Combinación de técnicas de big data analytics y web semántica para la detección de vocabulario de acoso escolar en internet

Combination of techniques of big data analytics and semantic web for the detection of vocabulary of harassment school in internet

Iván Castillo-Zúñiga[1], Francisco-Javier Luna-Rosas[1], Jaime Muñoz-Arteaga[3] y Jaime-Iván López-Veyna[4]

[1] Instituto Tecnológico de Aguascalientes (Méjico)
[2] Instituto Tecnológico el Llano Aguascalientes (Méjico)
[3] Universidad Autónoma de Aguascalientes (Méjico)
[4] Barcelona Supercomputing Center (España)

DOI: http://dx.doi.org/10.6036/8159

Interactuar con el mundo mediante redes sociales, buscar y recibir información, realizar transacciones, colaborar y compartir contenidos independientemente de nuestra situación geográfica e idioma se han convertido en actividades cotidianas que realizamos día a día a través de la Web. La gran cantidad de dispositivos conectados a Internet (*móviles, tabletas electrónicas o Pc's*), así como el crecimiento de la infraestructura de red, son algunos de los factores que han contribuido al éxito de la Web generando grandes volúmenes de datos (*Big Data*). Sin embargo el proceso para analizar la enorme cantidad de datos en la Web todavía debe establecerse, en este sentido existen Arquitecturas de Software poco definidas, algunas se han enfocado a la búsqueda y localización de información en la Web [1], al procesamiento distribuido en grandes volúmenes de datos [1], y a la recuperación de información con técnicas de Procesamiento de Lenguaje Natural (*PLN*) [1], implementando soluciones parciales que no han sido del todo aceptadas por la comunidad científica.

Para obtener un valor agregado de la información, es necesario llevar a cabo una administración eficaz en los datos y aplicar distintas técnicas de procesamiento que permitan manejar grandes cantidades de información en tiempo real (*velocidad*), asimismo utilizar procedimientos para analizar datos complejos semiestructurados y no estructurados como documentos, imágenes, vídeos, música, entre otros (*variedad*) [1]. Por otro lado se deben considerar problemas como la dificultad para la interpretación de palabras o significados (*semántica*), integración de información (*datos dispersos y sin relación*) y la recuperación de datos (*problemas de sinonimia, polisemia y multilingüismo*) [1].

El problema inicial de esta investigación reside en que la información se encuentra en formato no estructurado, en grandes cantidades de datos dispersa en sitios Web, con distintos protocolos de seguridad y comunicación, además, los textos presentan errores ortográficos, abreviaturas y sinónimos. Para iniciar el proceso de detección del vocabulario de un tema es necesario transformar la información a datos estructurados. Y para obtener conocimiento de la información, una de las opciones que recientemente han tomado auge, es la aplicación de técnicas del área de Aprendizaje Máquina (*Machine Learning*).

La problemática abordada en esta investigación considera los siguientes puntos:

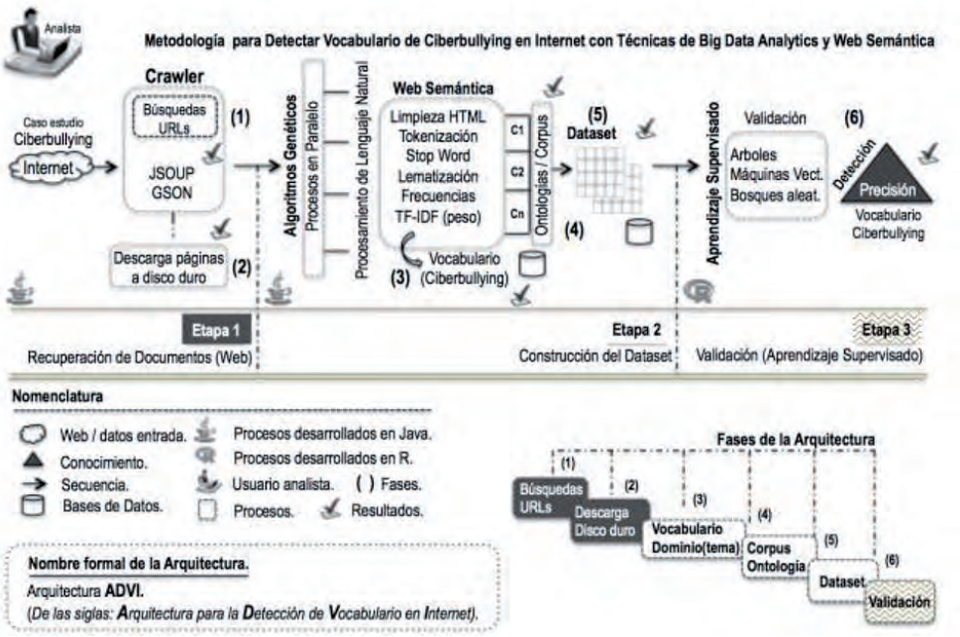

Fig. 1: Arquitectura (ADVI) para la Detección de Vocabulario de Ciberbullying en Internet combinando Técnicas de Big Data Analytics y Web Semántica. Fuente: Castillo, Luna, Muñoz, et al. [1]

