eib0965877708

Veinticinco años de investigación en redes sociales: evolución de temas entre 1997 y 2021 empleando el algoritmo Asignación Latente de Dirichlet

Juan-Antonio Martínez-Comeche*

Artículo recibido:
15 de marzo de 2023
Artículo aceptado:
1 de junio de 2023

Artículo de investigación

RESUMEN

El campo de las redes sociales ha sufrido importantes transformaciones en los últimos veinticinco años, en particular con la introducción de aplicaciones y plataformas digitales, así como la incorporación de estudios de otros campos del conocimiento que adoptan el enfoque de redes sociales en sus análisis. Este artículo ofrece una visión general de la evolución de los tópicos de investigación en este ámbito entre 1997 y 2021 a partir de la modelización de temas. El estudio parte de la producción académica que se recupera de la base de datos Scopus, considerando ventanas temporales de un año y utilizando el software Mallet. Se obtienen siete temas, cuya evolución en el tiempo se describe. Se concluye que los temas

* Departamento de Biblioteconomía y Documentación, Facultad de Ciencias de la Documentación, Universidad Complutense de Madrid, España
juaamart@ucm.es

relacionados con los medios de comunicación social, así como las redes sociales en línea son estudiados con especial intensidad en los últimos años.

Palabras clave: Redes sociales; Evolución temática; SCOPUS; LDA; Medios de comunicación social

Twenty-five Years of Research in Social Networks: Evolution of Topics between 1997 and 2021 Based on Latent Dirichlet Allocation (LDA)

Juan-Antonio Martínez-Comeche

ABSTRACT

The social network field has suffered significant transformations in the last 25 years, particularly with the introduction of social networks online, as well as incorporated studies from many other knowledge fields that adopt the social network approach in their analyses. This paper offers an overview of the evolution of research topics in this field between 1997 and 2021 based on topic modeling. The methodology used draws from the Scopus database, considering time windows of a year and using the software, Mallet. Seven topics are obtained, whose evolution over time is described. It is concluded that the topics related to social media and social networks online have been studied with special intensity in the last years.

Keywords: Social networks; Topic evolution; SCOPUS; LDA; Social media

INTRODUCCIÓN

Las Ciencias Sociales se interesan por el estudio de grupos sociales, concebidos en esencia como un conjunto de entidades que interrelacionan entre sí. En la década de 1930 Jennings y Moreno introducen los primeros sociogramas (Moreno y Jennings 1938), gráficos que muestran la red de relaciones existentes entre las unidades y la estructura social del grupo, como el de la *figura 1*.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

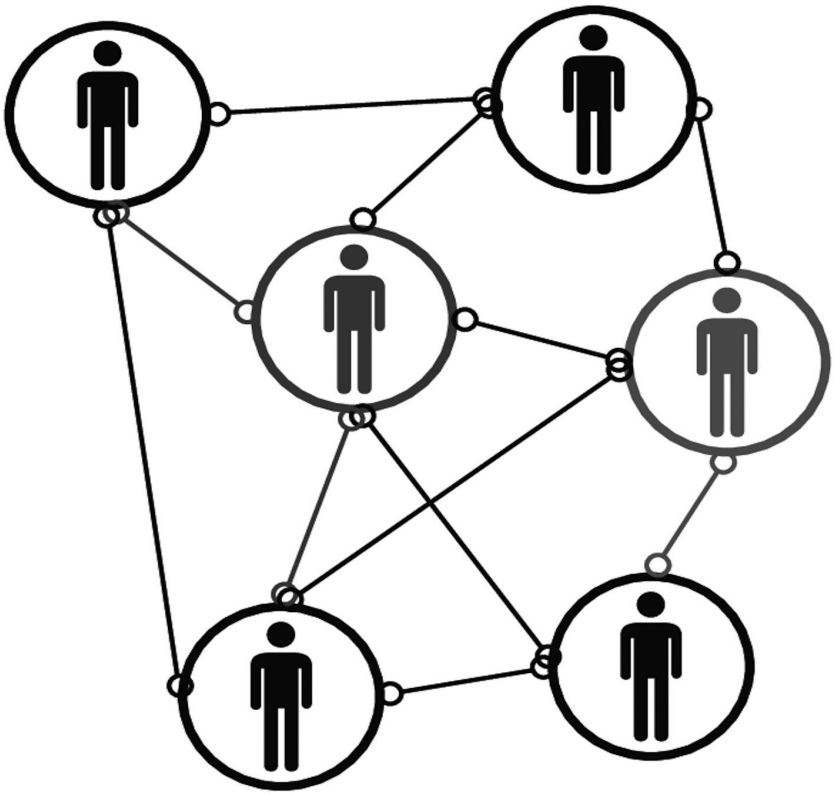

Figura 1: Estructura de relaciones de una red social
Fuente: Tomado de Ricci (2018).

En esta figura se muestra la estructura de una red social compuesta por un grupo de unidades o nodos (personas en este caso) que se relacionan entre sí por compartir un cierto objetivo común. Los vínculos (representados por líneas en la figura) expresan en sus extremos los nodos o unidades que se relacionan directamente. Entre estos se establece un flujo de información que es unidireccional o bidireccional, como en el caso de la figura, indicando que el flujo informativo se establece en ambos sentidos.

Moreno constituye uno de los pioneros que aborda un enfoque cuantitativo para describir la organización de relaciones y los roles de los individuos en un grupo, aportando la perspectiva de la Sociometría, bajo la influencia de la teoría psicológica de la Gestalt (Moreno 1937). Por su parte, Heider (1946) desarrolla la teoría de los grupos dinámicos que, junto a la teoría de grafos desarrollada por Harary (1969), representan los pilares del denominado análisis de redes sociales (Wasserman y Faust 1994).

A finales de 1970 la búsqueda de patrones en las interacciones entre individuos empieza a reconocerse como una especialidad científica emergente basada en la teoría de grafos y en modelos estadísticos y computacionales, iniciándose así la unificación de este campo. Tres hechos contribuyen a unir las hasta entonces dispersas aproximaciones: la fundación del primer boletín (*Connections*) por Barry Wellman (Freeman 2004), la publicación de los primeros textos con una perspectiva estándar (Berkowitz 1982; Knoke y Kuklinski 1982), y el desarrollo de los primeros programas informáticos, SONIS y UCINET (Pappi y Stelck 1987; Freeman 1988).

Muchas áreas diversas entre sí de conocimiento emplean el enfoque de redes sociales en sus análisis: desde la zoología (Sueur y Pelé 2016; Zanardo *et al.* 2018) hasta la química (Chodera y Pande 2011; Gore *et al.* 2021), pasando por la informática (Wang y Robinson 2002) y la tecnología de las comunicaciones (Zhang *et al.* 2009). Las disciplinas que emplean este enfoque se encuentran en constante transformación, a modo de ejemplo: la economía, la medicina, la geografía, las ciencias de la información, la educación o las redes de transportes (Otte y Rousseau 2002). Las comunidades virtuales se incorporan a este campo como un dominio nuevo, concebido como grupos de individuos que comparten intereses e intercambian entre sí contenidos generados por ellos mismos a través de aplicaciones informáticas basadas en la web 2.0 (Ridings *et al.* 2002).

Con el objeto de evitar confusiones terminológicas, en este trabajo se emplea «medios de comunicación social» para designar cualquier plataforma o servicio en línea utilizados para crear comunidades virtuales, desde aplicaciones de mensajería instantánea (WhatsApp), proyectos colaborativos (Wikipedia) e intercambio de fotos o vídeos (Flickr, YouTube), hasta redes sociales (Facebook) o incluso mundos virtuales como Second Life, dada su amplia significación que posee en la actualidad (Aichner y Jacob 2015; Statista *et al.* 2022). Se utiliza «redes sociales en línea» para especificar servicios basados en la web que permiten a las personas estructurar y hacer públicas o semi públicas sus redes sociales (Boyd y Ellison 2008), y «aplicaciones de mensajería instantánea» para designar servicios que permiten a los usuarios intercambiar contenidos privados entre dos o más miembros específicos y conocidos de un grupo (Taipale y Farinosi 2018).

Antecedentes

La evolución temática dentro de un campo se puede abordar desde dos enfoques principales: modelos bibliométricos y modelización de temas (Song *et al.* 2014). A su vez, los modelos bibliométricos se basan en diversos elementos como las palabras, los autores, las revistas o las citas en un corpus de textos (Han 2020; Onyancha 2018; Chang *et al.* 2015).

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

Los autores de los textos son de gran utilidad para descubrir la evolución temática de un área de conocimiento, considerando un tema como el conjunto de intereses comunes de un grupo de autores (Jung y Yoon 2020). Por su parte, el análisis de las revistas donde se publican los artículos conforma una técnica ampliamente utilizada con esta finalidad, mediante el estudio detallado de contenido con respecto a los tópicos abordados en una selección de revistas del área (Tuomaala *et al.* 2014; Armann-Keown y Patterson 2020).

Cuando se opta por las palabras como los elementos de los documentos que permiten dar a conocer los temas presentes en un corpus y su evolución temporal, se emplean diversas técnicas, desde las palabras clave más frecuentes hasta su aparición simultánea o su distribución por temas o en ciertos ciclos de tiempo (Onyancha 2018; Peset *et al.* 2020).

Las citas, por su parte, se utilizan ampliamente para observar diversos aspectos como la estructura intelectual o la evolución temática, de manera habitual mediante análisis de co-citación (Chen 2006; Chabowski y Samiee 2023) o acoplamiento bibliográfico (Yanhui *et al.* 2021). En ocasiones se ha superpuesto el análisis de redes sociales a dichos métodos bibliométricos (Yang *et al.* 2012). Con ayuda de cualquiera de estos elementos se dan a conocer diversos indicadores de la actividad científica y su evolución temporal en relación con la productividad, la colaboración, la circulación y el impacto, entre otros aspectos.

Sin embargo, las técnicas bibliométricas no son las más apropiadas cuando el volumen de datos de entrada aparece elevado. En circunstancias semejantes a este estudio —donde el número de documentos analizados ronda los 150 000—, resulta más efectivo el empleo de técnicas de modelización de temas (Jeong y Min 2014; Yau *et al.* 2014; Jung y Yoon 2020), basadas en un análisis estadístico de palabras en el corpus, el cual permite reunirlas en pocos grupos que constituyen los temas del área (Liu *et al.* 2020). No existe técnica única de modelado para la detección de temas. Unas se basan en las relaciones sintácticas entre las palabras del corpus (Ferrer *et al.* 2004), otras en las de semánticas entre las palabras (Arruda *et al.* 2016), e incluso algunas aportan el empleo de redes neuronales a estos modelos lingüísticos (Mikolov *et al.* 2013).

