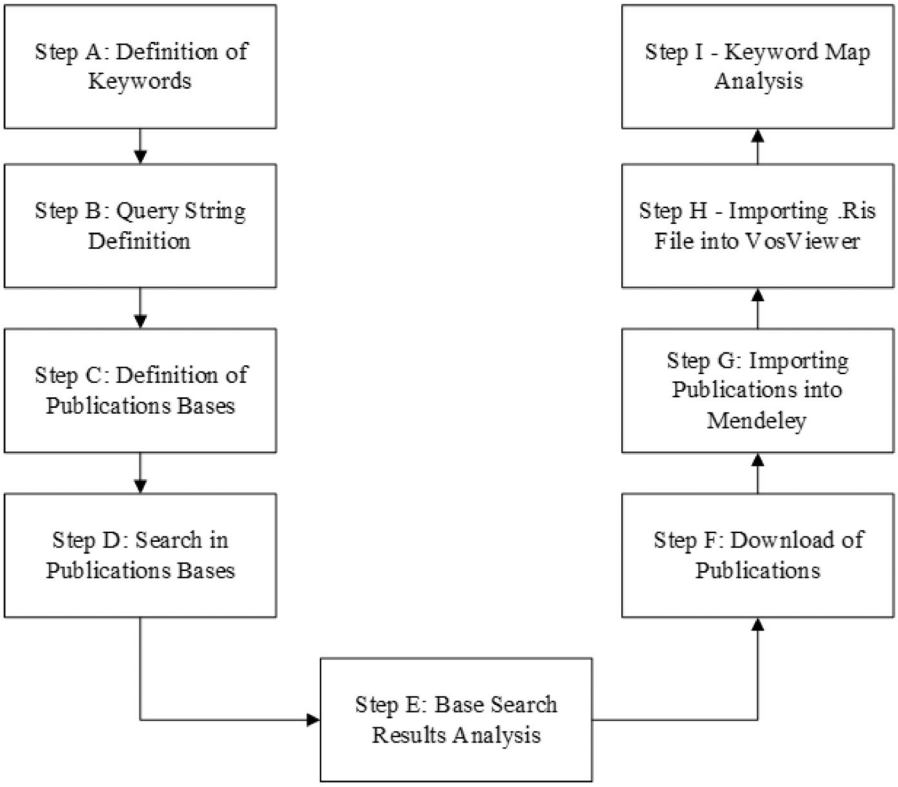

Figure 1. Systematic literature review 9-steps.

available in academic databases, without any bias of the researcher that may occur in a research subjective (Denyer & Tranfield, 2009). As for the type, this research is characterized as exploratory with a qualitative approach, as it seeks to understand the evolution of OSINT through the publications. According to Lakatos and Marconi (2003), exploratory research dealt with by systematic investigations whose objective is to apply a problem to develop hypotheses or increasing the researcher's familiarity with the research topic to understand his scenario for future studies or, to modify and clarify concepts.

This systematic literature review was divided into 9-steps. These 9-steps are shown in Figure 1.

The 9-steps of the systematic literature review are described below.

Step A—Definition of Keywords: Define the keywords to use them to search the databases.

Step B—Query String Definition: Define the search query with the keywords previously defined in Step A to search for publications.

Step C—Definition of Publications Bases: Define the bases that have publications in the area of computer science to find publications with a query string defined in step B.

Step D—Search in Publication Bases: Search in the databases of publications defined in step C with the search query defined in step B.

Step E—Base Search Results Analysis: Analysis of the results found in the search made in step D.

Step F—Download of Publications: Download the publications found in the search made in step D.

Step G—Importing publications into Mendeley: Imports of publications to Mendeley software (https://www.mendeley.com/) to generate a .Ris file with the information of the publications.

Step H—Importing .Ris File into VOSviewer: Importing the .Ris file with the publication information to the VOSviewer software (https://www.vosviewer.com/) to analyze the results.

Step I—Keywords Map Analysis: Analysis of the results produced by VOSviewer in step H.

5. Presentation and Discussion of Results

The following are the 9-steps developed in the research methodology.

Step A—Definition of Keywords: The first step consisted of determining the keywords. The keywords defined were: "OSINT" and "Open Source Intelligence" and "Inteligência de Fontes Abertas"—OSINT in Portuguese. We chose to search for the term OSINT in Portuguese to identify the national context of research on this item. The English term was chosen to try to find publications that address OSINT but do not use its acronym.

Step B—Query String Definition: In this step, the search query was defined with the previously defined keywords. The search query was defined as: "OSINT" OR "Open Source Intelligence" OR "Inteligência de Fontes Abertas." With this, we expected to find publications in databases for publications with OSINT in English and Portuguese, as well as publications containing the acronym.

Step C—Definition of Publications Bases: Next, the most relevant publication bases that address OSINT were defined. The bases defined were: ACM Digital Library, Emerald Insight, IeeeXplore – Digital Library, Portal Capes, Scielo, Science Direct, and Spell. Publications located on the Portal Capes include other bases, such as Scopus and Google Scholar. Set the bases for publications, it defined the following criteria: Be an article or chapter published in the previously defined bases, also the publication should address the issue of OSINT.

Step D—Search in Publication Bases: In this step, searches were made on the bases previously defined. The search was carried out in the period between 1990 and September 2019. The searches were carried out through

Table 2. Number of publications for bases.

Base	Selected	Disregarded	Total
ACM Digital Library	9	33	42
Emerald Insight	25	14	39
IEEEXplore – Digital Library	91	26	117
Portal Capes	46	11	57
Scielo	0	0	0
Science Direct	73	45	118
Spell	0	0	0
Total	244	109	353

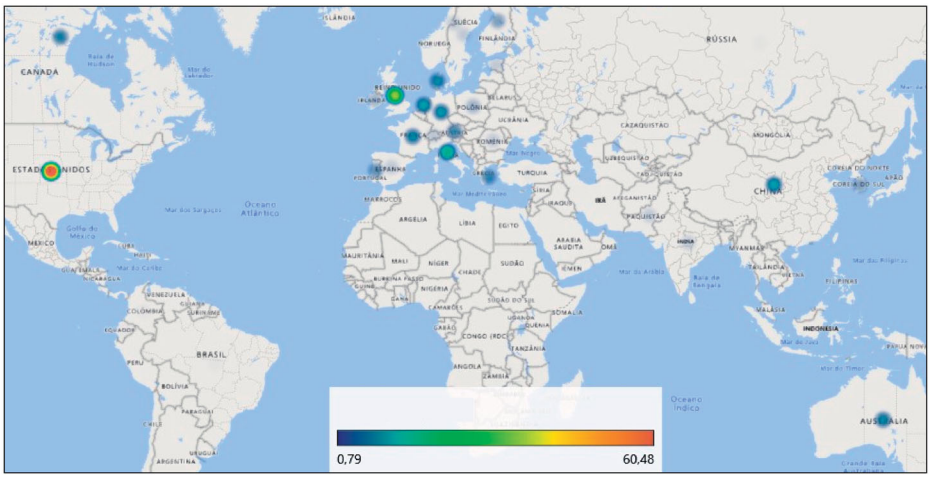

Figure 2. Number of publications for countries.

virtual access to the databases of journals. 353 publications were found using the research query defined in phase B.

