APLICACIÓN DE UN MODELO DE ANALÍTICA DE TEXTO PARA LA DETECCIÓN Y CLASIFICACIÓN AUTOMÁTICA DE TENDENCIAS DE INVESTIGACIÓN EN E-LEARNING (CASO UOC)

José López Ruiz y Guillem Garcia Brustenga
eLearn Center, Universitat Oberta de Catalunya

Justificación del informe

Como centro de investigación especializado en el aprendizaje en línea, el eLC de la UOC analiza permanentemente el espacio global del conocimiento para detectar y reflexionar sobre todas aquellas tendencias y acontecimientos que están transformando la educación superior en línea y la formación a lo largo de la vida en el mundo. Producto de esta tarea prospectiva, y como hiciera el año pasado con la difusión del "E-learning Research Report 2017" (eLRR'17) [1] publica periódicamente informes de tendencias entorno al ámbito del e-learning que requieren, por parte de sus expertos, reiteradas tareas de: extracción, análisis, síntesis, codificación y clasificación de cientos, y en algunos casos miles, de referencias bibliográficas. Dada la naturaleza sistemática de este proceso de análisis y el gran volumen de artículos y otras fuentes documentales que suelen revisarse, se pensó que el potencial de la tecnología de analítica de datos y las herramientas de analítica de texto, podrían cubrir eficazmente parte de este laborioso proceso.

Bajo esta hipótesis y apoyándose precisamente en el citado informe de investigación desarrollado por el centro, el eLC lleva a cabo este innovador proyecto de extracción y clasificación automática de artículos científicos. El modelo desarrollado permite extraer de manera ágil y sencilla la información relevante de cientos de artículos científicos de elearning y clasificarla automáticamente según las categorías temáticas elegidas (*topic* principal, categoría temática, metodología de investigación y etapa educativa).

El desarrollo del proyecto ha contado con la participación de la empresa MeaningCloud experta en analítica de textos en la nube, para el diseño y aplicación de un modelo automático de clasificación basado en reglas semánticas de categorización profunda.

El trabajo se divide en dos apartados:

1.- Desarrollo de un piloto para validar la aplicabilidad de herramientas de analítica de texto en la clasificación automática de artículos científicos.

2.- Realización de una prueba de clasificación automática con publicaciones de e-learning del 2018.

Introducción al modelo

El análisis de la literatura científica para detectar qué tendencias están dictando la evolución de una disciplina o de un ámbito concreto de investigación, es habitual entre investigadores que quieran conocer las iniciativas que están actuando como palancas de cambio del sector. Este conocimiento es a su vez indispensable en la toma de decisiones, para diseñar, reformular o abandonar proyectos que respondan a los intereses de cada institución. La revisión de artículos científicos y otras fuentes documentales para extraer información significativa, incluso cuando esta se halle en un formato estructurado, es larga y laboriosa, y puede representar un inmenso esfuerzo al equipo de analistas en función del *zoom* propuesto de cada búsqueda. Así pues, trabajos derivados de tal gestión del conocimiento, como la extracción, análisis y clasificación de la información, son sensibles al procesamiento automático.

En este sentido, la tecnología de *data analytics* y minería de texto (*text mining*) pueden jugar un papel fundamental para cubrir eficazmente parte de estas fases del proceso de investigación. Las técnicas basadas en minería de texto pueden definirse como la aplicación de métodos y técnicas computacionales sobre datos textuales para encontrar información relevante e intrínseca y conocimiento previamente desconocido (do Prado & Ferneda, 2007) [2].

Bajo esta perspectiva, se puso en marcha el piloto objeto de este proyecto. La prueba consistió en un ensayo de extracción y categorización automática de información relevante sobre la investigación mundial en e-learning recogida en cientos de artículos científicos de impacto. Se escogió el “E-Learning Research Report 2017. Analysis of the main topics in research indexed articles”, producido por investigadores del centro, para valorar la efectividad de la solución analítica, dado que para su elaboración se habían analizado esas mismas publicaciones. El citado trabajo permite conocer qué tendencias centraron el interés de los investigadores del aprendizaje en línea y donde se depositaron los esfuerzos y recursos de investigación durante el periodo analizado. Para preparar el análisis de texto de cara a la identificación de conceptos clave y su

posterior clasificación automática, se utilizó la codificación y categorías de agrupamiento previamente definidas en el estudio de investigación, con el objetivo de que los resultados reflejados por la máquina pudieran compararse y evaluarse contra los obtenidos por los autores del informe.