Sin embargo, las técnicas de modelado más utilizadas en el análisis de corpus grandes de carácter textual, como Asignación Latente de Dirichlet (LDA) (Blei *et al.* 2003), Análisis Probabilístico de Semántica Latente (PLSA) (Hofmann 1999) o Análisis de Semántica Latente (LSA) (Landauer *et al.* 2007), consideran los temas como grupos de palabras y se basan en procedimientos estadísticos de simple co-aparición de estas para su identificación. Aquí se empleó el modelo LDA porque ha mostrado ser un método efectivo en el descubrimiento de la estructura temática subyacente en corpus de temática dispersa en áreas de conocimiento distintas, como la de las redes sociales (Yau *et al.* 2014; Suominen y Toivanen 2015; Jeong y Min 2014; Banerjee y Basu 2007).

El algoritmo LDA aborda el descubrimiento de temas en un corpus a partir de las siguientes premisas: cada documento se representa mediante la co-ocurrencia de un conjunto de temas; cada tema se presenta a través de un conjunto de palabras co-ocurrentes. El propósito del algoritmo es generar un conjunto de palabras de alta probabilidad de ocurrencia para cada tema, expresión de una estructura semántica subyacente en el corpus más allá de palabras concretas empleadas en los documentos (Deerwester *et al.* 1990), de donde la inclusión de la palabra latente en la denominación del algoritmo (Wu *et al.* 2014).

Este procedimiento constituye un modelo de inferencia bayesiano en cuanto que genera esencialmente dos listados de probabilidades (Shen y Wang 2020): uno con las de aparición de todos los temas en cada documento del corpus y otro con las de aparición de todas las palabras en cada tema (Li y Lei 2021; Zhu *et al.* 2016; Kai *et al.* 2019). Además, constituye un algoritmo de aprendizaje no supervisado (Tdk Technologies 2020) en cuanto que la repetición de los cálculos de probabilidades de las palabras en grupos mejora los resultados en cada iteración sucesiva. Debe considerarse, sin embargo, la relación entre el tiempo invertido en un número alto de iteraciones y la mejora relativa en la precisión del modelo resultante.

De las varias implementaciones disponibles para ejecutar LDA, se optó por el paquete basado en Java Mallet (McCallum 2022a), conocido programa de código abierto para modelización de temas.

El análisis de la evolución de los temas se considera desde dos perspectivas: el desarrollo del contenido, que implica tanto cambios en las palabras distintivas de un tema como el aumento y disminución de su frecuencia con el tiempo; y el de la intensidad temática, que muestra cambios en la atención prestada por los investigadores al tema y la consiguiente mayor o menor presencia de este en los documentos del corpus (Shan y Li, 2010; Zhu *et al.* 2016). Ambos análisis se complementan y ofrecen un conocimiento profundo del desarrollo de los tópicos. En este estudio se aborda tanto la evolución del contenido como de la intensidad temática a partir del sometimiento del corpus de documentos al algoritmo LDA (Griffiths y Steyvers 2004).

Objetivos

El principal objetivo de este estudio radica en descubrir y analizar los principales temas de investigación en redes sociales y su evolución en contenidos e intensidad en los recientes 25 años (1997-2021). Así, la intensidad alude a la variación con el tiempo de la importancia relativa de los distintos temas en el corpus. En dicho análisis sobre las redes sociales se considera su naturaleza multidisciplinar

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

y la constante incorporación de nuevos dominios a este campo, entre los que destacan aquellas que se producen en línea, cuya primera aplicación (Six Degrees) aparece en 1997. Hasta donde se sabe, este tema no ha sido previamente abordado. La hipótesis parte de que el descubrimiento de los temas y su evolución en este período resulta posible mediante técnicas de modelización de temas, en concreto Latent Dirichlet Allocation (LDA) o Asignación Latente de Dirichlet, dado el amplio volumen de datos de entrada que dificulta la aplicación de métodos bibliométricos.

El segundo objetivo consiste en alcanzar una descripción detallada y al tiempo coherente desde el punto de vista temático. Existen técnicas de modelado de temas, como hLDA (Hierarchical Latent Dirichlet Allocation), que añade a la detección de estos su evolución temporal y la relación jerárquica entre ellos (Song *et al.* 2016). Sin embargo, se ha puesto de manifiesto que el resultado de estas técnicas, en principio más completas, suele ser de difícil interpretación (Ding y Chen 2014). Para superar el inconveniente, se opta por efectuar de manera independiente cada uno de los procesos necesarios, en vez de aplicar un solo programa o técnica, y poner en práctica la triangulación metodológica mediante el análisis textual de una muestra de los datos de entrada. Con ello se asegura la coherencia temática de los resultados, al tiempo que se detecta qué parámetros son decisivos para generar una descripción temáticamente congruente.

La originalidad del trabajo reside, además del campo objeto de estudio: las redes sociales en su conjunto –no únicamente las aplicaciones informáticas de éstas en línea–, también en dos aspectos metodológicos: uno, la aplicación de un enfoque cualitativo (análisis textual de 1 700 documentos) superpuesta a la técnica cuantitativa propia de la modelización de temas mediante LDA; dos, la aplicación de técnicas del Procesamiento del Lenguaje Natural (análisis morfológico, léxico y de reconocimiento de entidades) y de técnicas empleadas en la Recuperación de Información (ponderación TF.IDF o la similaridad coseno) para mejorar los resultados de la modelización de temas mediante LDA.

Métodos y materiales

En este estudio se utilizó una metodología mixta, cuantitativa y cualitativa, para alcanzar el objetivo propuesto. Las técnicas cuantitativas incluyeron el desarrollo de diversos programas en Java y distintos scripts en lenguaje R para realizar cada una de las tareas de naturaleza estadística involucradas en el estudio, dado el alto volumen de documentación gestionada, así como el empleo del paquete Mallet para efectuar el proceso de modelización de temas, también de carácter estadístico. La técnica cualitativa involucró el análisis textual de una muestra de

entre 10 y 20 documentos más destacados de cada uno de los 155 temas generados de manera inicial por el programa Mallet, lo que supuso alrededor de 1 700 documentos, esto es, algo más del uno por ciento del total de documentación manejada en la investigación.

En este apartado se explican en detalle los pasos seguidos en el estudio y las técnicas empleadas en cada uno de ellos, cuyo esquema se muestra en la *figura 2*.

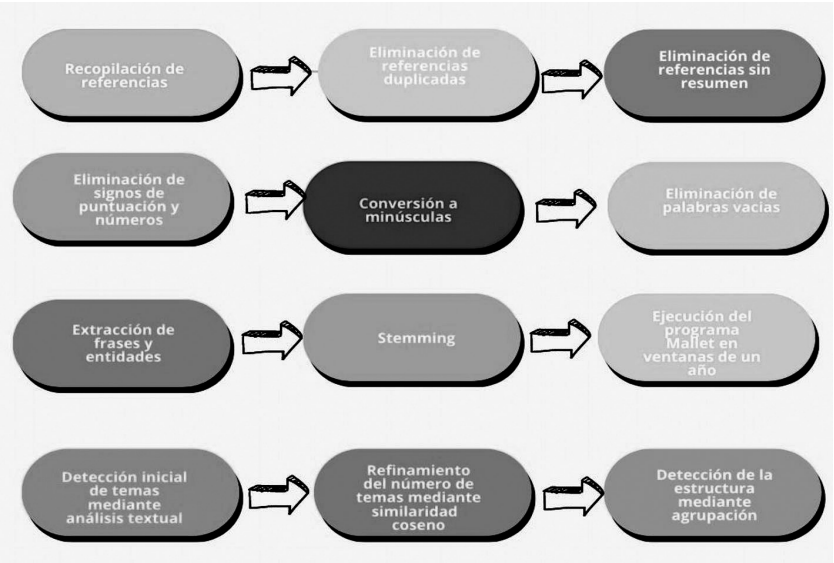

Figura 2: Esquema del proceso de análisis efectuado
Fuente: Elaboración propia.

Para la recopilación de los documentos sobre redes sociales de entre 1997 y 2021 se utilizó la base de datos Scopus, creada en 2004, publicada por Elsevier, siendo una de las bases con mayor volumen de documentación que abarca todas las ramas del conocimiento, empleando la consulta:

> TITLE(«social network*») OR KEY(«social network*») OR TITLE(«instant messag*») OR KEY(«instant messag*») OR TITLE(«facebook») OR KEY(«facebook») OR TITLE«myspace») OR KEY(«myspace») OR TITLE(«cyworld») OR KEY(«cyworld») OR TITLE«twitter») OR KEY(«twitter») OR TITLE(«qzone» OR KEY(«qzone») OR TITLE(«instagram») OR KEY(«instagram») OR TITLE(«ask.fm») OR KEY(«ask.fm») OR TITLE(«tiktok») OR KEY(«tiktok») OR TITLE(«kuaishou») OR KEY(«kuaishou») OR TITLE(«tumblr») OR KEY(«tumblr») OR TITLE(«whatsapp») OR KEY(«whatsapp») OR TITLE(«msn messenger») OR KEY(«msn messenger») OR TITLE(«wechat») OR KEY(«wechat») OR TITLE(«telegram») OR KEY(«telegram») OR TITLE(«tencent qq»)

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

OR KEY(«tencent qq») OR TITLE(«snapchat») OR KEY(«snapchat») OR TITLE(«hangouts») OR KEY(«hangouts») OR TITLE(«skype») OR KEY(«skype») OR TITLE(«viber») OR KEY(«viber») OR TITLE(«kakaotalk») OR KEY(«kakaotalk»)) AND PUBYEAR AFT 1996 AND PUBYEAR BEF 2022 AND LANGUAGE(english) AND (DOCTYPE(ar) OR DOCTYPE(bk) OR DOCTYPE(ch) OR DOCTYPE(cp) OR DOCTYPE(re) OR DOCTYPE(sh))

En esta consulta se limitó la lengua de los documentos al inglés, pues la presencia de dos lenguas complicaría en exceso el desarrollo de los programas informáticos empleados en el proceso seguido, además el volumen que corresponde a los de inglés es superior al español, así como el tipo de documento a las vías principales de comunicación de la investigación científica, esencialmente: artículos, libros, capítulos de libro y ponencias. En la consulta no se incluyó la aparición de las palabras en el Abstract porque se captaron documentos que utilizaban las redes sociales como metodología de obtención de datos, pero no analizaban las redes sociales como tales. Se obtuvieron un total de 154 582 documentos.

Antes de aplicar el algoritmo de modelización de temas fue necesario preparar adecuadamente los textos a someter al proceso. Las sucesivas fases en que constó el preprocesamiento de documentos tras la recopilación de datos fueron: 1) Eliminación de las referencias duplicadas; 2) Así también, de las referencias que careciesen de resumen; 3) Por igual, de los signos de puntuación y los números, convirtiendo las mayúsculas a minúsculas, mediante el analizador SimpleAnalyzer incorporado a la biblioteca Java de código abierto Lucene Core (Apache Software Foundation 2022; McCandless *et al.* 2010); 4) De manera similar, de las palabras vacías que constasen de tres o más caracteres.