Step E—Base Search Results Analysis: Of 353 publications found in step D, 109 results were disregarded because they are: Subjects that do not fit the OSINT theme, publications duplicated on different bases, indexes of publications or guides of publications and publications with the keywords only in the references. Table 2 shows the number of publications on the subject OSINT found in the publications base consulted.

Analyzing the results of Table 2, we can see that the bases with the highest concentration of publications on OSINT are based on Portal Capes, Science Direct and IEEExplore – Digital Library. Also, the ACM Digital Library database was the only consultant that had more disregarded publications than selected. This is because the database contains the largest number of publications ever contained in the other databases. Scielo and Spell do not have any OSINT publications in their database. Bases Scielo and Spell have mostly publications from South American countries. Because the South American continent is one of the least producing publications on OSINT, presented later in Figures 2 and 3, consequently, no research on the subject was found.

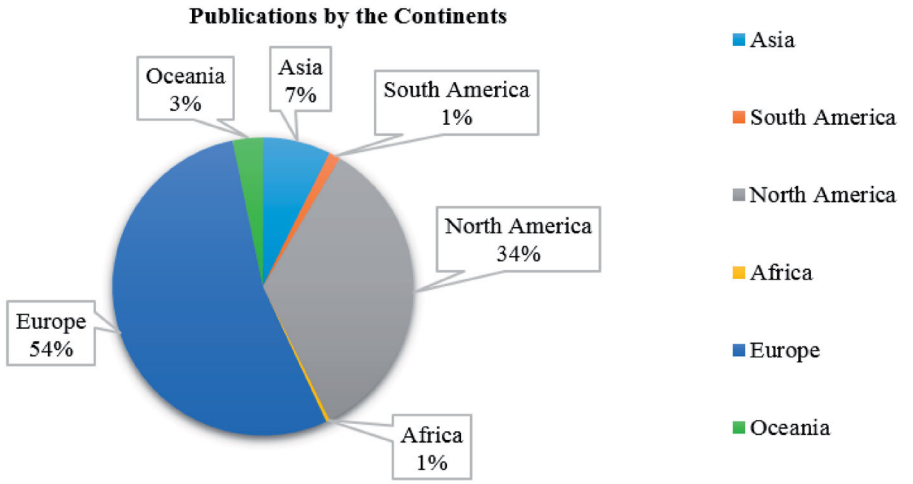

Figure 3. Number of publications per continents.

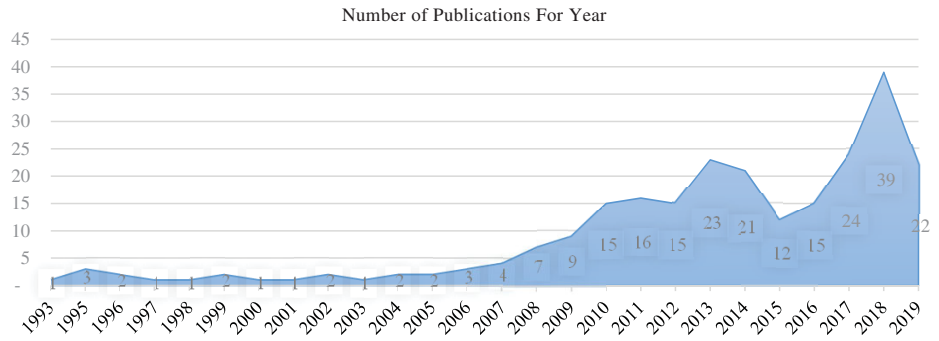

Figure 4. Number of publications per years.

Then, a temporal analysis was carried out, to be able to identify in which period the largest amount of publications on OSINT is found. It was verified in Figure 4 that the greatest concentration of publications occurred in the period between 2010 and 2019, and in 2018 is the largest amount of published works with the OSINT theme. With increased attacks, data leaks and other information security issues in 2018, note a large number of publications this year seeking alternatives to address these events.

Another important factor research over the past 10 years has been the rise of new technologies to store various types of data, like Data Warehouse and Big Data. Another point is the targeted volume of users and organizations signing up on new social networks and migrating their information to cloud storage, a trend that is growing steadily. Figure 4 illustrates the number of publications per year.

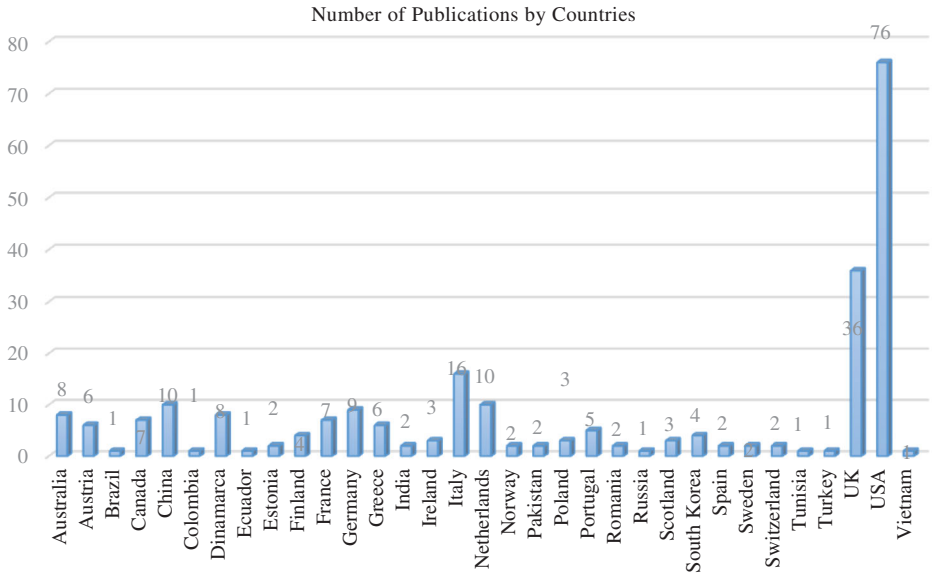

Figure 5. Number of publications per years.

In addition to the number of publications per year, the geographic classification was then performed to identify the concentration of publications on OSINT. For this classification by country, the origin of the first author responsible for the publication was verified in the publications to classify then. Thus, if the publication had authors from more than one country, the origin of the corresponding author would be selected. Figure 5 shows the number of publications in each country.

