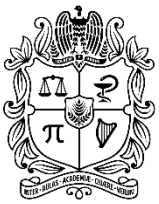

UNIVERSIDAD
NACIONAL
DE COLOMBIA

# Aplicación de técnicas de detección de temáticas emergentes para establecer prioridades de Investigación y Desarrollo en el área de Machine Learning

**Valentina Vásquez Hernández**

Universidad Nacional de Colombia

Facultad de Minas, Área Curricular de sistemas e Informática

Medellín, Colombia

2023

# Aplicación de técnicas de detección de temáticas emergentes para establecer prioridades de Investigación y Desarrollo en el área de Machine Learning

**Valentina Vásquez Hernández**

Tesis de Maestría presentada como requisito parcial para optar al título de:

**Magíster en Ingeniería – Analítica**

Director:

Ph.D., Juan David Velásquez Henao

Línea de Investigación:

Métodos computaciones para el análisis de datos

Grupo de Investigación:

Big Data y Data Analytics

Universidad Nacional de Colombia

Facultad de Minas, Área Curricular de Sistemas e Informática

Medellín, Colombia

2023

*Dedicatoria*

*A mi hermana, mi mamá y mis amigos,*
*Por su apoyo incondicional y desinteresado en cada paso.*

*A mi papá,*
*Quien falleció durante el desarrollo de este documento y nunca dejó de creer en que puedo lograr todo lo que me proponga.*

*A Tati,*
*Por su paciencia infinita.*

# Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

_____

Nombre

Fecha 28/05/2023

# Agradecimientos

Agradezco al profesor Juan David Velásquez, director de este trabajo de grado, quien siempre ha estado dispuesto a compartir conmigo incondicionalmente su conocimiento, experiencia y entusiasmo por tener un aprendizaje continuo y profundo. También le agradezco por haber acompañado mi trayectoria académica y profesional con mucho compromiso.

De igual manera a la Universidad Nacional de Colombia por brindarme la posibilidad de una vez más rodearme de infinito conocimiento por parte de mis compañeros y profesores.

# Resumen

**Aplicación de técnicas de detección de temáticas emergentes para establecer prioridades de Investigación y Desarrollo en el área de Machine Learning.**

La detección de temáticas emergentes es de gran relevancia para los equipos que se dedican a la Investigación y Desarrollo en Machine Learning, ya que les permite formular proyectos de investigación, generar nuevas oportunidades de negocio y aportar valor en términos de producto, tecnología y conocimiento. Sin embargo, estos equipos se enfrentan a varios obstáculos, como la rápida generación de información, la diversidad de fuentes de datos disponibles y la falta de implementaciones no comerciales escalables que permitan automatizar el análisis.

Con el objetivo de abordar esta necesidad, se propone una discusión sobre las metodologías actuales para la detección de temáticas emergentes. Además, se presenta una propuesta metodológica que combina tres técnicas de procesamiento de lenguaje natural: el Clasificador de Ontología de Ciencias de la Computación, los mapas temáticos y BERTopic. Una vez establecida esta metodología, se aplica a textos de artículos científicos en el campo del Machine Learning, lo que permite obtener una lista de temas prioritarios. Finalmente, se realiza una discusión de los resultados, contrastándolos con los cambios estructurales dentro del área.

Como resultado de este estudio, se logra identificar subáreas de investigación y temas específicos que se consideran emergentes en la actualidad. Se recomienda a los equipos de investigación aborden estos temas, ya que representan áreas de gran potencial y relevancia en el campo del Machine Learning.

**Palabras clave: Natural Language Processing, Machine Learning, Topic Modeling, BERTOPIC, CSO Classifier, Thematic Maps**

# Abstract

**Application of emerging topics detection techniques to establish Research and Development priorities in the field of Machine Learning**

The emerging topics detection is of great relevance for research and development teams in Machine Learning, as it empowers them to formulate research projects, generate novel business prospects, and contribute value in terms of products, technologies, and knowledge. Nonetheless, these teams encounter diverse challenges including the swift information generation, the abundance of data sources, and the scarcity of scalable non-commercial implementations that facilitate automated analysis.

