11th International Defense and Homeland Security Simulation Workshop
18th International Multidisciplinary Modeling & Simulation Multiconference

ISSN 2724-0363 ISBN 978-88-85741-67-6 © 2021 The Authors.
doi: 10.46354/i3m.2021.dhss.001

Enhancing the capacity of data collection tools to detect, prepare and respond to emerging CBRNEe threats through engaging with end-users

Roberto Mugavero[1,2,3*], Pietro Costanzo[2,3], William Thorossian[3]

1 University of Rome "Tor Vergata", Department of Electronic Engineering – DIE, 2 University of the Republic of San Marino, Center for Security Studies – CUFS, 3 Observatory on Security and CBRNEe Defense - OSDIFE, Via del Politecnico, 1 - 00133, Rome, Italy

*Corresponding author. Email address: r.mugavero@osdife.org

Abstract

An Intelligence Platform for Chemical, Biological, Radiological, Nuclear and explosives (CBRNe) Events and Asymmetric Threats has been developed as a pilot prototype to meet the needs of organizations, specialist, experts, professionals from the intelligence, law-enforcement, military, chemical, biological, radiological/nuclear and health domains.

The main goal was to provide a tool able to collect open source information in an asymmetric threats environment, with a focus on CBRNe events and terrorism, and with a pilot focus on COVID-19 related information, and to generate outcomes that can help analysis of trends, threats and intelligence sources, with application across security, academia and health fields.

The developed IT solution is a flexible and innovative instrument offering support to CBRNe risk and Asymmetric threat knowledge management by monitoring a wide range of information sources, using normalized terminologies, based on tuned ontology and able to enhance interaction and communication between different international entities (semantic interoperability). The experimentation aimed to provide a lite tool that can be adopted at different levels, including where less skills and economic resources are available (local units of complex organizations, public administrations in developing Countries, SMEs, ONGs, media).

The activities have been carried out by the Observatory on Security and CBRNe Defence OSDIFE - Italy, in cooperation with the University of Rome "Tor Vergata" - Department of Electronic Engineering - Italy, the State University of the Republic of San Marino - Center for Security Studies, the Flinders University - Australia and Expert AI - Italy.

Keywords: Asymmetric threats; Hazardous materials; Artificial Intelligence; ontology; semantic interoperability; CBRNe

1. Introduction

In recent decades, global transformations fostered the development of new knowledge and technologies to improve health and well-being.

At the same time, risk scenarios related to natural events, accidents, conflicts, instability, terrorism and public health threats evolved and called for new approaches. In particular, deliberate acts or threats involving the intentional release of hazardous substances to cause harm are becoming more common across the world. Hazardous substances can include chemicals, biological agents and radiological materials (CBRN), and can be delivered through a variety of mediums and mechanisms, where explosives represent a key factor for their exploitation on the field (thus, driving the use of the acronym CBRNe).

Current and future scientific and technical developments will have an impact on the protection of citizens, the environment and the strategic interests of Countries and international communities. In this context, Artificial Intelligence can be one of the main game changer to support analysts committed to deepening their knowledge of potential threats and the understanding of the future.

Moving from existing internal research activities, the Observatory on Security and CBRNe Defence OSDIFE - Italy, in cooperation with the University of Rome "Tor Vergata" - Department of Electronic Engineering - Italy, the State University of the Republic of San Marino - Center for Security Studies, the Flinders University - Australia and Expert AI – Italy, worked on the verticalization of cognitive computing technologies, based on Machine Learning ontologies and algorithms (pertaining to Cogito® technology), customizing the technological solutions to the domain of asymmetric risks, and with a particular experimental focus on the COVID-19 phenomenon.

The paper presents the evolution of an existing research and analysis activities that was integrated through the experimentation of an AI based technology, finally proposing a proof of concept of a lite tool that, with customised developments, can be adopted at different levels, including where less skills and economic resources are available (local units of complex organizations, public administrations in developing Countries, SMEs, ONGs, media).