Preparación y ejecución de la prueba piloto

Metodología

El diseño del modelo se inició el pasado año siguiendo la siguiente metodología:

- La prueba partió de los 855 artículos sobre e-learning extraídos de las principales bases de datos de publicaciones científicas: WoS y Scopus y analizados en el estudio eLRR '17. Como se explica en el mismo informe, esta muestra se obtiene después de un proceso de cribado de artículos repetidos en ambas BBDD o que no cumplían los requisitos de la búsqueda bibliográfica.

- Del total de este conjunto de artículos, se seleccionaron aproximadamente la mitad para trabajar la prueba piloto, se trataban de las publicaciones que contenían las categorías y topics más frecuentes dado que ello favorecía el entrenamiento y posterior evaluación del modelo. Esta muestra de entrenamiento se compuso con alrededor del 63% de los artículos escogidos.

- Utilizando la muestra de entrenamiento se desarrollaron tres modelos de clasificación profunda basados en reglas, que cubrían las taxonomías planteadas para cada uno de los niveles con los que se pretendía analizar el texto:

    - Nivel educativo
    - Metodología empleada
    - Clasificación temática

- Estos modelos, siguiendo la metodología habitual de entrenamiento, validación y test, se enfrentaron a la muestra test conformada por el resto de artículos (37%), lo que permitió calcular la precisión y la cobertura real de los tres modelos a evaluar. Estos mismos datos se enfrentaron también

al total de los casos seleccionados para el piloto, lo que permitió comprobar la similitud de la clasificación realizada a través de métodos automáticos y la clasificación manual realizada por el grupo de investigadores que desarrollaron el informe del 2017.

- Para dar respuesta a una de las cuestiones que se plantearon durante el desarrollo de los modelos de clasificación, referente a si era necesario utilizar el grueso del artículo para realizar la clasificación o era suficiente con analizar el título y el *abstract* de cada uno de ellos, se produjo una validación dual, título y abstract del artículo por un lado frente al título, abstract y texto completo de la obra por otro. Se llegó a la conclusión de que los resultados arrojados analizando únicamente título y abstract eran más precisos que los del texto completo (inferiores al 70% en relevancia), por lo que se optó por utilizar únicamente esta primera parte de las publicaciones para la clasificación automática de las categorías.

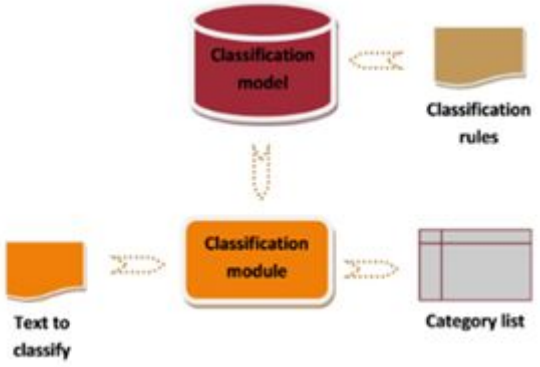
Reglas de clasificación

En la extracción y ordenación de la información relevante de los artículos de manera automática, se utilizó la minería de texto, en concreto, el motor de categorización profunda “Fig.1” de MeaningCloud que incorpora una tecnología de reglas avanzadas que abarca los niveles léxico, sintáctico y semántico del lenguaje natural. De esta manera, las reglas de clasificación se basaron en:

- Información morfológica que permite diferenciar forma de lema y marcar la categoría gramatical de un término.
- Información sintáctica
- Información semántica a través de una ontología propia y con opción de personalización.

Además de esto, también incorporan expresiones avanzadas, como expresiones regulares, de proximidad y operadores lógicos, lo que permite basar las reglas en el significado mismo de las expresiones y no en formas literales.

Figura 1. Diagrama proceso de clasificación textual seguido



El corpus de entreno del piloto se construyó en base a los siguientes parámetros:

Nivel educativo: 7 categorías - 60 reglas de clasificación.
Metodología empleada: 6 categorías - 135 reglas de clasificación.
Clasificación temática: 33 categorías - 230 reglas de clasificación.