To address this need, a discussion on current methodologies for detecting emerging topics is proposed. Additionally, a methodological proposal is presented, which combines three natural language processing techniques: Computer Science Ontology Classifier, thematic maps, and BERTopic. This methodology is applied to scientific articles in the Machine Learning domain, resulting in a prioritized inventory of topics. Lastly, a comprehensive discussion of the outcomes is conducted, contrasting them with the structural transformations occurring within the field.

As a result of this study, specific research subareas and topics that are considered emerging in the present time are identified. It is recommended that research teams address these topics, as they represent areas of significant potential and relevance in the field of Machine Learning.

**Keywords: Natural Language Processing, Machine Learning, Topic Modeling, BERTOPIC, CSO Classifier, Thematic Maps**

# Contenido

# Lista de figuras

**Pág.**

# Lista de tablas

**Pág.**

# 1. Introducción

Una de las tareas más importantes a las que se enfrentan constantemente los equipos de Machine Learning, conformados en la industria o en la academia, es la detección de las temáticas emergentes en el campo. Esto resulta relevante, ya que a través de ellas se pueden formular proyectos de investigación, generar nuevas oportunidades de negocio y, en general, aportar valor en términos de producto, tecnología o conocimiento.

La necesidad de detectar las temáticas emergentes surge tanto en el sector público como en el sector privado, los cuales interactúan dentro de los ecosistemas de innovación apuntando a diversos objetivos como el incentivo de las actividades de Investigación y Desarrollo (I+D) o la búsqueda de la conversión de conocimiento hacia capacidades de mercado. La cooperación entre todos los actores de los ecosistemas de innovación, y la correcta priorización de los temas emergentes en torno a la inversión de los recursos genera crecimiento económico.

Los criterios para identificar qué temática se considera emergente no es trivial debido a la velocidad con la que crece la información disponible, las diversas fuentes de las que proviene, y las pocas implementaciones no comerciales escalables que permiten abordar la automatización del análisis. En general, se cuenta con dos tipos de información disponible: la información descriptiva de los equipos involucrados en los ecosistemas, y la información resultante de las actividades a las que estos se dedican. Ambos tipos se almacenan en bases de datos administradas por instituciones de acuerdo con el país de origen, tipo de datos, idiomas, temas especializados, tipo de acceso, entre otras características, que finalmente limitan la integración y unificación de los datos y su potencial uso en los análisis.

En el presente trabajo, se propone la aplicación de técnicas de detección de temáticas emergentes para generar automáticamente la recomendación de prioridades en

Investigación y Desarrollo; esto permite que los equipos puedan enfocar sus recursos escasos en mantener a la vanguardia el conocimiento de su unidad de negocio o unidad académica**.** Para el desarrollo, se empleará una adaptación de la metodología CRISP-DM, la cual fue diseñada para guiar proyectos de minería de datos; esta metodología tendrá en cuenta el entendimiento del problema real, la definición del problema en términos de analítica, la preparación de los datos, el modelado y la evaluación de los resultados.

Este capítulo está organizado siguiendo el ciclo de la metodología mencionada. En la Sección 1.1, se abarcan los antecedentes del problema; luego en la Sección 1.2 se expone la definición del problema real, en donde se exploran las diferentes miradas de la problemática para los grupos de interés. Posteriormente, en la Sección 1.3 se presenta el problema desde los datos, fuentes y técnicas; seguidamente, en las Secciones 1.4 y 1.5 se presentan la hipótesis y los objetivos. Finalmente, en la Sección 1.6 se expone el mapa del documento.

# 1.1 Antecedentes

## 1.1.1 Metodología y conceptos

El término *temática emergente* ha sufrido transformaciones desde su aparición en la literatura en 1965. Ha tomado el nombre de tendencias emergentes, temas de investigación emergentes, temas en tendencia y tecnologías emergentes, entre otros [1] [2]. La etapa de exploración del término, que se desarrolló entre 1975 y 2015 aproximadamente, se caracteriza por realizar acercamientos conceptuales a través de la cienciometría y la bibliometría, que a su vez se implementaron a través de reglas heurísticas [3].