Figure 2 shows that the largest concentration of publications is in the USA with a total of 76 publications. The other countries with a large number of publications compared to the others are the UK, China, Italy, and the Netherlands. The caption in Figure 6 shows a scale ranging from 0.79 with a blue color to 60.48 with a red color. The blurred blue dots are countries with few publications, with at least one, while red shows the countries with the highest concentration of publications, with at least 60 publications.

The USA has the largest number of publications on a wide range of subjects with OSINT. At the outset of OSINT research or even the practice of collecting information from open sources, the publications focused on detecting and combating terrorism. Currently, OSINT research is less targeted at counter-terrorism and more focused on social media and information security. In the OSINT publications in the UK, there is great concern about using OSINT to combat terrorism and cyber-attacks, as well as work to improve the communication of government forces. As for China, most of the publications involve the automation of OSINT with platforms, frameworks, or even information systems

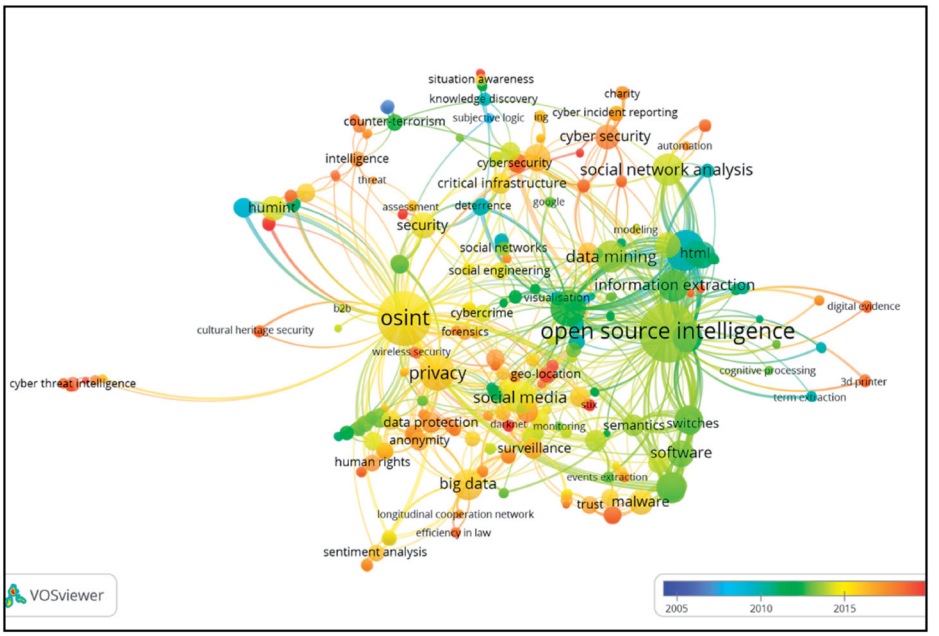

Figure 6. Keywords map.

supported by machine learning techniques for text mining, device classification, and pattern recognition.

As for publications in Italy, the country follows the same pattern as China, using machine-learning algorithms to improve OSINT's performance. The difference that the publications in Italy are focused on text mining of big data and in social media, as well as ontology and semantics. Finally, in the Netherlands, the vast majority of research involves OSINT with privacy on the internet, support for police investigations, and law enforcement. Although the country with the largest concentration of publications on OSINT is the USA, the continent with the most publications is the European continent with 54% of the publications, totaling 131 publications. Figure 3 shows the percentages of each continent.

The classification of publications by type of research was realized. For this classification of publications for your type in descriptive, exploratory, or explanatory, the abstract, methodology, and the results were verified. Publications that presented quantitative results or that addressed the application of OSINT in any scenario were classified as Descriptive. Publications that addressed a context more for discussion with hypotheses and analyzes were classified as Exploratory. Research that had the characteristics of trying to explain how OSINT happens or how tasks within OSINT happens was classified as Explanatory.

A great balance between descriptive and exploratory research is evident. While the exploratory totals 127 publications, the descriptive ones have 117

publications. No publication found took an explanatory approach. Descriptive approach publications deal with tools, applications, or even techniques for handling the information discovered by OSINT. The exploratory publications, however, deal more with the impact that OSINT can have when it is applied in a given environment in some circumstances.

Step F—Download of Publications: After analyzing the results in the periodic bases, each publication was downloaded and inserted into a directory 244 publications from the databases described in step C were downloaded.

Step G—Importing publications into Mendeley: Once the files were downloaded, they were imported into Mendeley software. With Mendeley, we exported a .Ris file containing the main information of each publication, such as Title, where it was published, authors, abstract, keywords, number of pages, year of publication, among other information.

Step H—Importing .Ris File into VOSviewer: After exporting the .Ris file, this file was imported into the VOSviewer software. Using VOSviewer, a relationship map was produced between the keywords contained in OSINT publications. The map is illustrated in Figure 6.

Step I—Keywords Map Analysis: Analyzing Figure 6, we can see that the keywords in the publications are: "OSINT" and "Open Source Intelligence." Both keywords are concentrated between 2014 and 2016. The oldest keywords come from publications up to 2005. From the first publication about collecting data to this period, OSINT publications are focused on extracting information and finding knowledge. Other older publications address this extracting information to using for combat and detect terrorism, but a large portion of these publications do not have keywords.

From 2012 to 2014, the publications are oriented in the analysis of social networks and social media, the first publications on OSINT in this period supported by techniques of text mining and sentiment analysis. Research began to address the visualization of data located on social media to extract knowledge. From these publications, the research begins by directing OSINT to the area of information security. Just as this information was being extracted to generate knowledge, it could be collected for malicious purposes.

From 2014 to 2016 the concentration of the publications in OSINT happens in the area of big data technology and information security. Now, machine learning algorithms and natural language processing are implemented to extract knowledge of large volumes of data and also to provide protection against data leakage, cyber-incidents, and social engineering, with a focus on information privacy and human rights.