Resultados del piloto

La evaluación se siguió confrontando en diferentes iteraciones, la muestra de entrenamiento previamente etiquetada por el equipo de autores del informe con los modelos de clasificación definidos. Fruto de estas iteraciones se fueron desarrollando los modelos de clasificación. De manera paralela, se confrontaron estos mismos modelos con la muestra de test, con la intención de calcular la eficiencia real de los modelos de clasificación, y con el total de artículos de la muestra para extraer conclusiones comparativas entre la clasificación automática y la etiquetada manualmente. La comparación se efectuó para cada una de las categorías (temática, metodología y etapa educativa) a niveles de título+abstract (ver tablas de clasificación 1-3) y del total del cuerpo del artículo. En cada una de ellas se midió la Precisión y Recuperación *(o exhaustividad)* de los resultados entendidos como cuán "útiles" son los valores recuperados en la búsqueda - *Precision*- y cuán completos son estos -*Recall*- [3].

A continuación se muestran las tablas comparadas de frecuencias efectuadas sobre los modelos de clasificación de: etapa educativa, metodología de investigación y temática de los artículos (ver eLRR'17). El análisis ha sido efectuado a nivel de título+abstract del total de artículos del piloto (entrenamiento+test).

Comparativa basada en la aplicación del modelo automático de clasificación vs. sin modelo (etiquetado UOC).

Tabla 1. Resultados de clasificación según categoría de nivel educativo: comparación de frecuencias

Categorías	Frec. sin modelo	Frec. con modelo	Diferencia
K12Secondary	*8.82%*	*9.31%*	*0.49%*
Undergraduate	*58.58%*	*56.13%*	*–2.45%*
Vocational	*0.98%*	*0.98%*	*0.00%*
Postgraduate	*4.90%*	*3.43%*	*–1.47%*
Professional	*12.50%*	*13.24%*	*0.74%*
Other	*14.22%*	*16.91%*	*2.70%*

Tabla 2. Resultados de clasificación según categoría de metodología: comparación de frecuencias

Categorías	Frec. sin modelo	Frec. con modelo	Diferencia
Design	*10.25%*	*11.00%*	*0.75%*
Evaluation	*16.00%*	*15.50%*	*–0.50%*
Descriptive	*41.50%*	*39.00%*	*–2.50%*
Review	*9.75%*	*9.75%*	*0.00%*
CaseStudy	*8.25%*	*9.50%*	*1.25%*
Experimental	*14.25%*	*12.00%*	*–2.25%*
Other	*0.00%*	*3.25%*	*3.25%*

Tabla 3. Resultados de clasificación según categoría temática: comparación de frecuencias

Categories	Frec. sin modelo	Frec. con modelo	Diferencia
Adoption	*11.00%*	*12.28%*	*1.28%*
Assessment	*7.42%*	*8.44%*	*1.02%*
Design Evaluation	*27.11%*	*28.90%*	*1.79%*
Disciplinary Research	*8.70%*	*10.23%*	*1.53%*
Innovation	*18.93%*	*17.14%*	*–1.79%*
Instructional Design	*8.44%*	*6.91%*	*–1.53%*
Interaction	*2.56%*	*3.32%*	*0.77%*
Learning Analytics	*7.67%*	*6.65%*	*–1.02%*
Student Performance	*3.07%*	*2.81%*	*–0.26%*
StudentPsychoVariables	*2.81%*	*1.28%*	*–1.53%*
TeacherRole	*2.30%*	*2.05%*	*–0.26%*

En relación a los resultados de las categorías examinadas, se llegó a la conclusión que **el análisis por título y abstract de los artículos parece suficiente para determinar con precisión cada clasificación sin necesidad de recurrir al texto completo**, por lo que se recomendó concentrar el análisis terminológico en estos apartados y descartar el segundo.

Se muestra pues oportuno, dadas las taxonomías establecidas en el informe, explotar estos apartados más reducidos de las publicaciones en busca de los conceptos clave que las construyen. En una comparativa porcentual global, las frecuencias obtenidas por cada categoría extraídas manualmente o a través del modelo automático, resultaron muy próximas (desviación típica menor del 2%) hecho que validaba a priori la aplicabilidad de utilizar este tipo de soluciones analíticas en la gestión del conocimiento.

