UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Revista Vínculos
http://revistas.udistrital.edu.co/ojs/index.php/vinculos

Revista VÍNCULOS
Ciencia, Tecnología y Sociedad: Un Enlace hacia el Futuro

A+T Actualidad Tecnologica

Inteligencia de fuentes abierta (OSINT) para operaciones de ciberseguridad. "Aplicación de OSINT en un contexto colombiano y análisis de sentimientos"

Open source intelligence (OSINT) as support of cybersecurity operations. "Use of OSINT in a colombian context and sentiment Analysis"

Ricardo Andrés Pinto Rico[1] Martin José Hernández Medina[2] Cristian Camilo Pinzón Hernández[3] Daniel Orlando Díaz López[4] Juan Carlos Camilo García Ruíz[5]

Para citar este artículo: R. A. Pinto, M. J. Hernández, C. C. Pinzón, D. O. Díaz y J. C. C. García, "Inteligencia de fuentes abierta (OSINT) para operaciones de ciberseguridad. "Aplicación de OSINT en un contexto colombiano y análisis de sentimientos"". *Revista Vínculos: Ciencia, Tecnología y Sociedad*, vol 15, n° 2, julio-diciembre 2018, 195-214. DOI: https://doi.org/10.14483/2322939X.13504.

Recibido: 21-04-2018 / Aprobado: 03-06-2018

Resumen

La Inteligencia de fuentes abiertas (OSINT) es una rama de la ciber inteligencia usada para obtener y analizar información relacionada a posibles adversarios, para que esta pueda apoyar evaluaciones de riesgo y ayudar a prevenir afectaciones contra activos críticos. Este artículo presenta una investigación acerca de diferentes tecnologías OSINT y como estas pueden ser usadas para desarrollar tareas de ciber inteligencia de una nación. Un conjunto de transformadas apropiadas para un contexto colombiano son presentadas y contribuidas a la comunidad, permitiendo a organismos de seguridad adelantar procesos de recolección de información de fuentes abiertas colombianas. Sin embargo, el verdadero aprovechamiento de la información recolectada se da mediante la implementación de tres modelos de aprendizaje automático usados para desarrollar análisis de sentimientos sobre dicha información, con el fin de saber la posición del adversario respecto a determinados temas y así entender la motivación que puede tener, lo cual permite definir estrategias de ciberdefensa apropiadas. Finalmente, algunos desafíos relacionados a la aplicación de técnicas OSINT también son identificados y descritos al respecto de su aplicación por agencias de seguridad del estado.

Palabras clave: Análisis de sentimientos, aprendizaje automático, ciber inteligencia, ciencia de datos, inteligencia de fuentes abiertas, perfilamiento de adversarios.

Abstract

Open source intelligence (OSINT) is a cyber-intelligence branch used to obtain and analyze information

1. Estudiante Ingeniería de Sistemas. Escuela Colombiana de Ingeniería Julio Garavito. Correo electrónico: ricardo.pinto@mail.escuelaing.edu.co
2. Estudiante Ingeniería de Sistemas. Escuela Colombiana de Ingeniería Julio Garavito. Correo electrónico: martin.hernandez@mail.escuelaing.edu.co
3. Estudiante Ingeniería de Sistemas. Escuela Colombiana de Ingeniería Julio Garavito. Correo electrónico: cristian.pinzon@mail.escuelaing.edu.co
4. Doctor en Informática; profesor asistente, Escuela Colombiana de Ingeniería Julio Garavito. Correo electrónico: daniel.diaz@escuelaing.edu.co
5. Especialista en Seguridad Informática; jefe División de Ciberdefensa, Dirección de Cibernética Naval. Armada Nacional. Correo electrónico: juan.garciaru@armada.mil.co

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

related to potential adversaries, so it can support risk assessments and help to prevent damages against critical assets. This paper presents a research about different OSINT technologies and how these can be used to perform cyber intelligence tasks of a nation. A set of transforms addressed to the Colombian context are presented, which were implemented and contributed to the community allowing to the law enforcement agencies to develop information gathering process from Colombian open sources. However, the real use of the information is given by the implementation of three machine learning models used to perform sentiment analysis over this information, in order to know the opinion of the adversary about certain topic and understand his motivation and, in this way, define proper cyber defense strategies. Finally, some challenges related to the application of OSINT techniques are identified and described regarding its use by state security agencies.

Keywords: Sentiment analysis, machine learning, cyber intelligence, data science, open source intelligence, adversary profiling.

1. Introducción

Having information has meant 'power' from a long time ago, in this way information from open sources has changed this paradigm because it is not initially restricted and can be accessed for many. So, the 'power' behind open source information is not related with the property (ownership) but with the knowledge about how to use it. Information from open sources can be recognized because is freely available in social networks, search engines, forums, photographs, wikis, online libraries, conferences, metadata, etc.

This paper will address cyber intelligence research conducted from open source intelligence, named OSINT (Open Source Intelligence). Through OSINT is possible to collect and process all types of information, which can be used for tasks such as conducting security profiling, psychological studies, market trend evaluations, security audits, review of digital and online reputation of a target, among others.

The more high-quality information is acquired, the more conclusions can be established in different processes. Since there is a vast amount of information on internet which does not have a good quality or is wrong, many false positives can arise in OSINT. An example of a misguided data collection process can occur when the collected data belongs to a namesake, i.e. someone with the same name, with the legitimate target.

Cyber intelligence should not be confused with criminal analysis, the latter seeks to obtain the "information necessary for the prosecution and repression of crimes (evidence)" i.e. a crime has already occurred. Instead the cyber intelligence applies in the pre-criminal scope of the threat and the risk [1].

OSINT is generally used by law enforcement agencies, companies or organizations with high value assets and even cybercriminals. Law enforcement agencies can use OSINT for example to research around the triangle of three aspects (Motive, Opportunity, Means) that must be established for a crime. In this way it could be possible to find out about the motives (the reason to develop an attack) behind an adversary, the opportunities (how much the asset is exposed or vulnerable) offered by the victim and the Means (capabilities required to perform the attack) that the adversary holds.

So, this paper aims the following objectives:

1. Identify different OSINT tools, most of them open source, that can be useful in cyber intelligence labors.
2. Implement data collectors (transforms) which allows to use an OSINT tool in a Colombian context.
3. Develop models for the processing of information collected from social networks to realize a sentiment analysis.

The achievement of these objectives will allow to build strategic, tactical and operational intelligence products. These products will be supported by all the collected and processed information regarding a target, such as: full names, identification, address, user names, emails, etc.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

This paper is composed as follows. Section 2 offers a review of OSINT tools, explains the OSINT architecture and presents the transforms developed for the Colombian context. Section 2 also shows at last a demonstration of use of transforms and OSINT tools to develop cybersecurity labors. Then, Section 3 presents three different machine learning models used to make sentiment analysis over collected data. Next, Section 4 makes a reflection on OSINT tools and techniques used Cyber security teams belonging to law enforcement agencies. Finally, some conclusions and future works are included.

2. Gathering information from open sources in a Colombian context

The collection of public information is a process that can be done in various ways using manual or automatic processes. This section makes a review of different OSINT tools which can support am automatic collection process and list a set of elements (transforms) that can be used to collect information e.g. emails, documents, domain, etc. in the Colombian context.

2.1 OSINT tools overview

Cyber intelligence labors over open information sources can be developed with several tools. Table 1 shows the most representative open source tools that can help in the development of researching and profiling. Most of these tools are complementary between them.

