UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

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

Actualidad Tecnológica

Avances y desafíos de métodos y modelos computacionales aplicados al análisis de información en redes sociales

Advances and challenges of computational methods and models applied to information analysis in social networks

Johan David Díaz Mendivelso[1], Marco Javier Suárez Barón[2]

Para citar: J. D. Díaz-Mendivelso, M. J. Suarez-Barón, "Avances y desafíos de métodos y modelos computacionales aplicados al análisis de información en redes sociales". *Revista Vínculos: Ciencia, Tecnología y Sociedad,* vol. 16, no. 2, julio-diciembre de 2019, pp. 298 -309. doi: https://doi.org/10.14483/2322939X.14714

Enviado: 13/10/19/ Recibido: 15/10/19/ Aprobado: 23/11/19

Resumen

Este artículo presenta la revisión de la literatura científica dirigida al estudio y análisis del estado actual de investigaciones relacionadas con la aplicación de métodos y modelos para el análisis social en entornos digitales, los cuales faciliten el descubrimiento de conocimiento a partir de la gestión de información contenida en redes sociales de tipo corporativo. El estudio explora temas relacionados con la extracción de información útil y análisis de cadenas textuales utilizando técnicas de indexación semántica latente apoyadas por el procesamiento del lenguaje natural (PLN). Para la revisión, se aplicó una metodología basada en el planteamiento de palabras clave que funcionen como insumo en la búsqueda de documentación en bases de datos indexadas y fuentes primarias; los documentos resultantes se filtran en un análisis detallado que se realiza individualmente, seleccionando así las mejores fuentes para plantear una revisión técnica. Por último, se plantean algunos resultados y trabajos futuros para garantizar el inicio de nuevas investigaciones.

Al realizar la revisión planteada, se detecta que investigaciones de este tipo establecen un camino apropiado para las organizaciones empresariales y sociales, dando estrategias computacionales para descubrir conocimiento a través de técnicas de visualización de patrones, los cuales ayudan a la toma de decisiones sobre I+D+i y permiten garantizar el desarrollo y avance de planes operativos. Se justifica la necesidad de realizar y llevar a cabo proyectos relacionados con temas de análisis de información que se encuentra en entornos virtuales como lo son las redes sociales, aplicando técnicas de PLN y modelos I+D+i; con lo anterior, se podría dar vía libre para el planteamiento de nuevos proyectos que pertenezcan al área de conocimiento.

Palabras Clave: análisis social, crawling, extracción de información, I+D+i, indexación semántica latente, procesamiento de lenguaje natural, redes sociales.

Abstract

This article presents the review of the scientific literature aimed at the study and analysis of the current state of projects and research related to the application of methods and models for social analysis in digital environments that facilitate the discovery of knowledge, based on information management contained in corporate social networks. The study explores topics related to the extraction of useful information and analysis of textual strings using latent semantic indexing techniques supported by natural language processing (PLN). For the review, a methodology based on the approach of keywords was applied, which function as input for the search of documentation in indexed databases and primary

1. Ingeniero de sistemas y computación, Universidad Pedagógica Tecnológica de Colombia. Docente investigador, Universidad Pedagógica y Tecnológica de Colombia. Correo electrónico: johan.diaz@uptc.udu.co.
2. Estudios de postdoctorado, Universidad de Viena. Profesor asociado, Universidad Pedagógica y Tecnológica de Colombia. Correo electrónico: marco.suarez@uptc.edu.co.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

sources; The resulting documents are filtered in a detailed analysis carried out individually in each selected document and thus select the best sources and propose a technical review. Finally, some results and future work are proposed, to guarantee the start of new investigations.
When conducting the proposed review, it is detected that research of this type establishes an appropriate path for business and social organizations, establishing computational strategies to discover knowledge through pattern visualization techniques, which help to make decisions about R&D. and that allow to guarantee the development and advancement of operational plans.

Finally, the need to carry out and carry out projects related to information analysis issues found in virtual environments such as social networks, applying PLN techniques and R + D + i models; with the above, it could be given free way for the approach of new projects that belong to the area of knowledge.

Keywords: social analysis, crawling, Information extraction, R&D, latent semantic indexing, natural language processing, social networks.

1. Introducción

El crecimiento exponencial de la internet ha traído cambios a la manera en que se comunican y comparten información las personas debido a la aparición de las redes sociales, lo cual generó un nuevo modelo de enlace y manejo de la vida social [1]; hoy día todo el mundo usa estos entornos virtuales como puente para controlar actividades personales, comerciales y educativas. Así, el uso concurrido de las redes sociales deja una gran cantidad de información que al ser operada puede generar análisis provechosos que ayuden al desarrollo de una organización [2].

Por lo anterior, surge la necesidad de crear herramientas que ayuden a analizar con mayor precisión y en un mínimo de tiempo grandes volúmenes de datos; de ahí emerge la expectativa de realizar una revisión bibliográfica de técnicas y herramientas que ayuden al tratamiento de la información encontrada dentro de entornos virtuales y físicos. En la actualidad, para el manejo de información física se aplican técnicas de ETL (extracción, transformación y carga), donde la correcta aplicación del proceso asegura la manipulación de información ya digitalizada usando algoritmos especializados. Por otro lado, la revisión de técnicas que ayuden a la gestión de información como lematización sintáctica e indexación semántica latente, pueden apoyarse en el procesamiento de lenguaje natural (PLN) con el fin de crear un sistema de capas para análisis de cadenas textuales extraídas de redes sociales, encontrando así fuentes que ayuden como soporte a nuevas investigaciones que identifiquen y descarten información borrosa y generen información útil; esto último puede funcionar como técnica para encontrar patrones que fortalecerán la toma de decisiones.