El análisis de artículos de la prueba también incluyó la experimentación con otra técnica de minería de datos (*clustering*) utilizada para la división de datos en grupos de objetos similares [4]. Este método se utilizó para descubrir nuevos topics relevantes (organizados como subtemas) que pueden guardar una similitud con el topic principal del artículo (ver Tabla 4). Esta opción fue posible dado que el modelo utilizado permite asociar niveles de relevancia a las categorías obtenidas en el análisis y descubrir relaciones, a priori desconocidas, entre ellas.

Tabla 4. Resultados de prueba multicategoría

Núm. artículos analizados	Umbral mínimo de relevancia sobre Title+Abstract	Umbral mínimo de relevancia sobre artículo completo
89	*30%*	*10%*
	Núm. medio de categorías obtenidas sobre Title+Abstract	Núm. medio de categorías obtenidas sobre artículo completo
	2	5

En un 93,07% de los casos se consideró relevante la categoría asignada por el motor en la clasificación subtemática o multicategoría del texto. Considerando estos resultados, **el análisis del artículo completo se convirtió en el más recomendado para la detección y extracción de subtemas relevantes de e-learning.**

Clasificación automática de publicaciones de 2018

Una vez comprobada en la prueba piloto la viabilidad de este tipo de herramientas de analítica de datos en la detección y clasificación de términos significativamente concurrentes de los artículos, se decidió completar el modelo con el total de taxonomías del informe eLRR ’17 y refinar las testeadas hasta cubrir el 100% de ellas. El propósito era concluirlo en su totalidad para ejecutar una clasificación automática completa con la información de los artículos de investigación en e-learning del periodo 2018. Así pues, para cada artículo, se han seguido distinguiendo los descriptores de:

- Topic principal y su categoría temática.
- Etapas educativas en la que se enmarcan las investigaciones.
- La metodología de investigación utilizada en cada publicación.

Todo el proceso se ha efectuado tomando como referencia el mismo marco de codificaciones y clasificaciones temáticas anteriormente desarrolladas, excluyéndose

del presente trabajo su revisión, actualización e identificación de nuevos topics no categorizados en el pasado informe. En este sentido, y a diferencia del eLRR '17 que era un trabajo de investigación, los datos recogidos en las siguientes tablas y gráficos, aunque ayuden a forjar una idea del estado de la investigación en e-learning durante el 2018, no deben considerarse en ningún caso exactos para definir sus tendencias.

Los datos mostrados a continuación, junto con el conjunto de resultados del proyecto, nos han ayudado a comprobar los beneficios del uso de la minería de texto para procesar rápidamente grandes volúmenes de información con los que llegar a conclusiones precisas de apoyo a las argumentaciones de nuestros estudios sobre aprendizaje en línea.

Resultados

Muestra

Para la elaboración del análisis automático se han procesado 1043 artículos sobre e-learning publicados en WoS y Scopus durante la anualidad del 2018. Para la obtención de las obras se ha utilizado los mismos criterios de búsqueda bibliográfica que en el informe eLRR '17.