The first tool that will be mentioned is "The Harvester"[2], which is focused to find emails addressed from a domain name. The emails are searched in servers such as Google, Bing, LinkedIn, etc. Emails can be useful as possible points of entry to the information of an organization e.g. identifying an email accounts with easy to guess passwords or validating the infection of accounts using cyberattacks that have as starting point a malicious email (phishing, trojans, spam, spear, whaling).

Another very useful tool for domain data collection is ReconNG [3]. This tool has a module for the scanning of a domain that collect a large amount of information about it. ReconNG can be used in a similar way to The Harvester obtaining emails from an organization, but as added value, ReconNG obtains information about the domain location, physical location of the domain server, name of the administrator and another "Who Is" data. This information can be used to profile an organization and its members.

When the target is a person, the information provided by ReconNG could allow to identify coworkers from the emails in the same domain. If the target is an organization, the information provided by ReconNG could deliver data about the employees and then perform a new OSINT iteration, but this time over one of the found employees.

Table 1. OSINT tools.

Tool	License	Input data	Platform
Maltego	MIT	Domain, username, url, email, image, DNS, IP, Location, phrase, etc.	Linux, Windows, Mac
Metagoofil	GNU 2.0	Url and type of file (extension), limit of results, etc.	Linux, Windows
The Foca	GPL 3.0	Type of file, domain, search engine, etc.	Linux, Windows
Shodan	MIT	Ip, country, port, keywords, hostname, DNS, protocol, url, etc.	Web
The Harvester	GPL 2.0	Domain, number of desired results, sources to search (Google, Bing, etc.)	Linux, Windows, Mac
Recon-NG	GNU 2.0	Domain, api-key, domain, special modules for gathering, etc.	Linux
Spiderfoot	GPL 2.0	Domain, username, files, url, email, etc.	Linux, Windows
Intel Techniques	N/A	Personal information (name, phone number, identification document, social network profile)	Web

Source: Own.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

FOCA [4] allows to find metadata in Open Office, Microsoft Office and PDF documents and correlate it to obtain relevant information. Metadata could indicate the file creation and modification date, in addition to the name of the user who made those changes.
Shodan [4] is a search engine used to find servers, routers or any type of device reachable by internet through protocols such as HTTP, SSH, Telnet, or others. Shodan uses a set of search filters like IP address, country name, TCP port, keywords, operating system, among others. For example, a Shodan search using the keyword "Password:" could list all devices that have between their reachable files the keyword as part of the content.
Finally, Metagoofil [5] helps the user to extract metadata from Microsoft Office and PDF documents, amongst others. With Metagoofil can be possible to make a search using a domain name and a type of file as parameter, which will give all the public found files from that domain. The use of this tool is very similar to The Foca since it also takes benefit of the metadata of documents.
Maltego, Spiderfoot and Intel Techniques are three powerful tools which allow to gather complementary data from an adversary. Due its technical features they can support widely a cyber intelligence process. A detailed description of these tools will be presented in the following sections.

2.1.1 Maltego

Maltego in one of the tools that facilitates to obtain information from different open and public sources. It has an enormous potential to find information about people, companies or organizations, supporting the recognition which is the first stage of an attack [6]. Maltego allows to search all information around a single initial point (Entity) and then pivot from it to make a new search and obtain more information. An entity is an abstract representation of any type of information that is found in real life, such as system user names, emails, full names, telephone numbers, addresses, social network accounts, IPs, geographical locations, and even phrases [6].
Relationship between Entities are called "Transforms" which develop the process of moving from one type of data "X" to another type of data "Y". For example, with a national ID number (Entity X) it is possible to obtain the full name of a person (Entity Y), or with an email address (Entity Z) it is possible to obtain the social network account (Entity A).
Maltego offers 15 categories of transforms, being between the most representative a transform that make extraction of document metadata which receives an input (domain or IP address) and generate an output (author, publishing date, country and other metadata from documents found as related to the domain). Another popular transform obtains information associated with an organization domain, which receive an input (organization domain) and produce an output (IPs, network range, server names, register name, registered phone number). Transform results can be represented in graphs which helps to illustrate relationships between entities as shown in Figure 1.
Maltego offers a very simple and easy to use interface. It introduces the concept of “machines” which are groups of associated and preconfigured transforms intended to obtain information regarding to a single type of entity. Maltego uses a TAS (Transform Application Server) server that contains the transforms executed by every request of the user. It is also possible to have a private TAS server with custom-made and not public transforms.[7]
A private TAS server would also avoid configuring a development environment on each Maltego workstation, since it could be installed in the TAS server. Additionally, a private TAS server maintain the privacy of transforms, the integrity of the source code and support a centralized control of versions transparent for Maltego users.

2.1.2 SpiderFoot

SpiderFoot is an open source intelligence tool that can be used offensively, as part of a black box penetration test to gather information about a target. Also, it can be used defensively by an organization to identify information that it freely provides and

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

that could be used by attackers. SpiderFoot uses more than fifty data sources such as search engines, web pages and public servers, to obtain the data. Information recovered by Spiderfoot is represented as a graph of nodes as shown in Figure 2.

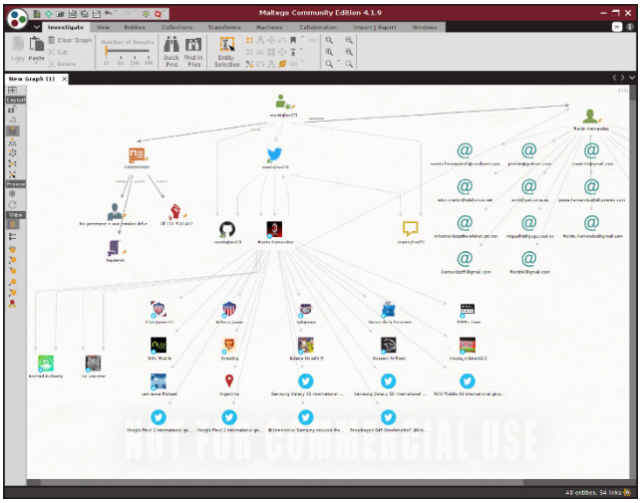

Figure 1. Graphs of entities in Maltego version 4.

Source: Own

Spiderfoot allows to perform different modes of search. “All mode” allows to get available target information from all Spiderfoot modules. “Footprint mode” makes a fast identification of what information the target exposes to internet. “Investigate mode” develops different secretive collection tasks useful when there is a suspect that the target is malicious. “Passive mode” is the most noiseless mode used to collect information and do not make the target suspect that is being investigated.

SpiderFoot can even access information hosted on the Deep Web using an "autonomous" TOR client and enabling control connections so SpiderFoot can manage it.