Finally, from 2016 to 2019, machine learning algorithms operate in more specific areas with OSINT. The publications generally address the use of

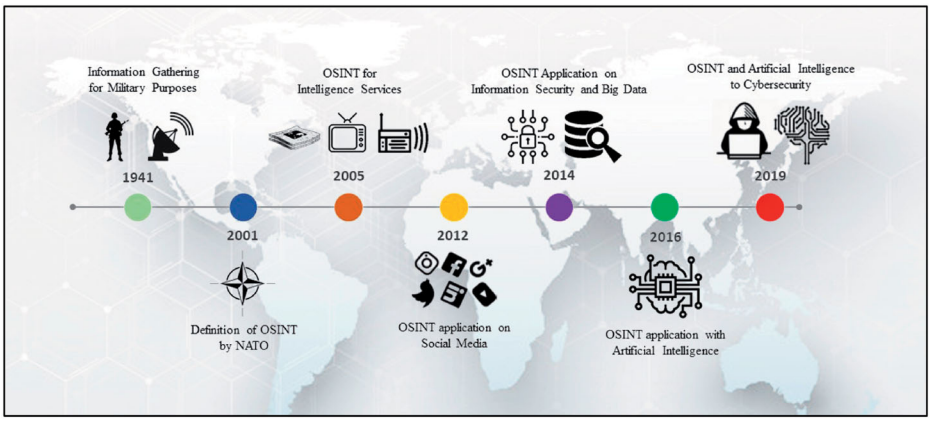

Figure 7. OSINT timeline.

OSINT tools with machine learning algorithms for the extraction of knowledge. The publications from 2016 to 2019 are focused on applying laws, search information in the cybersecurity area, more specifically about anonymity and Darknet, and using OSINT to obtain information such as geolocation, cyber threat intelligence, wireless security, and digital evidence.

Figure 7 presents a timeline about OSINT with the information discovered by the keyword map.

So, in the period between 2014 and 2019, the use of machine learning and natural language processing with OSINT grow up. To highlight this, Table 3 shows the publications that address the applications of AI with Natural Language Processing or Machine Learning with OSINT.

To evidence this, we searched in the Mendeley for publications that contained information about AI or any of its subareas. For this, the following search query was defined: "Artificial Intelligence" OR "Natural Language Processing" OR "Pattern Recognition" OR "Robotics" OR "Machine Learning." This search query was defined based on the study by Vijayakumar and Sheshadri (2019) referenced in this work.

We found 59 publications on OSINT that mention in their text these words used in the query. The citations of AI in publications ranging from applications in the execution of OSINT, in the analysis of the results, obtained, or even as a recommendation for future work. With a total of 59, this represents 24% of the 244 publications identified in this study, as shown in Figure 8.

Table 3 describes the title, authors and the year of each publication, in total there are 59 publications about OSINT with AI.

With the data in Table 3, we have produced a graph to understand in which area is the highest concentrion of OSINT publications with AI. This result is described in Figure 9.

Table 3. Publications with OSINT and Artificial Intelligence.

Title of publication	Applications area	Authors	Year
An investigation of using classification techniques in prediction of type of targets in Cyber attacks	Cybersecurity	Sina Pournouri, Shahrzad Zargari, Babak Akhgar	2019
BlackWidow: Monitoring the Dark Web for Cyber Security Information	Cybersecurity	Matthias Schafer, Markus Fuchs, Martin Strohmeier, Markus Engel, Marc Liechti, Vincent Lenders	2019
Cognitive security: A comprehensive study of cognitive science in cybersecurity	Cybersecurity	Roberto O Andrade, Sang Guun Yoo Facultad	2019
Design of a Classification Model for a Twitter-based Streaming Threat Monitor	Cybersecurity	Fernando Alves, Pedro M. Ferreira, Alysson Bessani	2019
Developing insights from social media using semantic lexical chains to mine short text structures	Social Media	Cecil Eng Huang Chua, Veda C. Storey, Xiaolin Li, Mala Kaul	2019
Enhancing Information Sharing and Visualization Capabilities in Security Data Analytic Platforms	Cybersecurity	Gustavo Gonzalez-Granadillo, Mario Faiella, Ibéria Medeiros, Rui Azevedo, Susana Gonzalez-Zarzosa	2019
Localising social network users and profiling their movement	Cybersecurity	Hector Pellet, Stavros Shiaeles, Stavros Stavrou	2019
Searching for Extremist Content Online Using the Dark Crawler and Sentiment Analysis	Military Purposes	Ryan Scrivens, Tiana Gaudette, Garth Davies, Richard Frank	2019
Turkish national cyber-firewall to mitigate countrywide cyber-attacks	Cybersecurity	Arif Sari	2019
A Supervised Machine Learning Based Approach for Automatically Extracting High-Level Threat Intelligence from Unstructured Sources	Cybersecurity	Yumna Ghazi, Zahid Anwar, Rafia Mumtaz, Shahzad Saleem and Ali Tahir.	2018
A survey on technical threat intelligence in the age of sophisticated cyber attacks	Cybersecurity	Wiem Tounsi, Helmi Rais	2018
Detecting Network Threats using OSINT Knowledge-based IDS	Cybersecurity	Ivo Vacas, Ibéria Medeiros, Nuno Neves	2018
Evaluating Automated Facial Age Estimation Techniques for Digital Forensics	Cybersecurity	Felix Anda, David Lillis, Nhien-An Le-Khac, Mark Scanlon	2018
Impact of AnonStalk (Anonymous Stalking) on users of Social Media: a Case Study	Cybersecurity	V. Kanakaris, K. Tzovelekis and D. V. Bandekas	2018
Is quantum computing becoming relevant to cyber-security?	Cybersecurity	Keegan Keplinger	2018
Managing cyber threat intelligence in a graph database	Cybersecurity	Seulgi Lee, Hyeisun Cho, Nakhyun Kim, Byungik Kim, Junhyung Park	2018
Modeling The Causes Of Terrorism From Media News: An Innovative Framework Connecting Impactful Events With Terror Incidents	Military Purposes	Truong Son Pham, Tuan-Hao Hoang	2018
Ontology population for open-source intelligence: A GATE-based solution	Languages and Translations	Giulio Ganino, Domenico Lembo, Massimo Mecella, Federico Scafoglieri	2018

(*continued*)

Table 3. Continued.