Finalmente, se desea revisar los marcos de referencia de innovación, desarrollo e investigación I+D+i para analizar los requisitos y pautas necesarias, crear y ejecutar proyectos correctamente que garanticen la generación de nuevas estrategias para el progreso de las organizaciones, obteniendo un punto de partida con el fin de iniciar proyectos de sistemas computacionales que garanticen resultados favorables sin importar el sector productivo donde se implementen.

2. Desarrollo del tema

Con el fin de completar el desarrollo de un proyecto centrado en el manejo de técnicas de lenguaje natural para el análisis sintáctico y semántico de cadenas textuales extraídas de redes sociales pertenecientes a centros o grupos de investigación que funcionan sobre estándares de innovación, desarrollo e investigación I+D+i, es necesario realizar una revisión de estado de arte para estudiar las investigaciones actuales y así garantizar lo inédito del proyecto y posibles investigaciones que ayuden como insumo para su ejecución.
La búsqueda de documentación se realizó en bases de datos indexadas usando palabras claves relevantes a los temas centrales del proyecto. Los términos principales fueron redes sociales *(social networks),* web social, fuentes de información,

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

crawling y extracción de información *(extraction of information)*. También se aplicaron búsquedas relacionadas con: técnicas para la extracción de datos, minería de texto y minería web, análisis sintáctico *(syntactic analysis),* análisis léxico *(lexical analysis),* lematización sintáctica *(syntactic lematization)* e indexación semántica latente *(latent semantic indexing)*. En cuanto a los temas enfocados a la estructura gramatical, se exploró alrededor de análisis de cadenas textuales, lenguaje natural *(natural language)* y procesamiento de lenguaje natural *(natural language processing)*. Por último, I+D+i, innovación *(innovation),* desarrollo *(development),* investigación *(research)* y R&D para cubrir el enfoque a marcos de referencia para la creación de proyectos I+D+i.

Los resultados de las búsquedas se pueden analizar en la Figura 1, donde se observa que varían dependiendo el tema y la base de datos en las que se encontraron los artículos usados para la conformación del estado de arte.

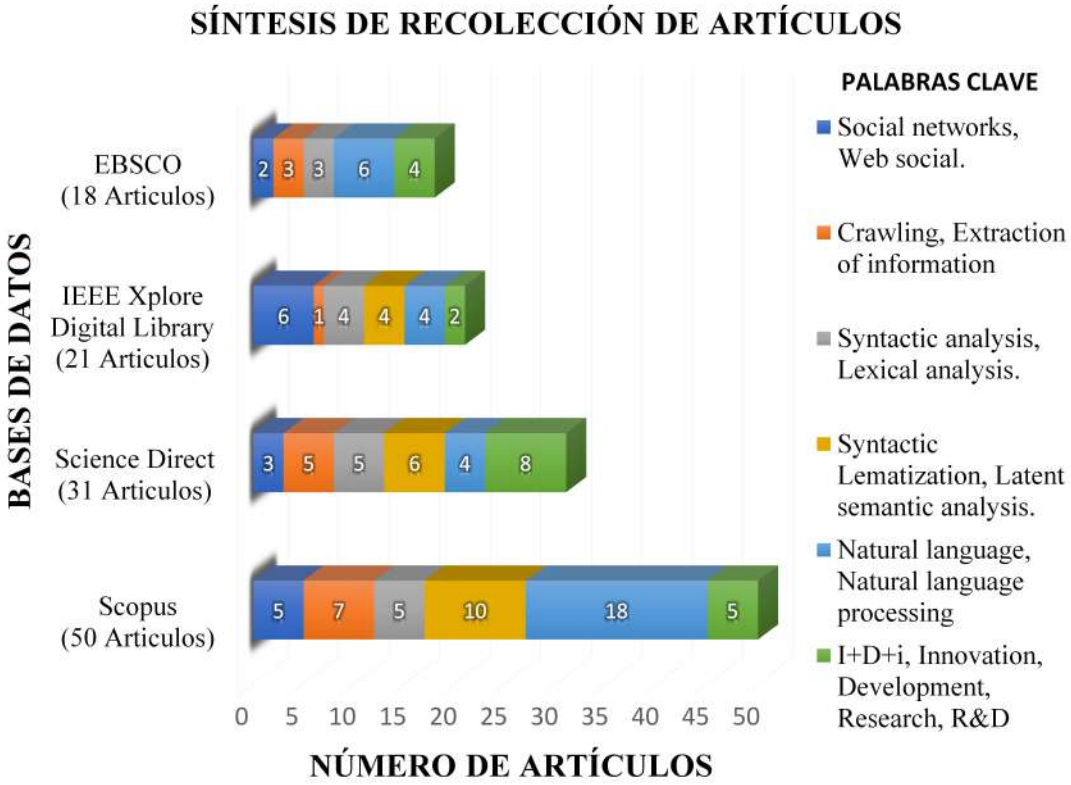

Figura 1. Síntesis de recolección de resultados de material científico.
Fuente: elaboración propia.

Con los artículos encontrados, se comienza a explorar tema por tema, seleccionando los productos más relevantes de cada uno y filtrando la información más relevante. Cabe aclarar que cada uno de los artículos referentes encontrados fueron revisados aplicando metodologías de revisión técnica.

2.1. Redes sociales como fuente de información

La evolución de los medios de comunicación ha obligado a las personas y organizaciones a migrar al uso de ambientes tecnológicamente más avanzados, ello partiendo de que la difusión de la información se ha realizado por medios físicos como centros de discusión o conversaciones persona a persona y que el progreso tecnológico ha generado que hoy en día se realice en entornos virtuales [1]. Esta nueva tendencia ha convertido el texto en un componente clave para la comunicación en la sociedad, consolidando este método como el más adecuado para el intercambio de información gracias a su elevada usabilidad digital [2]. Un ejemplo claro de esta migración

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

tecnológica son las redes sociales, al ser usadas se logró una mejor comunicación entre las personas, lo que converge entonces a un medio en que las organizaciones puedan usar este espacio como una estrategia de difusión publicitaria, *marketing* o planeación estratégica.