Además, para extraer la frecuencia de aparición de cada palabra en la muestra se utilizó el programa en Java desarrollado por Billy Joel Johnson (Johnson 2019). En la lista de palabras vacías se incluyeron solamente las 101 palabras con valor funcional o sintáctico que figuraban al menos en uno de los siguientes listados de palabras vacías en inglés: LingPipe (Alias-i 2016); la lista compilada por Kevin Bougé para el sistema de recuperación de información SMART (Brigadir 2022); la lista incorporada a la biblioteca NLTK (NLTK 2022); y la lista original de Snowball creada por Martin Porter (Snowball 2021). La eliminación de las palabras con 1 ó 2 caracteres se realizó posteriormente, pues el programa Mallet lleva a cabo esta tarea por defecto al preparar la colección antes de someterlo al proceso de detección de temas; 5) Extracción de frases y entidades. Para ello se emplearon todas las palabras clave del corpus compuestas de dos o más palabras (que incluyen también entidades geográficas, personales o temporales), convirtiéndolas en unitérminos (por ejemplo, *ecommerce*); 6) Stemming o reducción morfológica. Para lo cual se empleó el programa Snowball incorporado al proyecto OpenNLP, que

desarrolla el algoritmo Porter2 para el inglés (OpenNLP 2022); 7). Por último, importación de datos en Mallet, que por defecto elimina cualquier palabra de 1 ó 2 caracteres (Mallet 2022). El programa Mallet es un paquete basado en Java para el procesamiento estadístico del lenguaje natural, la clasificación de documentos, la agrupación o clustering y la extracción de información. Esta aplicación resultó de fácil instalación y manejo, siendo utilizada de modo amplio en el campo del aprendizaje máquina.

Dadas las peculiaridades de la colección durante el periodo que se analizó, se optó por considerar ventanas temporales de un año de duración, cada una de ellas acogiendo los documentos publicados el mismo año (Wu *et al.* 2014). Con ello se pretendió mejorar la adaptabilidad del algoritmo a los cambios que se han ido produciendo en estos años en un área tan dinámica como las redes sociales.

La aplicación del algoritmo LDA requirió la fijación previa de varios parámetros, además del número de periodos en que se dividió el corpus de entrada. El más importante fue el número de temas. Aunque requirió mucho más tiempo y carga de trabajo, no se basó en el cálculo de la perplejidad (Griffiths y Steyvers 2004), el consejo de expertos o en estudios previos para decidir un número aproximado que haya resultado de utilidad en áreas afines (McCallum *et al.* 2007), sino que se adoptaron los criterios que figuran en la propia documentación del programa Mallet (McCallum 2022b; Isasi 2022; Graham *et al.* 2021; Kumar 2022; Silge 2018). Así, tras aplicar el algoritmo LDA con varios números de temas, en una primera fase se seleccionaron los valores de temas con mejores rangos y valores medios de coherencia (*coherence*) y exclusividad (*exclusivity*); de entre ellos, en una segunda fase, los valores de temas con mejores rangos y valores medios del parámetro alpha; en una tercera y última fase, se eligieron para cada periodo anual el número de temas que presentó una mejor distribución de tales temas en la documentación correspondiente. En cuanto a otros parámetros, se fijaron 1 000 iteraciones para el desarrollo del modelo y 15 palabras destacadas por tema.

Una vez ejecutado el programa Mallet con cada ventana temporal de datos y obtenidos los temas correspondientes a cada año, el problema fue comprobar cómo se relacionaban los temas de años sucesivos. Este proceso de correlación de temas (Liu *et al.* 2020) se llevó acabo en dos fases: primero se efectuó el análisis textual de las palabras destacadas, títulos y resúmenes de al menos los 10 documentos más relevantes señalados por Mallet de cada tema anual, llegando en algunos casos hasta los 20 documentos más destacados hasta discernir con claridad a qué tema adscribir cada tema anual. En caso de advertir cualquier peculiaridad temática, se optó por considerarlo de manera inicial un tema distinto. Ello supuso el análisis de unos 1 600 documentos, en torno al 1 % del volumen total de la colección considerada en el estudio. En esta primera fase se detectaron 20 temas.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

A continuación, se procedió a refinar el número de temas y sus temas anuales componentes, mediante el cálculo de la similaridad coseno (Zhu *et al*. 2016). La similaridad coseno representó un procedimiento ampliamente utilizado en el área de la Recuperación de Información que calcula la similitud entre dos grupos de términos tomando en cuenta el número de términos comunes en ambos grupos y el peso o importancia de cada término para representar el grupo al que pertenece. La similaridad coseno más baja posible corresponde a cero, y la más alta a uno, en cuyo caso los grupos de términos fueron idénticos en número y representatividad.

Se partió de los resultados que proporciona Mallet para cada colección anual, en concreto, de la probabilidad de aparición de los términos de dicho año en cada uno de los temas generados en la mencionada anualidad. Considerando dicha probabilidad, como el peso de cada término en cada tema anual, y partiendo de los temas anuales del primer año 1997 como los temas iniciales, se procedió a calcular la similaridad entre cada uno de los temas anuales de los años sucesivos con todos los temas anteriores.

Para considerar un cierto tema anual, como el primero de un nuevo tema independiente, se impusieron las siguientes condiciones: en primer lugar, presentar una similaridad > 0,76 con los últimos temas anuales que componen cualquier tema detectado de manera previa; en segundo, poseer una temática distintiva y claramente definida, ajena al resto de temas obtenidos hasta ese momento, mediante el análisis textual de los 20 primeros documentos más relevantes del tema candidato; en tercero, la imposibilidad de crear un tema compuesto de forma exclusiva por un tema único anual; por último, el segundo tema anual del nuevo debe presentar una similaridad ≥ 0,76 con el primer tema anual del nuevo tema. En esta segunda fase los 20 temas iniciales se redujeron a 10.

Con la finalidad de detectar posibles relaciones entre los 10 temas, mejorando la precisión del análisis de la evolución temática, se examinó su estructura temática sometiéndolos a agrupación jerárquica. El procedimiento consistió en imponer un peso a los términos que aparecen en cualquiera de los temas anuales componentes de un tema, empleando ponderación TF.IDF, que consiste en imponer a cada término un número representativo de su importancia en el grupo de términos al que pertenece. Este número, denominado peso del término, se calcula multiplicando dos cantidades: una corresponde a la frecuencia de aparición del término en el conjunto (TF del inglés *Term Frequency*), y otra el inverso del número de grupos en los que aparece el término, indicativo de su valor para discriminar si el término surge en muchos o en pocos grupos del corpus (IDF del inglés *Inverse Document Frequency*). El cálculo se realizó de la siguiente manera:

1. En cuanto a TF (*Term Frequency*), cada vez que un término aparece en un tema anual, se suma +1 a su frecuencia. Para paliar el sesgo producido por temas con muchos temas anuales frente a otros con muy pocos temas anuales, se adoptó la siguiente fórmula para el cálculo del TF (Singhal 2001):

$$TF(term\ t\ in\ great\ topic\ d) = 1 + \log\left(1 + \log\left(tf_{t,d}\right)\right)$$

2. En cuanto a IDF (*Inverse Document Frequency*), se adoptó la siguiente fórmula para su cálculo, que evita penalizar en exceso los términos que aparecen en bastantes temas anuales de la colección en su conjunto (Singhal 2001):

$$IDF\ (term\ t) = log\frac{(N+1)}{n}$$

donde N es el número total de temas anuales en la colección (155), y donde n es el número total de temas anuales en la colección en los que aparece el término t. A continuación, se calcula la similaridad coseno entre cada pareja de los 10 temas y la matriz resultante se somete al algoritmo hclust del lenguaje R (R Core Team 2022).

Una vez observada la evolución temporal del contenido de los temas, el panorama general de la investigación en redes sociales se complementó con el análisis de la importancia relativa concedida por los investigadores a cada tema frente a los demás, aspecto denominado intensidad temática (Zhu *et al.* 2016) o popularidad temática (Zou 2018).

Se han utilizado diversos procedimientos para comparar la relevancia de cada tema en un campo de investigación. En ocasiones se han empleado palabras determinantes que muestran tendencias (Ding *et al.* 2001; Kleinberg 2003). Otras alternativas consisten en comparar el número de citas recibidas por cada tema (Ha *et al.* 2014), o utilizar modelos de aprendizaje máquina (Xu *et al.* 2019).

Como el algoritmo LDA contempla múltiples temas en un mismo documento, para el cálculo de la popularidad o intensidad temática se encontró la probabilidad media de cada tema en todos los documentos de cada año, distinguiéndose entre temas cuya intensidad resultó creciente, estable o decreciente. El aumento o disminución de popularidad de un tema se evaluó mediante el test no paramétrico de tendencia Mann-Kendall (Pohlert 2022), con un valor de $p < 0{,}001$ para considerar que fue estadísticamente significativo. Se incluyeron las gráficas de los temas con una intensidad creciente o decreciente.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

Resultados y discusión

De los 154 582 documentos iniciales obtenidos en Scopus, el preprocesamiento los redujo a 149 443 tratados finalmente por Mallet. Los cálculos para decidir el número de temas más adecuado en cada año supusieron el desarrollo de un total de 155 temas de alcance anual a lo largo de los 25 años del estudio. La *tabla 1* muestra el número final de documentos procesados y el número óptimo de temas para cada año.

Año	Número de temas	Número de documentos	Año	Número de temas	Número de documentos
1997	2	104	2010	7	5 226
1998	2	122	2011	7	6 487
1999	2	134	2012	7	7 997
2000	2	161	2013	8	9 520
2001	4	194	2014	8	11 740
2002	4	295	2015	8	1 3841
2003	6	634	2016	8	1 2086
2004	6	601	2017	8	1 2088
2005	6	894	2018	8	1 3404
2006	7	1 176	2019	8	1 4525
2007	7	1 768	2020	8	1 5345
2008	7	2 558	2021	8	1 4825
2009	7	3 718	TOTAL	155	14 9443

Tabla 1: Número de documentos y temas desarrollados por Mallet para cada año
Fuente: Elaboración propia.

Someter los 155 temas anuales a un proceso de agrupación mediante la similaridad coseno para observar cómo se relacionan temáticamente entre ellos a lo largo del tiempo requiere la imposición de un valor límite. Si no se impone un tope que impida la unión de dos temas, todos los temas anuales terminarían uniéndose en uno solo que abarcaría todo el corpus. En el caso extremo, si se impone un límite de similaridad muy alto para que dos temas anuales puedan unirse en un único tema, se tendría un número de temas cercano o igual a los 155 anuales de partida.