2. Preliminary considerations

The capacity and capability to counter the risk of a CBRNe event and deliver countermeasures can be aided by the collection and analysis of data. Examples of data collection tools include the Global Terrorism Database (GTD), Maryland University and the Incident and Trafficking Database (ITDB), developed by the International Atomic Energy Agency.

Both include systematic data and information, as well as periodical updates on terrorist events and/or events related to specific threats at the national and international level. However, neither database primarily or collectively focusses on CBRNe or informs the context in which CBRNe risks nor their impact can be analysed.

This research has been triggered by the need to improve the existing monthly "Report on CBRNe Events in the World" and database launched by the Observatory on Security and CBRNe Defence in 2014.

The Observatory on Security and CBRNe Defence (OSDIFE), in collaboration with its Italian and international partners, currently hosts a database which collates open source data related to current CBRNe incidents and their geographic distribution, providing summary newsletters and reports to end-users including security agencies, academia and international organisations.

Research teams at OSDIFE manually collate open source data (through the use of web search strings, and qualified sources monitoring) and enter the data into the database. The report is manually produced on a monthly basis by a team of analysts, before being distributed to subscribers.

Additional considerations were made in relation to the use of social media platforms (and its dual use potential in the CBRNe space), as the team recognized the limitations and functionality of the current methods used to curate the database and produce the monthly reports. Social media posts that contain possible CBRNe indicators may be nuanced, meaning that the true message may be difficult to detect, surveillance of more than 1 billion social media posts a day from only three social media platforms would not be possible or practicable. For this reason, specific attention was focused on creating a system able to collect content from open sources like Web, Blogs, RSS and Social Networks (Twitter, Facebook).

3. Materials and Methods

Within this context, the critical challenge is to automate reporting to support the analysis of massive amounts of data in a more effective and efficient way compared to traditional keyword based or statistical technologies.

The option adopted at the beginning was the use a series of taxonomy combined with software that crawls through the various social media platforms to search for nuanced word or terms. A taxonomy is defined as the science or technique of classification, or a classification into ordered categories. Having the technological means to create search terms based on a series of taxonomy, allows researchers to tailor a search of several open source social media platforms to suit organisational objective. The ability to create effective and efficient real time reports will provide organisations with a vital time advantage to neutralise or defuse potential threats.

In the specific case, OSDIFE team has sought to increase the capability of the database to collate, link and analyse data related to the potential nefarious use of CBRNE materials of concern to specifically generate reports that can provide analysis of trends, threats and intelligence sources, with application across academia, health and security fields. The team embarked on a study designed to assess the database end user experience and assess user needs as a means of informing a database upgrade.

The aim of the study was to inform the development of a crawler for structured and unstructured content based on the Cogito® Discover software, which when combined with various taxonomies would deliver the requirements of an automated system to meet individual organisational needs.

The team involved end-user participants based on organisations who historically receive the OSDIFE reports as international and national security agencies, non-government organisations and academia working across the health and security disciplines. The study used surveys to assess functionality and inform upgrades to the database, uncovering which specific improvements to the software could benefit end users in better tracking, anticipating and predicting events and trends which may not otherwise be evident. In order to assess the end-user experience and discern areas in which the database could be improved, a 20 questions survey was designed to gauge how the end user interacted with the database at present, ways they would like to see the database improved in order to provide more intelligence or create further linkages, and in which ways an improved database would augment their analysis of CBRNe threats.

The survey was completed in two stages. The first stage was completed during a meeting held by OSDIFE in Rome. The second stage involved a survey completed online by 27 participants using Qualtrics survey software. Results of both surveys were then aggregated by the team. Once the data from the initial pre-survey had been analysed, recommendations for increased database functionality were drafted. The team then worked on upgrading the software and capability of the database as per recommendations provided. A post survey was conducted after the database improvements have been developed and implemented to assess improved functionality. The post-survey measured end user satisfaction regarding the database upgrades and gauged whether improved useability for the database end user had been achieved by comparison with the initial responses provided in the pre survey.