Con la concurrencia de usuarios en tan poco tiempo, las redes sociales se han convertido en las fuentes más grandes de datos debido a su gran flujo de información; proyectos como [3] afirman que el análisis de los datos albergados en los servidores de las redes sociales se puede convertir en una buena estrategia de negocio. Información que se puede considerar relevante si son escenarios como lecciones aprendidas, casos de éxito y experiencias significativas, ya que pueden ser insumos para la gestión de conocimiento estratégico [4]. Estos aportes pueden darse por diferentes usuarios y grupos en la red en entornos como blogs (Blogger), redes sociales (Facebook, Twitter) o videos (YouTube, Vimeo) [5], es así que la gestión y análisis de esta información permitirá establecer la generación de ventajas competitivas.

Muchas de las redes sociales usan mecanismos que ayudan a detectar información o tendencias a partir del análisis del uso o rastros que dejan los usuarios al pasar por alguna publicación [6], estos análisis pueden partir desde comentarios que se realizan sobre algún tema relevante a el análisis de sentimientos que se demuestran en la actuación de las personas a la hora de interactuar con el sistema [5], o tan solo dando a conocer su opinión con un emoticón que demuestre el estado de ánimo en el que se encontraban.

We are social [7] es una organización mundial que tiene como objetivo incentivar el uso de redes sociales para ayudar a generar reconocimiento a organizaciones o mejorar la comunicación entre las personas que hacen uso de ellas; al tiempo que se crean nuevos estilos de conexión entre personas y organizaciones, se crea una nueva vía que contribuye con la externalización del conocimiento. Cabe aclarar que esta organización se centra en el estudio y publicación de los resultados sobre el uso del internet y redes sociales en el mundo, cada año es realizada una conferencia donde se exponen los datos recolectados, demostrando un claro aumento de los usuarios conectados a internet en el transcurso de los últimos años. We are social en su última conferencia muestra que de los 7593 millones de personas en el mundo, 4021 millones están conectados a internet, y de ellos 3196 millones pertenecen a redes sociales, dando a conocer un aumento del 13% de usuarios nuevos en el transcurso de un año [8]. Así, se afirma cada vez más que el internet y en particular el uso de las redes sociales se ha transformado en una herramienta de fácil acceso y de gran acogida por la población mundial, convirtiéndose en punto clave para la dilución o externalización de la información.

Toda la acogida de las redes sociales que se ha mencionado ha obligado a muchas personas, sociedades y organizaciones a ser parte de ellas, la información que se maneja es tan importante que muchas empresas en el mundo están dispuestas a adoptar estas tecnologías para ser parte de tal impacto [9]. Lo anterior se debe a que las redes sociales han evolucionado y ya no son solo usadas para compartir ocio, sino que también se convierten en una fuente que integra gran cantidad de información de innovación, tecnología, cultura, sociedad, entre muchos otros elementos de interés. Así, las empresas ven las redes sociales como un canal de libre acceso para crecer empresarialmente y mejorar sus relaciones comerciales o dar una mejor publicidad a sus productos [10].

Al hablar de toda la información que se maneja y se guarda en las redes sociales, se puede afirmar que es casi imposible que no exista algún tipo de información útil que brinde soluciones o nuevas ideas [11]. Algunas organizaciones como Coca-Cola Company, Hewlett-Packard y Walmart han usado estos medios para lograr una mejor acogida dentro de la sociedad, donde, apoyadas por centros de investigación, se desarrollan proyectos como: "How Large U.S. Companies Can Use Twitter and Other Social Media to Gain Business Value" [12]. En estos se demuestra que el uso de las redes sociales ayuda a la empresa a hacer publicidad de sus productos, a generar un nuevo reconocimiento de la marca y crear enlaces que ayuden a la elaboración de análisis y recolección de información a partir de los comentarios que dan los usuarios, encontrando información beneficiosa que ayude a solucionar

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

incógnitas o mejorar los productos. De igual manera, este artículo hace notar los resultados conseguidos por organizaciones que usan redes sociales, las cuales aseguran que la elección e interpretación de información es un buen punto de partida para la ayuda en la toma de decisiones.

2.2. Crawling

La extracción de información se ha convertido en una necesidad debido al crecimiento exponencial de las páginas web que se han creado en el transcurso de los años y por los datos que se alojan en ellas; se genera entonces un nuevo campo de investigación, el cual propone crear nuevos y mejores métodos de extracción de información a partir de múltiples fuentes heterogéneas [13]. Todo esto parte de la necesidad de manipular la información de entornos web de una manera local y no remota como lo son las consultas en internet y, para ello, es necesaria la implementación de nuevas técnicas.

Una de las técnicas más aplicadas es conocida como *crawling* [12]-[14], este método ayuda al rastreo y extracción de información de páginas web o entornos virtuales, debido a que son sistemas desarrollados sobre lenguajes de programación robustos y seguros para lograr crear motores de extracción de información de un dominio de internet determinado; así, al final se consiguen programas informáticos o *API'square* que al ser ejecutados logren realizar tal tarea automáticamente.