Tablas

Tabla 5. Clasificación 20 primeros topics 2018 por volumen de frecuencias

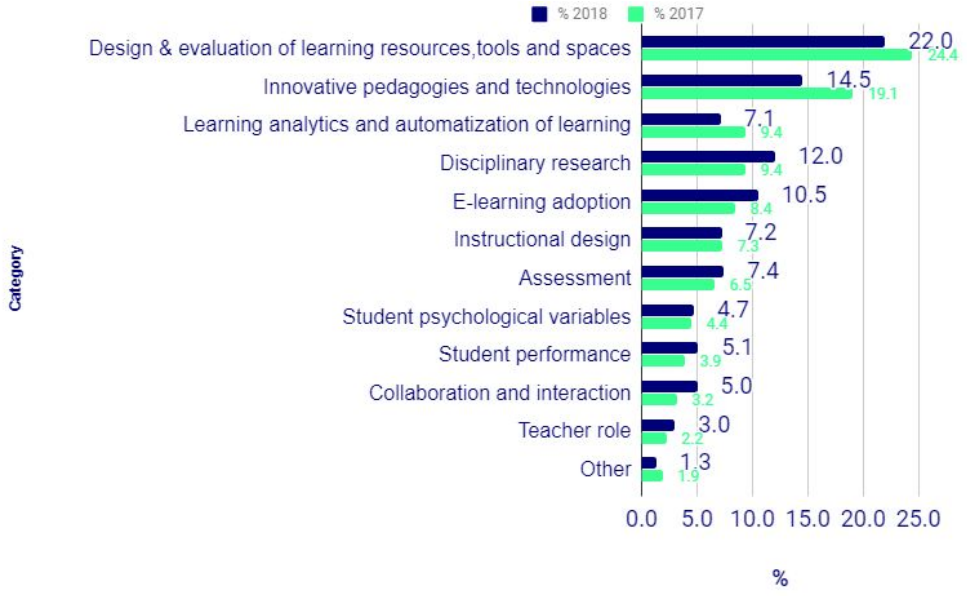
Posición en 2018	Categoría temática	Descripción	Variación desde 2017	Frecuencia en 2018	% total categoría en 2018
1	Assessment	Tipo de evaluación que posibilitan los entornos digitales (rúbricas, e-assessment, evaluación formativa, badges).	↑1	66	6.3
2	Mobile learning	Metodologías basadas en el aprendizaje móvil o ubicuo.	↓1	55	5.3
3	MOOC	Diseño, implementación y evaluación (efectividad, satisfacción) de MOOC.	0	55	5.3
4	Platform	Diseño, implementación, evaluación y selección de plataformas de formación en línea (LMS, VLE).	↑3	44	4.2
5	E-learning adoption	Implementación de cursos a distancia allí donde no había (comparación con la formación presencial, evaluación de la	0	42	4.0

		percepción o la eficiencia).			
6	E-health	Publicaciones sobre e-learning de formación disciplinar del ámbito de la salud y de promoción de la salud en la población en general.	↑12	38	3.6
7	Training	Formación profesional en línea en diferentes disciplinas (salud, empresa y emprendimiento, recursos humanos y formación a lo largo de la vida).	↑10	37	3.5
8	Communication	Publicaciones sobre la comunicación necesaria, entendida en sentido amplio (redes, estilos lenguaje, herramientas, etc.) para la enseñanza y aprendizaje en línea.	↑25	29	2.8
9	Gamification	Aplicación de juegos o de estrategias de juego para mejorar el aprendizaje (gamificación, *serious games*).	↓3	29	2.8
10	Virtual reality	Metodologías basadas en la realidad virtual (simulaciones, pacientes virtuales, mundos virtuales, 3D).	↓6	28	2.7
11	Blended learning	Publicaciones que evalúan el diseño de experiencias de transición de una educación presencial a una híbrida (*blended*), combinando presencialidad con no presencialidad.	↑8	26	2.5
12	Adaptive learning	Iniciativas de enseñanza y aprendizaje personalizado, itinerarios de aprendizaje adaptados (*learning paths*), y de retorno adaptado (*adaptive feedback*).	0	25	2.4
13	Acceptance	Estudios que miden la aceptación en la adopción de iniciativas y componentes de e-learning (cursos, herramientas, plataformas, etc.) por parte de sus actores.	↑12	24	2.3
14	Teacher	Estudiar el rol y acción docente en e-learning.	↑14	23	2.2
15	Student performance	Factores de impacto en el rendimiento de la actividad de aprendizaje del estudiante. Sistemas y mecanismos para medirlos y evaluarlos.	↑16	23	2.2
16	Instructional design	Diseño instruccional de secuencias didácticas, actividades y cursos.	↓7	22	2.1
17	Course assessment	Diseño, implementación y evaluación de cursos en línea específicos.	↓9	22	2.1
18	E-learning	Evaluación de las variables que miden el	↑22	21	2.0

	success factors	éxito y eficacia en la integración de sistemas de e-learning.			
19	Social media	Diseño e introducción de aplicaciones de *social media* para el aprendizaje basado en la interacción social.	↓4	21	2.0
20	Remote laboratory	Diseño, desarrollo y recursos la alrededor de espacios virtuales y/ o remotos de experimentación destinados al aprendizaje en línea.	↑7	21	2.0

