INTELLIGENT SYSTEM FOR DETECTION OF CYBERCRIME VOCABULARY ON WEBSITES

Iván Castillo-Zúñiga[1], Jaime Iván López-Veyna[2], Francisco Javier Luna-Rosas[3] y Gustavo Tirado-Estrada[1]

[1]TecNM/Instituto Tecnológico del Llano Aguascalientes, Profesor en departamento de Sistemas y Computación, Doctor en Ciencias de la Ingeniería, Carr. Ags-S.L.P. Km. 18. El Llano, Ags., México. Tel: +524492032068. ivan.cz@llano.tecnm.mx
[2]TecNM/Instituto Tecnológico de Zacatecas. Carr. Panam. entronque Guadalajara S/N. Domicilio la Escondida, Zac., México.
[3]TecNM/Instituto Tecnológico de Aguascalientes. Av. Adolfo López Mateos #1801 Ote. Fracc. Bona Gens, Ags., México.

Received: 4/Dic/2020 Reviewing: 13/Dic/2020--Accepted: 3/Mar/2020 - DOI: http://doi.org/10.6036/NT9589

TO CITE THIS ARTICLE
CASTILLO-ZUÑIGA, Ivan, LOPEZ-VEYNA, Jaime Ivan, LUNA-ROSAS, Francisco et al. INTELLIGENT SYSTEM FOR DETECTION OF CYBERCRIME VOCABULARY ON WEBSITES. DYNA New Technologies, January-December 2020, vol. 7, no. 1, [10 p.]. DOI: http://doi.org/10.6036/NT9589

ABSTRACT:

This article presents an intelligent system to detect Cybercrime lexicon on Web sites, to find knowledge about large amounts of information on the Internet in an acceptable response time. The proposed architecture uses a Web Scraper to locate and download information from the Internet. To obtain the linguistic corpus of Cybercrime, a parallel genetic strategy is executed, which distributes the processes of cleaning Web pages and the techniques for Natural Language Processing (tokenization, stop words, frequency of term, term frequency with inverse document frequency), together with lemmatization methods and synonyms. To obtain knowledge, a dataset was generated using a semantic ontology with the general characteristics of Cybercrime. To evaluate the efficiency of the model, supervised learning algorithms were used: Boosting, Neural Network and Random Forests in parallel. The results reveal 97.64% accuracy in the detection of Cybercrime vocabulary, which was verified by the LOOCV cross-validation technique, in addition, a time-saving was obtained in data recovery and knowledge search of 292% and 1220% respectively using parallel processing.

Keywords:

Cybercrime, Big Data Analytics, Web Mining, Semantic Web, Machine Learning, Intelligent Systems, Parallel Processing.

1. - INTRODUCTION

Although the benefits that new technologies have brought to our lives are undeniable, they have also caused new problems that were previously unknown, one of them is what is called Cybercrime. The phenomenon in which all can be victims to the extent that we carry out some type of usual activity on the Internet such as bank transactions, online purchases, communication on social networks, Internet searches, sharing information through email, among others, generating large volumes of data or better known as Big Data. These factors have led to new points of vulnerability, where the opportunity to commit a crime is present, making necessary an early detection and a quick response to this type of incident. Cyberspace is transforming into a battlefield, where Cybercrime represents a new danger and threat to people's safety. Information is essential for defending against this threat, success will be determined by the difference in information between victims and criminals [1].

Analyzing the large amount of information that circulates on the Web in order to obtain knowledge is a problem that is constantly developing, different authors have approached the problem from different perspectives, achieving some progress. Investigations such as Tao & Deokar [2], Sait et al. [3], Saraiba et al. [4], Han, Lee, & Kim [5], Stuart & Majewski [6], have developed software architectures and processes aimed at the treatment of large volumes of data, as well as the analysis of information in the semantic Web and Natural Language Processing (NLP). Various studies have focused on searching and locating information on the Web [4], distributed processing [5] in large volumes of data, and/or retrieving information using semantic Web and PNL techniques [2, 3, 4, 6], finally to the search for knowledge with machine learning techniques [2, 3, 6]. These investigations are summarized in Table I.

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345

Related research	Locate information in the Web		Big Data						Semantic Web	Natural Language Processing	Machine learning		
				Data source		Data processing						Processing	
	Crawler	Intelligent agent	Text document	Internet	Social network	Sequential	Parallel	Distributed			Supervised	Sequential	Parallel
Tao & Deokar [2] (2015)									✓	✓	✓	✓	
Sait et al. [3] (2015)									✓	✓	✓	✓	
Saraiba et al. [4] (2016)		✓		✓		✓			✓	✓			
Han, Lee & Kim [5] (2015)	✓				✓			✓	✓	✓			
Stuart, Majewski [6] (2015)			✓							✓	✓	✓	
Current research	✓		✓	✓	Do not Apply	✓	✓		✓	✓	✓	✓	✓

This investigation is not considered an intelligent agent, because a crawler is required to locate the information.

Considered as future work.