Crawling, debido a su funcionalidad principal, ha sido una técnica muy acogida dentro de los proyectos relacionados con la exploración de información en entornos web, ello debido a que es adaptable y reprogramable para que se centre en temas específicos y, al mismo tiempo, es compatible con otros métodos o técnicas que ayudan a la optimización del proceso de rastreo y extracción de información [15]. Existen varias herramientas ya desarrolladas como lo pueden ser *RCrawler* [16], un paquete de R que ayuda al rastreo de información; xCrawl [17] es una herramienta creada para la minería de datos sobre entornos web o Crawljax [18], una herramienta elaborada a partir de JavaScript que se centra en el seguimiento de los cambios que se hacen sobre las URL visitadas por los usuarios para crear análisis de interfaces.

Esta técnica se ha convertido en una de las más usadas por los buscadores de internet, uno de los proyectos más relevantes es la aplicación de *crawling* en el robot de Google "GoogleBot" [19], este robot o software es desarrollado a partir de las técnicas de rastreo y extracción de información de las páginas que están en la red, y dependiendo el tipo de contenido que ofrece, se ubican los resultados de un manera descendente dependiendo la calidad de información de cada uno, garantizando sitios web que contengan información confiable para el usuario.

2.3. Lematización sintáctica e indexación semántica latente (ISL)

El aumento de información manejada por los sistemas informáticos ha crecido sustancialmente, lo que genera la necesidad de crear nuevas herramientas que garanticen su correcto análisis. Un ejemplo claro es el seguimiento de las opiniones que se hacen en entornos virtuales, esto se puede convertir en pieza clave para generar nuevo conocimiento [20]; sin embargo, el problema actual es que muchas de las herramientas usadas no poseen una solución para todos los posibles problemas de lingüística que se puedan presentar [21], ya que el análisis de texto es aplicado a partir de técnicas de similitud, pero solo entre palabras contenidas en una oración, sin tener en cuenta el contexto en que se aplica. Esto último se convierte en un problema mayúsculo debido a cambios morfológicos que presentan los idiomas y jergas al usar el lenguaje de forma fonética o gramatical [22].

De esta manera, desde la necesidad de obtener informes o reportes que ayuden a la toma de decisiones, el analizar la información permite una mejor exploración de datos extraídos de la web, ya que para lograr un mayor aprovechamiento es necesario detectar la similitud entre los datos aplicando técnicas morfológicas, semánticas, sintácticas y léxicas [21]-[23]. Teniendo en cuenta lo mencionado, se puede analizar que es una de las maneras más precisas de obtener resultados favorables que ayuden a la ejecución de tareas o actividades como lo puede ser la minería de datos, la extracción de información provechosa y análisis

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

de sentimientos [24].

Una de las fuentes de información heterogénea más usadas para la extracción y análisis de información son los blogs y redes sociales, pero el análisis de esta información se convierte en un reto para los investigadores debido a que todo lo que es publicado es escrito sin cumplir ninguna regla gramatical ni léxica. De esta manera, tal información es declarada como ruidosa [25]. El anterior argumento demuestra la necesidad de crear nuevas herramientas que ayuden a analizar cadenas textuales que no cumplan con las reglas gramaticales estándar.
El problema descrito anteriormente da lugar a la aplicación de un método complementario conocido como técnicas de lematización [26]. La lematización complementa la semántica, sintáctica y la léxica jugando un rol importante en el análisis de información registrada en la web, un caso clave es el uso de tales técnicas en la investigación: "Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features" [27]. Esta se centra en el análisis de similitud de los estados de Twitter que publican los medios de comunicación árabe para poder crear paquetes con estados que hablen del mismo tema y así abordar toda posible noticia completamente.

Por otro lado, se encuentra la indexación semántica latente (ISL) o el análisis semántico latente (LSA, por sus siglas en inglés), técnicas lingüísticas plasmadas en programas informáticos que analizan las relaciones que pueden tener los textos comparando las letras, silabas, palabras u oraciones dependiendo el contexto en que se usen [28]; esta técnica también es empleada como una herramienta de recuperación de información [29], ya que compara palabras claves para encontrar textos con un gran porcentaje de similitud. La implementación de esta técnica es muy común debido a su variado campo de acción, ya que en la mayoría espacios es necesario manejar análisis y comparación de texto para recuperar datos valiosos. Por lo mencionado, la ISL se convierte en una de las técnicas de análisis de información más completas debido a que usa métodos de comparación multidimensionales, donde compara cada uno de los elementos de un bloque de información en una matriz para calcular el nivel de concurrencia [30]; de esta manera, la comparación y combinaciones entre los elementos es más precisa.

La indexación semántica latente ha contribuido en los últimos años a grandes avances en el análisis de textos, proponiendo varios modelos donde la caracterización de textos [31] ayuda a mejorar los análisis y los resultados conseguidos al manejar grandes lotes de información.

2.4. Procesamiento de lenguaje natural (PLN)

Con la presencia de alto contenido textual en la red, existen datos perdidos a los cuales no se les da ningún uso; para evitar esto y usar la información efectivamente es necesario la aplicación de métodos de aprendizaje a nivel de análisis morfológico, sintáctico y léxico. Una de las mejores estrategias es la aplicación de técnicas de procesamiento de lenguaje natural o PLN, ello debido a su fácil combinación con otros métodos de análisis de información como la tokenización, lematización, segmentación de palabras e indexación semántica latente, lo cual logra detectar qué información es útil para ciertos perfiles de usuario [32] e identificar palabras o conectores gramaticales que no son necesarios para centrarse en la información que el usuario necesita realmente [33].
PLN ha sido una de las técnicas que más ha evolucionado desde el inicio de los algoritmos de computador, los cuales se centraban en el objetivo de introducir lenguaje humano dentro de una máquina. De esta manera, son técnicas computacionales que interactúan con el lenguaje humano a partir de *software* especializado para ejecutar todo el proceso de análisis de una manera automática [34], garantizando un análisis de la información con resultados provechosos para los usuarios. Por lo mencionado, el manejo del lenguaje natural posee un gran campo de acción en cuanto a garantizar un análisis correcto de la información y, por esto, ha sido implementado sobre varias áreas de investigación como la recuperación y extracción de datos, resúmenes automáticos y sistemas de respuesta inteligentes [35].