Para superar este inconveniente, se procedió en primer lugar a un análisis textual de los 155 temas anuales, con el objetivo de conocer el número máximo

de estos que deberían considerarse en el corpus y sus características. Luego se procedió a calcular, en sucesivos pasos, los valores de similaridad coseno entre los temas anuales siguiendo el procedimiento expuesto en el apartado de Métodos y Materiales, poniendo especial cuidado ahora en no unir dos temas anuales cuando el análisis textual previo hubiese indicado que se trata de dos temas dispares que deben permanecer separados en grupos distintos.

De este proceso resultó la conveniencia de imponer un valor límite de 0,76, porque si se admitía la unión de temas con valores de similaridad de 0,74 y 0,75, cuatro de ellos tendrían que unirse en dos (topic A y topic E, por una parte; topic J y topic T, por otra, respectivamente), presentando los cuatro una temática claramente independiente. Como resultado de esta segunda fase, los 20 temas que fijaba como máximo el análisis textual, quedaron reducidos a los 10 que figuran en la *tabla 2*.

Tema	Número de temas anuales	Similaridad coseno media	Años de aparición	Años sin aparecer
A	25	0.913	1997-2021	--
B	7	0.747	1997-2004	2001
C	15	0.780	2001-2021	2005-2012 2016
D	29	0.749	2004-2021	--
E	23	0.865	2001-2021	--
I	10	0.868	2003-2012	--
J	16	0.878	2005-2020	--
K	20	0.836	2007-2021	--
M	3	0.511	2005-2008	2007
T	7	0.918	2016-2021	--

Tabla 2: Temas tras agrupación por similaridad coseno y años de aparición
Fuente: Elaboración propia.

Mallet proporciona una serie de palabras destacadas por cada uno de los temas anuales, a raíz de los cuales se deben escoger los descriptores adecuados para cada uno de ellos. El procedimiento seguido para seleccionar estos descriptores es el siguiente: primeramente se reúnen todas las palabras destacadas que proporciona Mallet de cada uno de los temas anuales que componen cada tema y se calcula su frecuencia; a continuación, se calcula el IDF de cada uno de dichos términos empleando la misma fórmula que se ha usado en el cálculo de la estructura temática, descrita en el apartado Métodos y Materiales; finalmente se calcula el

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

peso TF.IDF de cada uno de los términos. Los términos seleccionados finalmente para cada tema incluyen los que poseen un peso TF.IDF mayor, a los que se añaden aquellos considerados representativos del contenido del tema conforme al análisis textual efectuado, pero que se hallan en un rango de frecuencias media o baja. En la *tabla 3* se resumen los términos seleccionados para cada tema de entre los obtenidos por Mallet, junto con una breve descripción en inglés de su contenido con base al análisis textual efectuado.

Temas	Términos	Descripción textual
A	health, social network, age women, relation, social support, patient, family, friend, network, hiv, sexual, behavior, parent, contact	Documents about health, focusing on the social network (e.g., family, friends) of the patients. Factors as age, gender or race are usually considered. [Documentos sobre salud, centrándose en las redes sociales –familia, amigos– de los pacientes. Factores como la edad, el género o la raza son considerados habitualmente.]
B	develop, online, network, community, inform, support, structure, work, migrate, instant messaging, converse, email, collaborate, economics, learn	Documents about diverse online networks (e.g., students, workers) and physical communities (for political, religious or migration support), discussing their structure or information diffusion, and announcing the development of the first instant messaging and social network online applications. [Documentos sobre diversas redes en línea (por ejemplo, estudiantes, trabajadores) y comunidades físicas (de apoyo político, religioso o a la emigración), analizando su estructura o la difusión de información en ellas, y anunciando el desarrollo de las primeras aplicaciones de mensajería instantánea y de redes sociales en línea.]
C	web, system, design, application, data, privacy, security, social network, algorithm, process, big data, social media, semantic web, interface, mobile	Documents on technical aspects of the design of web-based applications, algorithms related to privacy, security, and integrity of information, efficient techniques for data storage, processing, and sharing, software for mobile communication systems, and their use in social media and online social networks. [Documentos sobre aspectos técnicos del diseño de aplicaciones basadas en la web, algoritmos relativos a la privacidad, seguridad e integridad de la información, técnicas eficientes para el almacenamiento, procesamiento y compartición de datos, programas para sistemas móviles de comunicación y su empleo en medios de comunicación social y en redes sociales en línea.]

D	tweet, Twitter, node, algorithm, graph, sentiment, event, detect, recommend, trust, network analysis, privacy, semantic web, Facebook, evolution	Documents on graphs, semantic web, social networks, and social media, addressing issues related to the structure of networks, their attributes (e.g., influential nodes, trust relationships), their evolution over time, and multiple aspects related to social media (e.g., user interests, recommendation algorithms, trends, sentiment analysis, fake news detection or information extraction). [Documentos sobre grafos, web semántica, redes sociales y medios de comunicación social, abordando aspectos relativos a la estructura de las redes, sus atributos (el concepto de influencia y relaciones de confianza, por ejemplo) o su evolución temporal, además de múltiples aspectos relacionados con los medios de comunicación social (intereses de los usuarios, algoritmos de recomendación, tendencias, análisis de sentimientos, detección de bulos o extracción de información.]
E	network, social, social network, community, group, politics, social network analysis, business, culture, economics, govern, immigrate, policy, social capital, science	Documents on economy, social capital, market, communities, and groups (e.g., immigrants, scientists), governance and policies, social, political, and cultural movements, all aspects analyzed from a social network perspective. [Documentos sobre economía, capital social, mercado, comunidades y grupos (emigrantes, científicos), gobernanza y políticas, movimientos sociales, políticos y culturales, analizados desde la perspectiva de las redes sociales.]
I	instant messaging, implement, application, base, service, platform, message, communicate, mobile, network, protocol, architecture, server, distribute, standard	Documents on text, voice, video and image processing and communication, protocols and mechanisms related to encryption, authentication, and access for mobile or wireless networks, and implementation of instant messaging and chat services. [Documentos sobre procesamiento y comunicación de texto, voz, vídeos e imágenes, protocolos y mecanismos de encriptación, autentificación y acceso en redes de móviles o sin cable, así como implementación de servicios de mensajería instantánea y de chat.]
J	online, course, learn, student, education, e-learn, social media, technology, virtual, teacher, academic, computer, skill, web, chat	Documents on e-learning, collaborative learning, and technology use in education, with special attention to virtual learning communities and the use of social media and social networks online in learning. [Documentos sobre aprendizaje en línea o virtual, aprendizaje colaborativo y la utilización de la tecnología en la educación, prestando especial atención a las comunidades virtuales de aprendizaje y al empleo de medios de comunicación social y de redes sociales en línea en la enseñanza.]

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

K	social network, online, social media, consume, use, privacy, behavior, factor, market, tag, weblog, tweet, game, student, public	Documents about social networks online (e.g., Facebook, Twitter) and social media (e.g., YouTube, Digg, Flickr), the user behavior and tagging patterns, with special emphasis on privacy issues and factors affecting consume and purchase decision related to brands accounts. [Documentos sobre redes sociales en línea –Facebook, Twitter, por ejemplo– y medios de comunicación social -como YouTube, Digg, Flickr-, el comportamiento del usuario y patrones de etiquetado, con énfasis en la privacidad y en los factores que afectan al consumo y decisión de compra de marcas comerciales.]
M	immigrate, migrate, culture, politics, local, community, network, economics, social, policy, Irish, refugee, China, ethnic, transnational	Documents about networks and communities of various kinds (immigrants, refugees, ethnic groups, organized crime or students), showing local cultural, political, economic or social aspects (in the host country) and those of the country of origin. [Documentos sobre redes y comunidades diversas (emigrantes, refugiados, grupos étnicos, crimen organizado o estudiantes), mostrando aspectos locales de la cultura, la política, la economía y de carácter social en el país de acogida frente a estos aspectos en el país de origen.]
T	Twitter, social media, discuss, politics, public, communicate, social, inform, engage, active, covid pandemic, pandemic, vaccine, practice, message	Documents on the use of social networks online and social media to disseminate political and religious propaganda, public information campaigns (e.g., gun control or health advice), and as a stage for protest movements (e.g., sexual harassment) and discussions on relevant news. [Documentos sobre la utilización de redes sociales en línea y medios de comunicación social para la propaganda política, religiosa o las campañas informativas de carácter público (control de armas o avisos sanitarios, por ejemplo), así como su empleo como plataformas para movimientos de protesta (sobre acoso sexual, por ejemplo) y como foros de discusión sobre noticias relevantes.]

Tabla 3: Términos obtenidos por Mallet y descripción textual de los tema
Fuente: Elaboración propia.

A fin de disponer de más información sobre las características de los temas, en especial para detectar posibles relaciones entre ellos, de ayuda en el análisis de la evolución temática, se examinó su estructura sometiéndolos a agrupación jerárquica. Siguiendo el procedimiento descrito en la sección anterior, se obtiene la estructura temática que muestra la *figura 3*.

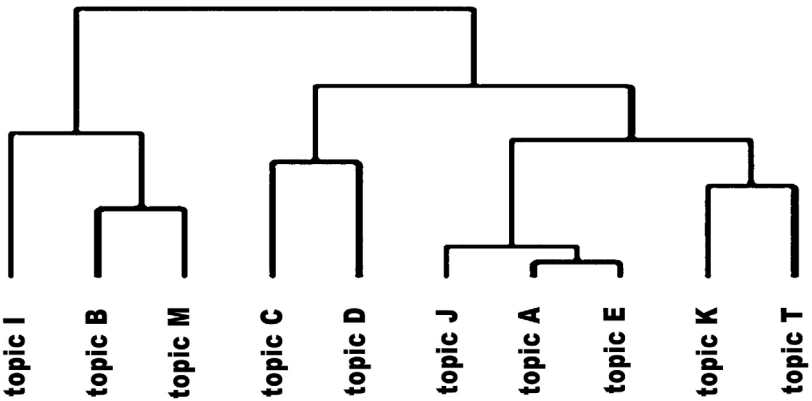

Figura 3. Estructura temática de los temas mediante agrupación jerárquica
Fuente: Elaboración propia.

