Enhancing Cyber Intelligence Capabilities through Process Automation: Advantages and Opportunities

Giorgi Iashvili [1], Maksim Iavich [1]

[1] *Caucasus University, Paata Saakadze st. 1, Tbilisi, 0102, Georgia*

Abstract

Information gathering through open sources is a crucial aspect of cybersecurity activities in today's digital landscape. This process relies on publicly available sources, such as social networks, websites, and blogs. It encompasses data mining, data gathering techniques, data extraction, and data analysis activities. Open source intelligence (OSINT) methods are now widely utilized across various domains. In the context of penetration testing activities, OSINT processes are typically executed manually and managed by human operators. The integration of OSINT with machine learning (ML) elements can be beneficial in various sectors, such as everyday users, offices, and industrial or corporate environments. Combining data collection processes with general security regulations can yield better results for both production and end users. Platforms like TensorFlow leverage artificial intelligence to analyze massive amounts of data, as seen in Gmail and Google Translate platforms. In addition, numerous approaches are employed in social networks for friend suggestions and video streaming platforms for content recommendations.

Utilizing artificial intelligence in OSINT activities can enhance the efficiency and quality of the search process, and consequently, the overall results of cybersecurity activities, including penetration testing. AI can automate various processes, such as web crawling, data collection, and pattern analysis. This opens up a wide range of new possibilities for building cybersecurity activities and exploring different vectors for data collection and its application in security events or penetration testing. Incorporating artificial intelligence in OSINT activities can also strengthen assistant mechanisms like Amazon Echo or Siri, enabling them to collect more relevant and rich data about a subject based on specific requirements.

Keywords

Machine learning, OSINT, open sources, intelligence cycle;

1. Introduction

Information gathering through open sources is a critical aspect of cybersecurity activities in today's digital landscape. This process relies on publicly available sources, such as social networks, websites, and blogs. It involves data mining and gathering techniques, as well as data extraction and analysis activities. Open source intelligence (OSINT) methods are widely employed across various domains. In the context of penetration testing activities, these processes are typically executed manually and managed by human operators.

Transitioning from manual to automated processes in OSINT is essential, particularly when dealing with real-world operations. A comprehensive system should be developed to facilitate automated open-

IVUS 2023: Information Society and University Studies 2023, May 12, 2023, Kaunas, Lithuania
EMAIL: giiashvili@cu.edu.ge (G. Iashvili); miavich@cu.edu.ge (M. Iavich);
ORCID: 0000-0002-1855-2669 (G. Iashvili); 0000-0002-3109-7971 (M. Iavich);

source-based activities, in conjunction with training simulations for machine learning. The architecture of the machine learning approach is contingent upon specific requirements.

- Based on our research we can obtain the following requirements:
- The data used in previous user activities;
- Gathering information using web crawlers and / or scrapers;
- Processing activities like pattern recognition and events detection;
- Vision of the automated mechanisms in the processes;
- Analytics based on matching the patterns and data visualization;
- Automation of the responses, error messages;

2. Use of OSINT in different sectors

Within the scope of our research, we employ an approach that focuses on the aforementioned aspects, which can be implemented using automated process mechanisms. Combining open source intelligence with machine learning elements can be advantageous in various fields, such as everyday users, offices, and industrial or corporate sectors. Integrating data collection processes with general security regulations can yield improved results for both production and end users.

We can highlight the following key fields from data gathering point of view:

- Cyber Intelligence groups;
- Law firms;
- IT security personnel;
- Special investigators;
- IT oriented corporations
- Financial and insurance sector;
- Ethical hackers;
- Black hat hackers;
- Hacktivist groups;

Open data: social media and sources:

- Potential hosts;
- Information about domains;
- Media lookup;
- Contact details;
- Files online;
- Location information;

Data collection automated methods

Automated data collection mechanisms are extensively employed across various sectors today. Even technology giants like Google utilize AI in their search mechanisms to generate predictive search results. Platforms such as TensorFlow leverage artificial intelligence to analyze vast amounts of data, as seen in Gmail and Google Translate platforms. Additionally, numerous approaches are implemented in social networks for friend suggestions and video streaming platforms for content recommendations. In these instances, user activity and preference data are collected and processed by AI algorithms, providing end users with improved recommendations and, consequently, enhancing the overall system's efficiency. Figure 1 illustrates the intelligence cycle and its stages [1].