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

El progreso de la tecnología ha generado necesidades en el manejo correcto de la información convirtiéndose en tarea de alta prioridad para las ciencias computacionales, ya que las técnicas tradicionales no cumplen con todos los objetivos deseados y generan pérdida de información y datos valiosos para las organizaciones [36]; al mismo tiempo, se debe considerar que toda la información que se encuentra en internet no es correcta en su totalidad, ya que existen datos borrosos que no contribuyen con ningún proceso organizacional.
Adicionalmente, se debe considerar el texto como uno de los estándares principales para la comunicación entre personas. PLN es una de las ramas de la inteligencia artificial que tiene gran ventaja para la creación de nuevos avances tecnológicos, y gracias a la acogida de las redes sociales las personas han vuelto a tomar ese gusto por escribir todo lo que sienten o tan solo comunicarse con los demás a partir de textos [11].

Uno de los mayores avances actuales del procesamiento de lenguaje natural es "WATSON", de los proyectos más ambiciosos de IBM, el cual implementa técnicas de PLN profundo para lograr una gran precisión en la lectura de textos o interpretación de audios [37], [38] y así evaluar de una mejor manera el contexto de frases, preguntas o textos, generando respuestas, análisis y resultados más provechosos y exactos. Estudios como este hacen ver que el procesamiento de lenguaje natural se convierte clave para lograr cumplir y entender mejor el funcionamiento y el procesamiento de cadenas textuales, el análisis de la lengua y el manejo correcto de la semántica, sintáctica y léxico, creando una necesidad de desarrollar aplicaciones que cubran todas las necesidades actuales que presenta el análisis de cadenas textuales.

En conclusión, el uso de técnicas de procesamiento de lenguaje natural se convierte en una de las mejores estrategias para poder explotar de una manera más óptima toda la información textual que es manejada día a día en fuentes heterogéneas.

2.5. Marcos de referencia I+D+i

Para la creación de proyectos computacionales es necesario la aplicación de normas y políticas en pos la gestión de calidad, lo que genera un nivel de confianza para los usuarios a los que van propuestos; un mecanismo viable y que ayude a cumplir los requisitos mencionados son los marcos de referencia de investigación, desarrollo e innovación (I+D+i–R&D), siendo pautas que ayudan a la ejecución de proyectos que garantizan el progreso de la sociedad. La aplicación de esquemas computacionales encargados de la explotación y análisis de información de internet en modelos de gestión de I+D+i se convierten en herramientas esenciales para la toma de decisiones, siendo cruciales para la definición de estrategias competitivas que ayuden a organizar las ideas y generar información provechosa [39].

Con lo mencionado, el promover al mejoramiento de los procesos que se imparten en las organizaciones como políticas de seguridad, gestión de calidad y medioambiental, deja a estas con la capacidad de implantar normas que ayuden dentro de sus organizaciones, como lo generalizan las normas de gestión de I+D+i [40], [41]; así, se dan resultados que funcionan para obtener soluciones e ideas que mejoren la calidad del negocio y que ayuden a orientar la organización.
Por otro lado, los marcos de referencia I+D+i ayudan a generar beneficios en las organizaciones creando puentes de conexión entre empresa, Estado y universidad o centros de investigación que apoyan a la externalización de conocimiento [42], ayudando a generar insumos para las nuevas organizaciones que desean emerger y crear innovación dentro de sus productos y procesos. La exploración y análisis de información externa como experiencias publicadas por otras organizaciones ayuda a romper barreras de conocimiento [43] e incentiva la creación de estrategias competitivas, lo que genera nuevos proyectos donde se aplican técnicas más apropiadas y óptimas.

Unos de los campos que más explotan los marcos de referencia I+D+i en los últimos años es la investigación farmacéutica, ello debido a que el desarrollo continuo de patentes [44] requiere un seguimiento para garantizar una correcta ejecución de la innovación y el desarrollo dentro de las organizaciones. El asegurar mecanismos de

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

transferencia tecnológica como lo son las patentes en todo un país o región ayuda a que se genere un nuevo progreso y desarrollo para todos los sectores productivos; China es uno de los países que más ha apostado a migrar sus normas tecnológicas para que cumplan con todos los parámetros I+D+i [45], demostrando que el cambio es lento, pero continuo. Con esto se ha ayudado a evolucionar varios sectores y la creación de nuevos proyectos, garantizando un crecimiento significativo en el registro de nuevas patentes.

El desarrollo de proyectos sobre estándares, marcos de referencia y modelos I+D+i aseguran que se generan fuentes provechosas para el desarrollo de nuevos productos, los cuales ayuden a la sociedad, organizaciones privadas, públicas o pymes para surgir de una manera más confiable y obtener logros a partir de la creación de productos o procesos nuevos e innovadores.

Con lo anterior, el cumplimiento de normas I+D+i ayudaría a que los modelos computacionales que se encargan de la extracción y análisis de información de redes sociales u otros medios digitales puedan surgir y generar nuevos mecanismos para que las organizaciones progresen.

3. Primeros resultados

Como se muestró en la Figura 1, la información recolectada fue extraída de fuentes primarias, lo que garantiza la veracidad de la información planteada durante toda la revisión bibliográfica. Con la revisión realizada se plantearon varios resultados que argumentan la falta de proyectos relacionados con los temas tratados. Por el lado de las redes sociales, fueron creadas desde de un comienzo para mejorar la manera de comunicación entre las personas, generando nuevos espacios de ocio que permitieran compartir información relevante sobre temas en común. La evolución de las redes sociales fue mejorando desde una perspectiva estratégica, creando una nueva red social que se adapta a las mejores utilidades de las anteriores para lograr un mayor alcance, el mejor ejemplo es Facebook que ha logrado la mayor cantidad de usuarios hasta 2019, como lo demuestra la Figura 2 donde están las redes ordenadas cronológicamente a partir de su creación. Los cambios se ven reflejados dependiendo de los avances de las redes sociales que se iban creando, queriendo obtener al final un espacio que cuente con todas las características que ya existieron o existen.