Table I. Characteristics of Architectures and processes related to the semantic Web, NLP and Big Data Analytics.

It is important to point out, that some of the cited investigations present partial solutions implementing different techniques and software developments, remaining pending the integration of other processes, among them, linguistic corpus, and datasets based on semantic ontologies combined with supervised learning algorithms. Given these drawbacks, in addition to the lack of consensus between the different approaches on methodology, as well as, processes used to analyze large volumes of data on the Internet to extract knowledge. This research proposes a solution based on the creation of an intelligent system that combines Big Data Analytics, Semantic Web and NLP techniques to detect on Web sites the words related to the subject of Cybercrime, taking into account the benefits that can be obtained from the information found on the Internet, discovering hidden patterns, unknown correlations, and additional information that can be very useful for organizations in decision making.

The scope of the study is oriented to the detection of Cybercrime words, considering the processes: a). Recovery of Web pages on the Internet, b). Dataset pre-processing, and c). Data analysis and classification, in addition to optimize the precision of our word detection algorithm with a parallel version on a multicore computer. The main contribution of the study focuses on a study model on data related to social problems on the Internet.

Cybercrime

Cybercrime is a current topic of great importance and interest, as it represents one of the greatest threats to society in the world. Cybercrime is aimed to achieving a mainly economic benefit, where the victim is the key element in the production of the criminal event on the Internet, since it determines its risk area by incorporating certain assets into cyberspace by interacting with others and particularly with strangers, and mainly by not using all possible self-protection measures [7]. A series of advantages derive from these aspects from which criminals, take advantages (anonymity, there are no borders, credibility over false businesses, simplicity due to computational ignorance, speed over data transmission and minimal investment to commit the crime) [8].

Medina & Molist [9], establishes that Cybercrime is a computer crime carried out through illicit operations through the Internet. From another perspective, Poveda [10] mentions that Cybercrime is any criminal conduct that uses computer technology to carry it out, whether as a method, means or end. Linked to the concept, Sánchez [11] set that, among the most common crimes of Cybercrime, we can consider computer fraud, theft of personal information, computer hacking, computer espionage, commercial piracy and other crimes against intellectual property, the invasion of privacy, the distribution of illegal content, incitement to prostitution and other crimes against morality and organized crime.

Investigations such as Leukfeldt [12], Al-Nemrat & Benzaid [13], Aguilar [14], Vlachos, et al. [15], reveal that different types of crimes are committed in cyberspace in various parts of the world. In Amsterdam, criminal groups use social media and social engineering as one of their tools to obtain information from their victims and commit scams [12]. In Jordan, a study reveals that users who visit Internet cafes are more likely to engage in risky behaviors online, making them more likely to be victims of Cybercrime [13]. In the United Kingdom, it has been detected that children between 8-17 years are more vulnerable to Cybercrime, so the government has implemented prevention and training actions in schools [14]. In Greece, the most common computer crimes where the user requests help from the government are financial fraud, cyberbullying and extortion [15].

Table II shows an analysis of characteristics of studies related to the detection of Cybercrime in digital media, where Sait, Bhandari, Khare, et al. [3] present a study focused on intrusion detection based on network anomalies. The process uses a dataset made up of

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345

labels that represent the usage logs of the local area network of the proxy server. For the classification of the information, they are based on the Bayes algorithm and for the detection of attacks in the time series algorithm. At the same time, Al-Nemrat & Benzaid [13] propose an approach based on the combination of decision trees and tree regression (CART) to detect cybercrime profiles. The results reveal that in four of six trees generated, Internet café users are more likely to be victims of Internet fraud, identity theft, piracy and cyberbullying. The tests were carried out in the R programming language with a dataset made up of surveys carried out by internet café users in Jordan. It should also be added that Alami & Elbeqqali [16], whose propose a method to detect and predict criminal activities in the messages of microblogs. The process breaks down the posts posted by users in terms and compares them using a distance of similarity, with the terms of suspicious Cybercrime behaviors stored in a database. The solution considers the synonymy and polysemy within the term analysis looking for better precision in the detection of Cybercrime. For their part, Wan, Ali & Mohamed [17], describe a strategy to detect malicious URLs based on a list of combined characteristics that can be used as patterns to define whether a website is benign or malicious. The detection process is carried out with statistical techniques using Java. Finally, Portnoff, Afroz, Durrett, et al. [18] describe a tool to analyze underground Internet markets that use forums to buy and sell a series of stolen items, data sets, resources, and criminal services. To evaluate they included eight different forums, achieving 80% accuracy in detecting publication categories, products, and prices. His technique integrates a machine learning approach, MSV and NLP.