3.1.1. Technology adaptation

On the basis of this consultation and according to technical advices and indications received by the industry partner, the team started the development and testing of a pilot platform that exploits advanced AI algorithms to simplify the acquisition and analysis of information and to increase the ability to find, categorize and correlate potentially relevant knowledge.

In this work, the human component remains central and irreplaceable, while it is possible to amplify the possibility of intercepting weak signals without underestimating false negatives.

From a technical point of view, effective extraction and categorization of unstructured data requires text analysis and taxonomy management rules that meet the needs of the organization involved and the sector where its operate, as well as the specific project requirements. The model process adopted considered the following elements:

- Development of custom taxonomies.
- Extraction rules and develop them automatically through machine learning algorithms.
- An integrated environment for text analysis project modelling and knowledge enrichment.
- Multi-language support and scalability.
- Project management with different levels of complexity.
- Importing thesauri and vocabulary.

A linguistic approach is adopted, based on the rules of Machine Learning (ML). The system can understand the relationship between the searched words and the domain of analysis, emulating "the way people write", and applying some of the cognitive processes that humans use to understand the text.

In this sense, the system "learns" through the knowledge base and / or the expansion of the linguistic rules that are created ad hoc.

4. Pilot technology development

The project allowed the development of a system focused to identify in advance the information whose evidence is not yet immediate (weak signals) and to scope data that is likely to be considered non-existent due to the impossibility of developing effective research activities, correlation and verification (false negatives). The OSDIFE Intelligence Platform for CBRNe Events and Asymmetric Threats, has been developed to meet the needs of organisations/specialist/experts/professionals from the intelligence, law-enforcement, military, chemical, biological, radiological/nuclear and health domains.

4.1.1. Platform development workflow

The development required several months, during which the necessary verticalizations had to be developed. The various phases planned were:

- Development of a "text processing engine" based on the customisation of the COGITO® semantic engine, initially for the BIOLOGICAL domain and then extended to the CBRNe (Chemical, Biological, Radiological, Nuclear, Explosive) domains.
- Extraction of semantic metadata in order to make strategic information contained within texts from both internal and external sources easily accessible (OSINT).
- Evaluation of the usability of the Cogito® Intelligence Platform in order to support Biological Analysts in the management of information

gathering and content analysis.

- Extension of the previous point to support analysts in the other domains: Chemical, Radiological, Nuclear, Explosive.

The proposed solution we have arrived at can be summarised in the following figure:

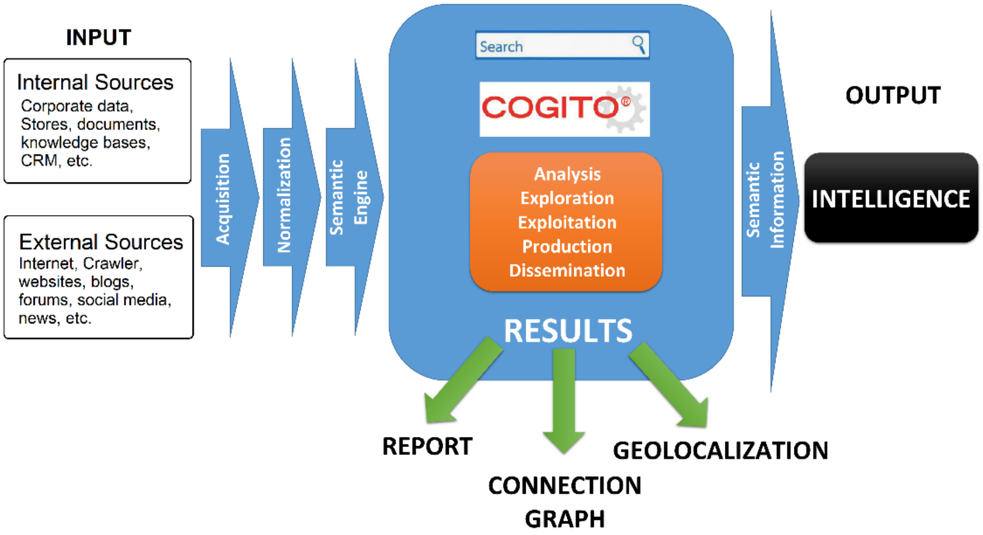


Figure 1. Figure 1. Pilot tool development process overview

The project phases were as follows:

- Exploration of the Cogito® Intelligence Platform in Standard version.
- Source configuration.
- Training.
- Study and analysis of linguistic verticalization.

Those phases were followed by in-depth project sessions in 3 phases:

- Release of the Cogito® intelligence Platform in Custom version.
- Sources configuration.
- Pilot tuning.

4.1.2. Logical architecture of the platform

The platform finds its maximum potentiality in the search of information from OSINT extrapolated through sources entered as input. The architecture of functioning, therefore, develops from these input data that the platform acquires through a Crawler. Then follows the phase of cataloguing and indexing.

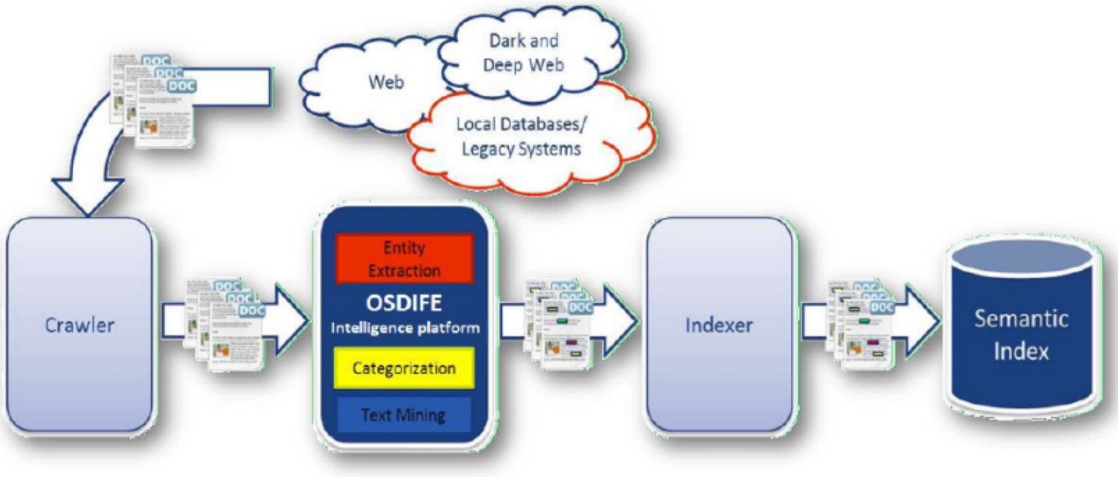


Figure 2. Input process

This information, catalogued and indexed, goes through a semantic index, in such a way that the information searched through: keywords, concepts, entities, categories, geo-localization, alerts and typology of the sources themselves; are recoverable and can be evaluated.

