DETECTION OF CYBERCRIME VOCABLES ON WEB PAGES WITH INTELLIGENT METHODS IN PARALLEL

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

[1]TecNM/Instituto Tecnológico del Llano Aguascalientes (México)

[2]TecNM/Instituto Tecnológico de Zacatecas (México)

[3]TecNM/Instituto Tecnológico de Aguascalientes (México)

DOI: http://dx.doi.org/10.6036/9755

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 the so-called Cybercrime. A phenomenon in which we can all 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 latent, making necessary early detection and a quick response to this type of incident. Cyberspace is transforming into a battlefield, where Cybercrime represents a new danger and a 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].

To offer a solution to the problem posed, Fig. 1 presents the methodology used in this investigation to detect Cybercrime lexicon on Web sites using intelligent parallel techniques.

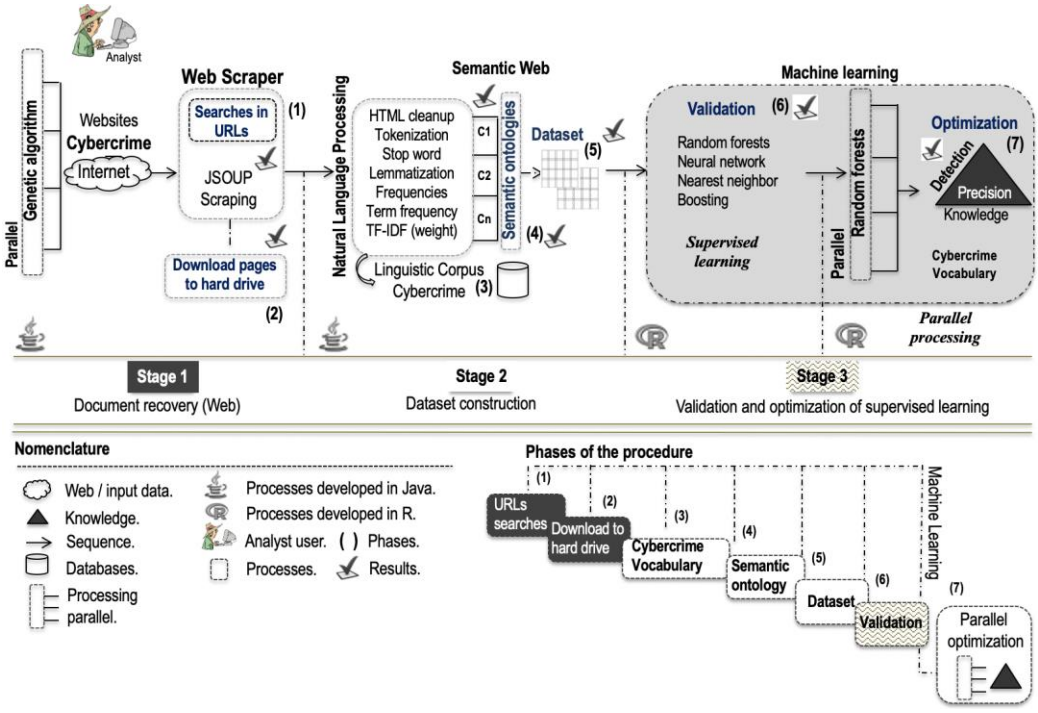

Figure 1: Expert system to detect Cybercrime words on Web sites with intelligent parallel methods. Source: Castillo-Zúñiga et al. [1].

The procedure is made up of a series of techniques, starting with obtaining Web pages and ending with the detection of words related to Cybercrime. To locate and download information from the Internet, a Web Scraper is used. To obtain the linguistic corpus of Cybercrime, a parallel genetic strategy is executed, in which the cleaning processes of Web pages and the techniques for Natural

Language Processing (NLP) are distributed, such as tokenization, stop words, frequency of term, frequency of term with an inverse frequency of the document, together with stemming methods and synonyms. To obtain knowledge, a dataset is generated that makes use of a semantic ontology with the general characteristics of Cybercrime. To evaluate the efficiency of the model, the supervised learning algorithms are used: boosting, neural network, and parallel random forests. The results show a 97.64% accuracy in detecting the Cybercrime vocabulary, which was corroborated using the LOOCV and K-Fold cross-validation techniques. Also, time savings in data retrieval and knowledge search of 292% and 1220% respectively were obtained using parallel processing. It should be noted that the Java and R programming languages were used to construct the algorithms.

The words related to the Cybercrime lexicon used for the tests in this investigation are supported in the books "Cybercrime" by Medina & Molist [2] and "Crimes on the Net" by Poveda [3]. The process begins by identifying sub-areas of the concept of Cybercrime, such as: 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, criminal, hackers, theft, scam, victim, piracy, sabotage, prostitution, among others; the complete vocabulary is pointed out in Castillo-Zúñiga et al [1]. It is important to mention that the literature does not report any Cybercrime ontology created previously.

The advantages of this study are focused on the combination of a series of techniques for the analysis of information that comes from the Internet, such as Big Data Analytics / Machine Learning (with supervised learning), NLP and Semantic Web (specifically semantic ontologies), proving to be effective in detecting Cybercrime vocabulary on Web sites. In addition, it is important to mention that considering intelligent methods in parallel, both for data recovery (genetic algorithm) and knowledge discovery (random forest), makes it easier to analyze and classify Web pages with an acceptable response time.

In conclusion, it is important to mention that the proposed methodology can be considered as a contribution to data analysis focused on harassment on the Internet. In the same way, conducting research aimed at cyberterrorism, cyberwarfare, hacktivism, workplace harassment, school harassment, among others.

REFERENCES

[1] Castillo-Zúñiga, I., Lopez-Veyna, J., Luna-Rosas, F., Tirado-Estrada, G. (2020). Intelligent System for Detection of Cybercrime Vocabulary on Websites. *DYNA New Technologies*, 7(1). [11 p.]. DOI: http://dx.doi.org/10.6036/NT9589

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

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