Related research	Orientation of the investigation	Data source	Statistical techniques	Machine Learning: Processing: Sequential: Distance (Similarity)	Machine Learning: Processing: Sequential: CART trees	Machine Learning: Processing: Sequential: Naive Bayes	Machine Learning: Processing: Sequential: K-nn	Machine Learning: Processing: Sequential: Decision Trees	Machine Learning: Processing: Sequential: VSM	Machine Learning: Processing: Sequential: Random Forests	Machine Learning: Processing: Parallel: Genetic algorithm	Machine Learning: Processing: Parallel: Random Forests	Information search (Web Scraper)	Language Processing Natural: TF-IDF	Language Processing Natural: Synonyms	Language Processing Natural: Tokenize	Language Processing Natural: Lemmatize	Language Processing Natural: Cleaning	Semantic Web: Ontologies	Data organization: Corpus	Data organization: Dataset (numeric)
Sait et al. [3] (2015)	LAN Network Attacks	Server Proxy		✓		✓								✓		✓	✓				✓
Al-Nemrat & Benzaid [13] (2015)	Cybercrime Profile	Cyber Cafe Users			✓			✓										✓			✓
Alami & Elbeqqali [16] (2015)	Microblogs	Twitter Hashtag		✓											✓					✓	✓
Wan, Ali & Mohmad [17] (2019)	Malicious Features	URLs	✓																	Table of Characteristics malicious	
Portnoff et al. [18] (2017)	Illegal sales In Internet	Forums on the Web							✓					Implement tool (CoreNLP toolkit)							✓
Current research	Cybercrime vocabulary	Internet Web Pages	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Stop Word	✓	✓	✓

In the present study, tokens are processed, so these algorithms oriented to the analysis of other types of data are not considered.

Table II. Characteristics of work related to the detection of Cybercrime.

2. – METHODOLOGY

Fig. 1 shows the methodology used in this investigation. It is made up of a series of techniques, starting with obtaining data on the Web and ending with the detection of Cybercrime words. The procedure is supported by Big Data Analytics, Semantic Web and NLP techniques.

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345

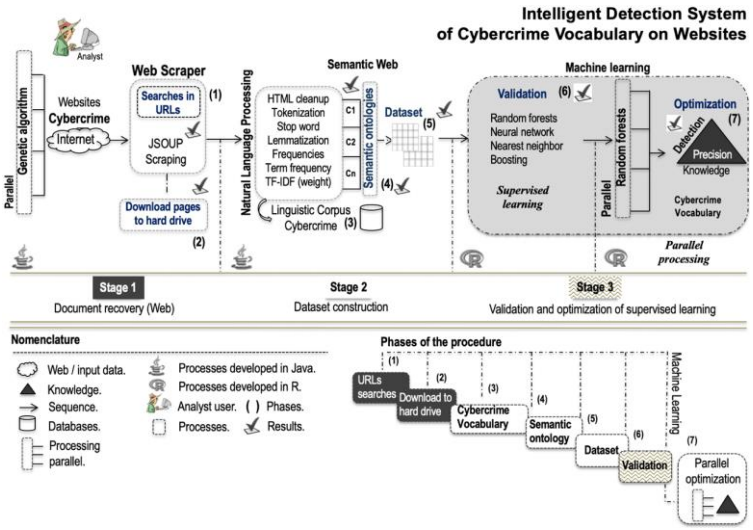

Fig. 1. Procedure to detect Cybercrime vocabulary on Web sites with Big Data Analytics, Semantic Web and NLP techniques, optimizing with parallel processing.

Stage 1 includes the recovery of Cybercrime documents. In ***phase 1***, a Web Scraper was developed for searching Web pages on the Internet with instructions based on open source JSOUP libraries (Java library for working with HTML code), combining ***phase 2*** with Web Scraping techniques (java libraries. net) to analyze the HTML code and obtain the name of the Web pages, Internet addresses and make a copy of the file to the hard disk (java.io libraries).

The construction of the dataset is developed in ***stage 2***, where the Cybercrime vocabulary is obtained using NLP techniques. ***Phase 3*** separates the HTML from the text and segments the words that come from the Web pages (tokenization) by storing the words in a database. Likewise, words without meaning (stop words) are eliminated, the number of vocabulary words and the number of times they are repeated (TF) is calculated, and a weight (TF-IDF) is also assigned to indicate the importance of each word over the files being scanned. Finally, the root of the word (stemming) is obtained with Porter's algorithm [19], in order to consider words in a search such as: "robbery, attaching robbery, robberies, steal, stealing", which allows us to include synonyms and obtain a more accurate result. It is important to mention, that a genetic strategy was developed to equally distribute the analysis of Web pages using parallel processing with the processor cores to optimize sequential processing. Among the techniques considered for optimization (hill-climbing, simulated recollection, and genetic algorithms), we opted for a genetic strategy to represent processes in parallel, forming a cluster with the processor core simulating a chromosome and its genes. The genetic process evolves until reaching the optimum in the distribution of Web pages, and to carry out the termination criterion with an adaptation function based on the mean. The evolution of the population is based on elitism, selection by tournament, the crossing of a point and mutation with random replacement. As a product of this phase, the linguistic corpus of Cybercrime in Spanish is obtained, the number of words recovered and the execution time for both the sequential process and the parallel process.