La separación entre los conceptos temáticas emergentes y tecnologías emergentes en 2018, se considera una discusión vital para el desarrollo de técnicas y metodologías posteriores. Esta discusión define que, si bien una tecnología emergente se puede considerar una temática de investigación emergente, esta última no se puede considerar en sí misma una tecnología emergente; esto se debe a que la aparición de una tecnología emergente no depende únicamente de la investigación en el área, sino que existen otros factores externos que influyen en el surgimiento de la tendencia [2].

El término aceptado por la comunidad para referirse al conjunto de técnicas de minería de texto aplicado a la detección de temáticas y tecnologías emergentes ha sido Tech Mining, introducido por el profesor Alan Porter en 2005. De forma paralela al concepto, Porter propone una metodología de toma de decisiones, enfocada en los actores que tienen interés en el estudio del área, que consta de tres pasos esenciales: inteligencia, diseño y análisis, y toma de decisiones. A su vez, cada uno de estos pasos contiene nueve actividades secuenciales: identificación del problema, selección de las fuentes de información, refinamiento de la búsqueda, limpieza de datos, análisis básicos, análisis avanzados, representación, interpretación y utilización [4] [5].

### 1.1.2 Indicadores

La discusión latente alrededor de los temas emergentes se desarrolla en el marco de la carencia de una definición clara, precisa y concreta del concepto *emergencia*, en términos conceptuales y prácticos. Cualquier estudio que se realice con el objetivo de detectar temáticas emergentes, debe brindar un acercamiento trasparente, directo y coherente de esta definición [2].

Desde el estudio de las tecnologías emergentes, se propone una medición de acuerdo con la fase de adopción en la que se encuentre la tecnología: *pre emergence, emergence o post-emergence* [1]. Este concepto se deriva del estudio sociológico de los procesos de innovación, introducido en 1962 por Everett Rogers a través de su libro *Diffusion of Innovations,* y se enfoca en el proceso de adopción generalizada de una tecnología en particular [6]. De acuerdo con esto, una tecnología se considera emergente si cumple con indicar novedad radical, crecimiento acelerado, coherencia, impacto notorio y, ambigüedad e incertidumbre. Estas se pueden poner en operación a través diferentes técnicas, las cuales deben ser pertinentes a la intención del indicador. Entre ellas se encuentran el análisis de tendencias, de citaciones, de agrupación simple de términos, espaciales y modelos híbridos que integran varias técnicas [1].

En contraste con la propuesta inicial, en 2018 se expone que, a diferencia de una tecnología emergente, una temática emergente debe cumplir con los criterios de novedad

radical, crecimiento relativamente rápido, coherencia e impacto científico [2]. Entre las técnicas mapeadas para este propósito, se encuentran el análisis de citaciones, correlaciones entre citaciones, el análisis de co-citación, las técnicas de agrupamiento no supervisadas, el análisis semántico y acercamientos híbridos como la combinación de resultados entre análisis basados en términos y los basados en citaciones [2].

En general, se han establecido 4 indicadores para medir la emergencia de una tecnología: novedad, persistencia, crecimiento y comunidad. Estos han sido utilizados en la creación de políticas públicas, como el programa FUSE en Estados Unidos (EU), vinculado a la oficina ARPA (Intelligence Advanced Research Project Activity), cuyo objetivo es la detección de tecnologías emergentes para el enfoque del gasto público. Estos indicadores fueron llevados a producción a través de la creación de un score de emergencia personalizado en el software VantagePoint [5].

Actualmente, otras perspectivas dentro del desarrollo del área que surgen luego de 2015 [3], toman como base los indicadores planteados por Wang en 2018 [2] realizando adaptaciones para la aplicación de técnicas de Machine Learning e Inteligencia Artificial. Entre ellos se encuentran el crecimiento acelerado, la novedad radical y el impacto prominente [7].