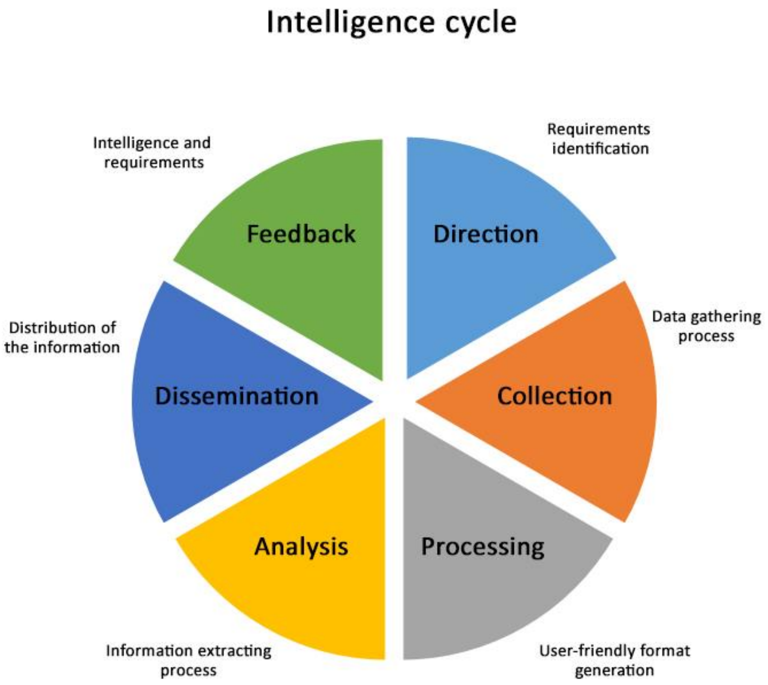
Intelligence cycle
Requirements identification
Direction
Data gathering process
Collection
User-friendly format generation
Processing
Information extracting process
Analysis
Distribution of the information
Dissemination
Intelligence and requirements
Feedback

Figure 1: Intelligence cycle

For better understanding of the process, we need to go through the entire intelligence process. It consist of the following steps:

1. Requirements identification: clarification of the requirements and search area. Choosing the right direction, working with the requirements is an important step to start the process correctly and not lose time to back to the beginning repeatedly.

2. Data gathering process: defining the vectors of the work and planning activities. On this stage should be defined the appropriate informational sources. Data collection process also can be performed in different manner of the activities, such active and passive data gathering.

3. User-friendly format generation: gathered information needs to be sorted and represented in understandable format. It can be special spreadsheet or graphical representation, or even database with prepared data.

4. Information extracting process: the collected data needs to be processed and exploitation activities should be performed based on the relevant information;

5. Distribution of the information: gathered information should be shared with the trusted parties to achieve more global results in the frame of the OSINT activities;

6. Intelligence and requirements: the final feedback should be check to meet the requirements. The accuracy of the entire process should be assessed.

Together with the possibilities of artificial intelligence today, there are different limitations when it goes to data gathering process. During such activities, we need to take into account the fact, that not everything is reachable using open sources and in some cases on the internet itself. During the building of AI-based OSINT working system, we need to consider the limitations and scenarios every time. As in cyber security and attacks almost everything in based on working environment. Together with the technical limitations, the ethical part is also should be considered.

Cyber intelligence process automation refers to the implementation of automated technologies and tools to gather, analyze, and disseminate cyber threat intelligence [2]. This allows organizations to improve their security posture by proactively identifying and addressing potential cyber threats. We can set the following points to highlight the advantages of cyber intelligence process automation:

Data Collection:

Automation allows for the efficient and continuous collection of data from various sources, such as open-source intelligence, social media, dark web, internal logs, and threat feeds. This reduces the workload of analysts and ensures a more comprehensive and diverse set of data.

Data Aggregation:

Automated tools can aggregate and normalize data from multiple sources, which improves the accuracy and relevance of the information. This helps organizations to gain a better understanding of the overall threat landscape.

Data Analysis:

Automated analysis of the collected data can reveal patterns, trends, and anomalies that may not be apparent to human analysts. Machine learning and artificial intelligence can be used to uncover hidden relationships, predict potential threats, and assess the likelihood of specific attacks.

Threat Prioritization:

Automation can help prioritize threats based on the potential impact on the organization and the level of risk involved. This enables security teams to focus on the most pressing issues and allocate resources accordingly.

Response and Remediation:

Automated systems can recommend appropriate countermeasures and response strategies based on the analyzed data. In some cases, they can even implement these measures, such as patching vulnerabilities, blocking malicious IP addresses, or updating firewall rules.