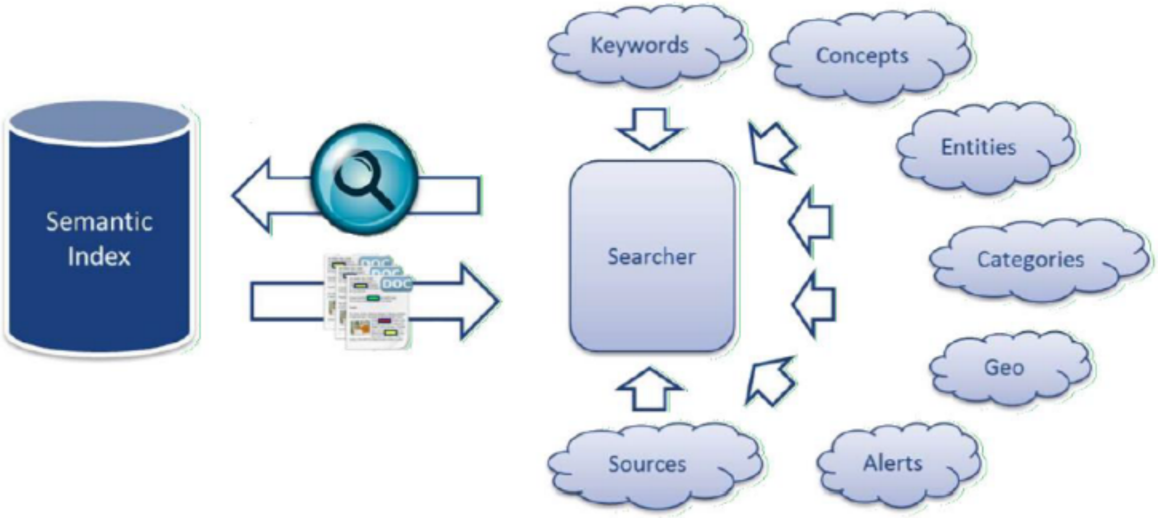
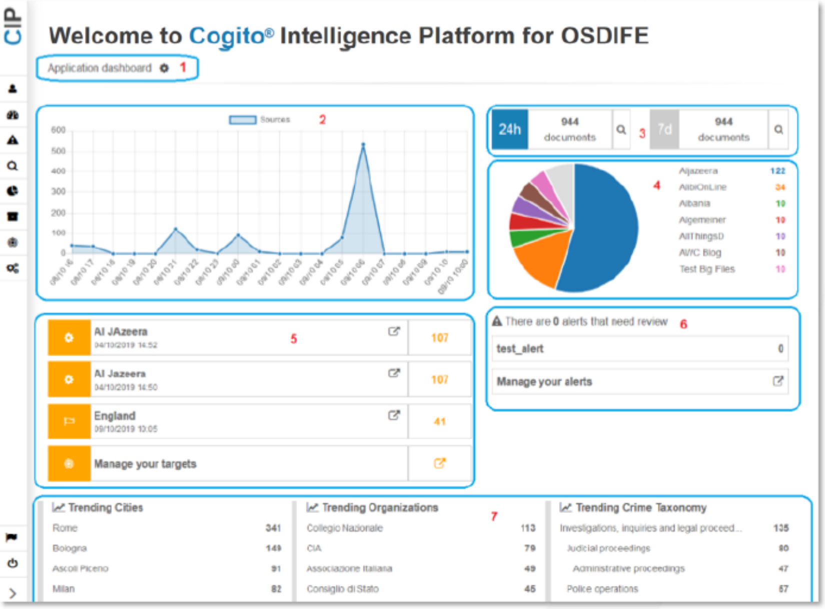


Figure 3. Indexing process

4.1.3. Platform dashboard and function

The platform is designed as a web platform, installed on a dedicated server, and can therefore be accessed from any browser with usual login process.



Welcome to Cogito® Intelligence Platform for OSDIFE

Aljazeera	122
AlbiOnLine	34
Albania	10
Algemeiner	10
AllThingsD	10
AWC Blog	10
Test Big Files	10

Al JAzeera 04/10/2019 14:52	107
Al Jazeera 04/10/2019 14:50	107
England 09/10/2019 10:05	41
Manage your targets	

There are 0 alerts that need review

test_alert	0
Manage your alerts	

Trending Cities		Trending Organizations		Trending Crime Taxonomy	
Rome	341	Collegio Nazionale	113	Investigations, inquiries and legal proceed...	135
Bologna	149	CIA	79	Judicial proceedings	80
Ascoli Piceno	91	Associazione Italiana	49	Administrative proceedings	47
Milan	82	Consiglio di Stato	45	Police operations	57

Figure 4. The pilot dashboard

The landing page of the platform shows some information to inform you about the behaviour of the system and to highlight the results of your activities.