TheGlobe y Sixdegrees se consideran las primeras redes sociales debido a que brindaron a las personas espacios en los cuales se podían crear perfiles que describían los intereses de cada uno, a partir de los mismos se podían generar conexiones con personas que compartieran los mismos gustos. Luego, LiveJournal, LastFm y Fotolog se guiaron por crear plataformas donde se pudieran encontrar usuarios con los mismos gustos, como lo puede ser el estudio, la música y la fotografía. A partir de la creación de la web 2.0 surgieron redes sociales como: FriendSter, MSN spaces, Hi5, MySpace, Linkedin, los que han generado los espacios con más cantidad de usuarios en comunidades virtuales, pero el secreto fue crear una red social que contuviera todo lo necesario para que un usuario cubriera todas las necesidades posibles.

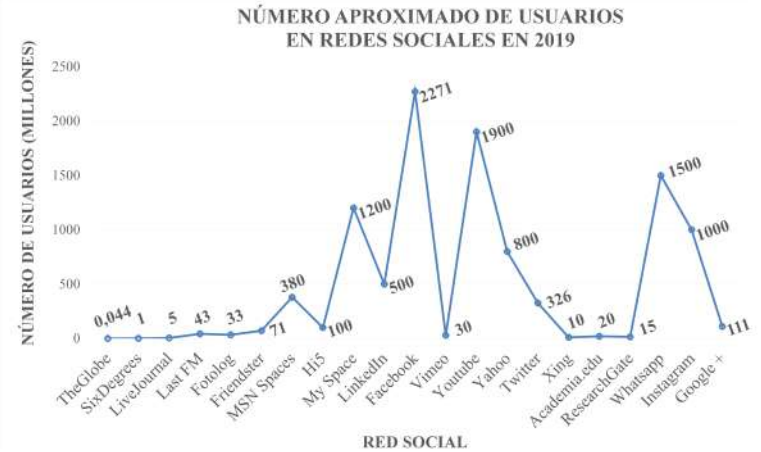

Figura 2. Creación de las redes sociales en el trascurso del tiempo **Fuente:** elaboración propia.

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

Al observar cómo se ha manejado la evolución de las redes sociales, se puede concluir que desde hace más de quince años el manejo de texto se ha convertido en pieza clave para la comunicación y seguirá siéndolo por muchos años más debido a la creación constante de plataformas que cubren la necesidad de estar conectados. Tener espacios en los cuales compartir se convierte en pieza clave para el desarrollo mundial, ello considerando que toda la información que se recolecta a medida del tiempo servirá como insumo para nuevas investigación y proyectos innovadores.
Por el lado de las investigaciones en procesamiento de lenguaje natural aplicando técnicas de lematización sintáctica y la indexación semántica latente, hay poco impacto sobre el análisis de información o cadenas textuales en idiomas como el español, ya que se ha centrado en los idiomas de los países que manejan la evolución tecnológica actual como el alemán [46], el árabe [27], el chino [22] y el inglés [47].

La carencia de estudios para la creación de herramientas computacionales que ayuden en el análisis del idioma español se convierte en un reto para centros de investigación, ello debido a que el español es declarado uno de los idiomas más usado en las redes sociales [48]. Lo anterior significa que existe una gran cantidad de información sin analizar por no contar con las herramientas óptimas que logren un análisis óptimo de texto; de esta manera, el desarrollo de un algoritmo basado en PLN y análisis léxico que permita el análisis de información en idioma español permitirá emerger en el estudio y análisis de una gran parte de la información que se alberga en internet.

4. Discusión y trabajo futuro

Como se mencionó en el transcurso del artículo, el uso de las redes sociales para la extracción de información útil es un tema de gran impacto por la cantidad de datos que se encuentra en ellas; desafortunadamente, las redes sociales que se dedican a compartir conocimiento investigativo como Academia.edu y ResearchGate se han dejado a un lado por no dedicarse a temas de ocio. Se recomienda el uso de estas por ser piezas clave para la innovación y el desarrollo de la sociedad, ya que la exploración de la información que se alberga en ellas asegura un porcentaje más alto de encontrar información que genere progreso y desarrollo dentro de las organización, ello gracias a los diferentes casos de estudio y experiencias que se han investigado en los distintos sectores productivos.

Desde la perspectiva de gestión de la información, se debe tener en cuenta que crear modelos o sistemas que se encarguen del manejo de información escrita en idiomas principales no es suficiente, pues se cuenta con una gran diversidad de idiomas en el mundo y eso significa que puede existir información relevante que no ha sido utilizada por no haber creado herramientas que cubran todos los modelos lingüísticos existentes; de tal manera, es necesario seguir explorando en la investigación de nuevos métodos que ayuden con el manejo de la información en los diferentes idiomas del mundo.

5. Conclusiones

La evolución tecnológica y el manejo de la información han convertido a las redes sociales en espacios propicios para la extracción y el análisis de información, ello con el fin lograr encontrar datos provechosos que ayuden a las personas y organizaciones a generar informes adecuados que funcionen como estrategia para la toma de decisiones e incentiven el progreso de nuevos proyectos.
El uso y aplicación de técnicas de PLN acompañadas de métodos de léxico pueden ser potenciales alternativas de solución a requerimientos de análisis e integración semántica de información contenida en redes sociales a partir de lecciones aprendidas. El manejo de estas técnicas ayudaría a organizar, entender y manejar la información recolectada para ser aplicada de forma correcta, realizando análisis que cumplan con las expectativas de los usuarios y ayude notablemente a grupos de investigación en el cumplimiento de sus objetivos.