Phase 4 describes an object-oriented semantic ontology model, where the class represents the domain of the main topic "Cybercrime", the subclasses represent the ontologies (concepts) to be analyzed, such as general words on the subject of Cybercrime, computer fraud, information theft, commercial piracy, organized crime, among others, and attributes represent each word of the subclass. The construction of ontologies is based on the methodology called "Methontology", which has its origins in a Chemical ontology of the Polytechnic College of Madrid [20], but which has been adopted as a mature methodology for the development of ontologies, since it is independent of the characteristics of the area. It should be noted that one of the advantages of this model is that it allows the analysis of Cybercrime through words and, in turn, defines necessary ontologies at any time, allowing (n) datasets to be built for testing.

Table III, shows the 107 words related to the Cybercrime lexicon used for the tests in this research, this set of words is based on the books "Cybercrime" author Medina & Molist [9] and "Crimes on the Net" author Poveda [10]. The process begins by identifying sub-areas of the concept of Cybercrime (who carries it out, to whom it is directed, purpose, means, type of crime, synonyms and legal aspects) followed by the words that make them up. It is important to mention that the literature does not report any Cybercrime ontology created previously.

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345

Number / Cybercrime term in Spanish

Quien lo realiza		Propósito						Tipo de delito		Sinónimos		Aspecto legal	
1.	Culpable.	21.	Injuria.	43.	Computarizado.	66.	Spam.	68.	Crakeo.	90.	Fraude.	100.	Legal.
2.	Delincuente.	22.	Difamar.	44.	Nube.	67.	Malware.	69.	Manipular.	91.	Falsificación.	101.	Política.
3.	Hackers.	23.	Abuso.	45.	Internet.			70.	Modificar.	92.	Destruir.	102.	Ley.
4.	Hackeo.	24.	Daño.	46.	Red.			71.	Pishing.	93.	Cibercrimen.	103.	Penal.
5.	Hackear.	25.	Robo.	47.	Correo.			72.	Transacción.	94.	Ciberdelito.	104.	Sanción.
6.	Sujeto.	26.	Hurto.	48.	Social.			73.	Bancaria.	95.	Ciberbullying.	105.	Protección.
7.	Persona.	27.	Chantaje.	49.	Web.			74.	Pornografía.	96.	Ilícito.	106.	Prevención.
8.	Individuo.	28.	Amenaza.	50.	Tablet.			75.	Prostitución.	97.	Crimen.	107.	Criminal.
9.	Habilidad.	29	Ataque.	51.	Smartphone.			76.	Infantil.	98.	Criminalidad.		
10.	Experto.	30.	Perdida.	52.	Laptop.			77.	Clonación.	99.	Delito.		
11.	Especialista.	31.	Estafa.	53.	Software.			78.	Tarjeta.				
12.	Cibercriminal.	32.	Riesgo.	54.	Sistemas.			79.	Cuenta.				
A quien va dirigido		Medio		55.	Programar.			80.	Crédito.				
13.	Dinero.	33.	TIC	56.	Windows.			81.	Magnético.				
14.	Efectivo.	34.	Tecnología.	57.	Linux.			82.	Cajero.				
15.	Capital.	35.	Comunicación.	58.	Mac.			83.	Automático.				
16.	Economía.	36.	Información.	59.	Digital.			84.	Piratería.				
17.	Víctima.	37.	Virtual.	60.	Online.			85.	Copia.				
18.	Gobierno.	38.	Informática.	61.	Telemática.			86.	Película.				
19.	Empresa.	39.	Dispositivo.	62.	Máquina.			87.	Música.				
20.	Conocimiento.	40.	Electrónico.	63.	Malicioso.			88.	Contenido.				
		41.	Computador.	64.	Troyano.			89.	Sabotaje.				
		42.	Ordenador.	65.	Virus.								

Table III. Sub-areas and words related to the Cybercrime lexicon in Spanish.

The construction of the dataset, ***phase 5***, is based on the "table testing", a training model which is the most common cross-validation training models [21]. This model consists of creating a set of tests that will be examined by the supervised learning algorithms to determine the precision of the model. The testing table is made up of the predictor variables, which integrate the words determined in the semantic ontology listed in Table III, to identify and classify the Web pages; and the variable to be predicted, which is identified with the name "Correct" and is defined with the values "Yes / No". This variable determines the algorithm response from the values of the predictor variables. To carry out this experiment, the value "Yes" was determined when the frequency of the sub-areas shown in Table III is greater than zero, and "No" otherwise. The variable to be predicted is controlled "supervised" by an external agent "an expert", from there the basic foundation of supervised learning [22]. Subsequently, the algorithm is trained with 70% of the sample (learning table), and tested with the remaining 30% (test table), to estimate the generalization error of the sample, obtaining, as a result, the percentage of precision in the Cybercrime vocabulary detection.

The validation and optimization of supervised learning, ***in stage 3***, focuses on machine learning techniques to automate the construction of analytical models to learn from data of repetitive way and optimize the algorithm with greater precision in detection Cybercrime.