Gráficos

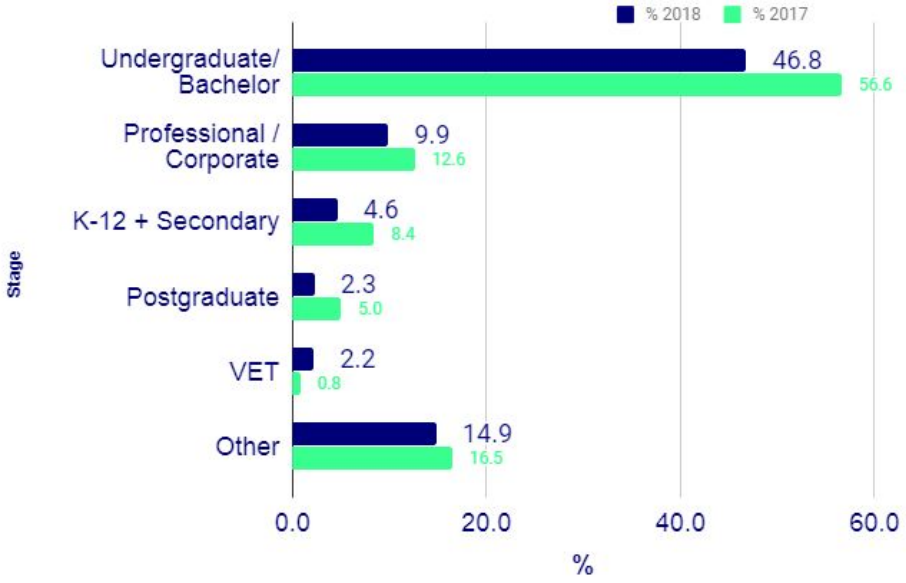
*Figura 2. Comparativa de frecuencias anuales según categoría temática**



**Excluidas 46 referencias a las que el sistema no ha podido asociar ninguna de las categorías temáticas estructuradas en el eLRR'17.*

Figura 3. Comparativa de frecuencias anuales según metodología

Figura 4. Comparativa de frecuencias anuales según etapa educativa



Clasificación por Subtemas

Como se introducía en el apartado del piloto, lo modelos de clasificación profunda, al ofrecer un análisis a nivel de fragmento, permiten extraer el tema principal de un texto y también subtemas, ordenados éstos por orden de relevancia. Así, se ha podido establecer una clasificación "multicategoría" mediante la extracción del topic principal de cada obra, para a continuación, extraer topics o "subtemas" con los que éste puede relacionarse de manera significativa. Esta característica puede resultar de utilidad para descubrir correlaciones temáticas que ayuden a conocer en mayor profundidad el contexto de una investigación en concreto, y junto la aplicación de otras técnicas computacionales, la configuración de dominios científicos completos.
A modo de ejemplo de este trabajo, tal y como se refleja en la siguiente tabla, en 3 artículos cuyo tema destacado fue "Training", el siguiente tema predominante detectado (con más de un 50% de relevancia) fue "eHealth". Si se analizan en detalle estos tres artículos se comprueba que abordan investigaciones alrededor de la capacitación profesional en el ámbito clínico. La inclusión de subtemas puede indicar las diferentes etiquetas relevantes con las que sería posible catalogar también el contenido de un artículo.

Tabla 6. Ordenación de topics multicategoría

1r topic relevante	2o topic relevante	Núm. de artículos
Active learning	Assessment	**3**
Adaptive learning	Student performance	**3**
Assessment	Course design	**6**
	Course evaluation	**5**
	MOOC	**4**
	Communication	**4**
	Satisfaction	**3**
	Platform	**3**
	Mobile learning	**3**
	Social media	**3**
	Skills	**3**
Course evaluation	Assessment	**4**

	Blended learning	3
Course design	Student performance	3
E-learning success factors	Platform	3
E-health	Communication	3
E-learning adoption	Acceptance	3
	Course Design	3
Gamification	Assessment	4
	Skills	3
	Student performance	3
	Virtual reality	3
Learning analytics	Student performance	3
Mobile learning	Social media	3
Platform	Communication	3
	Assessment	3
	Resources	3
Training	Teacher	4
	E-health	3

Reglas

Una vez completado el modelo con el total de categorías, las reglas utilizadas han sido de:

- Nivel educativo: 7 categorías - 93 reglas de clasificación.
- Metodología empleada: 6 categorías - 154 reglas de clasificación.
- Clasificación temática: 67 categorías - 380 reglas de clasificación.