### 1.1.3 Técnicas

El primer paso para tener en cuenta para la detección de temáticas emergentes es la selección de fuentes de información, la cual se basa en diversos criterios que deben estar alineados con los objetivos del análisis. De acuerdo con Porter, los registros que contienen más información son aquellos de los que se puede obtener atribución a un investigador o grupo de investigadores, donde se asegure la calidad de la información, que contenga con todos los campos requeridos, y que sea pertinente y especifico de acuerdo con el campo de conocimiento de interés [8]. Las técnicas de preprocesamiento de esta información han evolucionado conforme aumenta la cantidad de datos disponibles.

Ante la selección de fuentes, en 2015, Ying Huang et al [9], proponen una estrategia de búsqueda de información diseñada para el análisis de tecnologías emergentes. El objetivo de esta estrategia es realizar búsquedas escalables en las bases de datos y así, mejorar

la exhaustividad y precisión de los estudios. Esta se fundamenta en métodos como consultas basadas en léxico, consultas en revistas especializadas y análisis basados en citaciones.

Estudios como el anterior, se dieron en la fase exploratoria del término, en donde el contexto requería de bases de datos limitadas. Fue hasta 2018 cuando se introdujeron los primeros análisis con bases de datos grandes para la aplicación de técnicas de Machine Learning e Inteligencia Artificial [9]. Este enfoque se extiende hasta el día de hoy, en donde las técnicas de Machine Learning e Inteligencia Artificial marcan la tendencia de investigación el campo [10] [11].

Entre las técnicas de preprocesamiento más utilizadas se encuentran: aumento de datos (recopilar información complementaria a los artículos y patentes) [12], consolidación de términos [12], algoritmos Fold NLP Terms [5], stemming [12] [13], tokenization [7] [13], eliminación de stopwords [12] [5] [7] [13], fuzzy matching [12], clustering con dos palabras o con múltiples palabras [12], Netclus clustering [12] y Emergence Scoring Toutine en VantagePoint [12] [5].

Entre los modelos más utilizados se encuentran**:** LSTM-based neural network [30] y medias móviles con peso [12] para predecir la emergencia, análisis de influencia de los autores y citaciones como Weighting for journal influence (ISI Journal Citation Reports) y author's affiliation prestige (CWTS Leiden University Rankings) [12]; Topical N-Grams Model (TNG) [12] [7], Dynamic Influence Model (DIM) [13], Citation Influence Model (CIM) [12] [7] [13], y Multi-Task Least- Squares Support Vector Machine (MTLS-SVM) [13], para la detección de temáticas.

Entre los software destacados para realizar este tipo de análisis se encuentran: Cauliflower, Watson, RapidMiner, Google NLTK, MonkeySurvey, Derwent y VantagePoint.

## 1.2 Definición del problema real

En esta sección, se aborda el problema de la detección de temáticas emergentes y se discute el problema de innovación en las instituciones, y las dificultades que se presentan para medir y rastrear las actividades de Investigación y Desarrollo (I+D).

### 1.2.1 Problema de la innovación en las instituciones

En el contexto de ciencia y tecnología, la innovación se logra engranando recursos e instituciones para el aprovechamiento y apropiación de ideas existentes o para la generación de ideas completamente nuevas. Estas ideas, a su vez, se convierten en productos, líneas de negocio, patentes o proyectos de investigación de gran impacto [14].

Dentro de los ecosistemas de innovación interactúan diversos actores e instituciones que buscan generar una ventaja competitiva con respecto a sus pares, a través de la explotación y exploración de conocimiento. Un ecosistema de innovación se define como el organismo de tipo económico que coordina la interacción entre sus miembros, para permitir o facilitar el desarrollo tecnológico [15]. Los principales miembros de estos sistemas son el gobierno, el sector educativo (universidades y otras instituciones de educación superior) y el sector privado. La interacción entre ellos genera, entre otros efectos, la conversión de capacidades científicas en capacidades de mercado, alcanzando así ventajas competitivas.