A module in Spiderfoot is “equivalent” to a “transform” in Maltego. A module traduces one type of data or information to another very different using an existing relation. When a module discovers a piece of data, it is transmitted to all other modules that can be 'interested' who will use that data to develop its own data discovering processes. Data

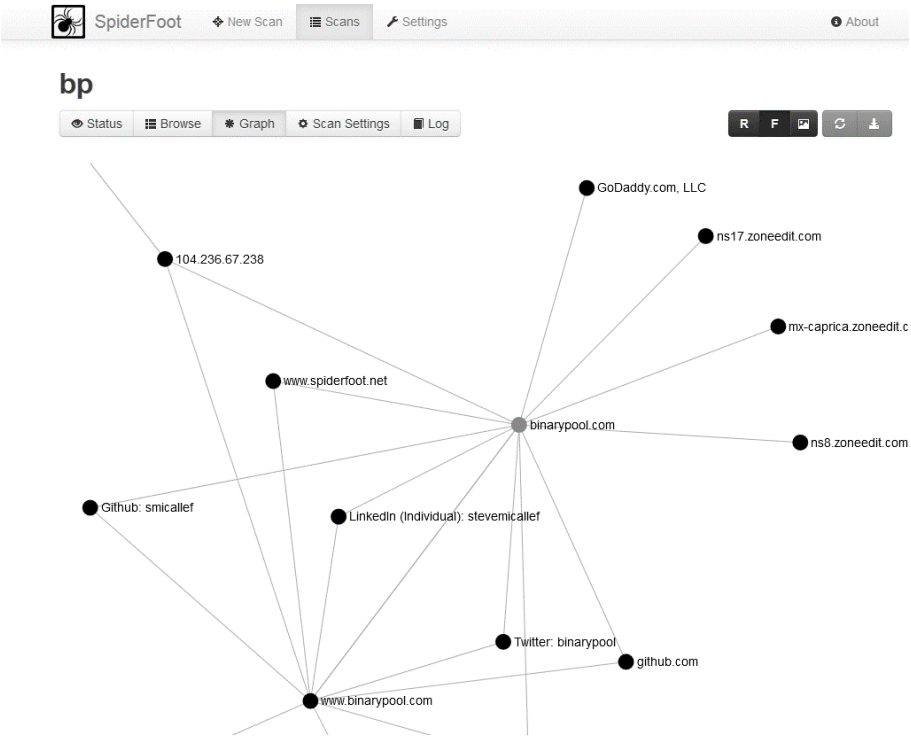

Figura 2. Estructura de las unidades de investigación.

Fuente: elaboración propia.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

discovered will also feed other modules that can be 'interested' and a new data discovering process starts, and so on successively. Table 2 summarizes some important Spiderfoot modules.

Module	Description
Accounts	Look for associated accounts on almost 200 websites (Ebay, Slashdot, reddit and others) Input: Email, domain name Output: Username, Account on External Site, User Account on External Site
badips.com	Check if a domain or IP is malicious according to badips.com Input: Internet name, IP, Affiliate - Internet Name, Affiliate - IP Address, Co-Hosted Site Output: Malicious IP, malicious internet name, malicious affiliate IP address, malicious affiliate, malicious co-host site
Base64	Identify Base64-encoded strings in any content and URLs, often revealing interesting hidden information Input: Linked URL - Internal, web content Output: Base64-encoded Data
Bing	Search Bing for hosts sharing the same IP Input: IP Address, Netblock Ownership Output: Co-Hosted Site, Search Engines Web Content

Table 2. SpiderFoot modules.

Source: Own

2.1.3 Intel Techniques

Intel Techniques is a website created by Michael Bazzel that provides a set of services oriented to OSINT. Between the services offered by Intel Techniques is possible to find modules that group a set of queries toward common web services e.g. Pipl, Facebook, among others, that can provide public information. Intel Techniques also offers links to books and conferences, access to forums, informative blogs and podcasts.

The use of each one of the modules provided by Intel Techniques depends on the available target data (Input parameter). The most useful modules offered by Intel Techniques are:

1. Pipl: This module uses the target username as an input parameter and develops a search to find accounts existing in popular websites that were created with the same username.
2. Facebook: This module allows to perform different searches using a Facebook username as input parameter (Figure 3). These searches bring public Facebook information related with the target.
3. Documents: This module supported by Google Hacking techniques can perform searches of documents such as docx, xlsx, pdf and others. These documents can contain data about the target such as organizations where had been affiliated, company where had worked, curriculum, documents created by the person, etc.
4. Image Reversal: This module uses Google artificial intelligence algorithms for facial recognition. It receives as input parameter an image URL and develop searches to find similar images published on internet.
5. UserSherlock: This module allows to search accounts in internet having a similar username than the target. It has more services where is possible to have an account than Pipl database.

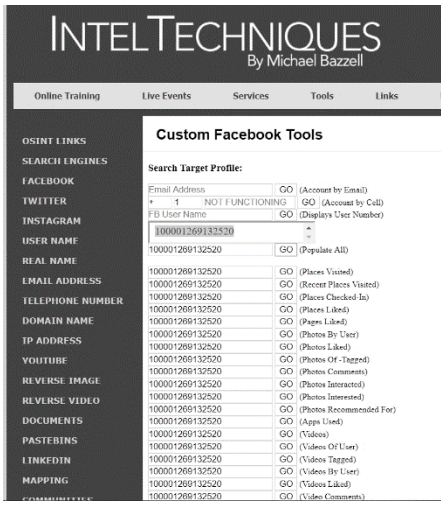

Figure 3. Facebook tools offered by Intel Techniques.

Source: Own

2.1.4 OSINT Process

The experiments developed in this paper follow the activities shown in Figure 4 which are part of an OSINT process. This process is composed by three important phases named gathering, processing and taking advantage.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

The gathering phase is the collection of all public information available around the target. This information may be in public servers, web pages, blogs, social networks and, in the context of this paper, refers to Colombian open source websites, e.g. Open Data initiative web page[1], National Civil Registry page[2], amongst others. As part of the gathering, some data correlation process can be developed that allows to complement an adversary profile using different kind of information, e.g. personal and professional data. This can allow us to obtain basic information about the person such as emails, names, addresses, aliases on websites and also national identification documents, helping us to know their possible location, motivations and behaviors. For this, a set of different Colombian sources are consumed by the set of transforms for the Colombian context mentioned in section 2.1.5. This phase requires a broad use and understanding of OSINT tools such as those mentioned in Table 1.

The processing phase refers to the actions applied over the collected data to make it usable. One of the first activities is to eliminate duplicate information that could have been collected by the different OSINT tools. It is also needed to discard no-related information that could had been collected by a non-proper filter set in an OSINT tool. Detect the false positives is also needed given that some collected information is not really related to the target (e.g. namesake) due the existence of similarities between the target and another person. All of the activities in this phase involve the analysis and correlation of no structured data to enrich the target profile with validated information. Another activity that can be done in this phase is the analysis of sentiment of the text produced by the target, such as comments on social networks, blogs, etc., regarding a specific topic, making it possible to estimate his thought or position. This analysis will determine a feeling whether negative, positive or neutral, through different prediction models as presented in section 3.1.