En esta gráfica se observa una clara separación entre los temas B, M e I frente al resto. Si se advierte en el grupo compuesto por los de B y M, se ve que comparten varias características: son los temas con un menor número de temas anuales; a excepción del de T, los temas con una menor duración temporal; además, los temas con un menor valor de similaridad media. Como entre 1997 y 2000 –los cuatro primeros años– solo hay dos temas, y el de A tiene un marcado carácter sanitario, necesariamente el tema B aúna todos los aspectos restantes, incluyendo los primeros artículos sobre las primeras aplicaciones tanto de mensajería instantánea como de plataformas de redes sociales en línea. El *topic* M comparte este mismo carácter genérico, al acoger redes y comunidades de diversa naturaleza, desde bandas de crimen organizado hasta refugiados. En consecuencia, los temas B y M pueden considerarse caracterizados por reunir aquellos aspectos temáticos vinculados con redes sociales o comunidades relegados por el resto de los temas en las anualidades donde aparecen.

En cuanto al tema I, comparte con los de B y M esa misma amplitud temática, aunque no centrada en grupos o redes sociales, sino en aspectos técnicos relacionados con el tratamiento de cualquier tipo de dato (texto, voz o imagen) de cara a su transmisión en redes móviles o inalámbricas. En resumen, los temas B, M e I se distinguen por su gran amplitud temática, y cuya aparición se justifica durante aquellos años de transición en que el resto de los temas no pueden acoger ese tipo de contenidos.

Con relación a la popularidad temática, el cálculo de la probabilidad media de cada tema en todos los documentos de cada año muestra –en general– una gran variabilidad en la intensidad o popularidad relativa de todos los temas, indicativo

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

de un área sometida a frecuentes cambios por la aparición de ramificaciones y facetas temáticas novedosas. El test no paramétrico de tendencia Mann-Kendall muestra que los de A y E tienen una tendencia decreciente a lo largo de los 25 años del estudio, aunque solo la variación del tema A es estadísticamente significativa. Del mismo modo, los de C, D, J, K y T presentan una tendencia general creciente en popularidad a lo largo de los años, pero solamente los temas K y T con significación estadística.

Por tanto, todos los aspectos relativos a los medios de comunicación social y de redes sociales en línea, en particular el comportamiento del usuario, o su utilización en las esferas política, económica, sanitaria, religiosa o social como escenario para la discusión pública, han sido investigados con especial intensidad en los últimos años. En las *figuras 4*, *5* y *6* se incluyen las gráficas de los temas con intensidad estadísticamente significativa de carácter decreciente o creciente.

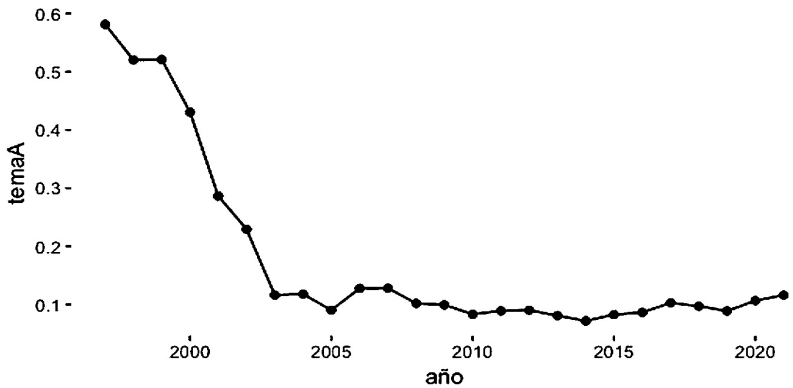

Figura 4. Evolución de la probabilidad media del tema A con intensidad decreciente
Fuente: Elaboración propia.

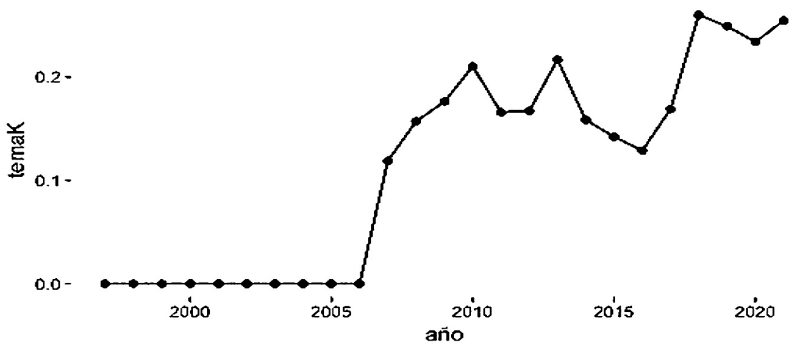

Figura 5. Evolución de la probabilidad media del tema K con intensidad creciente
Fuente: Elaboración propia.

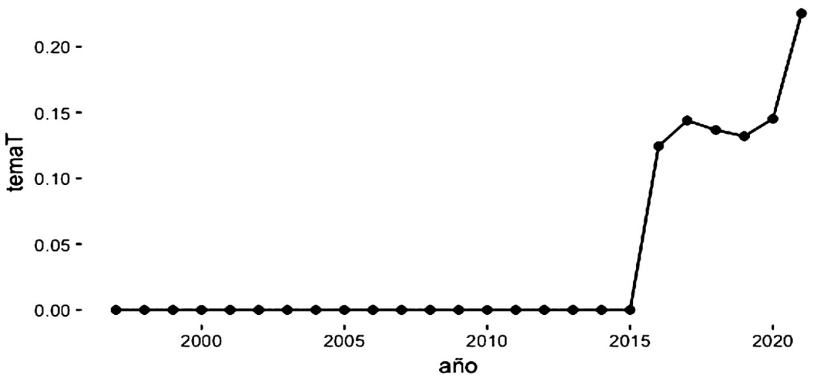

Figura 6. Evolución de la probabilidad media del tema T con intensidad creciente
Fuente: Elaboración propia.

El análisis sugiere la distinción de tres periodos o etapas en la investigación sobre redes sociales entre 1997 y 2021. En el primero (1997-2004), el estudio de las redes sociales como estructuras sociales basadas en el establecimiento de relaciones entre sus componentes (perspectiva sociológica) es la predominante hasta 2001 a través del tema A. Entre 2002 y 2004 se produce el progresivo ascenso y finalmente la preponderancia de los temas B y C, que incorporan nuevos enfoques de carácter más tecnológico, incluyendo las primeras aplicaciones de redes sociales en línea y medios de comunicación social. El tema A incide a lo largo de esta etapa inicial sobre todo en enfermedades psiquiátricas y de transmisión sexual, y en la labor que llevan a cabo las redes sociales compuestas por familia, amistades, compañeros o compañeras y parejas. El tema B aporta el estudio de comunidades virtuales de estudiantes así como de usuarios y de usuarias de sistemas y aplicaciones basados en la web, junto con la descripción de las primeras aplicaciones de mensajería instantánea (Keiretsu o Hubbub) y de medios de comunicación social (Twitter o Swisshouse). El tema C añade, en esta etapa, estudios técnicos sobre programas informáticos relacionados con sistemas móviles o interactivos (protocolos de Internet para comunicación móvil o sistemas interactivos basados en la web para el aprendizaje de idiomas).

En el segundo (2005-2012), desde la perspectiva sociológica, el tema E es el más abordado por los investigadores. Desde el enfoque que incorpora aspectos tecnológicos a los sociológicos, el tema D es el preponderante tras irrumpir con fuerza en esta etapa, con una popularidad en aumento hasta convertirse en el tema más popular desde 2010. En esta etapa el tema E analiza las redes sociales en relación con la economía en general (en la industria y en las empresas en particular), abordando también comunidades y grupos diversos (desde los étnicos

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

o criminales hasta los de trabajo o comunidades religiosas). Entre los aspectos que aborda el tema D en esta etapa destaca la aplicación del Análisis de Redes Sociales (SNA, por el término en inglés) a redes sociales de masas, la detección automática de relaciones y atributos en redes sociales (confianza o liderazgo, entre otros), así como la descripción y aspectos generales (comportamiento del usuario, etiquetado, detección de opiniones o temas) en medios de comunicación social y en redes sociales en línea (Last.fm, EventBrite, Facebook, Twitter, YouTube o Flickr, entre otros).

En el tercero (2013-2021) se consolida la disminución en la importancia relativa de los temas de enfoque sociológico frente a los de corte más tecnológico, tendencia que empezó a observarse a finales de la segunda etapa. El tema D se mantiene como el tema más popular hasta 2018, año en que cede esta primera posición a favor del tema K, posición que ostentará hasta el final de esta tercera fase. El tema K aborda principalmente el empleo de medios de comunicación social y de redes sociales en línea en diversos ámbitos como los medios de comunicación de masas (periódicos, televisión), los servicios públicos (bomberos, sanidad, administración local), la mercadotecnia y el consumo (imagen de marca, publicidad, persuasión del consumidor, intención de compra, comportamientos relacionados con la moda). Con menor frecuencia surgen aspectos generales relativos a las aplicaciones de mensajería instantánea (intención de continuación con el empleo de la aplicación, intención de abandono de esta, comportamiento al compartir información) y aspectos vinculados a la privacidad o la seguridad (suplantación de identidad, políticas de seguridad, malware). El tema T se especializa en la relación entre los medios de comunicación social o de las redes sociales en línea y el ámbito político (control ideológico, campañas electorales, identidad nacional), religioso (radicalización religiosa, actos de defensa del Islam), sanitario (desinformación durante la pandemia de Covid-19, campañas sanitarias), periodístico (difusión de noticias falsas, noticias sobre temas de actualidad) o policial (terrorismo, persecución de delitos). Con menor frecuencia también se abordan en este tema la utilización de medios de comunicación social y de redes sociales en línea en debates polémicos (anti y pro refugiados o migrantes), en movimientos de protesta o por el activismo (huelgas estudiantiles, violencia de género, discursos de minorías o activismo feminista).

Discusión y conclusiones

La literatura acerca de modelado de temas destaca, entre las múltiples herramientas disponibles, tres que incluyen tanto la detección de temas como su evolución temporal: Hierarchical Topic Evolution Model (HTEM), que organiza de manera

jerárquica los temas obtenidos y ofrece su evolución temporal (Song *et al.* 2016); Topic Over Time (TOT), que añade el factor tiempo al descubrimiento de los temas en un corpus (Wang y McCallum 2006); y Dynamic Topic Model (DTM), que muestra la evolución del contenido de los temas extraídos del análisis (Blei y Lafferty 2006). La revisión de estas técnicas destaca dos inconvenientes: se puede observar la evolución de cada tema a lo largo del tiempo, lo que mejora la percepción de sus características, pero no es posible descubrir la aparición o desaparición de nuevos temas a los inicialmente obtenidos (Blei y Lafferty 2006); además, ignoran las relaciones existentes entre los temas y la modificación temporal de tales vinculaciones entre temas, como la unión o la división de estos a lo largo del tiempo (Song *et al.* 2016).