Continuous Monitoring:

Automation enables continuous monitoring of the threat landscape, allowing organizations to adapt their security strategies as new threats emerge. This ensures a proactive and agile approach to cybersecurity.

Knowledge Sharing:

Automated tools can disseminate relevant and actionable intelligence to stakeholders throughout the organization, fostering a more informed and collaborative security environment.

Reducing Human Error:

Automation reduces the likelihood of human errors in the cyber intelligence process, such as missed threats, data entry mistakes, or incorrect analysis.

Cost Savings:

By automating repetitive tasks, organizations can reduce the need for additional staff, increase the efficiency of existing staff, and ultimately reduce costs related to cybersecurity.

Scalability:

Cyber intelligence process automation can be easily scaled to meet the changing needs and growth of an organization. As more data and resources become available, automated systems can adapt and expand accordingly.

From the points above we can say that automating the cyber intelligence process offers numerous advantages, such as improved efficiency, accuracy, and scalability. It enables organizations to proactively address potential threats, reduce human error, and save costs [3, 4].

3. Use of AI in OSINT

Use of artificial intelligence in OSINT activities can increase the efficiency and quality of the search process and consequentially the results of the entire cyber security activity, for example penetration testing. Using an artificial intelligence a lot of processes like web crawling, collection of the data, analysis of the patterns can be automate. This fact gives a wide range of new possibilities in building of cyber security activities and different vectors for data collecting and it’s usage for security events or during penetration testing. As artificial intelligence algorithms learn based on the previews experience, it might have an impact on the entire process. A good example would be the suggestions of the videos on modern video streaming services. Recommendation works based on different parameters, like userpreferred categories, music genre or timing. The process of the open source intelligence works as follows [5-7]:

1. Work with sources of information to get the relevant ones;
2. Data gathering process using appropriate informational sources;
3. Processing of the information based on the sources and search requirements;
4. Data collection and analysis based on multiple sources;
5. Generation of the results and reporting activities;

4. AI use in intelligence cycle

Switching form a manual processes to automation machine learning-oriented analysis is extremely important especially when we work with real-world operations. These processes relies on massive data collection and analysis to artificial intelligence powered mechanisms and simulations.

Based on the intelligence cycle, majority of the processes can be done using artificial intelligence approaches. The intelligence cycle can be improved on the following stages: automation of the data collection mechanisms; structuring of the data; automated alerts and reports; dissemination of the collected data. Based on this process the human involvement is needed on the first and last steps. As first the step the definition of the requirements is taken and the last step is work with the feedback to make sure, that the process was performed correctly and as much closer to the requirements as possible. OSINT is a great candidate to be improved with artificial intelligence because of a huge amount of the data to be filtered and sorted. Manually this process may take too much time, but if some main processes are automated, this job can be done in very short time.

A good real life example of OSINT processes improvement by AI is tracking the activities of some concrete groups of people. The system revives the information from open sources like social media about changes of the location of concrete group of people or individuals. Intelligence process automatically is generated in the form of user-friendly report with the full data about the individual or a group of people. Such information is better readable and can be used to gain more data.

Use of artificial intelligence in open source intelligence activities can power up the assistant mechanisms like Amazon echo, or Siri to make it possible to collect more relevant and rich data about the subject based on the requirements.

Open Source Intelligence (OSINT) refers to the gathering, analyzing, and disseminating of intelligence from publicly available sources. The future of AI use in OSINT will likely involve an increasingly integrated and sophisticated approach, leveraging advanced AI technologies to improve the quality, speed, and scope of intelligence gathering.

Enhanced Data Collection:

AI-powered web crawlers and scrapers will be able to efficiently collect data from a broader range of sources, including social media, news outlets, blogs, forums, and the deep/dark web. Natural language processing (NLP) and computer vision techniques will help extract information from text, images, and videos.

Multilingual and Multimodal Analysis:

AI algorithms will be capable of understanding and analyzing content in multiple languages, as well as extracting information from different types of media (text, images, videos, etc.), providing a more comprehensive view of the OSINT landscape.

Sentiment Analysis and Emotion Detection:

Advanced NLP and machine learning techniques will enable AI systems to analyze the sentiment and emotions of individuals or groups based on their online content. This can help identify trends, public opinion, and potential security threats.

Real-time Analysis and Event Detection:

AI-powered OSINT tools will analyze data in real-time, enabling the rapid identification of emerging events, trends, or threats. This will allow organizations to respond more proactively to potential issues.