El sector público, a nivel global, se encarga tanto de la asignación de recursos como de incentivar las actividades de I+D para lograr el crecimiento económico. Estudios recientes, ha encontrado que la relación entre el gasto gubernamental enfocado a dichas actividades, y el crecimiento de la economía por país a nivel mundial, sugiere una correlación positiva, además de generar un factor protector. Además, este factor protector causa que un decrecimiento en la asignación de recursos tenga un menor impacto en la desaceleración económica; en este mismo sentido, el aumento similar en la inversión causa un mayor crecimiento económico [16]. Esta situación se evidencia en países como Israel, Corea del

Sur, Japón, China y Estados Unidos (EU), quienes invirtieron más del 3.4% de su PIB en innovación, investigación y desarrollo en 2020 [17].

Otro de los actores a considerar son las universidades e instituciones de educación superior, cuya misión tradicional dentro del sistema ha sido educar a los próximos lideres e investigadores; sin embargo, en la estructura actual de innovación, se espera que, además, cumplan con un rol intermediario en el planteamiento y resolución de los problemas de la sociedad. De acuerdo con el informe de la Organización para la Cooperación y el Desarrollo Económicos (OECD) de 2017, en el que se detectan tendencias para crear políticas gubernamentales a nivel mundial, se recomienda generar alianzas entre universidades, compañías y ciudadanos, con el objetivo de conocer las necesidades de la sociedad que el gobierno pueda solucionar [18]. Para ello, cada región establece estrategias de gasto alineadas con el rol requerido; por ejemplo, Reino Unido (UK) en 2019 invirtió el 23,5% del Gasto Interno Bruto I+D (GERD) en educación superior, dándole un papel ejecutor a las universidades e instituciones de educación superior; mientras que, en países como EU, las universidades deben encargarse desde la obtención de recursos hasta la ejecución de los mismos [19]. Dentro de las tareas propias de estas instituciones, se destacan la coproducción de proyectos con el sector privado, la orquestación de encuentros de los diferentes actores y, en general, la gestión del intercambio de conocimiento.

Finalmente, el sector privado en la mayoría de los países cuenta con incentivos gubernamentales para destinar recursos a las tareas propias de innovación, generando empleos y garantizando la competitividad. Un aumento en dichas actividades incrementa la probabilidad de que se introduzcan nuevos productos o mejoras en los procesos [20]. Entre 2015 y 2017, más del 50% de las compañías en EU, pertenecientes a los sectores de equipos de comunicación, fabricación de software y procesamiento de datos (hosting y relacionados), reportaron innovaciones de producto o de proceso [21]. En 2019, las compañías en EU aportaron el 72,2% de los recursos invertidos en I+D, seguido por el gobierno federal con un 20,1%, los cuales fueron ejecutados en un 74,7% por los negocios y un 11,7% por universidades e instituciones de educación superior [22].

Existen industrias donde es intensiva la traducción de capacidades de ciencia y tecnología en capacidades de mercado. Esta conversión va desde de las universidades y laboratorios

gubernamentales hacia diferentes sectores de la industria. Se destacan la fabricación de software, ciencias biomédicas y ciencias de la salud. El valor agregado de las compañías que mantienen altos niveles de conversión, y por lo tanto una intensa actividad en I+D, se duplicó a nivel mundial entre 2002 y 2019, en regiones como EU y China [21].

Para lograr este crecimiento económico, los actores deben lograr ventajas competitivas, tal como ya se indicó. Estas se alcanzan por: primero, la coordinación entre la asignación de recursos entre el sector público y el privado; segundo, la transferencia de conocimiento de las instituciones científicas; y, tercero, el enfoque de las estrategias en los sectores industriales en crecimiento. Consecuentemente, la asignación de recursos escasos, los múltiples actores y el impacto que tienen las actividades I+D, resulta en un interés generalizado por el rastreo y medición de ellas.