- Dificultad en la búsqueda y recuperación de grandes cantidades de información en la Web.
- La información está mezclada con código HTML, abreviaciones y errores ortográficos.
- Alta complejidad en el preprocesamiento de datos debido a la sinonimia, polisemia y multilingüismo.
- Gran consumo de memoria RAM y procesamiento lento en grandes cantidades de información.
- Transformación de información a datos estructurados para su análisis.
- Dificultad para obtener conocimiento útil de la información (*valor en los datos*).

La principal aportación de esta investigación es una arquitectura de software que implementa una metodología inspirada en técnicas de *Big Data Analytics* (*Algoritmos Genéticos, Aprendizaje Supervisado (Machine Learning)*) y Web Semántica (*Ontologías, Vocabularios, Corpus Lingüísticos*) [1], con el objetivo de analizar grandes cantidades de información en la Web y obtener conocimiento que apoye en la toma de decisiones a organizaciones, empresas y sociedad en general.

A efectos de probar la arquitectura y sus métodos, como caso de estudio se analizaron páginas Web sobre el *Ciberbullying*. Descrito como fenómeno de acoso escolar que sufren los niños y adolescentes haciendo uso de Tecnologías de la Información para acosar, hostigar e intimidar a un individuo a través de ataques personales, divulgación de información confidencial o falsa [2].

La Fig. (1), muestra el esquema que representa la arquitectura de software planteada para la detección de vocabulario de *Ciberbullying* en Internet con técnicas de *Big Data Analytics*, Web Semántica y Procesamiento de Lenguaje Natural.

La arquitectura diseñada por nosotros [1], utiliza un *Crawler* para localizar y descargar la información en la Web. Para la recuperación del vocabulario se implementó una estrategia genética en paralelo que integra técnicas de Web Semántica (*ontologías*) y Procesamiento de Lenguaje Natural (*Limpieza HTML, Tokenización, Stop Word, Frecuencia de Término (TF) y Frecuencia de Término con Frecuencia Inversa del Documento (TF-IDF))*, métodos de lematización y sinónimos, con el propósito de recuperar más información. Para obtener conocimiento se utilizaron los algoritmos de Aprendizaje Supervisado, Arboles de Decisión, Máquinas de Soporte Vectorial y Bosques Aleatorios.

La arquitectura propuesta utiliza una metodología que integra un proceso completo que inicia obteniendo datos en la Web y termina con la detección de vocabulario (*conocimiento*), uniendo distintas técnicas de manera natural. Los resultados obtenidos en distintas fases del estudio en comparación con investigaciones similares como las de Kontostathis, Reynolds, Garrón et al. [3], muestran una mayor precisión, obteniendo un conjunto de datos que reporta porcentajes elevados en la detección del vocabulario del *Ciberbullying* logrando un 95% de precisión [1]. De igual manera se optimiza el proceso secuencial con una estrategia genética en paralelo reduciendo considerablemente el tiempo de procesamiento en un 302% [1]. Hay que mencionar, además, que el uso de ontologías semánticas facilitó el análisis de la información generando (n) conjuntos de datos con distintas perspectivas [1].

Por lo tanto, se puede mencionar, que nuestra arquitectura ha demostrado ser eficaz en la detección de vocabulario de *Ciberbullying* en Internet. Es importante resaltar que esta arquitectura puede representar una interesante aportación al análisis de datos sobre problemas sociales en Internet. Adicionalmente, cabe destacar que el agregar técnicas de *Big Data Analytics* en los procesos de recuperación de información, permite agilizar el análisis y clasificación de páginas Web originando información valiosa y conocimiento significativo para el usuario.

REFERENCIAS

[1] Castillo-Zúñiga I, Luna-Rosas F, Muñoz-Arteaga J, et al. "Architecture (ADVI) for the Detection of Cyberbullying Vocabulary in Internet Combining Techniques of Big Data Analytics and Semantic Web". Dyna New Technologies, Enero-Diciembre 2016, vol. 3, no. 1, p.0. DOI: http://dx.doi.org/10.6036/NT8032

[2] Ortega JI, González DL. "Análisis del impacto del Cyberbullying en el rendimiento académico de estudiantes de nivel medio superior". 1ª ed. 2015. Ed. Instituto Universitario Anglo Español. ISBN: 978-607-9003-22-7

[3] Kontostathis A, Reynolds K, Garron A, et al. "Detecting Cyberbullying: Query Terms and Techniques". ACM–WebSci. 2013, Paris, Francia, p. 195 - 204. ISBN: 978-1-4503-1889-1.