Aplicaciones de Procesamiento de Lenguaje Natural y Ciencia de Datos

Anabella De Battista, Soledad Retamar, Patricia Cristaldo, Esteban Schab, Carlos Casanova, Lautaro Ramos, Ramiro Rivera, Cristhian Richard, Lucas La Pietra, Lautaro Retamar, Franco Miret, Rodrigo Mignola, Juan Manuel Franzante
Grupo de Investigación en Bases de Datos, Departamento Ingeniería en Sistemas de Información, Fac. Reg. Concepción del Uruguay, Universidad Tecnológica Nacional, Entre Ríos, Argentina
{debattistaa, retamars, schabe, ramosl, riverar, richardc, cristaldop, lapietral, retamarl, miretf, mignolar, franzantejm}@frcu.utn.edu.ar

Norma Edith Herrera
Departamento de Informática, Universidad Nacional de San Luis, San Luis, Argentina
{nherrera}@unsl.edu.ar

Resumen

Cada día se generan grandes cantidades de datos de diversos tipos y vinculadas a diversas áreas de conocimiento (textos, imágenes, audios, videos, entre otros). La posibilidad de analizar estos enormes volúmenes de datos permite generar conocimiento que sirva como fundamento para la toma de decisiones en ámbitos tan diversos como la salud, la administración de distintos recursos, el deporte, las actividades turísticas, entre otros. En este contexto resultan de gran utilidad los modelos desarrollados en áreas de conocimiento como la Ciencia de Datos, el Aprendizaje Automático, la Minería de Textos por citar algunas. En el ámbito de análisis automático de textos (para aplicaciones de análisis de sentimientos, minería de opinión, entre otras) la mayoría de los algoritmos, herramientas y recursos disponibles han sido desarrollados para el idioma inglés por lo que no pueden aplicarse a textos escritos en idioma español.
En este artículo se presentan soluciones desarrolladas en el proyecto *Descubrimiento de Conocimiento en Bases de Datos*, basadas en aplicación de técnicas de Procesamiento de Lenguaje Natural y Ciencia de Datos. En particular, se presentan aplicaciones web para seguimiento de la evolución de casos de Dengue y COVID-19 en Argentina, otra aplicación que presenta una aplicación de Topic Modelling a noticias en lenguaje español, otro caso de aplicación de Análisis de Sentimientos a entidades vinculadas al turismo y finalmente un caso de generación de estadísticas avanzadas a partir del procesamiento de conjuntos de datos históricos de básquet.
Palabras clave: ciencia de datos, minería de textos, bases de datos, descubrimiento de conocimiento, idioma español.

Contexto

Las aplicaciones que se presentan en este trabajo se desarrollan en el ámbito del proyecto de investigación *Descubrimiento de conocimiento en Bases de Datos (Código 5109)* del Grupo de Investigación en Bases de Datos, perteneciente al Departamento Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay.

1. Introducción

En los últimos años han ocurrido importantes avances en lo que respecta al desarrollo de soluciones basadas en aplicaciones de Ciencia de Datos. Las organizaciones en general han adoptado este tipo de herramientas para mejorar sus procesos de toma de decisiones y disminuir la incertidumbre en base al descubrimiento de patrones o reglas que, de otro modo, permanecerían ocultas en sus bases de datos. El Descubrimiento de Conocimiento en Bases de Datos consiste en el proceso de identificar

patrones válidos, novedosos y potencialmente útiles y comprensibles a partir de los datos, e involucra áreas de conocimiento como inteligencia artificial, aprendizaje automático, estadística, sistemas de gestión de base de datos, técnicas de visualización de datos y medios que apoyan toma de decisiones.

La Ciencia de Datos es un área de estudio multidisciplinar que integra campos de conocimiento variados como estadística, aprendizaje automático y análisis de datos, y que se puede aplicar en muchos dominios de conocimiento [1].

La Minería de Textos tiene como objetivo la extracción de patrones relevantes a partir de un gran conjunto de textos con el propósito de obtener conocimiento. Abarca una serie de tareas que se enfocan en distintos aspectos del análisis de textos. Algunas de las más relevantes:

- Recuperación de información (Information Retrieval, IR): es la tarea de encontrar material de naturaleza no estructurada (generalmente textos) proveniente de grandes colecciones que satisfagan determinadas necesidades de información [2]. Una tarea crucial para un sistema de IR es indexar la colección de documentos para hacer que sus contenidos sean accesibles de manera eficiente. Generalmente la indexación se realiza sobre una representación lógica del documento, que puede consistir en un conjunto de palabras clave o términos relevantes que aparezcan en el texto [3].
- Procesamiento del Lenguaje Natural: es un campo de las ciencias de la computación que combina Inteligencia Artificial y conceptos lingüísticos con el fin de hacer que oraciones o palabras escritas en lenguaje natural puedan ser interpretados por programas de computadoras [1, 4].
- Extracción de Información (*Information Extraction*, IE): es una subdisciplina de la Inteligencia Artificial que se aboca a la identificación, y consecuente clasificación y estructuración en grupos semánticos, de información específica que se encuentran en fuentes de datos no estructurados, como el texto en lenguaje natural, lo que hace que la información sea más adecuada para las tareas de procesamiento de la información [5].
- Resumen de textos (*Text Summarization*): es la tarea de producir un resumen conciso y fluido preservando el contenido clave de la información y el significado general de una colección de textos [6].
- Métodos de Aprendizaje Supervisado y No Supervisado: los métodos de aprendizaje supervisado son técnicas de aprendizaje automático relacionadas con entrenar un modelo, por ejemplo, de clasificación, utilizando un conjunto de datos de entrenamiento para realizar predicciones sobre datos desconocidos de antemano. Existe una amplia gama de métodos supervisados, como clasificadores de vecinos más cercanos, árboles de decisión, clasificadores basados en reglas y clasificadores probabilísticos.

Los trabajos sobre minería de textos en español que se presentan en la actualidad se enfocan principalmente en Análisis de Sentimientos o Minería de Opinión, en los cuales se evidencian dos enfoques: uno basado en el empleo de lexemas, y otro en técnicas de *Machine Learning*, para identificar los sentimientos expresados en los textos. En la gran mayoría de estos trabajos se utilizan recursos traducidos de forma automática generados para otros idiomas, como el inglés, lo cual manifiesta una escasez de recursos genuinos para el lenguaje español.

Existen también trabajos sobre perfilado de autor, en los que se menciona la dificultad de encontrar colecciones de textos adecuadamente etiquetados y con poco ruido [7]. A partir de eso se han producido trabajos tendientes al desarrollo de conjuntos de textos en español específicos para esta tarea [8,9].

Las técnicas de minería de datos de texto se han propuesto para tareas de estudios bibliográficos, simplificación de textos y etiquetado semántico. Se ha propuesto la aplicación de algoritmos de clasificación [10] para identificar automáticamente el dominio disciplinar de un nuevo texto académico en un repositorio bibliográfico mediante la construcción de lexemas compartidos en cada disciplina.

2. Líneas de Investigación, Desarrollo e Innovación

La línea de trabajo principal de nuestro proyecto de investigación es el estudio, análisis y comparación de técnicas de minería de datos para el tratamiento de textos en español, el análisis de su desempeño en distintos escenarios y la propuesta de modificaciones o mejoras a las técnicas de minería de textos existentes para incrementar la calidad de los resultados en el tratamiento de textos en español. También está previsto realizar una evaluación del funcionamiento de técnicas de minería de datos sobre conjuntos "tradicionales" enriquecidos con atributos provenientes de textos relacionados. Durante el año 2020 se realizaron también trabajo en el ámbito de Analítica de Datos, aplicados al seguimiento de la pandemia de COVID-19 y a la epidemia de Dengue, lo que incrementó la visibilidad del trabajo realizado por el grupo de investigación y permitió concretar acuerdos para trabajar, por ejemplo, en analítica de datos aplicada al deporte.

2.1. Topic Modelling

El término Agenda Setting hace referencia a la influencia que tienen los medios de comunicación en la fijación de temas en la opinión pública [14]. Se desarrolló una aplicación web que despliega la aplicación de Topic Modeling a un corpus de texto en español, generado a partir de la recolección de noticias sobre COVID-19 del mes de mayo de 2020 de los periódicos digitales argentinos Clarín, Infobae y La Nación. El objetivo propuesto fue investigar los patrones subyacentes en los documentos y determinar las temáticas principales de las noticias publicadas por los medios durante esa etapa de la pandemia. Se realizó una comparativa de dos herramientas que implementan el algoritmo LDA y se realizaron las pruebas con la herramienta que obtuvo mejores resultados. Se desarrolló además una aplicación web en la que se presentan los hallazgos del proyecto a través de diferentes visualizaciones. Para la publicación de resultados se desarrolló una aplicación web accesible en la URL https://nlp-noticiascorona.herokuapp.com y que permite visualizar distintos aspectos resultantes del procesamiento del conjunto de noticias como la cantidad de noticias clasificadas recolectadas por cada diario, la cantidad de noticias clasificadas en cada tópico, los bi-gramas y n-gramas de palabras que aparecen con más frecuencia en las noticias.