To evaluate the dataset in ***phase 6***, it was determined to consider the related works by selecting the algorithms that present the greatest similarity with this research: Neural network [6], Boosting [23] and Random forests [24], which according to their characteristics it is not surpassed in precision by current algorithms. After the tests, an analysis is carried out, indicating the algorithm that achieve the best results, and the conditions for detecting the Cybercrime vocabulary.

Phase 7 performs the optimization process by selecting the algorithm with the highest percentage in the prediction of cybercrime, and improves it with parallel processing, obtaining execution time and precision percentage, comparing it with sequential processing.

The construction of the parallel random forests algorithm is based on symmetric multiprocessing parallel environment (SMP), effective in a multicore computer, which distributes in parallel the workload among the system cores called calculation nodes, sharing the same memory, disk space, and operating system. The R language provides the doMC, foreach, and multicore libraries for this purpose, which work only on Unix-fork operating systems. The algorithm begins by defining a cluster with the number of processors cores and establishes the total number of trees to generate, specifying their distribution by each core. Finally, the results are combined in a single forest and a percentage of precision and execution time are obtained.

Parallel processing of random forests is carried out using a master/slave architecture which communicates with the processor cores, establishing a parallel backend with doMC and parallelizing the code with foreach. The process begins with the selection of individuals at random using replacement sampling to create different datasets. Subsequently, a large number of independent decision trees are generated, in this research, we determined to generate at least 200 trees per cores at random to select the best variable in each node of the tree. To find the best division in the tree, the classification error, the Gini index, and entropy are used within the Hunt algorithm with a top-down approach (divide and conquer) which is one of the most common methods [25], in this method, each tree is evaluated independently. In the end, the prediction is carried out based on the average of the result of the trees or by majority vote in the classification.

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345

Materials

The software tools used in this study include the Java programming language (version 8.0) and the MySQL database manager (version 5.7), which were used to build the search processes and download Web pages (phase 1 and 2); for the construction of the genetic algorithm with the NLP techniques, as well as for the construction of the linguistic corpus (phase 3); for the design of semantic ontologies based in object (phase 4); and to obtain the dataset (phase 5). The R version 3.2.2 language was used for the construction of the supervised learning algorithms (parallel processing) with the purpose to carry out the prediction tests (phase 6 and 7).

The computer equipment used in the experiment was a MacBook Pro with a 2GHz Intel Core-i7 processor, 8GB memory and a 250GB flash hard drive, connected at an Internet speed of 50 MB for searching and downloading web pages.

3. - RESULTS

Testing procedure

The proposed test method allows the analysis of small and large data sets, and was performed following the procedure described below:

- The semantic ontology was established with the words of Cybercrime and their structures were built with metadata using automatic SQL statements.
- The set of Cybercrime websites was located using the Web Scraper for testing, ***stage 1.***
- The cybercrime linguistic corpus was obtained through data pre-processing with NLP and semantic Web techniques, ***stage 2.***
- The structure of the semantic ontology was linked to the databases of the linguistic corpus and the dataset was obtained for the Cybercrime detection tests.
- Supervised learning was used in the prediction tests, ***stage 3***, supporting the machine learning with the algorithms: Neural network (perceptron), Boosting (AdaBoost) and Random forests (randomForest).
- An analysis was performed with the aforementioned algorithms, indicating the one that showed better results in the percentage of precision and the conditions for detecting the vocabulary of Cybercrime.
- Finally, the precision percentage of the algorithm was improved and the response time was optimized with parallel processing, comparing it with sequential processing.

Results

Within the results framework, it is specified that the sample of 1,326 URLs located and downloaded to hard disk represent the different Cybercrime Web pages in Spanish, since the search for the Web Scraper on the Internet showed repeated URLs, mostly after 1000 The selection criteria was to collect all the Web sites that were related to the Cybercrime topic using a list of synonyms as a search seed.

Regarding the pre-processing time to obtain the linguistic corpus of the Cybercrime, eight tests were carried out in the investigation using 1,326 URLs. The first test considered the sequential process obtaining the longest time with 57.23 minutes implementing a computer core. In the same sense, with the genetic strategy, the remaining evaluations were carried out combining two to eight cores on one chromosome, as shown in Fig. 2. The results discover a trend in decreasing time, where the shorter time is 19.57 minutes, generating a 292% saving when executing a chromosome with 8 processes in parallel, notably improving the response time in the results. The idea of gradually increasing processor cores in tests demonstrates the hypothesis: "the more parallel processes are executed, the greater the response time gains".