Este monitoreo de las actividades de I+D puede presentar diferentes enfoques que dependen del grupo de interés y de la finalidad del análisis. Por un lado, se encuentran las agencias gubernamentales interesadas en cómo medir el progreso o retroceso de sus inversiones, o cómo construir mejores metodologías de desarrollo de proyectos que garanticen impacto social. De manera complementaria, el sector privado muestra interés hacia la conversión del conocimiento académico en innovaciones comerciales o retornos financieros. En ambos casos, se requiere la priorización de temáticas o proyectos que ayuden a enfocar los recursos escasos en el máximo retorno.

### 1.2.2 Problema de la medición y el rastreo de las actividades

El rastreo y monitoreo de las actividades de I+D, y por consiguiente el análisis de temáticas emergentes, representa un reto para todos los involucrados en los ecosistemas de innovación, debido a la cantidad de información disponible y la diversidad de fuentes. Existen organizaciones gubernamentales y privadas que recopilan datos y realizan análisis descriptivos, pero estos resultan generales ante la priorización de temas en un sector especializado.

Los datos necesarios para realizar los análisis se pueden clasificar en dos categorías: de entrada, o de salida. Los datos de entrada caracterizan a los equipos e instituciones

involucrados, por ejemplo, número de investigadores, recursos invertidos, redes de colaboración, investigadores destacados, entre otros. Los datos de salida se refieren a la evidencia documental que retornan las tareas, tales como artículos científicos, registros o solicitudes de patentes, memorias de conferencia, planes de negocio, libros, etc [8].

Las revistas y artículos científicos son considerados los formatos obligados para la publicación de los avances en ciencia y tecnología. De acuerdo con el Centro Nacional de Estadísticas para Ciencia e Ingeniería (NCSES), en 2020 se publicaron 2.940.807 de artículos científicos revisados por pares, en el área de ciencia y tecnología (no médica) a nivel mundial, con un crecimiento medio anual de 4% desde 1996. Entre 2019 y 2020, aumentó la producción en un 7%, respondiendo a la pandemia por COVID-19 en donde la comunidad científica fue protagonista [23].

En áreas del conocimiento en auge, se evidencian altas tasas de crecimiento en relación con otras áreas consolidadas. Entre esas áreas en auge, se encuentran las Ciencias de la Computación y específicamente la subárea de Inteligencia Artificial, que para el año 2021, alcanzó un total de 334.497 publicaciones, con un crecimiento medio anual del 13% entre 2018 y 2021; el 51,5% de estas publicaciones son artículos de revista y el 21,5% artículos de conferencia. Los campos que más contribuyeron a esta tendencia fueron el Reconocimiento de Patrones y el Machine Learning con un incremento del 12% y 30%, respectivamente [24]. Este último supera en más de dos veces el crecimiento promedio de toda la subárea.

La solicitud y registro de patentes es un indicador de la transferencia de conocimiento desde el campo científico hacia el mercado, medido por las citaciones incluidas en los documentos de solicitud. Según la Organización Mundial para la Propiedad Intelectual (WIPO), en 2020 se solicitaron 3.276.700 patentes a nivel mundial, de las cuales el 45,7% se realizaron en China [25]. En Inteligencia Artificial, se presentaron 141.241 solicitudes de patentes en 2021 cuya tasa de aceptación ha variado entre aproximadamente el 14% y el 43% desde 2011 hasta 2021 [24].

Teniendo en cuenta la cantidad de datos disponibles y su crecimiento constante, resulta un reto para los involucrados detectar qué temas podrían ser foco de su investigación. A pesar de que algunas instituciones realizan mediciones (tales como la OECD, el Servicio

de Información para la Comunidad de Investigación y Desarrollo (CORDIS), el Programa para el Monitoreo, Evaluación, investigación y Aprendizaje en Innovación (MERLIN), Centro Nacional de Estadísticas para Ciencia e Ingeniería (NCSES)), estas resultan ser generales, y no muestran la tendencia por campos del conocimiento fuera de sus herramientas comerciales, ni reportan el crecimiento real de temáticas individuales por área, sobre todo en sectores emergentes y de gran potencial como la inteligencia artificial, el desarrollo de software, energías renovables, cambio climático, salud pública, entre otros.