2.2. Análisis de Sentimientos

Mediante un convenio con la Dirección de Producción del Municipio de Concepción del Uruguay, se trabaja en un estudio consistente en la aplicación de técnicas de minería de textos sobre las opiniones que turistas que visitan la ciudad vierten en la plataforma digital de turismo TripAdvisor (www.tripadvisor.com.ar). Para realizar el análisis se implementó una tarea de categorización de textos conocida como Análisis de Sentimientos que consiste en la determinación de la orientación positiva o negativa que un escritor expresa hacia algún objeto [15]. Estas técnicas pueden aplicarse a revisiones de películas, libros, productos o servicios en la web, textos editoriales o políticos, entre otros. El resultado de su aplicación permite identificar el sentimiento hacia el objeto o sujeto bajo análisis, lo que las convierte en aplicaciones de gran relevancia en ámbitos como experiencia de usuarios, marketing o política. La aplicación de Análisis de Sentimientos a comentarios de turistas que han visitado hoteles, restaurantes y lugares turísticos de la ciudad de Concepción del Uruguay, permite conocer la percepción sobre los lugares visitados y obtener información detallada que sirva como soporte para la toma de decisiones en aspectos relacionados con el turismo como el diseño de estrategias de marketing orientadas según el tipo de público que visita la ciudad, acciones de fortalecimiento orientadas a determinados rubros, entre otras.

2.3. Visualización de datos

La generación y el almacenamiento de grandes volúmenes de información hacen que el mismo pase desapercibido y muchas veces se pierda la oportunidad de encontrar valor en ella. La visualización de datos es el proceso de representación de datos, en formato gráfico, de una manera clara y eficaz. Se convierte en una herramienta poderosa para el análisis e interpretación de datos grandes y complejos, volviéndose un medio eficiente en la transmisión de conceptos en un formato universal [16, 17].

Se desarrollaron dos aplicaciones que, mediante la aplicación de técnicas de analítica de datos, permiten visualizar la evolución de la pandemia de COVID-19 en Sudamérica, y con mayor granularidad en Argentina y sus provincias, y otra aplicación que permite comparar la evolución de la epidemia de Dengue en Argentina en los últimos años.

La aplicación para seguimiento de COVID-19 se desarrolló con el objetivo de aportar información sobre la evolución de la pandemia. Se seleccionaron fuentes de datos oficiales y mediante la aplicación de técnicas de analítica de datos y visualización de información, se presentan distintos análisis que permiten comprender la evolución de casos. La aplicación web está accesible en la URL https://gibd.github.io/covid [18].

En el caso de la aplicación de seguimiento de Dengue se procedió inicialmente a la identificación de fuentes oficiales de información, accediendo a sitios web de los Ministerios de Salud de las distintas provincias argentinas. El principal inconveniente encontrado fue que no todas las provincias publicaban información sobre esta epidemia, por lo que no era posible generar un conjunto de datos homogéneo. Se realizó un análisis de los datos publicados por el Ministerio de Salud, específicamente en el Boletín Integrado de Vigilancia (BIV), y se decidió tomar esa información para generar un conjunto de datos apropiado para aplicar análisis estadístico. Para realizar el análisis de la evolución de casos en Argentina en los últimos años se trabajó con la información publicada en el sitio web de la Organización Panamericana de la Salud. Posteriormente se trabajó en la generación de datasets a partir de la información recolectada de las fuentes antes mencionadas se generaron dos conjuntos de datos. Finalmente se realizó el análisis exploratorio de datos: en esta etapa se trabajó en la generación de visualizaciones empleando la librería Plotly de Python. Se adoptaron distintos enfoques para el análisis desde el punto de vista geográfico, generando gráficas a nivel país, a nivel provincia y a nivel región. Para el análisis a nivel país se realizó una comparativa de cantidad de casos y fallecimientos para el período 2014-2020. Para el análisis de la evolución de casos durante el año 2020 se realizaron visualizaciones de: Casos autóctonos, Casos importados, Casos en Investigación, Notificados y Fallecimientos, a nivel Provincia, Región y País. Además, se generó un mapa con la Cantidad Acumulada de casos por provincia hasta la última Semana Epidemiológica para la cual se publicaron datos. La aplicación web está accesible en la URL http://denguegibd.herokuapp.com [19].

2.4. Aplicación de metodología para la gestión de proyectos de Minería de Datos

En la gestión de las actividades de cada una de las líneas de investigación y desarrollo del proyecto se emplean fundamentos de metodologías ágiles. Partiendo de la propuesta metodológica de CRISP-DM [25] se realizó una adaptación empleando dichos fundamentos ágiles. Se formalizó dicha adaptación como una propuesta de metodología ágil para la gestión de proyectos de ciencia de datos [26, 27]. Actualmente se trabaja en definir y desarrollar métricas e indicadores para la evaluación transversal de metodologías de gestión de proyectos, a partir de expresiones textuales de la descripción de un proyecto. Se busca especificar un marco de medición que permita, a través de las métricas consignadas, evaluar las metodologías, obteniendo un resultado métrico que exprese el grado de aplicabilidad, eficacia y eficiencia en

diferentes contextos. Por otro lado, se busca medir la eficiencia de las metodologías en gestión de proyectos a partir de la descripción del proyecto en base a los hallazgos estadísticos, y aprendizaje automático con las métricas e indicadores propuestos.