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345

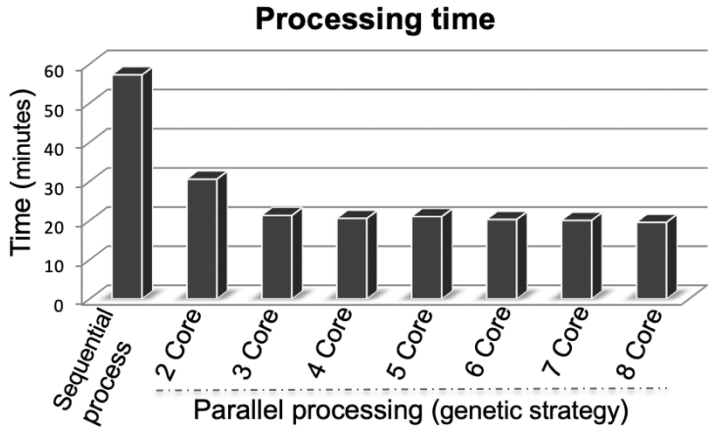

Fig. 2. Genetic strategy for obtaining the Cybercrime vocabulary.

In the procedure, 1,321,258 words were extracted and 519,862 stop words were eliminated, obtaining a linguistic corpus of 801,396 words. For the construction of the dataset, 143,208 queries were needed on the databases of the linguistic corpus with the predictor variables of Table III, and the variable to be predicted "Correct", forming an array of 108 words per 1326 web pages, exporting the data to an open format (CSV) in the form of a table that supports the R language for testing.

The process to evaluate the model is based on the supervised learning method, which determines the precision and accuracy with which the dataset is built, having a classifying variable that indicates whether or not the analyzed web page belongs to Cybercrime. Based on this criterias, we select some algorithms of machine learning to learn to classify Web pages that belong or not belong to cybercrime. Using 70% of the dataset to learn how to classify, and the remaining 30% to test its effectiveness. The results of those tests are shown in Table IV. It should be noted that 32 iterations were performed in each algorithm with different groups of Web pages based on the central limit theorem, which establishes that after test 32 the results do not vary [26]. We found that, the algorithms: Boosting and Random Forests, obtained the average of the 32 executions, the best precision with 94% and 95%, respectively, while the Neural Network algorithm was the algorithm presented the lowest percentage with 92% precision.

Iteration number	Web page groups	Neural Networks	Boosting	Random Forest
1	50	89.94%	93.45%	**93.96%**
2	100	90.93%	95.72%	**95.46%**
3	150	93.21%	92.19%	**94.97%**
4	200	91.43%	93.21%	**94.72%**
5	250	91.20%	92.19%	**95.22%**
6	300	90.93%	94.22%	**92.21%**
7	350	92.21%	92.71%	**95.72%**
8	400	90.95%	93.96%	**94.96%**
9	450	92.46%	94.72%	**95.97%**
….		….	….	….
32	1326	92.21%	95.78%	**95.46%**
Average		**92%**	**94%**	**95%**

Table IV: Detection of Cybercrime vocabulary.

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345

Optimization starts with a process to rescue the numerical variables of the dataset, so that none influences another as a predictor when building predictive models. Furthermore, it is verified that there is an equal proportion between correct and incorrect web pages for the training and validation set.

The optimization results of the random forest algorithm (selected), are obtained from the techniques of Cross-Validation leaving one out (LOOCV) and Cross-Validation of K iterations (K-Fold). For each LOOCV iteration, one record was considered for testing and the rest for training. In the experiment, 1326 records were used, leaving one out, so 1326 repetitions were generated. The result of the average of the precisions is high, with 97.64% constant in 32 different iterations, but it has a high computational cost due to the number of iterations that may be required depending on the total number of records. In the same sense, the K-Fold Cross Validation technique was considered to compare the precision results and the error of the classification techniques. For 10 groups that we were used, one group of the total data was used for testing and the remaining nine groups for training. It was performed 32 times to verify the variability of the results, obtaining an average precision of 96.52% close to the LOOCV technique, but with a low computational cost. The tests were performed using functions from the Caret and e1071 libraries of the R language.

Table V shows the behavior of the random forest algorithm with parallel processing, using different groups of cores with different sets of Web Pages for the detection of Cybercrime vocabulary. In the description, the runtime is included and the sequential process is compared with the parallel processing. It is important to comment that, because of the articles included in Tables 1 and 2, none report parallel processing times, it was decided to compare the parallel processing time against the sequential processing time, to show the benefits of the proposed strategy.

		Execution time																	
		I1	I2	I3	I4	I5	Average	I1	I2	I3	I4	I5	Average	I1	I2	I3	I4	I5	Average
Sequential process		1.13	1.04	1.08	1.08	1.11	1.16	2.55	2.65	2.58	2.55	2.61	2.57	5.38	5.50	5.37	5.35	5.32	**5.37**
Parallel Process Core groups	2 Core	0.30	0.27	0.67	0.28	0.32	0.34	0.31	0.31	0.32	0.30	0.30	0.31	0.4	0.4	1.7	0.4	0.4	**0.94**
	3 Core	0.29	0.42	0.37	0.28	0.30	0.35	0.35	0.36	0.37	0.36	0.35	0.35	0.44	0.43	0.44	0.47	0.44	**0.44**
	4 Core	0.31	0.32	0.31	0.31	0.30	0.30	0.35	0.38	0.38	0.37	0.37	0.37	1.22	1.20	1.19	0.49	1.19	**0.98**
	6 Core	0.58	0.55	0.58	0.53	0.37	0.55	0.78	0.41	0.43	0.42	0.41	0.56	1.79	1.79	1.18	1.16	1.79	**1.54**
	8 Core	0.59	0.59	0.59	0.57	0.38	0.60	0.47	0.44	0.45	0.49	0.82	0.75	2.44	2.38	2.38	2.36	1.79	**2.26**
		400 records (Web Pages)						**800 records** (Web Pages)						**1326 records** (Web Pages)					

* The first 5 iterations of the 32 performed due to space issues are shown in the table.