Lo anterior ha provocado que el modelo Asignación Latente de Dirichlet (LDA) atraiga cada vez más adeptos y sea la técnica utilizada más ampliamente en la actualidad, aunque no incorpora el análisis temporal (Ballester y Penner 2022). Por estos motivos, además de por su facilidad de empleo y por el hecho de aportar resultados más interpretables humanamente que otros modelos (permitiendo el etiquetado de los temas, como se ha hecho aquí), se ha decidido su utilización en el presente estudio, tratando al tiempo de que sirva para evaluar la pertinencia o utilidad de estas herramientas en determinadas circunstancias en análisis propios de nuestra disciplina.

El descubrimiento de las relaciones entre los temas se basa en técnicas automáticas de agrupamiento (*clustering* en inglés), empleando para ello medidas matemáticas de similaridad. Entre estas destacan dos: symmetric Kullback-Leibler divergence, la más frecuente entre los programas de modelado de temas que incorporan esta faceta (Zhou *et al.* 2017); y la similaridad coseno, habitual en estudios bibliométricos, en estudios basados de programas de modelado como el presente que no incluyen el agrupamiento, y en el campo de la recuperación de información (Zhang *et al.* 2018; Leydesdorff 2008; Cacheda Seijo *et al.* 2011).

En este estudio no se emplea, sin embargo, la fórmula clásica de TF e IDF para calcular los pesos que figuran en la fórmula de la similaridad coseno. Se introduce una variante novedosa, apropiada cuando los documentos tratados (los resúmenes de los textos) presentan longitudes muy distintas, como sucede en la colección utilizada aquí. Con ello se evitan sesgos debido a que los documentos más largos tienden a contener mayor número de términos y con mayor frecuencia que los textos cortos, alterando al alza los valores de similaridad que presentan con otros documentos.

Por tanto, en el estudio se sigue una de las aproximaciones más recientes mediante el empleo de LDA y la detección posterior de los temas a raíz del agrupamiento de los temas detectados en las ventanas temporales seleccionadas (temas anuales en este caso). Una diferencia, sin embargo, se ha introducido en

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

este análisis. El procedimiento habitual consiste en efectuar LDA con todo el corpus en su integridad, y mediante similaridad coseno descubrir la modificación, unión o división de los temas globales (correspondientes a la colección completa de documentos) mediante el incremento o disminución del número de temas anuales conectados al mismo tema global (Chen *et al*. 2017).

En el presente estudio se introduce un procedimiento novedoso, que no exige la selección previa del número de temas considerados. Se calcula el número óptimo de temas en cada periodo de un año, y mediante similaridad coseno entre los temas anuales se descubre el número de temas que deben considerarse en la totalidad del periodo analizado, además de su evolución temporal. Sea cual sea el procedimiento seguido, la existencia de una correlación fuerte entre los temas involucrados se detecta cuando la similaridad entre temas traspasa un valor umbral (Gaul y Vincent 2017), dependiente en primera instancia del área de conocimiento analizado. En el caso de las redes sociales se obtiene un valor límite o umbral de 0,76, muy semejante al valor 0,75 obtenido en el campo de la recuperación de información (Chen *et al*. 2017), pero distante del valor 0,85 detectado al analizar modelos mentales (Ma *et al*. 2023).

Otro aspecto novedoso del presente estudio radica en el procedimiento de selección del número de temas en cada año de manera independiente. La *perplexity* consituye la métrica más utilizada para la elección del número de temas en un corpus, que indica el grado de incertidumbre sobre la pertenencia de un documento a un tema. Se pone de manifiesto, sin embargo, que esta medida es adecuada siempre que el número de temas se sitúe entre 80 y 100 (Chen *et al*. 2017). Este resulta un número demasiado grande en nuestro caso, porque se pretende en principio descubrir amplios temas de carácter general. Por este motivo se opta por aplicar los criterios y medidas sugeridos por el programa Mallet empleado en el estudio: la coherencia (*coherence*), la exclusividad (*exclusivity*), el parámetro 'alpha' y la distribución lo más homogénea posible del tema en el corpus correspondiente. Este modo de proceder es mucho más laborioso, pero permite detectar un número de temas variable por año, el más adecuado en cada una de estas ventanas temporales. De este modo se facilita el análisis de la evolución temática detallada en un área con grandes cambios en el periodo considerado.

En relación con el preprocesamiento de los documentos, se siguen aquí las fases habituales, con una única excepción relativa a la extracción de bigramas o trigramas. Suele realizarse habitualmente la detección de bigramas o trigramas en todos los documentos del corpus (Buehling 2021), mientras que en el presente estudio se opta por emplear las palabras clave compuestas de dos o más palabras seleccionadas por los autores en la descripción de los documentos de la colección. De esta manera se consigue extraer no solo términos relevantes de naturaleza temática, sino también entidades geográficas, personales o temporales de interés.

En este trabajo se analizan las investigaciones llevadas a cabo sobre cualquier tipo de relación establecida entre los miembros de un grupo, sean cuales sean sus objetivos y su naturaleza (humana o no), entre 1997 y 2021. En consecuencia, se afronta el estudio del área de conocimiento de las redes sociales en su integridad, no solo el de las plataformas de redes sociales en línea como Twitter o Facebook, temática no abordada previamente –hasta donde conocemos–. Se han localizado escasos estudios generales sobre la historia de las redes sociales en línea (Boyd y Ellison 2008), así como diversos análisis de la evolución del empleo de las redes sociales en línea en ámbitos concretos como el sanitario (Cho *et al*. 2020), el histórico (Bunnenberg *et al*. 2021), el económico (Agarwal *et al*. 2021) o el educativo (Yu *et al*. 2023), pero ningún análisis comparable al que se realiza aquí.

Las contribuciones de este enfoque se resumen en los siguientes aspectos principales: 1) Incardinar la investigación en redes sociales en línea en su área de conocimiento originario, lo que aporta al área de las Ciencias de la Información una perspectiva teórica más amplia que enriquece los métodos empleados; 2) Evaluar la utilidad –cuando las características del tema lo demanda– de ciertas herramientas informáticas en estudios propios de la disciplina; y 3) Aportar elementos de juicio más ajustados a la realidad sobre el verdadero alcance e importancia relativa del estudio de las redes sociales en línea frente a cualquier otra temática relacionada con las redes sociales en general, desde su aparición en 1997 hasta la actualidad.

Los resultados muestran la pertinencia de distinguir tres fases en la investigación sobre las redes sociales en su integridad: 1997-2004, 2005-2012 y 2013-2021. Los estudios consultados sobre la evolución de las plataformas de redes sociales y medios de comunicación social refuerzan esta distinción, al considerar que a partir de 2003-2004 se inicia una nueva fase en la evolución de las redes sociales en línea (Boyd y Ellison 2008), que a partir de 2013 se distingue una nueva etapa en la literatura sobre medios de comunicación social (Shibuya *et al*. 2022), o que en 2012 acaba un periodo y en 2013 se inicia otro nuevo en la investigación sobre medios de comunicación social (Gálvez 2019).

En relación con el primer periodo (1997-2004), Boyd y Ellison (2008) constatan que entre 1997 y 2001 surgen las primeras aplicaciones de redes sociales en línea de carácter personal, profesional y social; y que en 2001 se inicia la preponderancia de redes profesionales y el carácter económico de las mismas. Estas características coinciden en buena medida con los términos extraídos por el programa Mallet para el tema B (*develop*, *community*, *network*, *work*, *economics*) y con la descripción textual a partir del análisis cualitativo del corpus («Documentos sobre diversas redes en línea –por ejemplo estudiantes, trabajadores–) […] anunciando el desarrollo de las primeras aplicaciones de mensajería instantánea y de redes sociales en línea»). Al tiempo, el análisis desde la perspectiva de las redes sociales

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

en general indica que los aspectos sanitarios son los más populares en esta etapa, y que los aspectos tecnológicos –entre los que se incluyen las primeras aplicaciones de redes sociales en línea, pero que también abarca la investigación sobre sistemas móviles, por ejemplo– presentan un progresivo ascenso en esta fase, mostrando que las redes sociales en línea representa una temática en desarrollo que necesita de un paralelo avance en las herramientas informáticas que las hacen posible.

En el segundo (2005-2012) se destaca la irrupción de las tecnologías y de las aplicaciones de redes sociales en línea (Gálvez 2019). Estas características se reflejan en el surgimiento del tema I entre 2003 y 2012, que desde un enfoque general engloba los aspectos más técnicos del procesamiento de texto, voz e imágenes en móviles y la implementación de servicios de mensajería instantánea y aplicaciones de chat. Por su parte, estudiando la evolución de la mercadotecnia, Sharma y Verma (2018) destacan entre 2001 y 2013 el auge de investigaciones relativas a los usuarios de medios de comunicación social como herramienta de mejora de la relación entre las marcas y sus consumidores. Este aspecto forma parte de los contenidos del tema K, que pone el acento en los factores que afectan al consumo y la decisión de compra de marcas comerciales por los usuarios de redes sociales en línea. El enfoque general que se adpota aquí enriquece el alcance de estas características al mostrar que en esta etapa el tema E (que aborda las redes sociales en relación con la economía en su integridad) resulta más popular, y que el tema K presenta un progresivo auge durante estos años, incorporando el estudio de aspectos como los estados emocionales de los consumidores o la credibilidad de las valoraciones. Incluso desde el punto de vista tecnológico el tema I no es el preponderante en la investigación en estos años, sino el tema D, que pone el acento en la detección automática de relaciones y atributos en redes sociales, así como la descripción del comportamiento del usuario –etiquetado, opiniones– en medios de comunicación social y en redes sociales en línea.

En relación con el tercero (2013-2021), se advierte entre 2013 y 2017 un incremento notable de estudios sobre la conversión de las redes sociales en línea en herramientas influyentes en la política y en instrumentos de movimientos sociales (Gálvez 2019). Estos aspectos están incorporados al tema T, pues en él se aborda la utilización de los medios de comunicación social en la propaganda política, en movimientos de protesta y en foros de discusión pública. También se destaca que entre 2013 y 2016 surgen con fuerza los estudios sobre aspectos económicos relacionados con las redes sociales (Shibuya *et al.* 2022). El enfoque general que se adopta aporta matices de interés, pues el tema E en estos años se interesa –entre otros aspectos– por las características de las plataformas de redes sociales que son de utilidad económica (como la exploración de información social que facilite las recomendaciones del sistema a los usuarios o la reducción de

rumores negativos). En cuanto al tema T, durante estos años también aborda el empleo de las redes sociales en línea en los ámbitos religioso, sanitario, periodístico o policial.

En resumen, aspectos reseñados en estudios anteriores tienen su correlato en los temas que se analizan aquí, aunque en el presente trabajo se incluyen otros temas no reseñados con anterioridad, además de aspectos no descritos en los temas involucrados, lo que enriquece el análisis y aporta una visión más detallada y completa de la investigación durante estos años.