3. Resultados obtenidos y esperados

Con este proyecto se espera lograr aplicaciones novedosas de técnicas y herramientas de minería de textos para textos en español, en particular en áreas de estudio como bibliometría y la teoría de establecimiento de agenda. Además, se continúa trabajando en analítica de datos aplicada al básquet. Estas iniciativas se desarrollan mediante la aplicación de la metodología ágil para proyectos de ciencias de datos propuesta.

4. Formación de Recursos Humanos

En el marco del proyecto se está desarrollando una tesis de maestría y unade las investigadoras está cursando el Doctorado en Informática. Se cuenta con un becario graduado con beca de iniciación a la investigación y dos becarios alumnos de la carrera Ingeniería en Sistemas de Información que inician su formación en la investigación. Está prevista la realización de al menos una Práctica Supervisada de Ingeniería en Sistemas de Información en el marco del proyecto.

5. Referencias

[1] Wang, R., Hu, G., Jiang, C., Lu, H., & Zhang, Y. (2020). Data Analytics for the COVID-19 Epidemic. 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), 1261–1266. https://doi.org/10.1109/COMPSAC48688.2020.00-83

[2] Stefan Büttcher, Charles L.A. Clarke, Gordon V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines. Published July 23rd 2010 by MIT Press (MA)

[3] Allahyari, Mehdi et al. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. 2017. arXiv:1707.02919v2

[4] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to information retrieval. Cambridge University Press, 2008.

[5] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, and S. Quarteroni, The Information Retrieval Process. In Web Information Retrieval, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–26.

[6] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges," Aug. 2017.

[7] Moens, Marie-Francine. Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer Netherlands, 2006.

[8] M. Allahyari et al. Text Summarization Techniques: A Brief Survey. Jul. 2017.

[9] M. J. Garciarena Ucelay, M. P. Villegas, L. Cagnina, and M. L. Errecalde. Cross domain author profiling task in spanish language: an experimental study. *JCS&T*, vol. 15, no. 02, pp. p. 122-128. Nov. 2015.

[10] M. P. Villegas, M. J. Garciarena Ucelay, M. L. Errecalde, and L. Cagnina. A Spanish text corpus for the author profiling task. 2014.

[11] M. J. Garciarena Ucelay and M. P. Villegas. Determinación del Perfil del Autor de Documentos en español. Universidad Nacional de San Luis, 2015.

[12] R. Venegas. Clasificación de textos académicos en función de su contenido léxico-semántico. Rev. signos, vol. 40, no. 63, pp. 239–271, 2007.

[13] SCOPUS. http://www.scopus.com Accedido 02/2021.

[14] Scopus API. https://goo.gl/mqpFpA Accedido 03/2021.

[15] Jurafsky, Daniel and Martin, James H. Speech and Language Processing. 2019. https://web.stanford.edu/~jurafsky

[16] M. McCombs and D. Shaw. The agenda-setting function of mass media. Public opinion

quarterly, 36(2):176–187, 1972.

[17] Yeooul Kim, Suin Kim, Alejandro Jaimes, and Alice Oh. A computational analysis of agenda setting. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion), 323-324, 2014.

[18] Grupo de Investigación en Base de Datos. Aplicación para seguimiento de COVID-19. https://gibd.github.io/covid. Accedido en 03/2021

[19] Grupo de Investigación en Base de Datos. Aplicación para seguimiento de Dengue. http://denguegibd.herokuapp.com Accedido en 03/2021

[25] Ken Schwaber and Jeff Sutherland. The scrum guide. Scrum Alliance, 2011, vol. 21.

[26] Cristaldo, Richard, Rivera, Schab, Anabella De Battista. Propuesta Metodológica de Enfoque "Híbrido" para la Gestión de Proyectos de Minería de Datos. SABTIC 2018. ISSN 2237-2970

[27] Cristaldo, Schab, Richard, Rivera, Anabella De Battista, Retamar, Herrera. Adecuación de una Propuesta Metodológica de Enfoque "Híbrido" para la Gestión de Proyectos de Ciencia de Datos. 6to CoNaIISI. 29 y 30 de noviembre de 2018 – Universidad CAECE – Mar del Plata, Bs. As., Argentina.