Geospatial and Temporal Analysis:

AI systems will be able to combine OSINT data with geospatial and temporal information to provide a more contextualized understanding of events and threats, improving situational awareness and decision-making.

Network and Relationship Analysis:

By leveraging machine learning and graph-based algorithms, AI can uncover hidden relationships, patterns, and networks within OSINT data, helping to identify key influencers, threat actors, or potential collaborators.

Automated Threat Assessment and Prioritization:

AI systems will automatically assess and prioritize threats based on factors such as impact, likelihood, and relevance to the organization, helping security teams focus their efforts on the most pressing issues.

Predictive Analytics:

Advanced AI models will be able to predict future events, trends, or threats based on historical data and current patterns, providing organizations with valuable foresight and strategic planning capabilities.

Customization and Personalization:

AI-powered OSINT platforms will offer tailored intelligence feeds based on an organization's specific interests, needs, and risk profile, ensuring that relevant and actionable information is readily available.

Improved Collaboration and Knowledge Sharing:

AI-driven OSINT tools will facilitate better collaboration and knowledge sharing among analysts, organizations, and other stakeholders, resulting in a more informed and unified approach to security and intelligence.

In conclusion, the future of AI use in OSINT will likely involve a more advanced, integrated, and efficient approach to intelligence gathering and analysis. This will empower organizations to better understand and navigate the complex and ever-evolving threat landscape, making more informed decisions and taking proactive measures to ensure their security [8,9].

5. Artificial intelligence methods use in open source intelligence

A practical example of AI use in OSINT is sentiment analysis for tracking public opinion on a specific topic, such as political events or product launches. Sentiment analysis is a technique that uses NLP and machine learning to identify and classify the sentiment of text data into positive, negative, or neutral categories. One common mathematical model used in sentiment analysis is the Naïve Bayes classifier.

We can describe the example of how the Naïve Bayes classifier can be used for sentiment analysis in OSINT:

Data Collection:

Gather text data from public sources such as social media, news articles, and forums related to the topic of interest.

Preprocessing:

Clean and preprocess the data by removing irrelevant information (e.g., URLs, special characters), converting text to lowercase, and tokenizing words.

Training Data and Feature Extraction:

Split the dataset into a training set and a testing set. Manually label a portion of the training set with the sentiment categories (positive, negative, or neutral). Extract features from the text, such as word frequencies or n-grams, to create a feature matrix.

Naïve Bayes Classifier:

The Naïve Bayes classifier is a probabilistic model that calculates the probability of a given sentiment category (e.g., positive) based on the features extracted from the text. The model is based on Bayes' theorem, which is formulated as:

$$P(C\_k|X) = P(X|C\_k) * P(C\_k) / P(X) \tag{1}$$

where:

- P(C_k|X) is the posterior probability of class (sentiment category) C_k given the features X.
- P(X|C_k) is the likelihood of the features X given the class C_k.
- P(C_k) is the prior probability of class C_k.
- P(X) is the probability of the features X.

The Naïve Bayes classifier makes a "naïve" assumption that the features are conditionally independent given the class, which simplifies the likelihood calculation. For text data, the likelihood is typically calculated using the bag-of-words representation:

$$P(X|C\_k) = P(x\_1|C\_k) * P(x\_2|C\_k) * \ldots * P(x\_n|C\_k) \tag{2}$$

Model Training:

Using the labeled training data, calculate the likelihoods and prior probabilities for each sentiment category. This will form the basis for the Naïve Bayes classifier.

Prediction:

Apply the trained Naïve Bayes classifier to the testing set to predict the sentiment of the unlabeled text data. Choose the sentiment category with the highest posterior probability as the final prediction.

Evaluation and Analysis:

Evaluate the performance of the classifier by comparing its predictions to the actual sentiment labels. Analyze the results to gain insights into public opinion on the topic of interest.

In this example, the Naïve Bayes classifier is used as the mathematical model for sentiment analysis in OSINT. This practical application demonstrates how AI can be utilized to gather and analyze public opinion data from open sources, providing valuable insights for decision-making and strategic planning [10-12].

6. Processing of textual data

For textual data processing we can use Python library for processing textual data - TextBlob. It provides a simple API for common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. TextBlob is built on top of the NLTK (Natural Language Toolkit) library, making it an easy-to-use interface for working with textual data.