The taking advantage phase aims to use the collected and processed information to support cyber intelligence objectives. One of the possible uses of this information is to predict an attacks or crime. Prediction could be possible through the determination of the triangle of three aspects (Motive, Opportunity,

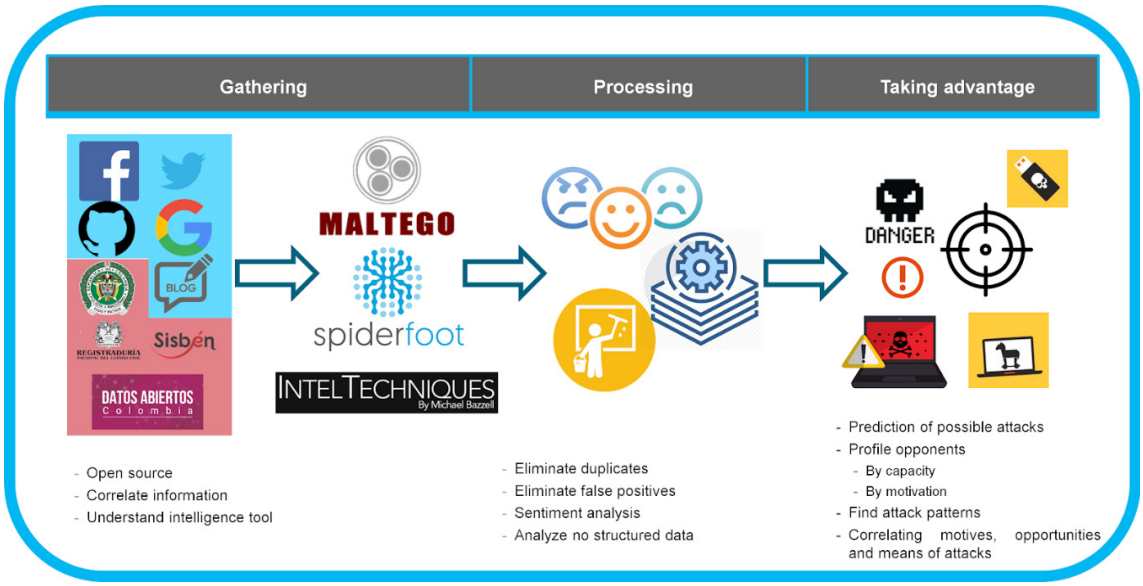

Figure 4. Open source intelligence phases.

Source: Own

1. https://datos.gov.co/
2. https://www.registraduria.gov.co/

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Means) that must come together to develop a crime. With the collected information could be possible for example to determine the means, e.g. capabilities and skills of the adversary obtained from the curriculum, and the motives, e.g. reasons to perform a crime supported by the feelings expressed by the adversary.

2.1.5 OSINT Transforms for the Colombian context

The collected information in the open source intelligence phases can be useful to develop cyber security labors. However most of the OSINT tools do not consult Colombian open sources so there is a big amount of useful information that a Colombian law agency can miss. In this paper a set of components (transforms) are developed and proposed to allow the collection of information from Colombian open sources.
For this purpose, a set of Colombian open sources offering information useful in OSINT processes were identified. Then, a set of transforms were developed using Python 2.7.x as programming language[3]. One of the main Python libraries used in the transforms construction was "Selenium" that allows to interact with web pages emulating a browser client to access information. Another's important Python libraries used in the transforms were "Pandas" and "Sodapy", the first one used to handle information in different modes (datagram, xml, html) and the second one to make connections to API, e.g. Open Data API.
Table 3 shows the identified main Colombian open sources, the service being consulted and the description of the developed transform.

2.1.6 OSINT process using transforms in a Colombian context

An intelligence exercise was done over a person with job in the Colombian state. The target name was renamed as Adam Goldstein and its personal collected data were changed to not expose the privacy of the person. The input parameter was the personal email adam.goldstein@hotmail.com.

Using the personal email as input parameter in Intel Techniques was possible to obtain the curriculum and then his national ID number. The national ID number was then used as input parameter for the transforms shown in Table 3.
The transforms that were applied are following:

- National Civil Registry Transform: This transform uses the national ID number to consult citizen information available in the National Civil Registry website. The obtained information (Polling place and place of issue of the Colombian identification card) becomes a Maltego entity.
- Colombia National Police Transform: This transform accesses the judicial background check service available in the website of the Colombian National Police and through Selenium library interact with the web page to enter the national ID number and submit the service request. Obtained information (Legal background of a citizen) is used to construct an entity with the data recovered from the Colombian National Police database.
- National Army of Colombia Transform: This transform uses the national ID number as input parameter in the website of the National Army of Colombia. Using Selenium library makes the interaction with the website and recover the citizen data that is used to build two entities, one with the name of the person and another with the Military status of the citizen which includes the military district where the military card was issued.
- Open data transform: This transform generates request to the API offered by the web services of Open Data. It uses the national ID number as input parameter and the response obtained is used to build two entities. One entity contains politic information, in this case the target is an elected councilor, so the recovered information refers to the municipality that represents, the region where is located and the name. The other entity has a single property that refers to the Colombian political party where the target is affiliated.

3. https://gitlab.com/ricardopinto08/OSINT

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Table 3. Colombian open sources and transforms.

Colombian open source	Service	Transform Input / Output
Policía Nacional de Colombia / Colombia National Police	Judicial background check	Input → Colombian identification number Output → Legal background of a citizen
Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales (SISBEN) / Identification System for Potential Beneficiaries of Social Programs	Affiliation and score check	Input → Colombian identification number Output → Socio-economic stratum of a citizen and SISBEN score
Administradora de los Recursos del Sistema General de Seguridad Social en Salud (ADRES) / Administrator of Resources of the General System of Social Security in Health	General System of Social Security affiliation	Input → Colombian identification number Output → Affiliation of a citizen to a health service provider
Registro Único Nacional de Tránsito (RUNT) / National Registry of Traffic	Traffic infractions check	Input → Colombian identification number Output → Traffic infractions of a citizen
Ejército Nacional de Colombia / National Army of Colombia	Military service check	Input → Colombian identification number Output → Military status of a citizen
Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior (ICETEX) / Colombian Institute of Educational Credit and Abroad Technical Studies	Affiliation check	Input → Colombian identification number Output → Educational program where a citizen is enrolled
Registraduría Nacional del Estado Civil / National Civil Registry	Information associated with the Colombian identification card	Input → Colombian identification number Outputs → Polling place and place of issue of the Colombian identification card
Procuraduría General de la Nación / Attorney General Office	Criminal background check	Input → Colombian identification number Output → Criminal background of a citizen
Servicio Nacional de Aprendizaje (SENA) / National Learning Service	Affiliation Check	Input → Colombian identification number Output → SENA educational program where a citizen is enrolled
Datos Abiertos / Open Data	Consult open data published by government organizations	List of councilors of the municipality of San Andres, Santander Input → Colombian identification number, Output → Job position, Political affiliation, Full names, Gender, Municipality
		Government secretaries of municipalities of Valle del Cauca 2016-2019 Input → Full name, Email Output → Phone number, Location
		Mayors of Municipalities of Antioquia 2016-2019 Input → Email Output → Phone number, Location, Full name, Period
		Senate employees Input → Phone number Output → Email, Full name, Location
		Delegates from the Ministry of Defense of Santa Marta Input → Email Output → Full name, Location, Phone number

Source: Own.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

2.1.6 OSINT process using transforms in a Colombian context

An intelligence exercise was done over a person with job in the Colombian state. The target name was renamed as Adam Goldstein and its personal collected data were changed to not expose the privacy of the person. The input parameter was the personal email adam.goldstein@hotmail.com.
Using the personal email as input parameter in Intel Techniques was possible to obtain the curriculum and then his national ID number. The national ID number was then used as input parameter for the transforms shown in Table 3.