En conclusión, el estudio muestra que es posible abordar la evolución de la investigación en un campo en permanente transformación durante un amplio lapso y con un volumen elevado de datos, como es el caso de las redes sociales, mediante modelización de temas. Los problemas derivados en cuanto a precisión y coherencia de los resultados se superan mediante un procesamiento más extenso de los datos de entrada. Así, para alcanzar una descripción detallada, se disminuye la amplitud de las ventanas temporales, reducidas a un año en nuestro caso, seleccionando el número de temas óptimo para cada una de dichas ventanas. Ello conlleva, a su vez, la necesidad de someter estos datos fragmentarios a procesamientos adicionales para descubrir las conexiones entre dichas temáticas parciales, aportando aspectos poco frecuentes en el tradicional análisis de modelización de temas.

El análisis muestra la existencia de siete grandes temas en el campo de las redes sociales, que se describen sucintamente como los aspectos: sanitarios, económicos, educativos, técnicos (seguridad, almacenamiento, comunicación o procesamiento de la información), sociales (influencia, relaciones entre los componentes de la red, intereses de los miembros), redes sociales en línea (comportamiento de los usuarios, privacidad, patrones de etiquetado) y la utilización de las redes sociales en línea como medio de comunicación social de ideas políticas o religiosas. En cuanto a las fases que pueden discernirse en estos veinticinco años, el análisis sugiere la distinción de tres periodos o etapas en la investigación sobre redes sociales entre 1997 y 2021, cuyas características se describen. Cabe enfatizar que en el tercer periodo (2013-2021) el estudio de las redes sociales en línea (tema K) y su utilización en distintos ámbitos como el político o el religioso (tema T) representan los aspectos investigados con especial intensidad.

Referencias

Agarwal, T., Arya, S. y Bhasin, K. (2021). The evolution of internal employer branding and employee engagement: The temporal role of internal social media usage. *Journal of Information and Knowledge Management*, 20(1), 2150012. https://doi.org/10.1142/S021964922150012X

Aichner, T. y Jacob, F. H. (2015). Measuring the degree of corporate social media use. *International Journal of Market Research,* 57(2), 257-275. https://doi.org/10.2501/IJMR-2015-018

Alias-I (2016). *LingPipe Home*. Alias-i.
http://www.alias-i.com/lingpipe/

Apache Software Foundation (2022, 25 de octubre). *Welcome to Apache Lucene*. ASF.
https://lucene.apache.org

Armann-Keown, V. Y Patterson, L. (2020). Content analysis in library and information research: An analysis of trends. *Library & Information Science Research*, 42(4), art. 101048.
https://doi.org/10.1016/j.lisr.2020.101048

Arruda, H. F., Costa, L. F. y Amancio, D. R. (2016). Topic segmentation via community detection in complex networks. *Chaos (Woodbury, N.Y.)*, 26(6), 063120.
http://dx.doi.org/10.1063/1.4954215

Ballester, O. y Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics*, 16 (1), 101224.
https://doi.org/10.1016/j.joi.2021.101224

Banerjee, A. y Basu, S. (2007). Topic Models over Text Streams: A Study of Batch and Online Unsupervised Learning. En *Proceedings of the seventh SIAM international conference on Data Mining* (pp. 431-436).
https://doi.org/10.1137/1.9781611972771.40

Berkowitz, S. D. (1982). *An introduction to structural analysis: The network approach to social research*. Butterworths.

Blei, D. M. y Lafferty, J. D. (2006). Dynamic topic models. En *Proceedings of the 23rd International Conference on Machine Learning* (ICML '06, pp. 113-120). ACM Press.
https://doi.org/10.1145/1143844.1143859

Blei, D. M., Ng, A. Y. y Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5), 993-1022.
https://jmlr.org/papers/volume3/blei03a/blei03a.pdf

Boyd, D. M. y Ellison, N. B. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210-230.
https://doi.org/10.1111/j.1083-6101.2007.00393.x

Brigadir, I. (2022, 25 de octubre). *Default English stopwords lists from many different sources*. Github.
https://github.com/igorbrigadir/stopwords

Buehling, K. (2021). Changing research topic trends as an effect of publication ranking: The case of German economists and the Handelsblatt Ranking. Journal of *Informetrics*, 15(3), 101199.
https://doi.org/10.1016/j.joi.2021.101199

Bunnenberg, C., Logge, T. y Steffen, N. (2021). Social Media History. *Historische Anthropologie*, 29(2), 267-283.
https://doi.org/10.7788/hian.2021.29.2.267

Cacheda Seijo, F., Fernández Luna, J. M. y Huete Guadix, J. F. (coords.) (2011). *Recuperación de información: un enfoque práctico y multidisciplinar*. Ra-Ma.

Chabowski, B. R. y Samiee, S. (2023). A bibliometric examination of the literature on emerging market MNEs as the basis for future research. *Journal of Business Research*, 155, art. 113263.
https://doi.org/10.1016/j.jbusres.2022.08.027

Chang, Y.-W., Huang, M.-H. y Lin, C.-W. (2015). Evolution of research subjects in Library and Information Science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics*, 105(3), 2071-2087.
https://doi.org/10.1007/s11192-015-1762-8.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

Chen, B., Tsutsui, S., Ding, Y. y Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. Journal of *Informetrics*, 11(4), 1175-1189. https://doi.org/10.1016/j.joi.2017.10.003

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377. https://doi.org/10.1002/asi.20317

Cho, S. M., Park, C. y Song, M. (2020). The evolution of social health research topics: A data-driven analysis. *Social Science & Medicine*, 265, 113299. https://doi.org/10.1016/j.socscimed.2020.113299

Chodera, J. D. y Pande, V. S. (2011). The social network (of protein conformations). *Proceedings of the National Academy of Sciences of the United States of America*, 108(32), 12969-12970. https://doi.org/10.1073/pnas.1109571108

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. y Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Ding, W. y Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. *Journal of the Association for Information Science and Technology*, 65(10), 2084-2097. https://doi.org/10.1002/asi.23134

Ding, Y., Chowdhury, G. y Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing y Management*, 37(6), 817-842. https://doi.org/10.1016/S0306-4573(00)00051-0.

Ferrer, R., Solé, R. V. y Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review* E, 69(5), 051915. http://dx.doi.org/10.1103/PhysRevE.69.051915.

Freeman, L. C. (1988). Computer programs in social network analysis. *Connections*, 11(2), 26-31. https://www.researchgate.net/publication/239060164_Computer_Programs_and_Social_Network_Analysis

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science*. Empirical Press. https://www.researchgate.net/publication/238341375_The_Development_of_Social_Network_Analysis_A_Study_in_the_Sociology_of_Science

Gálvez, C. (2019). Evolución del campo de investigación de los social media mediante mapas de la ciencia (2008-2017). *Communication & Society*, 32(2), 61-76. https://doi.org/10.15581/003.32.2.61-76

Gaul, W. y Vincent, D. (2017). Evaluation of the evolution of relationships between topics over time. *Advances in Data Analysis and Classification*, 11, 159-178. https://doi.org/10.1007/s11634-016-0241-2

Gore, D. J., Schueler, K., Ramani, S., Uvin, A., Phillips, G., McNulty, M., Fujimoto, K. y Schneider, J. (2021). HIV response interventions that integrate HIV molecular cluster and social network analysis: A systematic review. *AIDS and Behavior*, 26(6), 1750-1792. https://doi.org/10.1007/s10461-021-03525-0

Graham, S., Weingart, S. y Milligan, I. (2021, 3 de septiembre). *Getting started with topic modeling and Mallet*. https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Griffiths, T., Steyvers, M. (2004). Finding scientifics topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(suppl. 1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Ha, I., Park, H. y Kim, C. (2014). Analysis of Twitter research trends based on SLR. En *16th International Conference on Advanced Communication Technology* (pp. 774-778). IEEE. https://doi.org/10.1109/ICACT.2014.6779067

Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model. *Scientometrics*, 125(3), 2561-2595. https://doi.org/10.1007/s11192-020-03721-0

Harary, F. (1969). *The Graph Theory*. Addison-Wesley Publishing Company.

Heider, F. (1946). Attitudes and cognitive organization. *The Journal of Psychology*, 21, 107-112. https://doi.org/10.1080/00223980.1946.9917275

Hofmann, T. (1999). Probabilistic latent semantic indexing. En *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57). Association for Computing Machinery. https://doi.org/10.1145/312624.312649

Isasi, J. (2022, 15 de noviembre). *Modelado de temas con Mallet*. https://repositories.lib.utexas.edu/handle/2152/72737

Jeong, D. H. y Min, S. (2014). Time gap analysis by the topic model-based temporal technique. *Journal of Informetrics*, 8(3), 776-790. https://doi.org/10.1016/j.joi.2014.07.005

Johnson, B. J. (2019, 23 de febrero). *Contar todas las palabras diferentes en un archivo de texto*. LWP, lawebdelprogramador. https://www.lawebdelprogramador.com/foros/Java/1685229-Contar-todas-las-palabras-diferentes-en-un-archivo-de-texto.html

Jung, S. y Yoon, W. C. (2020). An alternative topic model based on Common Interest Authors for topic evolution analysis. *Journal of Informetrics*, 14(3), 101040. https://doi.org/10.1016/j.joi.2020.101040

Kai, H., Qing, L., Kunlun, Q., Siluo, Y., Jin, M., Xiaokang, F., Jie, Z., Huayi, W., Ya, G. y Qibing, Z. (2019). Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. *Information Processing and Management*, 56(4), 1185-1203. https://doi.org/10.1016/j.ipm.2019.02.014

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. *Data Mining y Knowledge Discovery*, 7(4), 373-397. https://doi.org/10.1145/775047.775061

Knoke, D. y Kuklinski, J. H. (1982). *Network analysis*. Sage.

Kumar, K. (2018, 3 de mayo). *Evaluation of topic modeling: Topic coherence*. https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/

Landauer, T. K., McNamara, D. S., Dennis, S. y Kintsch, W. (2007). *Handbook of latent semantic analysis*. Taylor y Francis Group. https://doi.org/10.4324/9780203936399

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

Leydesdorff, L. (2007). On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1), 77-85. https://doi.org/10.1002/asi.20732

Li, X. y Lei, L. (2021). A bibliometric analysis of topic modelling studies (2000-2017). *Journal of Information Science*, 47(2), 161-175. https://doi.org/10.1177/0165551519877049

Li, Y., Wu, Y. y Chen, Y. (2021). A review of enterprise social media: visualization of landscape and evolution. *Internet Research*, 31(4), 1203-1235. https://doi.org/10.1108/INTR-07-2020-0389.

