Hacia el análisis de tesis de grado de carreras informáticas de la UM mediante minería de textos

Gabriel Mariuz[1] Iris Sattolo[1], Marisa Panizzi[1]

[1]Escuela Superior de Ingeniería, Informática y Ciencias Agroalimentarias Universidad de Morón. Cabildo 134 (B1708JPD), Partido de Morón, Argentina.
gmariuz91@gmail.com, iris.sattolo@gmail.com, marisapanizzi@outlook.com

Resumen. La categorización de documentos de textos es una aplicación de la minería de textos que pretende extraer información de texto no estructurado o semi estructurado. La justificación de su aplicación se debe a que se estima que alrededor del 80% de los datos de las organizaciones son no estructurados. El presente trabajo de tesis de la carrera Licenciatura de Sistemas de la UM pretende analizar los títulos de las tesis realizadas en la cátedra para categorizarlas según su área temática mediante minería de textos y evaluar la eficacia de la técnica utilizada al hacerlo. Antes de comenzar con la construcción de modelos de minería de textos, se construyó el estado del arte mediante un mapeo sistemático de la literatura (en inglés, *systematic mapping study* o SMS). Se presentan los resultados logrados mediante el desarrollo del SMS y se describen las actividades definidas para la finalización del trabajo de tesis.

Palabras claves: Minería de textos, categorización, aprendizaje automático, tesis de grado, carreras de informática.

1 Introducción

La cantidad de documentos de diversos tipos disponibles en una organización o establecimiento es enorme y continúa creciendo cada día. Estos documentos son a menudo un repositorio fundamental del conocimiento de la organización, pero a diferencia de éstas la información no está estructurada. La minería de textos tiene como objetivo extraer información de texto no estructurado, tal como entidades (personas, organizaciones, fechas, cantidades) y las relaciones entre ellas. La categorización de documentos de texto es una aplicación de la minería de texto que asigna a los documentos una o más categorías, etiquetas o clases, basadas en el contenido.

El enfoque tradicional para la categorización de textos en que los expertos en el dominio de los textos definían manualmente las reglas de clasificación ha sido reemplazado por otro basado en técnicas de aprendizaje automático, o en combinaciones de éste con otras técnicas [1].

Actualmente, las cátedras de tesis del área de Informática en la Universidad de Morón cuentan con un archivo en formato xls que contiene los datos referidos a las tesis realizadas en las carreras informáticas desde el año 2004 hasta la actualidad. Este archivo cuenta, entre otros datos, el título de la tesis, su resumen, el área temática a la que

corresponde cada tesis basada en las áreas propuestas en los workshops del Congreso CACIC[1] organizado por la RedUNCI[2]. Con el objetivo de categorizar cada una de las tesis según su área temática se planteó utilizar técnicas de minería de textos para generar patrones, además esto permitirá validar si la categorización automatizada es similar a la realizada manualmente y determinar su eficacia.

Con el fin de conocer el estado del arte con respecto al uso de la minería de textos, en el dominio educativo, es que se ha realizado un mapeo sistemático de la literatura de acuerdo con los directrices propuestas por Kitchenham *et al.* [2].

El artículo se estructura de la siguiente manera: en la Sección 2 se describe la planificación del SMS, en la Sección 3 se describe su ejecución. Los resultados se presentan en la Sección 4. En la Sección 5 se exponen las conclusiones y trabajos futuros.

2 Planificación del SMS

Se presenta la definición del protocolo de revisión del SMS: preguntas de investigación (PI), estrategia de búsqueda, selección de los estudios, criterios y proceso de selección, formulario de extracción y el proceso de síntesis de los datos.
El objetivo de este SMS es responder la siguiente pregunta de investigación (PI):
¿Qué trabajos existen de la aplicación de minería de textos para la categorización de tesis?

Esta pregunta principal se descompone en un conjunto de sub-preguntas (PI1-4), las cuales se presentan en la Tabla 1 junto con su motivación.

Tabla 1. Preguntas de investigación (PI) y su motivación.

Pregunta de investigación (PI)	Motivación
PI1: ¿Qué técnicas y algoritmosson los más utilizados en minería de textos?	Identificar las técnicas y algoritmos más usados en minería de textos para categorizar documentos.
P2: ¿Con qué herramientas y lenguajes de programación se trabaja en la minería de textos?	Identificar las herramientas y lenguajes de programación más utilizados para la categorización dentro de la minería de textos.
PI3: ¿Qué metodologías y procesos son utilizadas en la minera de textos?	Identificar las metodologías y procesos más utilizados en minería de textos.
PI4: ¿Qué tipos de investigación se encuentra en los artículos?	Identificar los tipos de investigación realizada en los estudios primarios de acuerdo con la clasificación propuesta por Wieringa *et al.* [3].

[1] CACIC: Congreso Argentino en Ciencias de la Computación.
[2] RedUNCI: Red de Universidades con carreras informáticas. Disponible: https://redunci.info.unlp.edu.ar/

Se definieron para la búsqueda de artículos las siguientes librerías, plataformas y repositorios digitales: SEDICI[3], *Scielo, Dialnet*, *Google Schoolar* y *ScienceDirect*, considerando artículos de congresos y artículos de revistas. El período comprendido definido ha sido entre julio de 2005 hasta marzo de 2022.
La cadena de búsqueda resultante es:

(Minería de datos) OR (Data Mining) OR (Minería de texto) OR (Text Mining) (Minería de texto en la educación) OR (Text Classification) OR (Educational Text Mining) OR (Clasificación de texto)

En la Tabla 2, se presentan los criterios de inclusión y exclusión utilizados para el proceso de selección de artículos.

Tabla 2. Criterios de inclusión y exclusión.

Criterios de inclusión	Criterios de exclusión
Artículos que respondan a las preguntas de investigación.	Artículos que no estén accesibles para su lectura completa.
Artículos publicados a partir de julio de 2005 hasta marzo de 2022.	Literatura gris, tesis doctorales, presentaciones en PowerPoint.
Artículos preferentemente en español ya que se busca conocer el estado y las investigaciones realizadas en aquellos países de habla hispana.	

El proceso de selección de los estudios consistió en realizar la búsqueda en las fuentes definidas aplicando la cadena en el título y/o en el resumen, para luego eliminar los artículos duplicados y aplicar los criterios de inclusión y exclusión en el título, resumen y palabras clave, después se aplicaron los criterios de inclusión y exclusión al texto completo.

Para dar respuesta a cada una de las preguntas de investigación (PI) se definió un esquema de clasificación que junto con el formulario de extracción de datos se presenta en un apéndice por restricciones de espacio [4].

3 Ejecución y resultados del SMS

Se encontraron 27 artículos de los cuales se analizaron 15 estudios primarios que se encuentran en el apéndice [4]. Los resultados del SMS para dar respuesta a las preguntas de investigación en base a la literatura analizada mediante gráficos.

PI1: ¿Qué técnicas y algoritmos son los más utilizados en minería de textos?

Los algoritmos más utilizados en minería de textos para la categorización de documentos son las redes neuronales artificiales, siendo además de las más precisas, comparadas a las redes bayesianas y a las máquinas de vectores de soporte (SVM). (Ver Figura 1).

[3] SEDICI: Repositorio Institucional de la Universidad Nacional de La Plata. Disponible en: http://sedici.unlp.edu.ar/

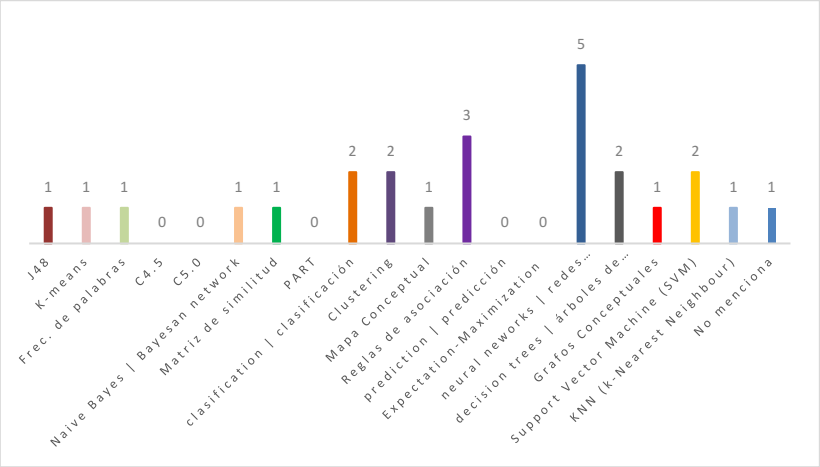

Figura. 1. Técnicas y algoritmos utilizados en la minería de textos.

P2: ¿Con qué herramientas y lenguajes de programación se trabaja en la minería de textos?

Se ha observado que existe variedad en el uso de herramientas y se halló que el uso de una u otra dependerá de cada usuario según ciertos valores como la usabilidad, potencia y versatilidad que pueda ofrecer dicha herramienta. Se evidenció que weka es la herramienta más usada mientras que Phyton y R los lenguajes más utilizados. Sin embargo, existe otra opción poco explorada debido a su reciente aparición como es GPT-3. Es un tipo de red neuronal que emplea aprendizaje automático y está enfocada en producir texto que simula la redacción humana y que, dada la cantidad de información disponible para su entrenamiento, tiene el potencial de ser usada para otras tareas para las que no fue pensada originalmente. Los resultados hallados se presentan en la Figura 2.

PI3: ¿Qué metodologías y procesos son utilizadas en la minera de textos?

En general, se puede evidenciar que la metodología más utilizada en la minería de textos es la metodología KDT (*Knowledge Discovery in Text*), una variante de KDD enfocada en el proceso de descubrimiento de conocimiento en texto. (Ver Figura 3).

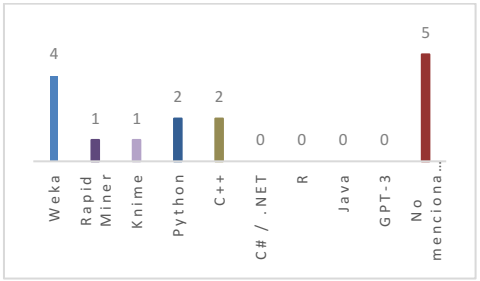

Figura. 2. Herramientas y lenguajes de programación utilizados en la minería de textos.

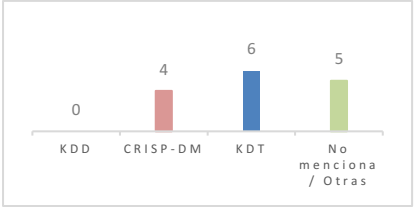

Figura. 3. Metodologías y procesos utilizados en la minería de textos.

PI4: ¿Qué tipos de investigación se encuentra en los artículos?
La mayor cantidad de los estudios analizados presentan la evaluación de los resultados de aplicar minería de texto (6 estudios) e informar las experiencias personales obtenidas al utilizarlo sobre un conjunto de datos (3 estudios). (Ver Figura 4).

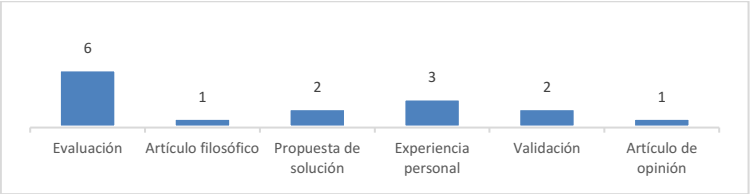

Figura. 4. Tipos de investigación de los artículos [4].

4 Conclusiones y trabajos futuros

Se logró construir el estado del arte respecto a la aplicación de la minería de textos para la categorización de las tesis de grado mediante el desarrollo de un SMS. Se analizaron 15 estudios primarios y se concluye que:

- En la mayoría de los estudios analizados se presentan principalmente propuestas de evaluación y en menor medida de informar una experiencia como tipo de investigación.
- Los algoritmos más utilizados son las redes neuronales.
- Las herramientas o lenguajes de programación más usados son Weka y Rapid Miner, mientras que, en menor medida, para los lenguajes de programación, son R y Phyton.
- La metodología más utilizada es KDT.

Como futuro trabajo para continuar el desarrollo de la tesis, se realizará: 1) Experimentación con la herramienta GPT-3 para conocer el alcance de sus capacidades, 2) Uso de la metodología KDT y, por último, 3) Evaluación de la categorización automática obtenida con la minería de textos para contrastarla con la categorización manual.

Referencias

[1] Abelleira, M., Cardoso, A. Categorización automática de documentos. XII Argentine Symposium on Artificial Intelligence (ASAI), 20-31. (2011).
[2] B. Kitchenham, D. Budgen y P. Brereton, Evidence-Based Software Engineering and Systematic Reviews, USA: CRC Press (2015).
[3] Wieringa, R., Maiden, N., Mead, N., Rolland, C. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. Requirements Engineering, 11(1), pp. 102-107 (2006).
[4] Mariuz G., Sattolo I., Panizzi M. Apéndice. Disponible en: https://doi.org/10.6084/m9.figshare.20514666.v1 (2022).