La externalización de conocimiento y creación de ideas innovadoras se convierten en mecanismos que ayudan a que el proceso de desarrollo

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

tecnológico continúe su curso adecuadamente; de esta manera, la creación de proyectos guiados por marcos de referencia I+D+i resulta ser la forma adecuada de seguir generando conocimiento que ayude al proceso organizacional de las empresas.

Referencias

[1] O. Yagan, D. Qian, J. Zhang y D. Cochran, "Conjoining speeds up information diffusion in overlaying social-physical networks", IEEE Journal on Selected Areas in Communications, vol. 31, pp. 1038-1048, 2013. https://doi.org/10.1109/jsac.2013.130606

[2] D. Mladeni y M. Grobelnik, "Automatic text analysis by artificial intelligence", Informatica, vol. 37, pp. 27-33, 2013.

[3] W. Tan, M. B. Blake, I. Saleh y S. Dustdar, "Social-network-sourced big data analytics", IEEE Internet Computing, vol. 17, pp. 62-69, 2013. https://doi.org/10.1109/mic.2013.100

[4] M. J. Barón, "Semantic analysis over lessons learned contained in social networks for generating organizational memory in centers R&D", En The Sixth International Conference on Computer Science, Engineering and Information Technology, 2016. https://doi.org/10.5121/csit.2016.60621

[5] Á. Garcia-Crespo, R. Colomo-Palacios, J. M. Gómez-Berbis y B. Ruiz-Mezcua, "SEMO: a framework for customer social networks analysis based on semantics", Journal of Information Technology, vol. 25, pp. 178-188, 2010. https://doi.org/10.1057/jit.2010.1

[6] D. Zhang, B. Guo y Z. Yu, "The emergence of social and community intelligence", Computer, vol. 44, pp. 21-28, 2011.

[7] We Are Social Web, "We Are Social USA". https://wearesocial.com

[8] We Are Social, "Digital in 2017: global overview". https://wearesocial.com/special-reports/digital-in-2017-global-overview

[9] J. K. Sinclaire y C. E. Vogus, "Adoption of social networking sites: an exploratory adaptive structuration perspective for global organizations", Information Technology and Management, vol. 12, pp. 293-314, 2011. https://doi.org/10.1007/s10799-011-0086-5

[10] H. Delerue y J. L. Hopkins, "Can Facebook be an effective mechanism for generating growth and value in small businesses?", Journal of Systems and Information Technology, vol. 14, pp. 131-141, 2012. https://doi.org/10.1108/13287261211232153

[11] R. A. Frost, "Realization of natural language interfaces using lazy functional programming", ACM Computing Surveys (CSUR), vol. 38, n.° 4, 2006. https://doi.org/10.1145/1177352.1177353

[12] M. J. Culnan, P. J. McHugh y J. I. Zubillaga, "How large U.S. companies can use Twitter and other social media to gain business value", MIS Quarterly Executive, vol. 9, pp. 243-259, 2010.

[13] H. A. Sleiman y R. Corchuelo, "A survey on region extractors from web documents", IEEE Transactions on Knowledge and Data Engineering, vol. 25, pp. 1960-1981, 2013. https://doi.org/10.1109/tkde.2012.135

[14] C. Olston y M. Najork, "Web crawling", Information Retrieval, vol. 4, n.° 3, pp. 175-246, 2010.

[15] A. I. Saleh, A. E. Abulwafa y A. Rahmawy, "A web page distillation strategy for efficient focused crawling based on optimized naïve bayes (ONB) classifier", Applied Soft Computing, vol. 35, pp. 181-204, 2017. https://doi.org/10.1016/j.asoc.2016.12.028

[16] S. Khalil y M. Fakir, "RCrawler: An R package for parallel web crawling and scraping", SoftwareX, vol. 6, pp. 98-106, 2017. https://doi.org/10.1016/j.softx.2017.04.004

[17] K. Shchekotykhin, D. Jannach y G. Friedrich, "xCrawl: a high-recall crawling method for Web mining", En Eighth IEEE International Conference on Data Mining, 2008. https://doi.org/10.1109/icdm.2008.121

[18] A. Mesbah, A. Van Deursen y S. Lenselink, "Crawling Ajax-based web applications through dynamic analysis of user interface state changes", ACM Transactions on the Web (TWEB), vol. 6, n.° 3, 2012. https://doi.org/10.1145/2109205.2109208

[19] P. Sexton, "The googlebot guide". https://varvy.com/googlebot.html

[20] A. Devitt y K. Ahmad, "Is there a language of sentiment? An analysis of lexical resources for sentiment analysis", Language resources and evaluation, vol. 47, pp. 475-511, 2013. https://doi.org/10.1007/s10579-013-9223-6

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

[21] R. Ferreira, R. D. Lins, S. J. Simske, F. Freitas y M. Riss, "Assessing sentence similarity through lexical, syntactic and semantic analysis", Computer Speech and Language, vol. 39, 2016. https://doi.org/10.1016/j.csl.2016.01.003

[22] Z. Li, M. Zhang, W. Che, T. Liu y W. Chen, "Joint Optimization for Chinese POS Tagging and Dependency Parsing", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, n.° 1, pp. 274-286, 2014. https://doi.org/10.1109/taslp.2013.2288081

[23] D. Bollegala, Y. Matsuo y M. Ishizuka, "A web search engine-based approach to measure semantic similarity between words", IEEE Transactions on Knowledge and Data Engineering, vol. 23, n.° 7, pp. 977-990, 2011. https://doi.org/10.1109/tkde.2010.172