Liu, H., Chen, Z., Tang, J., Zhou, Y. y Liu, S. (2020). Mapping the technology evolution path: a novel model for dynamic topic detection and tracking. *Scientometrics*, 125(3), 2043-2090. https://doi.org/10.1007/s11192-020-03700-5

Ma, J., Wang, L., Zhang, Y.-R., Yuan, W. y Guo, W. (2023). An integrated latent Dirichlet allocation and Word2vec method for generating the topic evolution of mental models from global to local. *Expert Systems With Applications*, 212, 118695. https://doi.org/10.1016/j.eswa.2022.118695

Mallet (2022, 12 de julio). *Importing data Mallet*. Mallet. https://mimno.github.io/Mallet/import.html

McCallum, A. K. (2022a, 15 de noviembre). *MALLET: A Machine Learning for Language Toolkit*. https://mallet.cs.umass.edu/index.php/Main_Page

McCallum, A. K. (2022b,15 de noviembre). *Topic model diagnostics*. https://mallet.cs.umass.edu/diagnostics.php

McCallum, A., Wang, X. y Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on Enron and academic email. *Journal of Artificial Intelligence Research*, 30, 249-272.

McCandless, M., Hatcher, E. y Gospodnetic, O. (2010). *Lucene in action*. Manning.

Mikolov, T., Chen, K., Corrado, G. y Dean, J. (2013). Efficient estimation of word representations invector space. *arXiv*, 1301.3781 [cs.CL]. https://doi.org/10.48550/arXiv.1301.3781

Moreno, J. L. (1937). Inter-personal therapy and the psychopathology of inter-personal relations. *Sociometry*, 1(1-2), 9-76. https://doi.org/10.2307/2785258

Moreno, J. L., Jennings, H. H. (1938). Statistics of social configurations. *Sociometry*, 1(3-4), 342-373. https://doi.org/10.2307/2785588

NLTK (2022, 25 de octubre). *Natural Language Toolkit*. NLTK. https://www.nltk.org

Onyancha, O. B. (2018). Forty-five years of LIS research evolution, 1971-2015: An informetrics study of the author-supplied keywords. *Publishing Research Quarterly*, 34(3), 456-470. https://doi.org/10.1007/s12109-018-9590-3

OpenNLP (2022, 5 de junio). *SnowballStemmer (Apache OpenNLP Tools 1.8.0 API)*. OpenNLP Tools. https://opennlp.apache.org/docs/1.8.0/apidocs/opennlp-tools/opennlp/tools/stemmer/snowball/SnowballStemmer.html

Otte, E. y Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6), 441-453. https://doi.org/10.1177/016555150202800601

Pappi, F. U. y Stelck, K. (1987). Ein Databanksystem zur Netzwerkanalyse. En Pappi, F. U. (ed.), *Methoden Netzwerkanalyse* (1st ed., pp. 253-265). Oldenberg.

Peset, F., Garzón-Farinos, F., González, L. M. *et al.* (2020). Survival analysis of author keywords: An application to the library and information sciences area. *Journal of the Association for Information Science and Technology*, 71(4), 462-473. https://doi.org/10.1002/asi.24248

Pohlert, T. (2022, 26 de marzo). *Non-parametric trend tests and change-point detection*. R project. https://cran.r-project.org/web/packages/trend/vignettes/trend.pdf

R Core Team (2022, 15 de noviembre). *Hclust function: Hierarchical Clustering.* https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust

Ricci, R. (2018). Movimentos e mobilizações sociais no Brasil: de 2013 aos dias atuais. *Saúde em Debate*, 42, 90-107. https://doi.org/10.1590/0103-11042018S308.

Ridings, C. M., Gefen, D. y Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *The Journal of Strategic Information Systems*, 11(3-4), 271-295. https://doi.org/10.1016/s0963-8687(02)00021-15

Shan, B. y Li, F. (2010). A survey of topic evolution based on LDA. *Journal of Chinese Information Processing*, 24(6), 43-50.

Sharma, S. y Verma, H. V. (2018). Social media marketing: Evolution and change. En G. Heggde y G. Shainesh (eds.). *Social Media Marketing: Emerging Concepts and Applications*, pp. 19-36. Springer.

Shen, X. y Wang, L. (2020). Topic evolution and emerging topic analysis based on open source software. *Journal of Data and Information Science*, 5(4), 126-136. https://doi.org/10.2478/jdis-2020-0033

Shibuya, Y., Hamm, A. y Pargman, T. C. (2022). Mapping HCI research methods for studying social media interaction: A systematic literature review. *Computers in Human Behavior*, 129, 107131. https://doi.org/10.1016/j.chb.2021.107131

Silge, J. (2018, 8 de septiembre). *Training, evaluating, and interpreting topic models*. https://juliasilge.com/blog/evaluating-stm/

Singhal, A. (2001). Modern information retrieval: a brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4), 35-43. http://www1.cs.columbia.edu/~gravano/Qual/Papers/singhal.pdf

Snowball (2021, 25 de octubre). *Snowball*. Snowball.
https://snowballstem.org

Song, M., Heo, G. E. y Kim, S. Y. (2014). Analyzing topic evolution in bioinformatics: Investigation of dynamics of the field with conference data in DBLP. *Scientometrics*, 101(1), 397-428. https://doi.org/10.1007/s11192-014-1246-2

Song, J., Huang, Y., Qi, Y., Li, Y., Li, F., Fu, K. y Huang, T. (2016). Discovering Hierarchical Topic Evolution in Time-Stamped Documents. *Journal of the Association for Information Science and Technology*, 67(4), 915-927. https://doi.org/10.1002/asi.23439

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

Statista, We Are Social, Hootsuite y DataReportal (2022, 26 de enero). *Most popular social networks worldwide as of January 2022, ranked by number of monthly active users (in millions).* https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Sueur, C. y Pelé, M. (2016). Social network and decision-making in primates: A report on Franco-Japanese research collaborations. *Primates*, 57(3), 327-332. https://doi.org/10.1007/s10329-015-0505-z

Suominen, A. y Toivanen, H. (2015). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, 67(10), 2464-2476. https://doi.org/10.1002/asi.23596.

Taipale, S. y Farinosi, M. (2018). The big meaning of small messages: The use of WhatsApp in intergenerational family communication. En J. Zhou, J. y Salvendy, G. (eds.), *Human aspects of IT for the aged population. Acceptance, Communication and Participation* (pp. 532-546). Springer. https://doi.org/10.1007/978-3-319-92034-4_40

Tdk Technologies (2020, 12 de noviembre). *Topic modeling explained: LDA to Bayesian Inference.* https://www.tdktech.com/tech-talks/topic-modeling-explained-lda-to-bayesian-inference/

Tuomaala, O., Järvelin, K. y Vakkari, P. (2014). Evolution of library and information science, 1965-2005: Content analysis of journal articles. *Journal of the Association for Information Science and Technology*, 65(7), 1446-1462. https://doi.org/10.1002/asi.23034

Wang, G. y Robinson, R. (2002). An architecture for web-enabled engineering applications based on lightweight high-performance CORBA. En Williams, A. D. (ed.), *Proceedings of the 6th International Enterprise Distributed Object Computing Conference* (pp. 249-257). IEEE Computer Society. https://doi.org/10.1109/EDOC.2002.1137714

Wang, X. y McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. En *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '06, pp. 424-433). ACM Press https://doi.org/10.1145/1150402.1150450

Wasserman, S. y Faust, K. (1994). Social network analysis in the social and behavioral sciences. En *Social network analysis: Methods and applications* (Structural Analysis in the Social Sciences, pp. 3-27). Cambridge University Press. https://doi.org/10.1017/CB09780511815487.002

Wu, Q., Zhang, C., Hong, Q. y Chen, L. (2014). Topic evolution based on LDA and HMM and its application in stem cell research. *Journal of Information Science*, 40(5), 611-620. https://doi.org/10.1177/0165551514540565

Xu, S., Hao, L., An, X., Yang, G. y Wang, F. (2019). Emerging research topics detection with multiple machine learning models. Journal of *Informetrics*, 13(4), 100983. https://doi.org/10.1016/j.joi.2019.100983

Yang, C., Tang, X., Kim, S. Y. y Song, M. (2012). A trend analysis of domain-specific literatures with content and co-author network similarity. En Chen, H. H. y Chowdhury, G. (eds.), *The 14th International Conference on Asia-Pacific Digital Libraries* (ICADL 2012, pp. 73-76). Springer. https://doi.org/10.1007/978-3-642-34752-8_10

Yanhui, S., Lijuan, W. y Junping, Q. (2021). A comparative study of first and all-author bibliographic coupling analysis based on Scientometrics. *Scientometrics*, 126(2), 1125-1147. https://doi.org/10.1007/s11192-020-03798-7

Yau, C. K., Porter, A., Newman, N. y Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, 100(3), 767-786. https://doi.org/10.1007/s11192-014-1321-8

Yu, Z., Sukjairungwattana, P. y Xu, W. (2023). Bibliometric analyses of social media for educational purposes over four decades. *Frontiers in Psychology*, 13, 1061989. https://doi.org/10.3389/fpsyg.2022.1061989

Zanardo, N., Parra, G. J., Diaz-Aguirre, F., Pratt, E. A. L. y Möller, L. M. (2018). Social cohesion and intra-population community structure in southern Australian bottlenose dolphins. *Behavioral Ecology and Sociobiology*, 72(9), 1-13. https://doi.org/10.1007/s00265-018-2557-8

Zhang, J., Chen, H., Chan, H. C. B. y Leung, V. C. M. (2009). PUCS: Personal unified communications over heterogeneous wireless networks. En Ramasubramanian, S. y Aracil-Rico, J. (eds.), *Proceedings of the 2009 6th International Conference on Broadband Communications, Networks and Systems*, BROADNETS 2009 (article number 5336353). https://doi.org/10.4108/ICST.BROADNETS2009.7851

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H. y Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics*, 12(4), 1099-1117. https://doi.org/10.1016/j.joi.2018.09.004

Zhou, H., Yu, H. y Hu, R. (2017). Topic evolution based on the probabilistic topic model: a review. *Frontiers of Computer Science*, 11(5), 786-802. https://doi.org/10.1007/s11704-016-5442-5

Zhu, M., Zhang, X. y Wang, H. (2016). A LDA based model for topic evolution: Evidence from Information Science journals. En *Proceedings of 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications* (pp. 49-54). Atlantis Press. https://doi.org/10.2991/msota-16.2016.12

Zou, C. (2018). Analyzing research trends on drug safety using topic modeling. *Expert opinion on drug safety*, 17(6), 629-636. https://doi.org/10.1080/14740338.2018.1458838

Para citar este texto:

Martínez-Comeche, Juan-Antonio. 2023. "Veinticinco años de investigación en redes sociales: evolución de temas entre 1997 y 2021 empleando el algoritmo Asignación Latente de Dirichlet". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 37 (96): 145-177.

http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2023.96.58777