Table V: Time optimization with parallel processing of the random forest algorithm.

Among the findings, it is found that using the dataset of 1326 Web Pages, the smallest time with 0.44 seconds is identified when using a cluster formed by 3 cores in parallel and the longest time with 5.37 seconds with the sequential process, obtaining 1220% savings of time with parallel processing optimizing the response time in the results considerably.

Fig. 3 shows the results graphically over the test execution time, using three sets of web pages. Another discovery indicates that the algorithm shows stability from 3 parallel processes in the three data sets.

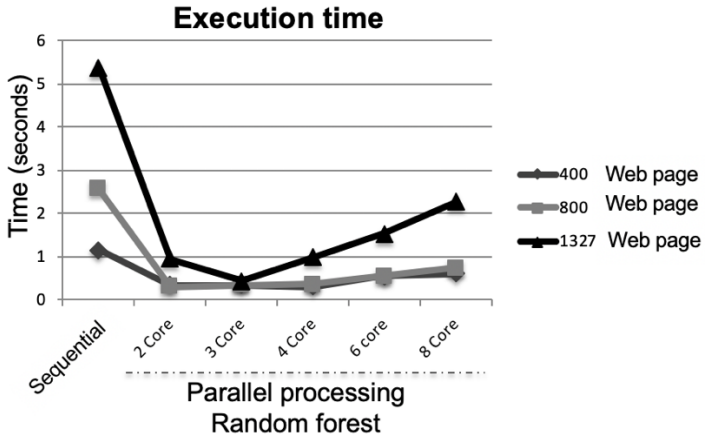

Fig 3. The behavior of random forests with sequential and parallel processing.

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345

4. – CONCLUSIONS AND FUTURE WORK

This research project shows an intelligent system that combines different techniques for the analysis of information on the Internet (case study: Cybercrime Lexicon), integrating a Web Scraper for locating and downloading Web pages; HTML code cleaning processes and removal of stop words; NLP techniques to build the language corpus; creating semantic ontologies for the construction of the dataset; the use of machine learning for search to knowledge with the algorithms: Neural network, Boosting and Random forests; Big Data Analytics to optimize data recovery and knowledge acquisition using parallel processing with genetic algorithms and random forests, proving to be effective in detecting cybercrime vocabulary on the Internet.

It is important to highlight that this model can represent an interesting contribution to the analysis of data on social problems on the Internet. Additionally, it should be noted that adding Big Data Analytics techniques in the information retrieval processes and in the discovery of knowledge makes it possible to streamline the analysis and classification of Web pages for the user with the acceptable response time.

On the other hand, the results report high percentages in the detection of the Cybercrime vocabulary, achieving a 97.64% accuracy with the LOOCV technique using random forests. In addition, the response time is optimized by 1220% when using parallel processing and 292% in saving time in the recovery of words with the genetic strategy, both percentages compared to the sequential process.

As future work in this research, it has been considered to use this methodology for the analysis of feelings in social networks and microblogs. Similarly, future tests will be considered on massive parallel processing environments (MPP), which involve a cluster of networked computers, and/or GPU processors, considering R's doMPI library, and CUDA [27,28].

REFERENCES

[1] Schenone F. Una visión sobre el Ciberterrorismo. 1ª ed. 2014. *Editorial Amazon (Versión Kindle)*. ASIN: B006QQ08BI.

[2] Tao J, & Deokar A. Semantics-based event log aggregation for process mining and analytics. *Springer (Inf-Syst-Front)* 2015. p.1209-1226. DOI: http://dx.doi.org/ 10.1007/s10796-015-9563-4

[3] Sait S, Bhandari A, Khare S, et al. Multi-level anomaly detection: Relevance of Big Data Analytics in Networks. *Springer 2015.* p.1737-1767. DOI: http://dx.doi.org/10.1007/s12046-015-0416-0.

[4] Saraiba C, Fusco E, Santarem J, et al. Ontological Semantic Agent in the Context of the Big Data: a tool applied to information Retrieval in Scientific Research. *Springer 2016*. p.307-316. DOI:10.1007/978-3-319-31232-3_29.

[5] Han Y, Lee H, Kim Y. A Real-time Knowledge Extracting System from Social Big Data using Distributed Architecture. *ACM 2015.* p.74-79. DOI: http://dx.doi.org/10.1145/2811411.2811481.