TextBlob's sentiment analysis functionality is based on a pre-trained Naïve Bayes classifier. This classifier has been trained on a dataset of movie reviews, with the goal of determining whether the reviews are positive or negative. The sentiment analysis in TextBlob returns two values: polarity and subjectivity. Polarity is a float between -1 and 1, representing the sentiment's negativity or positivity, while subjectivity is a float between 0 and 1, representing the level of subjectivity or objectivity of the text.

The following pseudocode represents the Python script for gathering and analyzing public opinion data from Twitter using TextBlob's sentiment analysis:

```
Initialize Twitter API with credentials (consumer_key, consumer_secret, access_token, access_token_secret)
Define search_query and num_tweets
Create a connection to the Twitter API
Fetch num_tweets tweets containing search_query
Initialize a dictionary to store sentiment counts: sentiments
For each tweet in the fetched tweets:
  Extract tweet text
  Create TextBlob object with tweet text
  Analyze sentiment polarity using TextBlob sentiment analysis
  If sentiment polarity > 0:
    Increment positive sentiment count
  Else if sentiment polarity == 0:
    Increment neutral sentiment count
  Else:
    Increment negative sentiment count
Calculate sentiment percentages
Display sentiment analysis results
```

This pseudocode outlines the main steps for fetching tweets based on a search query and analyzing their sentiment using TextBlob's sentiment analysis feature. The results are then displayed as percentages of positive, neutral, and negative sentiments [13,14].

Enhancing cyber intelligence capabilities through process automation has emerged as a critical aspect of contemporary cybersecurity strategies. While it offers numerous advantages in terms of scalability, speed, and efficiency, there are several potential difficulties and challenges associated with this endeavor. This paper aims to discuss these challenges, which include:

Complexity of cyber threats: The constantly evolving nature of cyber threats, along with their increasing complexity, poses a challenge to the effectiveness of process automation. Developing automated systems that can keep pace with the rapid advancements in threat techniques necessitates continuous adaptation and updates.

Integration challenges: A seamless integration of process automation tools with existing cyber intelligence systems is crucial for optimal performance. This may prove to be difficult, given the wide array of systems, platforms, and applications used by organizations. The integration process may be time-consuming and costly, further complicating the adoption of process automation.

False positives and negatives: The automated systems may generate false positives and false negatives, potentially leading to missed threats or unnecessary alerts. This challenge necessitates the development of sophisticated algorithms that can accurately identify true threats while minimizing false detections.

Human factor: The reliance on automation may lead to a reduction in human involvement, potentially undermining the role of human expertise in cyber intelligence. A balance between automated and human-driven processes is essential to ensure that valuable insights from human analysts are not lost.

Legal and ethical considerations: As process automation tools become more sophisticated, the potential for misuse and abuse increases. Ensuring that these tools are deployed ethically and in accordance with relevant laws and regulations is a critical challenge that needs to be addressed.

Data privacy and security: Process automation tools may require access to vast amounts of sensitive data, which raises concerns about data privacy and security. Ensuring that the data used by these tools is stored and processed securely is a vital challenge.

Skills shortage: Implementing process automation in cyber intelligence requires a workforce skilled in both cybersecurity and automation technologies. The current shortage of such skilled professionals may hinder the widespread adoption of process automation.

Cost considerations: The development, implementation, and maintenance of process automation tools can be costly. Organizations must weigh the benefits of automation against the financial investment required to implement and maintain these systems.

7. Conclusion

In conclusion, the future of AI in the OSINT sphere appears promising and poised for significant advancements. As AI technologies, such as natural language processing, machine learning, and computer vision, continue to improve, their applications in OSINT will become more sophisticated, accurate, and efficient. The integration of AI in OSINT will lead to the following key developments:

Enhanced data collection and analysis: AI-powered tools will be able to gather and analyze data from a broader range of sources, such as social media, news outlets, blogs, forums, and the deep/dark web. This will provide a more comprehensive and diverse dataset for organizations to analyze and respond to.

Real-time monitoring and response: AI-driven OSINT platforms will enable real-time analysis of data, allowing organizations to identify emerging events, trends, and threats rapidly. This will facilitate more proactive responses to potential issues and improve overall situational awareness.

Improved accuracy and efficiency: AI algorithms will be able to identify patterns, trends, and anomalies more accurately, reducing false positives and false negatives in the analysis process. Automation will also increase efficiency, allowing organizations to process large amounts of data more quickly and effectively.

Predictive analytics: Advanced AI models will enable organizations to predict future events, trends, or threats based on historical data and current patterns. This will provide valuable foresight and support strategic planning and decision-making.