[24] G. Qiu, B. Liu, J. Bu y C. Chen, "Opinion word expansion and target extraction through double propagation", Computational Linguistics, vol. 37, pp. 9-27, 2011. https://doi.org/10.1162/coli_a_00034

[25] M. M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments", Expert Systems with Applications, vol. 40, pp. 4241-4251, 2013. https://doi.org/10.1016/j.eswa.2013.01.019

[26] R. G. Diaz, "La lematización en español: una aplicación para la recuperación de información". https://www.trea.es/books/la-lematizacion-en-espanol-una-aplicacion-para-la-recuperacion-de-informacion

[27] A. L. S. Mohammad, Z. Jaradat, A. L. A. Mahmoud y Y. Jararweh, "Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features", Information Processing and Management, vol. 53, pp. 640-652, 2017. https://doi.org/10.1016/j.ipm.2017.01.002

[28] J. Botana, "La técnica del Ánalisis de la Semántica Latente (LSA/LSI) como modelo informático de la comprensión del texto y el discurso: una aproximación distribuida al análisis semántico", Tesis doctoral, Universidad Autónoma de Madrid, Madrid, 2010.

[29] T. Jaber, A. Amira y P. Milligan, "Enhanced approach for latent semantic indexing using wavelet transform", IET Image Processing, vol. 6, pp. 1236-1245, 2012.

[30] C. Aswani Kumar, M. Radvansky y J. Annapurna, "Analysis of a vector space model, latent semantic indexing and formal concept analysis for information retrieval", Cybernetics and Information Technologies, vol. 12, pp. 34-48, 2012. https://doi.org/10.2478/cait-2012-0003

[31] H. Elghazel, A. Aussem, O. Gharroudi y W. Saadaoui, "Ensemble multi-label text categorization based on rotation forest and latent semantic indexing", Expert Systems with Applications, vol. 57, 2016. https://doi.org/10.1016/j.eswa.2016.03.041

[31] M. A. Tayal, M. M. Raghuwanshi y L. G. Malik, "ATSSC: Development of an approach based on soft computing for text summarization", Computer Speech and Language, vol. 41, pp. 214-235, 2017. https://doi.org/10.1016/j.csl.2016.07.002

[31] S. Sun, C. Luo y J. Chen, "A review of natural language processing techniques for opinion mining systems", Information Fusion, vol. 36, pp. 10-25, 2017. https://doi.org/10.1016/j.inffus.2016.10.004

[31] E. Cambria y B. White, "Jumping NLP curves: a review of natural language processing research [review article]" IEEE Computational Intelligence Magazine, vol. 9, n.° 2, pp. 48-57, 2014. https://doi.org/10.1109/mci.2014.2307227

[31] M. B. Hernández y J. M. Gómez, "Aplicaciones de Procesamiento de Lenguaje Natural", Revista Politécnica, vol. 32, 2013.

[31] J. Kacprzyk y S. Zadrozny, "Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural-language generation", IEEE Transactions on Fuzzy Systems, vol. 18, n.° 3, pp. 461-472, 2010. https://doi.org/10.1109/tfuzz.2010.2040480

[37] G. R. Gendron, "Natural Language Processing: A Model to Predict a Sequence of Words", MODSIM World, n.° 13, 2015.

[38] R. High, "La Era de los Sistemas Cognitivos: Una mirada al interior de IBM Watson y ¿Cómo funciona?". https://www.redbooks.ibm.com/redpapers/pdfs/redp4955_es.pdf

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.

[39] I. Nonaka y H. Takeuchi, Die organisation des wissens: wie japanische unternehmen eine brachliegende ressource nutzbar machen, Campus Verlag, 2012.
[40] ICONTEC, "NORMA 5801, Gestión de la investigación, desarrollo e innovación (I+D+i). Requisitos del sistema de gestión I+D+i". https://es.slideshare.net/racape/ntc-5801
[41] AEN/CTN 166, "Gestión de la I+D+i: Requisitos del sistema de gestión de la I+D+i," Madrid, 2008.
[42] OCDE y EUROSTAT, "Manual de Oslo". http://www.itq.edu.mx/convocatorias/manual deoslo.pdf
[43] R. Filieri y S. Alguezaui, "Knowledge sourcing and knowledge reuse in the virtual product prototyping: an exploratory study in a large automotive supplier of R&D", Expert Systems, vol. 32, pp. 637-651, 2015. https://doi.org/10.1111/exsy.12101
[44] S. M. Paul et al., "How to improve R&D productivity: the pharmaceutical industry grand challenge", Nature reviews Drug discovery, vol. 9, pp. 203-214, 2010.
[45] X. Cui, B. Chen y Y. Chang, "Transnational R&D centers and national innovation systems in host countries: empirical evidence from China" Canadian Public Policy, vol. 43, n.° S2, pp. 107-121, 2017. https://doi.org/10.3138/cpp.2016-075
[46] G. Müller, "On deriving CED effects from the PIC", Linguistic Inquiry, vol. 41, pp. 35-82, 2010. https://doi.org/10.1162/ling.2010.41.1.35
[47] R. Hervás et al., "Integration of lexical and syntactic simplification capabilities in a text editor", Procedia Computer Science, vol. 27, pp. 94-103, 2014. https://doi.org/10.1016/j.procs.2014.02.012
[48] Instituto Cervantes, "El español: una lengua viva". https://www.cervantes.es/ima

ISSN 1794-211X • e-ISSN 2322-939X • Vol 16, No 2 (julio-diciembre 2019). pp.298-309.Universidad Distrital Francisco José de Caldas-Facultad Tecnológica.