[6] Stuart K.D, & Majewski M. Intelligent Opinion Mining and Sentiment Analysis Using Artificial Neural Networks. *Springer (ICONIP)* 2015. Part IV, p.103-110. DOI: http://dx.doi.org/10.1007/978-3-319-26561-2_13.

[7] Miro F. La victimización por cibercriminalidad social. Un estudio a partir de la teoría de las actividades cotidianas en el ciberespacio. *Rev. Española de Investigation Criminológica*. 2013. Art. 5. Vol. 11. p.1-35. ISSN:1696-9219.

[8] Moise A. Some Considerations on the Phenomenon of Cybercrime. *Journal of Advanced research in Law and Economics*. 2014. Vol. 5 Summer. 1(9). p.38-43. DOI: http://dx.doi.org/10.14505/jarle.v5.1(9).04.

[9] Medina M, y Molist M. Cibercrimen. 1ª ed. 2015. *TIBIDABO EDICIONES*. ISBN:978-84-1620-482-3.

[10] Poveda M. Delitos en la Red. 1ª ed. 2015. *Editorial Fragua*. ISBN:978-84-7074-682-6.

[11] Sánchez G. Cibercrimen, Ciberterrorismo y Ciberguerra: Los nuevos desafíos del siglo XXI. *Revista Cenipec*. 2012. Vol. 31. p.239-267. ISSN:0798-9202.

[12] Leukfeldt E. Cybercrime and social ties. *Springer (Trends Organ Crim)*. 2014. Vol. 17. p.231-249. DOI: http://dx.doi.org/10.1007/s12117-014-9229-5.

[13] Al-Nemrat A, & Benzaid C. Cybercrime profiling: Decision Tree, Induction, Examining Perceptions of Internet Risk and Cybercrime Victimization. *IEEE (Trustcom/BigDataSE/ISPA)* 2015. p.1380-1385. DOI: http://dx.doi.org/10.1109/Trustcom.2015.534.

[14] Aguilar M. Cibercrimen y cibervictimización en Europa: instituciones involucradas en la prevención del ciberdelito en Reino Unido. *Revista Criminalidad*, 2015. 57(1). p.121-135.

[15] Vlachos V, Minou M, Assimakopouos V, et al. The landscape of Cybercrime in Greece. *ACM (Information Management & Computer Security)* 2011. Vol. 19 p.113-123. DOI: http://dx.doi.org/10.1108/09685221111143051.

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345

[16] Alami S, & Elbeqqali L. Cybercrime Profiling: Text Mining Techniques to detect and predict criminal activities in microblog posts. *IEEE.* 2015. p.1-5. ISBN:978-1-4799-7560-0.

[17] Wan, W., Ali, A., & Mohmad, M. Characterizing Current Features of Malicious Threats on Website. Ed. 2019. *Editorial Springer*. p.210-218. DOI: http://dx.doi.org/10.1007/978-3-030-00979-3_21.

[18] Portnoff R, Afroz S, Durrett G, et al. Tools for Automated Analysis of Cybercriminal Markets. 2017. *ACM (International World Wide Web Conference)*. p.657-666. DOI:http://dx.doi.org/10.1145/3038912.3052600.

[19] Porter MF. An Algorithm for suffix stripping. *Emerald Insight* 2006, Vol. 40 p.211-218. DOI: http://dx.doi.org/10.1108/00330330610681286.

[20] Gómez-Pérez, A., M. Fernández-López, & Corcho, O. Ontological Engineering: with examples from the areas of knowledge management, London: *Springer*-Verlan, 2004.

[21] Rodríguez O. Validación cruzada (cross-validation) y Bootstrapping. [Online]. Disponible:

http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentación_-_CV.293124233.pdf

[22] Rodríguez O. Aprendizaje Supervisado. [Online]. Disponible:

http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentación_-_KNN.20085205.pdf.

[23] Biba M, & Gjati E. Boosting Text classification through Stemming of Composite Words. 2014 *Springer*, p.185-194. DOI:10.1007/978-3-319-01778-5_19.

[24] Breiman L, & Cutler A. "Random Forest". Recuperado en octubre 2016 de

http://www.stat.berkeley.edu/~breiman/RandomForest/cc_home.htm

[25] Han J, & Kamber M. Data Mining: Concepts and Techniques. 2ª ed. 2006. *Morgan Kaufman*. ISBN:978-1-55860-901-3.

[26] Fischer H. A history of the Central Limit Theorem from classical to modern probability Theory. 2010. *Springer* ISBN:978-0-387-87856-0.

[27] McCallum Q.E & Weston S. Parallel R. Versión Kindle. 2012. *Editorial O'REILLY*. ISBN:978-1-449-30992-3.

[28] Hothorn T & Everitt B.S. A Handbook of Statistical Analyses Using R. 3ª Ed. 2014. *Editorial CRC Press*. ISBN:978-1-4822-0458-2.

ISSN: 2254-2833 / DYNA New Technologies Vol 7 nº1 DOI: https://doi.org/10.6036/NT9345