Title of publication	Applications area	Authors	Year
Open source intelligence (OSINT) as support of cybersecurity operations. "Use of OSINT in a colombian context and sentiment Analysis"	Cybersecurity	Ricardo Andrés Pinto Rico, Martin José Hernández Medina, Cristian Camilo Pinzón Hernández, Daniel Orlando Díaz López, Juan Carlos Camilo García Ruíz	2018
Security OSIF: Toward Automatic Discovery and Analysis of Event Based Cyber Threat Intelligence	Cybersecurity	Ke Li, Hui Wen, Hong Li, Hongsong Zhu, Limin Sun	2018
Using Deep Neural Networks to Translate Multi-lingual Threat Intelligence	Languages and Translations	Priyanka Ranade, Sudip Mittal, Anupam Joshi and Karuna Joshi	2018
Applying fuzzy logic for sentiment analysis of social media network data in marketing	Social Media	Karen Howellsa, Ahmet Ertugan	2017
Classification of Colloquial Arabic Tweets in real- time to detect high-risk floods	Languages and Translations	Waleed Alabbas, Haider M. al-Khateeb, Ali Mansour, Gregory Epiphaniou, Ingo Frommholz	2017
Cloud security issues and challenges: a survey	Cybersecurity	Ashish Singh, Kakali Chatterjee	2017
Extracting Cyber Threat Intelligence From Hacker Forums: Support Vector Machines versus Convolutional Neural Networks	Cybersecurity	Isuf Deliu, Carl Leichter, Katrin Franke	2017
Toward a breakthrough Speaker Identification approach for Law Enforcement Agencies: SIIP	Languages and Translations	Khaled Khelif, Yann Mombrun, Gerhard Backfried, Farhan Sahito, Luca Scarpato, Petr Motlicek, Srikanth Madikeri, Damien Kelly, Gideon Hazzani, Emmanouil Chatzigavriil	2017
Utility and potential of rapid epidemic intelligence from internet-based sources	Social Media	S.J. Yan, A.A. Chughtai, C.R. Macintyre	2017
A problem shared is a problem halved: a survey on the dimensions of collective cyber defense through security information sharing	Cybersecurity	Florian Skopik, Giuseppe Settanni, Roman Fiedler	2016
Automating social network analysis: A power tool for counter-terrorism	Military Purposes	Leslie Ball	2016
Building Document Treatment Chains Using Reinforcement Learning and Intuitive Feedback	Industry	Esther Nicart, Bruno Zanuttini, Hugo Gilbert, Bruno Grilhères, Frédéric Praca	2016
How to Apply Privacy by Design in OSINT and big Data Analytics?	Cybersecurity	Jyri Rajamäki, Jussi Simola	2016
Sampling Labeled Profile Data for Identity Resolution	Social Media	Matthew Edwards, Stephen Wattam, Paul Rayson and Awais Rashid	2016
A Systematic Survey of Online Data Mining Technology Intended for Law Enforcement	Cybersecurity	Matthew Edwards, Awais Rashid, And Paul Rayson	2015
Social Opinion Mining : an approach for Italian language	Social Media	Vito Santarcangelo, Giuseppe Oddo, Maria Pilato, Fabrizio Valenti, Claudio Fornaro	2015
CAPER: Crawling and Analyzing Facebook for Intelligence Purposes	Social Media	Carlo Aliprandi, Antonio E. De Luca, Giulia Di Pietro, Matteo Raffaelli, Davide Gazzè, Mariantonietta N. La Polla, Andrea Marchetti, Maurizio Tesconi	2014

(*continued*)

Table 3. Continued.

Title of publication	Applications area	Authors	Year
Crawling Open-source Data for Indicators of Human Trafficking	Military Purposes	Ben Brewster, Timothy Ingle, Glynn Rankin	2014
Foraging Online Social Networks	Social Media	Gijs Koot, Mirjam A.A. Huis in 't Veld, Joost Hendricksen, Rianne Kaptein, Arnout de Vries, Egon L. van den Broek	2014
NLP as an Essential Ingredient of Effective OSINT Frameworks	OSINT	Sandra Noubours, Albert Pritzkau, Ulrich Schade	2014
OSINT for B2B platforms	Business	V.F. Pais, D.S. Ciobanu	2014
Semantic Crawling: an Approach based on Named Entity Recognition	Cybersecurity	Giulia Di Pietro, Carlo Aliprandi, Antonio E. De Luca, Matteo Raffaelli, Tiziana Soru	2014
The Big Data Imperative Air – Force Intelligence for the Information Age	Military Purposes	Col Shane P. Hamilton, Lt Col Michael P. Kreuzer	2014
TheGame: an evaluation on Self Organization & Engagement by semantic analysis	Languages and Translations	Giovanna Ferrari, Nicoletta Magnetti, Paolo Marianib, Federico Neri	2014
Can we trust this user? Predicting insider's attitude via YouTube usage profiling	Social Media	Miltiadis Kandias, Vasilis Stavrou, Nick Bozovic, Lilian Mitrou, Dimitris Gritzalis	2013
Cluo: Web-Scale Text Mining System For Open Source Intelligence Purposes	Languages and Translations	Przemysław Maciołek, Grzegorz Dobrowolski	2013
Massively Scalable Near Duplicate Detection in Streams of Documents using MDSH	Web Texts and Documents	Paul Logasa Bogen II, Christopher T. Symons, Amber McKenzie, Robert M. Patton, Robert E. Gillen	2013
Proactive Insider Threat Detection Through Social Media: The YouTube Case	Social Media	Miltiadis Kandias, Vasilis Stavrou, Nick Bozovic, Dimitris Gritzalis	2013
Automatic Exploitation of Multilingual Information for Military Intelligence Purposes	Military Purposes	Sandra Noubours, Matthias Hecking	2012
Hybrid model of content extraction	Web Texts and Documents	Pir Abdul Rasool Qureshi, Nasrullah Memon	2012
Challenges in Open Source Intelligence	Military Purposes	Clive Best	2011
Data mining with LinkedIn	Cybersecurity	Danny Bradbury	2011
LanguageNet: A Novel Frameworlk for Processing Unstructured Text Information	Web Texts and Documents	Pir Abdul Rasool Qureshi , Nasrullah Memon, Uffe Kock Wiil	2011
Desktop Text Mining for Law Enforcement	Languages and Translations	Jonathan Brett Crawley, Gerhard Wagner	2010
Detecting Terrorism Evidence in Text Documents	Military Purposes	Pir Abdul Rasool Qureshi, Nasrullah Memon, Uffe Kock Wiil	2010
Using Term Extraction Patterns to Discover Coherent Relationships from Open Source Intelligence	Cybersecurity	William L. Sousan, Qiuming Zhu, Robin Gandhi, William Mahoney, Anup Sharma	2010
WISDOM from Light-Weight Information Retrieval	Social Media	David B. Bracewell, Steven Gustafson, Abha Moitra and Gregg Steuben	2010
Near Real Time Information Mining in Multilingual News	Languages and Translations	M. Atkinson, E. Van der Goot	2009
Web Mining for Open Source Intelligence	Languages and Translations	Clive Best	2008
Toward an interoperable dynamic network analysis toolkit	Social Media	Kathleen M. Carley, Jana Diesner, Jeffrey Reminga, Maksim Tsvetovat	2007
Textually Retrieved Event Analysis Toolset	Military Purposes	John Palmer	2005

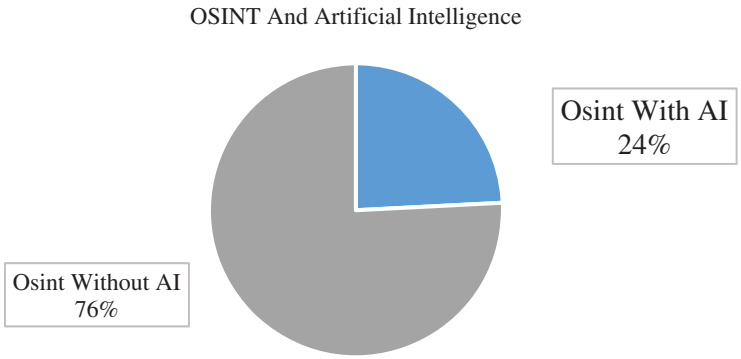

Figure 8. OSINT and Artificial Intelligence.

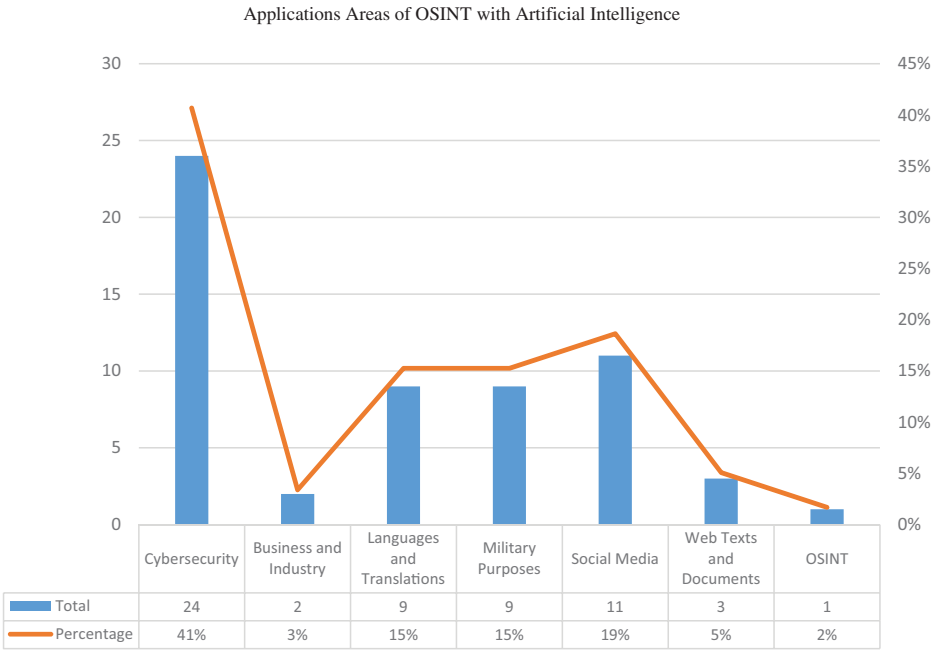

Figure 9. Applications areas of OSINT with Artificial Intelligence.

Analyzing Figure 9, it is possible to identify which are the areas where OSINT is being applied with AI. Of the 59 publications, only one is specific to the operation of OSINT, without addressing another area objectively. The largest concentration is in the Cybersecurity area, with a total of 24 publications, representing 41% of the total publications. This is more than double that of Social Media, the second application area with the highest concentration of publications, with 19%.

The areas of languages and translations appear in third place together with applications for military purposes, each one with 15% of publications. Finally, we have the Business and Industry areas and the Texts and Web Documents area with the lowest concentration of publications, representing

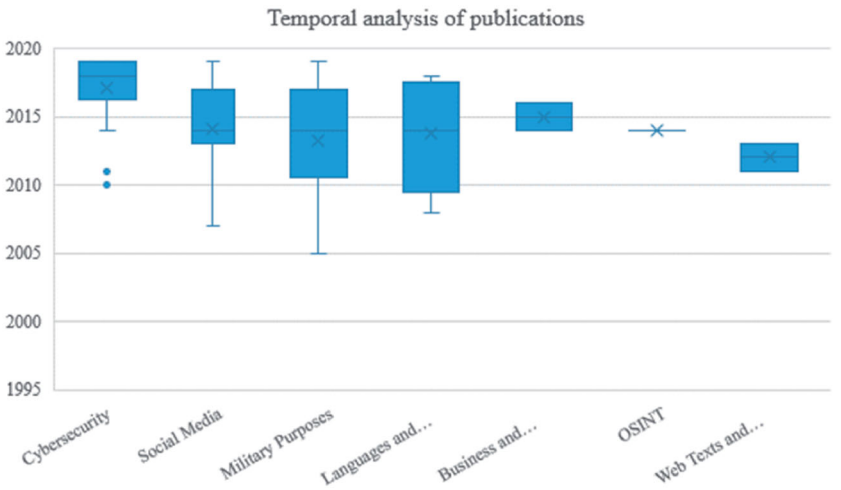

Figure 10. Applications areas of OSINT with Artificial Intelligence.

a total of 8%. In addition to this information, a temporal analysis was performed to understand how these OSINT publications with AI are distributed. This analysis is shown in Figure 10.

Analyzing Figure 10, it can be seen which years are the highest concentrations of publications for each topic. The works from 2005 to 2010 are works that only recommend AI as a solution, while the later works show effectively its application. Themes like Web Texts and Documents and Business and Industry have papers until 2016 only. Publications addressing social media, military purposes, and Languages and translations have the greatest disparity and advance a little further until 2017, representing the largest range of time. However, the area with publications in the most recent period, reaching 2019 is the area of Cybersecurity.

Considering that the volume of OSINT publications with AI presented interesting results, we performed again the steps G, H and I to which the keywords relate. It then split these publications into a directory and imported them into Mendeley software. With Mendeley, a .Ris extension file was exported, which was later inserted into VOSviewer to create the relationship map between keywords. These results are presented in Figure 11.

An important point to note is that the largest concentration of these publications is from 2010 to 2018, where publications address information security issues. This shows that AI has been studied as a solution to support the application of OSINT in the area of information security. Before 2010, OSINT publications only recommended AI as a possible solution to resolve some problems. The internet was beginning to rise and it was necessary to look for ways to collect this information distributed in different countries and concentrate it to extract knowledge.

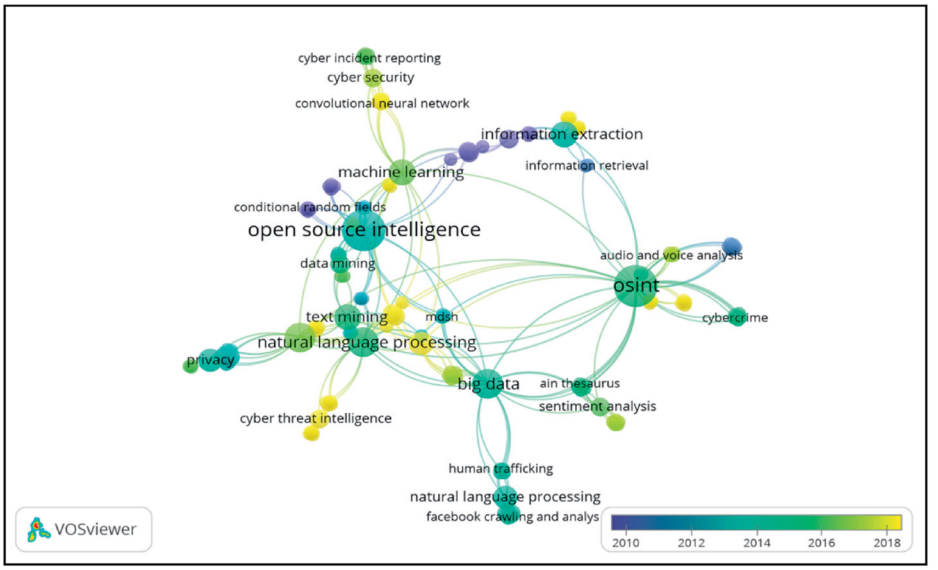

Figure 11. Keywords map about OSINT and Artificial Intelligence.

From 2010 to 2014, OSINT publications with an AI feature the use of natural language processing to extract knowledge from social media and Big Data to find important information about privacy, human trafficking, and cybercrimes. From 2014 to 2019 publications address not only natural language processing but also machine learning for sentiment analysis, text mining and media file analysis as a solution to cyber incident problems and cyber threat intelligence.

Therefore, analyzing the keywords it can be seen that OSINT has been working with the following areas: In the area of Information Security with words cybersecurity, cybercrime, and cyber threat intelligence. In the area of AI and Data Science with the words: Information Extraction, Natural Language Processing, Big Data, and Sentiment Analysis; and also Digital Marketing and Social Media Analysis with the words: Data Mining, Big Data, and Facebook Crawling & Analysis.

6. Summary and Conclusion

In this work, a systematic literature review was approached to investigate the applications of OSINT with AI. Analyzing the 244 publications found, it was a discovery that OSINT came up with military purposes for the discovery of publicly available information to support intelligence services. With the rise of the internet, OSINT started to be used to search the network to find specific information, that has some meaning and that can be used for its pre-defined purposes.

This systematic literature review shows which are the main article bases that have the most OSINT publications. The bases with the largest volume of publications are ScienceDirect with 74 publications and IeeeXplore – Digital Library with 91 publications. Also, it was possible to understand the highest concentrations of publications on OSINT. The country with the largest number of publications about OSINT is the USA, while the continent with the largest amount of publications on OSINT is Europe.

It was also found that there is a balance between OSINT publications as to the type of publication. While the descriptive and exploratory researches present similar numbers, explanatory publications about OSINT were not found in the researched literature, thus evidencing a research gap to be filled.

As for the application of OSINT to search for information on the internet, the main areas of application are Information Security, Digital Marketing and AI. The first OSINT publications dealt with intelligence, combat, and detection of terrorism, communications, and military technology. Regarding the execution of OSINT in conjunction with AI, the publications describe the use of machine learning algorithms and natural language processing to deal with the high volume of information present on the internet. The union between OSINT and AI happens through platforms, models, frameworks or systems. While machine learning algorithms are focused on running OSINT for performance and speed gains, while natural language processing is more used in analyzing the results discovered by OSINT.

In the area of information security, publications with OSINT search for specific information that can be used to generate new knowledge. With this knowledge, it is possible to improve processes and tasks involving subjects such as privacy, social engineering, cyber threat intelligence, law enforcement, collection of digital evidence, penetration tests, data leakage and security in wireless networks. As for the application of OSINT with AI and Information Security, publications seek ways to extract more information from the results obtained with OSINT and even, automate OSINT to obtain performance gains.

For the Digital Marketing area, publications with OSINT use analysis of information shared in social media to generate new knowledge. In this area, there is also the application of AI, mainly natural language processing to perform sentiment analysis to understand users' opinions. Other applications involve issues such as privacy and truth, as well as attempts to reduce the occurrence of "Fake News."

It was also analyzed how OSINT publications with AI are distributed. It turned out that as of 2015, it started to suggest the application of OSINT with AI. From then on, it started to be applied in Texts and documents

made available on the Web to extract some information. With the positive results, OSINT and AI began to be applied in the area of Languages and Translations, Military Purposes and Social Media and then reached its highest concentration: The Cybersecurity area.

The Cybersecurity area represents 41% of OSINT publications with AI, that is, almost half of the publications. This is an interesting value if we analyze that the growth of these publications starts to gain strength from 2016. Thus, only in these 4 years, going from 2016 to 2019, the publications of OSINT with AI for the area of Cybersecurity represent the area with the highest concentration of applications, a trend in recent years.

Analyzing these results, it is concluded that the application of a systematic literature review can show the application of OSINT with AI, and your trends, that is, the areas where OSINT is being applied on the world stage. As future work, it is recommended to investigate other bases of publications and to carry out analyses regarding the information of the authors of OSINT publications to understand the relationships and possible co-authorship projects. For the development of the next studies, it is recommended to investigate the application of OSINT with the main areas identified in this work: Information Security, Digital Marketing and AI.

Acknowledgments

We thank the Universidade Nove de Julho for the encouragement and support this work.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ORCID

João Rafael Gonçalves Evangelista http://orcid.org/0000-0003-3541-8354
Renato José Sassi http://orcid.org/0000-0001-5276-4895
Márcio Romero http://orcid.org/0000-0001-7595-9249
Domingos Napolitano http://orcid.org/0000-0001-5840-6757

References

Cantu-Ortiz, F. J. (2014). Advancing Artificial Intelligence research and dissemination through conference series: Benchmark, scientific impact and the MICAI experience. *Expert Systems with Applications*, *41*(3), 781–785. https://doi.org/10.1016/j.eswa.2013.08.008

Chen, M., & Décary, M. (2018). *A cognitive-based semantic approach to deep content analysis in search engines* [Paper presentation]. Semantic Computing (ICSC), 2018 IEEE 12th International Conference on (pp. 131–139). IEEE.

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. A. Buchanan & A. Bryman (Eds.), *The SAGE handbook of organizational research methods* (pp. 671–689). SAGE Publications Ltd.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Vigneswara Ilavarasan, P., Janssen, M., Jones, P., Kumar Kar, A., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … , Williams, M. D. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Edwards, M., Robert, L., Benjamim, G., Awais, R., & Baron, A. (2017). Panning for gold: Automatically analysing online social engineering attack surfaces. *Computers & Security*, *69*, 18–34. https://doi.org/10.1016/j.cose.2016.12.013

Evangelista, J. R. G., Gatto, D. D. O., & Sassi, R. J. (2019). Classification of web history tools through web analysis. *Leacture Notes in Computer Science*, *11594*, 266–276.