The platform has a dedicated module called "Source Management", which is used to create new data flows to acquire information for the analysis process.

The dashboard gives access to all the functions needed to carry out analytical research on the sources entered. The user can use the search bar at the top to perform a search based on semantic elements or perform natural language queries based on the semantic engine.

There are several parameters to refine the analysis and search, based on categorisation and semantic processes of entity extraction that vary according to the selected perspectives. The search is based on help through 'word stemming': the elements searched for are then reduced to their root form to include more relevant results or to choose from those proposed.

5. Results

The main goal was to provide a tool able to collect open source information with a CBRNe/Terrorism/Crime/Cyber context, strengthen in particular the data around biological incidents, and generate outcomes that can help analysis of trends, threats and intelligence sources, with application across security, academia and health fields

The developed IT solution offering support to CBRNe risk and Asymmetric threat knowledge management by monitoring a wide range of information sources. In this regard, normalized terminologies, based on tuned ontology and able to enhance interaction and communication between different international entities (semantic interoperability), have been created.

A Visual analytics platform, allows to:

- analyse systematically and continuously online and offline information sources such as: Surface/Deep Web, Social Networks and Data Base;
- use a variety of functions (glossary, taxonomy, georeferencing, filters and correlation) in order to evaluate events, scenarios, threats and their evolution in space and time;
- use semantic attributes in order to discover contents, elements, targets and topics of interest by categories, entities, relations or cluster concepts (not only keywords);
- navigate geographical extracted information and relations between entities;
- dispatch automatically "early warnings".

In particular, the use of Artificial Intelligence advanced algorithms and rules (both deep semantic and machine learning) allows to:

a. extract relevant categories coming from different taxonomies (CBRNe, Health, Intelligence Terrorism, Crime, Cyber);
b. extract "standard" entity types (people, organizations, places, dates);
c. extract "domain" entity types linked to CBRNe, Health, Intelligence, Terrorism, Crime and Cyber environments;
d. extract data and information related to CBRNe, Health, Intelligence, Terrorism, Crime categories and link them to people, organizations, places and dates;
e. extract relations between standard and domain entities;
f. provide sentiment and emotion detection;
g. create cluster of equal or similar contents.

Algorithms and Text Mining works together on process of analysis of texts written in natural language and extract high-quality information from text. It involves looking for interesting patterns in the text or to extract data from the text to be inserted into a No-Sql-database. Text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities). Developers have to prepare text using lexical analysis, POS (Parts-of-speech) tagging, stemming and other Natural Language Processing techniques to gain useful information from text.

6. Conclusions and way forward

High level and complex tools are suitable for «top-down» information gathering and analysis. They are the «state of the art» and the leading solutions, that can be mainly adopted by central administrations and main companies. On the other side, lite tools can be adopted at different levels, including where less skills and economic resources are available (local units of complex organizations, public administrations in developing Countries, SMEs, ONGs, media). With regards to this latest operational context, monitoring and alert systems (e.g. sentiment analysis on social media, monitoring of local media and social media sources) can be a game changer for local, small and medium players.

Flexible tools can be quickly customized to work on both the web and static (structured and unstructured) data, and such spreading of lite tools can reinforce the investigation capacity of the central level, providing reliable «bottom-up» support.

The IT solution can allow our analytical team to prepare reports related to specific research and analysis questions, related to threat assessment, identification of trends, periodic monitoring of set of information.