Customization and personalization: AI-powered OSINT platforms will offer tailored intelligence feeds based on an organization's specific interests, needs, and risk profile. This will ensure that relevant and actionable information is readily available to stakeholders.

Enhanced collaboration and knowledge sharing: AI-driven OSINT tools will foster better collaboration and knowledge sharing among analysts, organizations, and other stakeholders, resulting in a more informed and unified approach to security and intelligence.

The adoption of AI in the OSINT sphere will continue to grow as technologies evolve and organizations recognize the value of AI-driven intelligence gathering and analysis. The future of AI in OSINT will be characterized by increased efficiency, accuracy, and effectiveness in processing and understanding vast amounts of publicly available information, ultimately leading to better-informed decision-making and more robust security strategies.

8. Acknowledgements

This work was supported by Shota Rustaveli National Foundation of Georgia (SRNSFG)[NFR-22-14060].

9. References

[1] Iashvili, G., Avkurova, Z., Iavich, M., Bauyrzhan, M., Gagnidze, A., Gnatyuk, S. (2021). Content-Based Machine Learning Approach for Hardware Vulnerabilities Identification System. In: Hu, Z., Petoukhov, S., Dychka, I., He, M. (eds) Advances in Computer Science for Engineering and Education IV. ICCSEEA 2021. Lecture Notes on Data Engineering and Communications Technologies, vol 83. Springer, Cham. https://doi.org/10.1007/978-3-030-80472-5_10.

[2] G. Backfried, C. Schmidt, M. Pfeiffer, G. Quirchmayr, M. Glanzer and K. Rainer, "Open Source Intelligence in Disaster Management," 2012 European Intelligence and Security Informatics Conference, 2012, pp. 254-258, doi: 10.1109/EISIC.2012.4.

[3] S. Gong, J. Cho and C. Lee, "A Reliability Comparison Method for OSINT Validity Analysis," in IEEE Transactions on Industrial Informatics, vol. 14, no. 12, pp. 5428-5435, Dec. 2018, doi: 10.1109/TII.2018.2857213.

[4] Casanovas, P. (2017). Cyber Warfare and Organised Crime. A Regulatory Model and Meta-Model for Open Source Intelligence (OSINT). In: Taddeo, M., Glorioso, L. (eds) Ethics and

Policies for Cyber Operations. Philosophical Studies Series, vol 124. Springer, Cham. https://doi.org/10.1007/978-3-319-45300-2_9.

[5] Aikaterini Kanta, Iwen Coisel, Mark Scanlon, A survey exploring open source Intelligence for smarter password cracking, Forensic Science International: Digital Investigation.

[6] Volume 35, 2020, 301075, ISSN 2666-817, https://doi.org/10.1016/j.fsidi.2020.301075;

[7] James L. Regens. (2019) Augmenting human cognition to enhance strategic, operational, and tactical intelligence. Intelligence and National Security 34:5, pages 673-687.

[8] G. Backfried, C. Schmidt, M. Pfeiffer, G. Quirchmayr, M. Glanzer and K. Rainer, "Open Source Intelligence in Disaster Management," 2012 European Intelligence and Security Informatics Conference, 2012, pp. 254-258, doi: 10.1109/EISIC.2012.42.

[9] C. Best, "Web Mining for Open Source Intelligence," 2008 12th International Conference Information Visualisation, 2008, pp. 321-325, doi: 10.1109/IV.2008.86.

[10] Ball, L., Ewan, G., & Coull, N. (2012). Undermining: social engineering using open source intelligence gathering. In A. Fred, & J. Filipe (Eds.), Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (Vol. 1: KDIR, pp. 275-280). Scitepress Digital Library. https://doi.org/10.5220/0004168802750280.

[11] Darren Quick, Kim-Kwang Raymond Choo, Digital forensic intelligence: Data subsets and Open Source Intelligence (DFINT+OSINT): A timely and cohesive mix, Future Generation Computer Systems, Volume 78, Part 2, 2018, Pages 558-567, ISSN 0167-739X, https://doi.org/10.1016/j.future.2016.12.032.

[12] Igor Kononenko, Semi-naive bayesian classifier, European Working Session on Learning, EWSL 1991: Machine Learning — EWSL-91 pp 206–219.

[13] Sona Taheri and Musa Mammadov, Learning the naive Bayes classifier with optimization models, International Journal of Applied Mathematics and Computer Science, Volume 23 (2013) - Issue 4 (December 2013).

[14] S SAKTHI VEL, Pre-Processing techniques of Text Mining using Computational Linguistics and Python Libraries, 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS).