Cod. 9755 | Ciencia de los ordenadores | 1203.04 Inteligencia artificial

Detección de vocablos de cibercrimen en páginas web con métodos inteligentes en paralelo

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] y 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

Para ofrecer una solución a la problemática planteada, la Fig.1, presenta la metodología utilizada en esta investigación para detectar léxico de Cibercrimen en sitios Web utilizando técnicas inteligentes en paralelo.

El procedimiento está compuesto por una serie de técnicas, iniciando con la obtención de páginas Web y terminando con la detección de vocablos relacionados con Cibercrimen. Para ubicar y descargar la información de Internet, se utiliza un Web Scraper. Para obtener el corpus lingüístico de Cibercrimen, se ejecuta una estrategia genética en paralelo, en la cual se distribuyen los procesos de limpieza de páginas Web y las técnicas para el Procesamiento de Lenguaje Natural (PLN), como: tokenización, stop words, frecuencia de término, frecuencia de término con frecuencia inversa del documento, en conjunto con métodos de lematización y sinónimos. Para obtener conocimiento se genera un dataset que hace uso de una ontología semántica con las características generales del Cibercrimen. Para evaluar la eficiencia del modelo se utilizan los algoritmos de aprendizaje supervisado: potenciación, red neuronal y bosques aleatorios en paralelo. Los resultados arrojan un 97.64% de precisión en la detección del vocabulario de Cibercrimen, los cuales fueron corroborados mediante las técnicas de validación cruzada LOOCV y K-Fold. Además, se obtuvo un ahorro de tiempo en la recuperación de datos y búsqueda de conocimiento del 292% y 1220% respectivamente usando procesamiento paralelo. Cabe resaltar que, para la construcción de los algoritmos se utilizaron los lenguajes de programación Java y R.

Las palabras relacionadas al léxico de Cibercrimen utilizadas para las pruebas en esta investigación son sustentadas en los libros "Cibercrimen" autor Medina & Molist [2] y "Delitos en la red" autor Poveda [3]. El proceso inicia identificando subáreas del concepto de Cibercrimen, como: quien lo realiza, a quien va dirigido, propósito, medio, tipo de delito, sinónimos y aspectos legales, seguido de las palabras que los conforman, delincuente, hackers, robo, estafa, víctima, piratería, sabotaje, prostitución, entre otras; el vocabulario completo se puntualiza en Castillo-Zúñiga et al [1]. Es importante mencionar que la literatura no reporta ninguna ontología de Cibercrimen creada con anterioridad.

Las ventajas del presente estudio, se centran en la combinación de una serie de técnicas para el análisis de información que proviene de Internet, Big Data Analytics / Machine Learning con aprendizaje supervisado, PLN y Web semántica (en específico ontologías semánticas), demostrando ser eficaz en la detección de vocabulario de Cibercrimen en sitios Web. Además, es importante mencionar que considerar métodos inteligentes en paralelo, tanto para la recuperación de datos (algoritmo genético), como para el descubrimiento de conocimiento (bosques aleatorios), permite agilizar el análisis y clasificación de las páginas Web con un tiempo de respuesta aceptable.

En conclusión, es importante mencionar que la metodología propuesta pueda considerarse como una contribución para el análisis de datos enfocado al acoso cibernético en Internet. De la misma manera dirigirse a casos de estudio orientados al ciberterrorismo, ciberguerra, hacktivismo, acoso laboral, acoso escolar, entre otros.

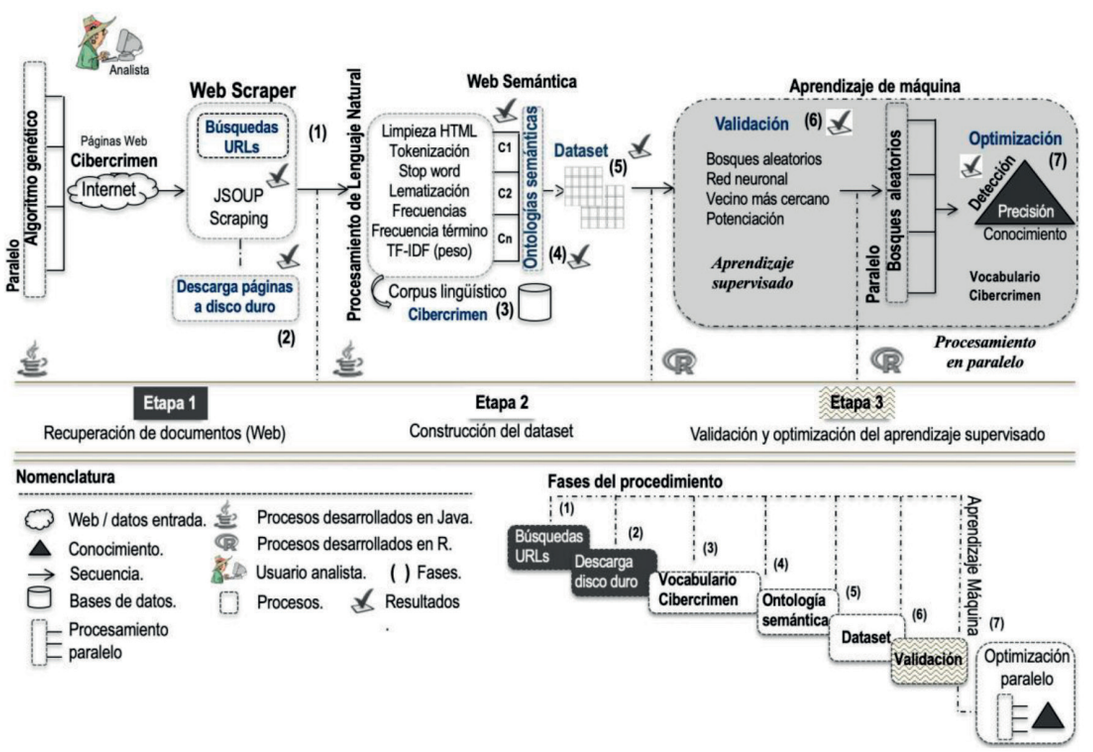

Figura 1: Sistema experto para detectar vocablos de Cibercrimen en sitios Web con métodos inteligentes en paralelo. Fuente: Castillo-Zúñiga et al. [1]

Si bien resultan innegables los beneficios que las nuevas tecnologías han traído a nuestra vida, estos también han provocado nuevos problemas que antes eran desconocidos, uno de ellos es el denominado Cibercrimen. Fenómeno en el cual todos podemos ser víctimas en la medida en que realizamos algún tipo de actividad usual en Internet como: transacciones bancarias, compras en línea, comunicación en redes sociales, búsquedas en Internet, compartir información a través del correo electrónico, entre otros, generando grandes volúmenes de datos o mejor conocidos como Big Data. Estos factores han propiciado nuevos puntos de vulnerabilidad, donde la oportunidad de cometer un crimen es latente, haciendo necesaria una detección temprana y una respuesta rápida ante este tipo de incidentes. El ciberespacio se esta transformando en un campo de batalla, en donde el Cibercrimen representa un nuevo peligro y amenaza para la seguridad de las personas. La información es esencial para la defensa contra esta amenaza, el éxito estará determinado por la diferencia de información entre víctimas y criminales [1].

REFERENCIAS

[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.