Conclusiones

Los datos recogidos en el proyecto, demostraron la viabilidad de utilizar técnicas de minería de texto y modelos de reglas de clasificación automática en la detección y clasificación de tendencias en un ámbito de investigación concreto. Especialmente, como en este caso, cuando las temáticas del campo de estudio a explorar se encuentran ya estructuradas y categorizadas mediante trabajos previos de investigación, lo que facilita enormemente la captación y ordenación de la información existente. El uso de la minería de datos con tales fines de investigación u otros de

inteligencia de negocio, no es un fenómeno nuevo, hace tiempo se aplica en numerosos campos y disciplinas del conocimiento ya consolidadas.

Debemos considerar, sin embargo, que el campo del eLearning no se trata precisamente de uno de estos dominios consolidados, ya que esta categoría temática aún no existe como tal en el sistema mundial de publicaciones científicas (Tibaná-Herrera et al., 2018) [5]. Como explican estos autores, se trata de un dominio fragmentado donde, por ejemplo, en términos bibliométricos, los artículos sobre aprendizaje electrónico suelen transitar entre enfoques técnicos y educativos y bajo categorías de publicación relacionadas con las Ciencias Sociales y de la computación principalmente. Observamos pues cómo en el área del e-learning se pueden presentar dificultades desde el proceso inicial de búsqueda y recogida de datos.

Respecto al tratamiento de los contenidos, es importante señalar que, a pesar de que la taxonomía diseñada ha facilitado la clasificación automática de los artículos, ésta no se hizo pensando en “lenguaje máquina”, hecho que introduce en el procesamiento algunas ambigüedades y solapamientos terminológicos que pueden influir en la exhaustividad y precisión de los resultados. Poder enseñar al sistema situaciones de contexto en las que sepa diferenciar que el objetivo final de un SNA (*Social Network Analysis*) pueda ser el de evaluar aprendizajes en liderazgo, distinguir si la cita a un aprendizaje profundo se efectúa en términos de IA o de competencias, o que tras los ensayos de unos modos de comunicación docente se esconde el análisis de una estrategia pedagógica, es imprescindible para mejorar la calidad en las clasificaciones.

Mientras se cohesiona el consenso en la singularidad del ámbito de aprendizaje en línea, este podría ser un contexto propicio para seguir completando y ajustando un modelo de captación y categorización automática de tendencias como el testeado en la UOC. Implementaciones orientadas a la actualización y adaptación de su diccionario al procesamiento automático. Acciones también encaminadas a profundizar en las técnicas de análisis de datos y complementar una capa cognitiva que permita precisar aún más sus clasificaciones, a la vez que le brinda ayuda para descubrir, entre la multitud de la literatura científica, nuevos y significativos descriptores temáticos dentro y fuera de la disciplina.

Agradecimientos

MeaningCloud™ (https://www.meaningcloud.com) ha sido utilizado para analítica de texto en el desarrollo y validación de esta prueba piloto.

Referencias

[1] eLearn Center (2018). *E-Learning Research Report 2017. Analysis of the main topics in research indexed articles*. Barcelona: eLearn Center (UOC). Recuperado el 17 de julio de 2019 de http://hdl.handle.net/10609/75705.

[2] Do Prado, H.A., Ferneda, E. (2007), *Emerging Technologies of Text Mining: Techniques and Applications*. Hershey, New York: Information Science Reference.

[3] Precision and recall (n.d.). In *Wikipedia.* Recuperado el 17 de julio de 2019 de https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=913652928

[4] Berkhin P. (2006), A Survey of Clustering Data Mining Techniques. In: Kogan J., Nicholas C., Teboulle M. (Ed.) *Grouping Multidimensional Data*. Berlin/Heidelberg: Springer.

[5] Tibaná-Herrera, G., Fernández-Bajón, M.T., De Moya-Anegón, F. (2018). Categorization of E-learning as an emerging discipline in the world publication system: a bibliometric study in SCOPUS. *International Journal of Educational Technology in Higher Education 15*, 21.

[6] Sharma, D., Kumar B., Chand, S. (2018). Trend Analysis in Machine Learning Research Using Text Mining. Greater Noida (UP), India: *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN),* pp. 136-141.