The transforms that were applied are following:

- National Civil Registry Transform: This transform uses the national ID number to consult citizen information available in the National Civil Registry website. The obtained information (Polling place and place of issue of the Colombian identification card) becomes a Maltego entity.
- Colombia National Police Transform: This transform accesses the judicial background check service available in the website of the Colombian National Police and through Selenium library interact with the web page to enter the national ID number and submit the service request. Obtained information (Legal background of a citizen) is used to construct an entity with the data recovered from the Colombian National Police database.
- National Army of Colombia Transform: This transform uses the national ID number as input parameter in the website of the National Army of Colombia. Using Selenium library makes the interaction with the website and recover the citizen data that is used to build two entities, one with the name of the person and another with the Military status of the citizen which includes the military district where the military card was issued.
- Open data transform: This transform generates request to the API offered by the web services of Open Data. It uses the national ID number as input parameter and the response obtained is used to build two entities. One entity contains politic information, in this case the target is an elected councilor, so the recovered information refers to the municipality that represents, the region where is located and the name. The other entity has a single property that refers to the Colombian political party where the target is affiliated.

All entities created by transforms are included in Maltego workspace as shown in Figure 5.

Figure 5. Maltego workspace with entities derived by transforms execution.

Source: Own

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

3. Developing models to understand collected information

In this section we will talk about how machine learning and one of the branches of data science, i.e. the sentiment analysis, can support cyber intelligence labors. Three specific descriptive models will be presented which can be used to make sentiment analysis of phrases written in spanish.
The training set used to train the models was obtained from a repository published as part of the Feeling Analysis Workshop of the Spanish Society for the Processing of Natural Language (SEPLN)[4]. The training set contains approximately 60,000 tweets, each of them labeled with a certain feeling (positive or negative). Models were trained with the 70% of the training set and were tested using the 30% residuary percentage.

3.1 Machine learning and Sentiment analysis

Machine learning is an area of study derived from Data Science that has been presented for many decades but has acquired a high popularity in last years for its multiples application possibilities in many fields, including cybersecurity.

Machine learning refers to the practice of teaching a computer how to detect patterns and make connections by showing it a massive volume of data. Another definition of machine learning tells about the nontrivial extraction of implicit, previously unknown and potentially useful information from data. The large amount of data managed by fields such as commerce, banking or healthcare and the power of new computers to perform data processing operations give impetus to machine learning technologies. For the development of this paper, two models were mainly considered: descriptive and predictive.

3.1.1 Descriptive models

Descriptive models look for interpretable patterns to describe data. These models include: clustering, discovery of association rules and discovery of sequential patterns [8].
Permiten establecer relevancia / irrelevancia de factores y si aquélla es positiva o negativa respecto a otro factor o variable a estudiar.

3.1.2 Predictive models

Predictive models employ some variables to predict future or unknown values of another variables.

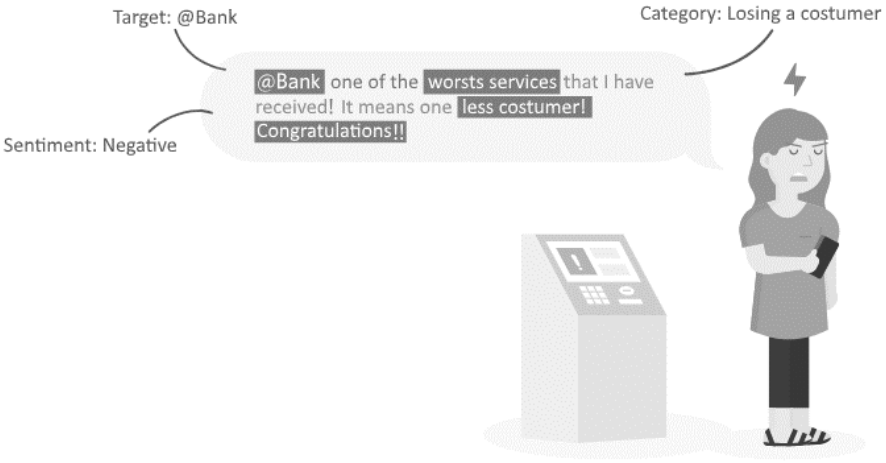

Figure 6. Example of sentiment analysis for a twit. [8-11]

4. https://github.com/roramas/sentitext

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

These models include: classification, regression and deviation detection [8].
This paper uses predictive models to make sentiment analysis of information published by an adversary in a social network (Twitter) around a specific topic.

3.1.3 Sentiment analysis

Sentiment analysis [9] is related with Opinion Mining [10] and aims to: extract subjective information from data, perform massive classification of the positive, negative or neutral connotation of a text, and try to determine the attitude of a writer regarding a topic.
Human supervision is required in sentiment analysis exercises due these models have some limitations that only a human can overcome, e.g. sentiment analysis models cannot analyze historical trends.
An example of sentiment analysis can be seen in Figure 6 where a client is upset due the poor service provided by a bank, and post a twit expressing its nonconformity. The twit can be processed to determine positive or negative words and identify a sentiment.
Sentiment analysis can be developed through one or more of the following techniques for text processing [11] : keyword location, lexical affinity, statistical methods, concept level.

- Keywork location.

This technique classifies the text into effect categories based on the presence of unambiguous affection words, such as happy, sad, frightened and bored [12] .

- Lexical affinity.

This technique assigns to arbitrary words (birth, growth, speed, etc.) a probable "affinity" with particular emotions [13].

- Statistical methods.

This technique takes advantage of machine learning elements such as latent semantics analysis, support vector machines, bag of words, among others [14] .

- Concept level.

Conceptual level approaches employ elements for knowledge representation, e.g. semantic networks. Conceptual level detects semantics that are expressed in a subtle manner, e.g. through the analysis of concepts that do not explicitly convey relevant information but which are implicitly linked to other concepts that do so [15].

3.2 Model 1: Bayes Naïve with a derivation of bag of words

Model 1 was implemented using Bayes Naive with a variation of bag of words methodology [8]. The analysis of sentiment for twits written by an adversary regarding a specific topic must go through a series of preprocessing steps to finally enter into Bayes Naive classification model. Next, each of these steps will be described which are necessary for the analysis of sentiments [16][17].

3.2.1 Tokenization

This step includes deletion of characters, change of uppercase to lowercase and separation of a phrase in list. Deletion of characters eliminates any character that is not important and that does not change the feeling of the phrase such as: @ {} []? ¡¨ *? ¡! "#! &% $ ¬ | | + * -. Phrases containing questions, irony or exclamation were not taken in account at the moment and will be considered as future work.
Change of uppercase to lowercase pass all the words of the phrase that are in uppercase to lowercase to avoid having repeated words that are considered as different only because of having capital letters. For example, words "bad" and "Bad" have the same feeling so should not be set as different words. Separation of a phrase in list separates the phrase into a list of words for easy handling of the model and training phase.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

3.2.2 Context

This step gives context to the phrase making an analysis by word to determine if it "empower" another word. For example, "very" is a feeling empower word, so the words "very bad" have to be considered together instead of considering them as individual words.

3.2.3 Stop words

This step eliminates from the list of words previously processed the "Stop words". Stop words are recognized because they do not alter the feeling of the phrase, such as: an, a, over, all, also, in addition, where, some, etc.

3.2.4 Training

Analysis of sentiment is done using Bayes Naive equation derived from the bag of word (Equation 1). Each word *di* of the phrase is selected and *count(di,C)* counts how many times word *di* exists in positive and negative sentences of the training set. Then, the probability associated with the word *di* is calculated by dividing the number of occurrences of the word *di* in the training set, (*count(di,C)*), on the total number of words belonging to class *C* (Positive or negative) existing in the training set, *V(Cn)*.
Probability of each word is cumulated in the total probability of the phrase $\sum_{i=0}^{n}$. Finally, the classification of the phrase will be determined by the accumulated probability for each class (positive, negative).

$$P = \frac{p(C)}{V_{C^n}} \sum_{i=0}^{n} count(d_i, C) \quad (1)$$

Where

- P is the final probability of the phrase
- C is the phrase class which can be positive or negative
- p(C) is the probability that a word is part of a positive or negative phrase
- count (di, C) is the number of occurrences of the word *di* for each class C (Positive or negative)
- Vc is the total number of words belonging to class C (Positive or negative) existing in the training set
- n is the total number of words in the training set

3.2.5 Testing

As mentioned previously accuracy calculation (testing) was done using a testing set composed by the 30% of the training set. The trained model received the testing set and calculated the classification of each twit. The area under the curve method was used using the Python "sklearn" library to compare the classification of each twit with its original and correct classification. Bayes Naïve model obtained a score of 0.59.

3.3 Model 2: Support Vector Machine

Model 2 uses Support Vector Machines (SVM) [18] for the analysis of sentiment of sentences written by an adversary. As indicated in [19] the phrases (twits) to be classified has to go through a series of steps to finally be entered into the SVM classification model. Next, each of these steps will be described which are necessary for the analysis of sentiments [20].

3.3.1 Tokenization

This step includes deletion of characters, change of uppercase to lowercase and separation of a phrase in list. Deletion of characters eliminates any character that is not important and that does not change the feeling of the phrase such as: @ {} []? ¡¨ *? ¡! "#! &% $ ¬ | | + * -. Phrases containing questions, irony or exclamation were not taken in account at the moment and will be considered as future work.
Change of uppercase to lowercase pass all the words of the phrase that are in uppercase to lowercase to avoid having repeated words that are considered as different only because of having capital letters. For example, words "bad" and "Bad" have the same

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

feeling so should not be set as different words. Separation of a phrase in list separates the phrase into a list of words for easy handling of the model and training phase.

3.3.2 Stop words

This step eliminates from the list of words previously processed the "Stop words". Stop words are recognized because they do not alter the sentiment of the phrase, such as: an, a, over, all, also, in addition, where, some, etc.

3.3.3 Stemming

This step reduces a word to its root or stem, so that the model finds it easy to train and calculate the sentiment of a word. This is due the sentiment can be the same for a word in plural or singular.

3.3.4 Vectorization

This step changes the phrase representation toward a matrix where the columns are the words processed by previous steps, and the row is the number of occurrences that word appears in the phrase.

3.3.5 Training

Analysis of sentiment is doing using SVM model which has been trained with multiple vectorized and tagged phrase placed in a multi-plane. As part of the training phase, the SVM model define a single plane that separate negative and positive phrases. Unclassified phrases follow the preprocessing steps defined previously and enter to the model to be classified.

The single plane that separates positive and negative phrases is chosen according to Equation 2 where *Xi* and *Xj* represent points that define the plane that classifies the different vectors in the multiplane. At last, SVM model determines the sentiment of a phrase according to its spatial location in the multi-plane.

$$K(X_i, X_j) = \begin{cases} X_i \times X_j & Linear \\ (\gamma X_i \times X_j + C)^d & Polynomial \\ \exp\left(-\gamma (X_i - X_j)^2\right) & RBF \\ \tanh(\gamma X_i \times X_j + C) & Sigmoid \end{cases} \quad (2)$$

Where:

- C is the constant of capacity
- Core is used to transform input data to the function space
- γ and d represents parameters for the management of non-separable data

3.3.6 Testing

As mentioned previously accuracy calculation (testing) was done using a testing set composed by the 30% of the training set. The trained model received the testing set and calculated the classification of each twit. The area under the curve method was used using the Python "sklearn" library to compare the classification of each twit with its original and correct classification. SVM model obtained a score of 0.823.

3.4 Model 3 Bernoulli

Model 3 uses Bernoulli model for the analysis of sentiment. It is important to understand that all the models presented in this paper for analysis of sentiment are oriented to calculate the polarity of a phrase and not just a word. These models serve as a starting point to develop other studies related to the disambiguation of polarity presented by phrases or documents[21].

As mentioned in[22], there are preprocessing steps that must be performed to prepare the phrase to entry to the Bernoulli model. Those steps are tokenization and vectorization. Additionally, it is required to build a pipeline element that is used to serialize the trained model to store it and streamline the classification of phrases.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

3.4.1 Tokenization

This step includes deletion of characters Deletion of characters eliminates any character that is not important and that does not change the feeling of the phrase such as: @ {} []? ¡¨ *? ¡! "#! &% $ ¬ | | + * -. Phrases containing questions, irony or exclamation were not taken in account at the moment and will be considered as future work.

3.4.2 Stop words

This step eliminates from the list of words previously processed the "Stop words". Stop words are recognized because they do not alter the sentiment of the phrase, such as: an, a, over, all, also, in addition, where, some, etc.

3.4.3 Stemming

This step reduces a word to its root or stem, so that the model finds it easy to train and calculate the sentiment of a word. This is due the sentiment can be the same for a word in plural or singular.

3.4.4 Count vectorizer

Count vectorizer counts the occurrences of a word in positive and negative tweets belonging to the training set. This count is used to determine the probability of the phrase of being positive or negative.

- Training

The Bernoulli model is of binomial type, which means that it mainly analyzes successes and failures, in a sentiment context phrases whose connotation are negative or positive are analyzed. It uses a formula where the occurrences of a certain word are counted within the training set and also infer the phrase classification from the number of occurrences of the same in negative or positive sentences. Bernoulli model uses the word counter to calculate the probability related with the classification. Equation 3 shows how to calculate the probability.

$$P(t|c) = \frac{T_{ct} + 1}{\sum_c t + 2} \quad (3)$$

Where

T_{ct} is the number of occurrences of t in the training set

$\sum_c t$ is the number of elements that the class C has within the training set

3.4.5 Testing

Different tests were carried out with phrases with a well mark polarity and the response of the model for all these cases was correct. In addition, accuracy calculation (testing) was done using a testing set composed by the 30% of the training set. The trained model received the testing set and calculated the classification of each twit. The area under the curve method was used using the Python "sklearn" library to compare the classification of each twit with its original and correct classification. Bernoulli model obtained a score of 0.81.

3.5 Comparison of results

The SVM model was the one with the highest accuracy (0.82) in comparison with Bernoulli model (0.81) and Bayes Naïve model (0.59). The reliability of the models depends on several aspects like the preprocessing stage, which were applied in a different way. For example, SVM and Bernoulli models share the Vectorization/Count Vectorizer step, however Bayes Naive does not. Also, SVM and Bayes Naive models implement a stop words step, however Bernoulli does not.

The accuracy is probably higher for the SVM model because of a preprocessing stage that prepare phrases (twits) in the proper way with nor excess in data cleaning neither scarcity. A scarcity of data preprocessing can avoid that the model identifies

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

really meaningful keywords, but an extreme cleaning can also produce that the phrase loses its original meaning. Bernoulli model having just a few preprocessing steps has the lowest accuracy.

Another possible reason that justifies the highest accuracy of the SVM model is its linear statistical nature that does not apply the classic binomial and multinomial paradigms applied by the other models. Instead, SVM opts for an alternative mathematical modeling where the proximity between words is measured using a hyperplane.

4. OSINT for Colombian Law Enforcement Agencies

In an interview with the former FBI director, Robert Mueller, he was asked: "3000 lives were lost on 9/11, what is your worry from a cyber-perspective that would be catastrophic like that?" Muller answered: "The way of looking at the power grids and our infrastructure, and financially because it would cripple us if there was a substantial attack on wall street on the exchanges. It also could lead to a loss of lives to the extent that we can have our command and control knocked in Afghanistan or Iraq where we are at war, and somebody would be ahead in terms of intelligence and technology." [23]. This indicates that one of the greatest concerns of nations is the cyberspace, called the fifth war domain, is generally included in the national security strategy.

Thus, on July 14, 2011, the Colombian State, through the National Department of Planning (DNP), set the guidelines for the Policy of Cybersecurity and Cyberdefense of Colombia with the document of the National Council of Economic and Social Policy (CONPES 3701) [24]. This document presents the road map and defines the roles of each of the organism in charge of cybersecurity, cyberdefense and cyberincidents management in Colombia. The aggrupation of these organisms is called the Intersectoral Commission (Figure 7) and is responsible for provide technical assistance, coordinate incident management, offer emergency assistance, development operational capabilities, provide cyberintelligence information and advice and support cyberdefense. Today, 7 years after creation of CONPES 3701, the Colombian state security agencies in charge of cybersecurity are in growing its cyber capacities and face the following challenges:

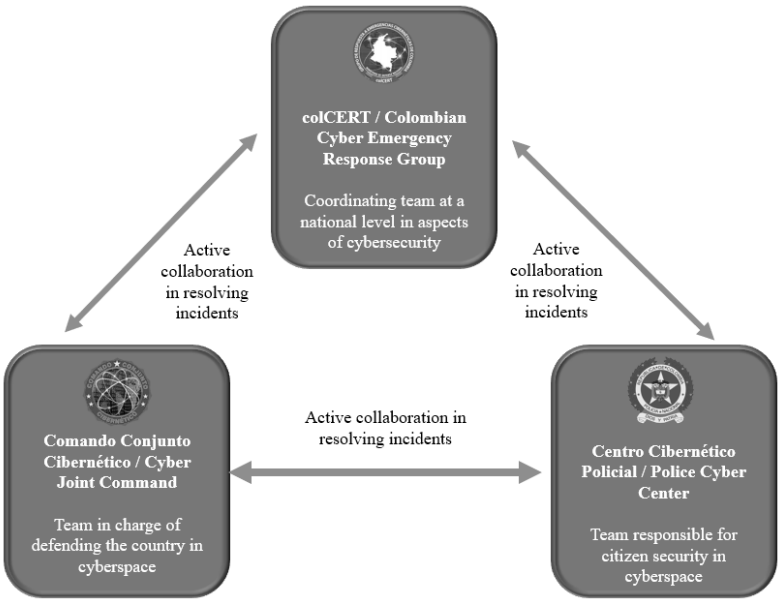

Figure 7. Intersectoral commission defined in CONPES 3701.

Source: Own.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

I. Cyber terrorism representing an asymmetric warfare tool since it is cheap and has a high destructive power[25], since the attacker e.g. terrorist, adversary state or guerrilla, does not need to buy large arsenals but with few resources can develop a cybernetic weapon, transport it, e.g. using a USB and generate an impact over a cybernetic infrastructure.
II. Application of soft power [26] by foreign countries which use cultural, ideological and diplomatic means to influence national politics or society, e.g. the use of social networks with fake news presidential campaigns in behalf of one candidate to obtain geostrategic advantage in a foreign country.
III. Increase of Internet connections, mainly due systems appear to be interconnected: security, defense, commercial, energy, health, communication, transportation, banking, librarians, etc. These highly connected systems make internet crucial and vital for the most advanced societies[27].
IV. Growing technological dependency, which becomes one of the biggest challenges since a cyberattack can produce chaos in state and society affecting delivery of essential services like water, electricity and banking.
V. Secure Operation Technologies (TO) like industrial control systems designed to be functional but not safe. TO systems can exist for example in energy companies with different connected electrical substations which can be adversely controlled to alter the energy supply. Stuxnet [28] is other representative case where an uranium enrichment plant was vulnerated leaving it useless and delaying the Irani project to produce energy by means of radioactive elements.
VI. Internet of Things technologies like Smart TVs, IP cameras, smart toys, among others, which lately has been used by hackers to carry out denial of services attacks against technological infrastructure.
VII. Large amount of people connected to different networks voluntarily or involuntarily, sharing personal data, e.g. photos, biographical data, mobile numbers and emails, through mobile applications and web services. This situation becomes a challenge for law enforcement agencies who must carry out investigations around a person, since they require proper tools and methodologies to support cyber intelligence labors over big amount of available personal data.

Many of these challenges are related to large flows of information available on the Internet that must be analyzed by the state security agencies to support cybersecurity and cyberdefense objectives defined in CONPES 3701. The collection, processing and analysis of all this public information can be supported by open source intelligence tools and methodologies. State security agencies in charge of cybersecurity and cyber defense uses OSINT to perform intelligence which represents a strategic advantage that allows to anticipate a possible attack mitigating the risks and conduct an investigation after an attack or crime.

This paper presented in section 2 an OSINT architecture applicable to state security agencies supported in a set of transforms applicable to the Colombian context. This contribution represents an advance in the prevention of attacks coming from internal and external agents allowing in many cases to develop a strategic anticipation to the adversary. In addition, the transforms improve the focus that the cyber analyst must have to deliver an intelligence product with accurate information, since there is contextualized. Additionally, this paper presents in section 3 an application of machine learning models to make sentiment analysis, which allows cyber analyst to generate early warnings regarding illicit acts and anticipate sabotage campaigns.

Contributions included in Section 3 and 4 can be used not only for cybersecurity purposes, but it could be used to prevent suicides, human traffic, citizen security perception, citizen complaints expecting to be attended, drug traffic, among others. Definition of public policies could also be aided from information collected and analyzed by transforms and models, due this could help to build a society profile that

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

allow to design policies accordingly. In this way is possible to analyze, identify, counteract, mitigate and prevent threats and even predict possible attacks. These contributions allow state security agencies to prospectively anticipate the threats new wars defined by George Friedman[29] in where he predicts that the upcoming wars will be based on a combination of computer sabotage and high-precision attacks against targets strategic to generate chaos.

Among the most popular OSINT solutions are: Voyager[5] from Voyager Labs (Israel), iSIHT[6] from Fireye (USA), iNSIGTH[7] from Checkpoint (Israel) and Flashpoint[8] from Flashpoint-intel (USA). The mentioned solutions are proprietary with high costs in licensing and maintenance. Additionally, in many cases these solutions are offered as Software as a Service (SaaS), which can be a disadvantage since highly confidential information is hosted in the cloud. When a proprietary solution is contracted, some intelligence labors can be developed by people not belonging to state security agencies, which could decrease the development of internal capacity and expertise. In addition, data collected by foreign solutions generally belongs to American and European sources but not Latin American or Colombian, which difficult the cyberintelligence tasks due Colombian security agencies requires mainly data within a Colombian context.

When the aforementioned solutions are not available, or these do not provide enough information, the cyber analyst proceeds to carry out search and collection in a manual way according to his knowledge of the threat which could generate subjectivity and vagueness. Additionally, manual activities increase the time required to develop a cyber intelligence cycle, making difficult to carry out other cyber security tasks in parallel.

The high ownership costs of proprietary OSINT solutions, the confidentiality of data and the collection of information from a Colombian or Latin American context pose the necessity of develop own cyberintelligence solutions. For these purpose, it is important to joint efforts between state, industry and academia around research, development and innovation in cybersecurity solutions supported using open source frameworks.

5. Conclusions

Any blog, web page, online newspaper, social network, forum and even free datasets can become a great source of information, which can be accessed to collect data used in cyber intelligence labors. Despite of this, privacy of data is a very serious issue, so it is also mandatory to know and recognize the difference between violating privacy and collecting information in reason of the protection of critical assets.

The analysis of sentiments can represent a great tool for law agencies to face crime, because it allows to determine and analyze the position of a criminal regarding a specific subject. It could allow to identify the reasons to carry out an attack on a person or organization. The identification of a possible adversaries serves to design and implement a cyber defense strategy that prevent future attacks.

One of the most valuable skill for cyber intelligence labors is to know how to look for and find information regarding a target. Law agencies and cyber intelligence organizations values this skill because it represents an advantage against adversaries that can be extremely useful to handle national security incidents. A final suggestion for organizations or individuals is to be aware of all the information that is shared or published on social networks or any web page. This is due through OSINT is possible that collect information which could be used by a criminal to achieve an attack.

5. http://voyagerlabs.co/
6. https://www.fireeye.com/solutions/isight-cyber-threat-intelligence-subscriptions.html/
7. https://www.insight.com/en_US/buy/partner/checkpoint.html
8. https://www.flashpoint-intel.com/

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

6. Future Works

We state as future work the automation of captcha resolution present in the services consulted by the transforms presented in this paper, so the user does not have to solve it manually. Different techniques can be explored like optical character recognition algorithms (OCR). Development of new transforms could also be considered as future work, which can be able to develop advanced searches, like the ones done by Intel Techniques. These new transforms could be integrated in an open source tool like Maltego looking for the integration of OSINT tools [30]. Transforms able to analyze unstructured information like Microsoft documents could also be useful to decrease the manual review and achieve more efficient cyber intelligence processes. Additionally, looking for information in the deep web, e.g. through TOR network, could be useful. Finally, improve the accuracy of descriptive machine learning models should also be considered for example using better training data sets which can be customized for the specific variations of a language (Colombian Spanish) or for specific contexts (professional or informal).

Acknowledgment

This work has been supported partially by the Colombian School of Engineering Julio Garavito (Colombia) through the project "Cyber Security Architecture for Incident Management", funded by the Internal Research Opening 2017.

References

[1] M. Glassman and M. J. Kang, "Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT)", *Computers in Human Behavior*, vol. 28, no. 2, pp. 673–682, 2012, https://doi.org/10.1016/j.chb.2011.11.014

[2] L. Brotherston and A. Berlin, "Defensive security handbook: best practices for securing infrastructure". O'Reilly Media, 2017.

[3] W. Alcorn, C. Frichot, and M. Orrù, "The Browser hacker's handbook", New Jersey: John Wiley and Sons, 2014.

[4] M. Gregg, "Certified Ethical Hacker (CEH) Version 9 Cert Guide" London: Pearson Education, 2017.

[5] P. Engebretson, "The basics of hacking and penetration testing" Syngressr Publishing, 2013.

[6] D. Bradbury, "In plain view: open source intelligence", *Computers in Human Behavior*, no. 4, pp. 5–9, 2011.

[7] B. de S. G. Rodrigues, "Open-source intelligence em sistemas SIEM" Lisboa: Universidade de Lisboa, 2015.

[8] C. Pérez, "Minería de datos: técnicas y herramientas" Paraninfo Cengage Learning, 2007.

[9] G. Subramanian, "R Data analysis projects: build end to end analytics systems to get deeper insights from your data", Birmingham: Packt Publishing, 2017.

[10] L. Zhang and B. Liu, "Sentiment Analysis and Opinion Mining". in Encyclopedia of Machine Learning and Data Mining, Boston: Springer, 2017, pp. 1152–1161, https://doi.org/10.1007/978-1-4899-7687-1_907

[11] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis", *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013, https://doi.org/10.1109/MIS.2013.30

[12] A. Ortony, G. L. Clore, and A. Collins, "The cognitive structure of emotions" Cambridge: Cambridge University Press, 1988, https://doi.org/10.1017/CBO9780511571299

[13] R. A. Stevenson, J. A. Mikels, and T. W. James, "Characterization of the Affective Norms for English Words by discrete emotional categories", *Behavior Research Methods*, vol. 39, no. 4, pp. 1020–1024, 2007, https://doi.org/10.3758/BF03192999

[14] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", In Pro*ceedings of the 40th Annual Meeting of the Association for*

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Computational Linguistics (ACL), Philadelphia, july 2002, pp. 417-424.

[15] S. M. Kim and E. Hovy, "Identifying and Analyzing Judgment Opinions", Association for Computational Linguistics Stroudsburg, pp. 200–207, 2006, https://doi.org/10.3115/1220835.1220861

[16] Liangxiao Jiang, H. Zhang, and Zhihua Cai, "A Novel Bayes Model: Hidden Naive Bayes", *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 10, pp. 1361–1371, 2009, https://doi.org/10.1109/TKDE.2008.234

[17] Y. Yang and G. I. Webb, "A Comparative Study of Discretization Methods for Naive-Bayes Classifiers", *J. Res.*, vol. 2, p. 267-324, 2007.

[18] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines", *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998, https://doi.org/10.1109/5254.708428

[19] F. Sebastiani, "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 1999, https://doi.org/10.1145/505282.505283

[20] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proceedings of ACL, pp. 271-278, 2004, https://doi.org/10.3115/1218955.1218990

[21] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis", Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347–354, 2005, https://doi.org/10.3115/1220575.1220619

[22] H. Wang, D. Can, A. Kazemzadeh, F. Bar and S. Narayanan, "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycl,". In 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, july, 2012.

[23] C-SPAN, "Robert Mueller on Cybersecurity" [En línea] Disponible en: https://www.c-span.org/video/?319726-3/robert-mueller-cybersecurity&start=1876

[24] Departamento Nacional de Planeación, "CONPES 3701 - Lineamientos de Política para Ciberseguridad y Ciberdefensa. Colombia". Consejo Nacional de Política Económica y Social, 2011.

[25] R. Rodríguez, "Guerra Asimétrica". [En línea]. Disponible en: https://dialnet.unirioja.es/descarga/articulo/4602435.pdf

[26] J. Nye, "Bound to Lead: The Changing Nature of American Power" Hachette U. Basic Books, 2016.

[27] G. S. Medero, "Ciberespacio y el crimen organizado. Los nuevos desafíos del siglo XXI", *Revista Enfoques*, vol.10, no. 16, pp. 71–87, 2012.

[28] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon", *IEEE Security and Privacy*, vol. 9, no. 3, pp. 49–51, 2011, https://doi.org/10.1109/MSP.2011.67

[29] G. Friedman, "The next 100 years: a forecast for the 21st century", Knopf Doubleday Publishing Group, 2009, pp. 193–212.

[30] R. Steele, "Handbook of Intelligence Studies" London: Routledge, 2007.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 15, N° 2 (julio-diciembre 2018). pp. 195-214. Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.