Fu, X. (2019). *Application of Artificial Intelligence technology in medical cell biology* [Paper presentation]. 2019 International Conference on Robots & Intelligent System (ICRIS) (pp. 401–404). IEEE. https://doi.org/10.1109/ICRIS.2019.00106

Ghazi, Y., Anwar, Z., Mumtaz, R., Saleem, S., & Tahir, A. (2018). *A supervised machine learning based approach for automatically extracting high-level threat intelligence from unstructured sources* [Paper presentation]. 2018 International Conference on Frontiers of Information Technology (FIT) (pp. 129–134). IEEE. https://doi.org/10.1109/FIT.2018.00030

Glassman, M., & Kang, J. (2012). Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT). *Computers in Human Behavior*, *28*(2), 673–682. https://doi.org/10.1016/j.chb.2011.11.014

Haqaf, H., & Koyuncu, M. (2018). Understanding key skills for information security managers. *International Journal of Information Management*, *43*, 165–172. https://doi.org/10.1016/j.ijinfomgt.2018.07.013

Haufe, K., Colomo-Palacios, R., Dzombeta, S., Brandis, K. & Stantchev, V. (2016). Security management standards: A mapping. *Procedia Computer Science*, *100*(100), 755–761. https://doi.org/10.1016/j.procs.2016.09.221

Hayes, D. R., & Cappa, F. (2018). Open-source intelligence for risk assessment. *Business Horizons*, *61*(5), 689–697. https://doi.org/10.1016/j.bushor.2018.02.001

Howells, K., & Ertugan, A. (2017). Applying fuzzy logic for sentiment analysis of social media network data in marketing. *Procedia Computer Science*, v. *120*, 664–670. https://doi.org/10.1016/j.procs.2017.11.293

Koops, B. J., Hoepman, J. H., & Leenes, R. (2013). Open-source intelligence and privacy by design. *Computer Law & Security Review*, *29*(6), 676–688. https://doi.org/10.1016/j.clsr.2013.09.005

Krombholz, K., Hobel, H., Huber, M., & Weippl, E. (2015). Advanced social engineering attacks. *Journal of Information Security and Applications*, *22*, 113–122. https://doi.org/10.1016/j.jisa.2014.09.005

Lakatos, E. M., & Marconi, M. A. (2003). *Fundamentos de metodologia científica* (5th ed.). Atlas.

Lee, S., & Shon, T. (2016). Open source intelligence base cyber threat inspection framework for critical infrastructures. *Future Technologies Conference (FTC)*, *2016*, 1030–1033.

Maciołek, P., & Dobrowolski, G. (2013). Cluo: Web-scale text mining system for open source intelligence purposes. *Computer Science*, *14*(1), 45–62. https://doi.org/10.7494/csci.2013.14.1.45

Magalhães, A., & Magalhães, J. P. (2018). *TExtractor: An OSINT tool to extract and analyse audio/video content* [Paper presentation]. International Conference on Innovation, Engineering and Entrepreneurship (pp. 3–9). Springer.

McKeown, S., Buivys, M., & Azzopardi, L. (2016). *InfoScout: An interactive, entity centric, person search tool* [Paper presentation]. Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1113–1116). ACM.

Mckinnel, D. R., Dargahi, T., Dehghantanha, A., & Choo, K. K. R. (2019). A systematic literature review and meta-analysis on artificial intelligence in penetration testing and vulnerability assessment. *Computers & Electrical Engineering*, *75*, 175–188. https://doi.org/10.1016/j.compeleceng.2019.02.022

Medkova, J. (2018). Composition attack against social network data. *Computers & Security*, *74*, 115–129.

Naveen, G., Naidu, M. A., Rao, B. T., & Radha, K. (2019). A comparative study on artificial intelligence and expert systems. *International Research Journal of Engineering and Technology*, *6*(2), 1980–1986.

Nicart, E., Zanuttini, B., Gilbert, H., Grilhéres, B., & Praca, F. (2016). *Building document treatment chains using reinforcement learning and intuitive feedback* [Paper presentation]. 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 635–639). IEEE. https://doi.org/10.1109/ICTAI.2016.0102

North Atlantic Treaty Organization (2001). *NATO Open Source Intelligence handbook*. NATO.

Noubours, S., Pritzkau, A., & Schade, U. (2013). NLP as an essential ingredient of effective OSINT frameworks [Paper presentation]. 2013 Military Communications and Information Systems Conference (pp. 1–7). IEEE.

Quick, D., & Choo, K. K. R. (2018). Digital forensic intelligence: Data subsets and Open Source Intelligence (DFINT + OSINT): A timely and cohesive mix. *Future Generation Computer Systems*, *78*, 558–567. https://doi.org/10.1016/j.future.2016.12.032

Settanni, G., Skopik, F., Shovgenya, Y., Fiedler, R., Carolan, M., Conroy, D., Boettinger, K., Gall, M., Brost, G., Ponchel, C., Haustein, M., Kaufmann, H., Theuerkauf, K., & Olli, P. (2017). A collaborative cyber incident management system for European interconnected critical infrastructures. *Journal of Information Security and Applications*, *34*, 166–182. https://doi.org/10.1016/j.jisa.2016.05.005

Talwar, R., & Koury, A. (2017). Artificial intelligence–the next frontier in IT security? *Network Security*, 14–17. https://doi.org/10.1016/S1353-4858(17)30039-9

US Department of the Army. (2012). *Open-Source Intelligence*. US Department of the Army.

Vijayakumar, S., & Sheshadri, K. N. (2019). Applications of artificial intelligence in academic libraries. *International Journal of Computer Sciences and Engineering*, *7*, 136–140.

Watters, P. A., & Layton, R. (2016). *Automating Open Source Intelligence: Algorithms for OSINT*. Syngress.

Yang, H. C., Lee, C. H. (2012). *Mining open source text documents for intelligence gathering* [Paper presentation]. 2012 International Symposium on Information Technologies in Medicine and Education (pp. 969–973). IEEE.

Yates, A., & Zvegintzovi, N. (1999). *A Siberian reality check on open source information* [Paper presentation]. ASLIB Proceedings (pp. 175–186). MCB up Ltd. https://doi.org/10.1108/EUM0000000006976

Zhang, X., Zhang, S., Liu, J., Cai, L., & Wang, J. (2019, July). *Artificial Intelligence recruitment analysis* [Paper presentation]. The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (pp. 434–442). Springer.

Zunino, R., Surlinelli, R., & Sangiacomo, F. (2013). An analyst-adaptive approach to focused crawlers [Paper presentation]. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013) (pp. 1073–1077). IEEE. https://doi.org/10.1145/2492517.2500328