Within this framework, the actual use of the platform is subject to a previous contact between the final user and the analytical team, to define the need. In fact, in order to provide a dedicated reporting service, the team will define, together with the user, the effort required to adapt the ontologies, to instruct the platform with semantic know how needed for the requested domain of study, to feed the platform with the needed sources.

Thus, the future of the platform sees the implementation of two steps:

– the search for funding allowing the industrialization and commercialization of the pilot prototype;

– the proposition of the platform as a software provided centrally as a service via the Internet, i.e. as a Saas service, "Software as a service". The disadvantages and possible risks of this model are largely of limited impact, also considering that the Saas model is spreading rapidly and competitive pressure is contributing to the continuous improvement of data security and performance issues.

Funding

All the research activities as well as testing and developing operations were funded by OSDIFE with internal funds.

References

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

Ruan, D., Chen, G., Kerre, E. E., & Wets, G. (Eds.). (2005). Intelligent data mining: techniques and applications (Vol. 5). Springer Science & Business Media.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA.

Serrano, L., Bouzid, M., Charnois, T., Brunessaux, S., & Grilheres, B. (2013, November). Events extraction and aggregation for open source intelligence: From text to knowledge. In 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (pp. 518-523). IEEE.

Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139084789

Costanzo, P., D'Onofrio, F., & Friedl, J. (2015). Big data and the Italian legal framework: Opportunities for police forces. In Application of big data for national security (pp. 238-249). Butterworth-Heinemann.

Lindstrom, M. (2016). Small Data: The Tiny Clues that Uncover Huge Trends. St. Martin's Press.

Bergamaschi, S., Cappelli, A., Circiello, A., & Varone, M. (2017, June). Conditional random fields with semantic enhancement for named-entity recognition. In Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics (pp. 1-7).

Garcia-Silva, A., Palma, R., & Gomez-Perez, J. M. (2017, October). Semantic Technologies and Text Analysis in Support of Scientific Knowledge Reuse. In 2017 IEEE 13th International Conference on e-Science (e-Science) (pp. 454-455). IEEE.

Deliu, I., Leichter, C., & Franke, K. (2017, December). Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 3648-3656). IEEE.

Ranade, P., Mittal, S., Joshi, A., & Joshi, K. (2018, November). Using deep neural networks to translate multi-lingual threat intelligence. In 2018 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 238-243). IEEE.

Ghazi, Y., Anwar, Z., Mumtaz, R., Saleem, S., & Tahir, A. (2018, December). 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) (pp. 129-134). IEEE.

VoPham, T., Hart, J. E., Laden, F., & Chiang, Y. Y. (2018). Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. Environmental Health, 17(1), 1-6.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. International journal of information management, 39, 156-168.

Kim, N., Lee, S., Cho, H., Kim, B. I., & Jun, M. (2018, January). Design of a cyber threat information collection system for cyber attack correlation. In 2018 International Conference on Platform Technology and Service (PlatCon) (pp. 1-6). IEEE.

Wang, M. H., Tsai, M. H., Yang, W. C., & Lei, C. L. (2018, April). Infection categorization using deep autoencoder. In IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) (pp. 1-2). IEEE.

Wang, R., Ji, W., Liu, M., Wang, X., Weng, J., Deng, S., ... & Yuan, C. A. (2018). Review on mining data from multiple data sources. Pattern Recognition Letters, 109, 120-128.

de la Torre-Abaitua, G., Lago-Fernández, L. F., & Arroyo, D. (2019). A compression based framework for the detection of anomalies in heterogeneous data sources. arXiv preprint arXiv:1908.00417.

Pellet, H., Shiaeles, S., & Stavrou, S. (2019). Localising social network users and profiling their movement. Computers & Security, 81, 49-57.

Choraś, M., Pawlicki, M., Kozik, R., Demestichas, K., Kosmides, P., & Gupta, M. (2019, August). SocialTruth project approach to online disinformation (fake news) detection and mitigation. In Proceedings of the 14th International Conference on Availability, Reliability and Security (pp. 1-10).