## 1.3 Definición del problema de analítica

La información necesaria para el análisis de temáticas emergentes se aloja en bases de datos documentales, donde la estructura de los datos, la cantidad de información disponible y los servicios de análisis adicionales, son particulares de cada una de las fuentes. Un usuario que desee conocer las tendencias de un área específica deberá considerar estos aspectos para escoger la mejor fuente de datos además de contar con herramientas computacionales para su escalamiento.

Actualmente, los datos de salida de las actividades de I+D se alojan en bases de datos especializadas encargadas de recopilar y exponer información documental de manera digital. Los formatos más comunes son libros, artículos científicos, patentes, memorias de conferencia, reportes de investigación y tesis. Los tipos de datos almacenados pueden ser texto, imágenes, audio o video.

Se distinguen diferentes tipos de bases de datos de acuerdo con la clase de información que almacenan: base de datos bibliográficas, de texto completo y de metadatos. Las bases de datos bibliográficas recopilan únicamente la citación de los trabajos científicos, creando un catálogo disponible para encontrar literatura especifica. Por su parte, las bases de datos de metadatos contienen fracciones del texto estandarizadas y de mayor contenido como resumen, autores, instituciones, y título, entre otros; estas brindan información precisa, pero sin permitir el acceso a todo el texto. Por último, las bases de texto completo ofrecen el artículo para realizar búsquedas, sin que esto implique que el usuario pueda descargar

o leer todo el contenido; las búsquedas sobre estas suelen ser amplias y poco precisas [26].

En ciencia y tecnología, el formato más utilizado para publicar los avances son los artículos científicos. Desde su aparición en el siglo XVII, con la primera publicación de la revista Philosophical Transactions, asociada a la Sociedad Real de Londres, las revistas científicas se han consolidado como el principal acceso multidimensional sobre descubrimiento y nuevas investigaciones [27]. La principal característica de las publicaciones científicas, ya sean especializadas en un área o de conocimiento general, es la recopilación de diversos autores y la publicación periódica de contenido validado. Para este fin, es necesario velar por la calidad y veracidad de la información, por lo que cada revista cuenta con procedimientos estandarizados para la validación y aceptación de artículos. La estrategia más común y de mayor aceptación es la revisión por pares académicos.

Cada revista establece si el acceso al contenido será gratuito, con suscripción o híbrida. Gracias a la democratización del conocimiento, cada vez es más común encontrar revistas de acceso gratuito y completamente digital; sin embargo, entre más especializada sea el área, aumenta la probabilidad de que el acceso sea por suscripción y se restrinja la cantidad de documentos completos que se puedan consultar. Consecuentemente, si un usuario está interesado en conocer los últimos avances en materia de investigación, debe enfocar su búsqueda en bases de datos que recopilen información de revistas especializadas.

Por otro lado, si el foco es el registro de patentes, es importante tener en cuenta que estas dependen de las leyes de protección de propiedad intelectual por país; por lo tanto, en la mayoría de los casos, si una patente es registrada en una oficina nacional, solo tendrá validez en el territorio especifico. Así, la consolidación de la información dependerá de la oficina de patentes de interés. En este sentido, Scopus, es una de las principales bases de datos indexadas, y cuenta con información de la Organización Mundial para la Propiedad Intelectual (WIPO, Oficina Europea de Patentes (EPO), Oficina de Patentes de Estados Unidos (USPTO), Oficina Japonesa de Patentes (JPO) y la Oficina de Propiedad Intelectual de Reino Unido (IPO.GOV.UK), pertenecientes a los países con mayor registro de patentes a nivel mundial [28].

Además de recopilar información de revistas especializadas y abarcar la mayoría de las oficinas de patentes, la elección de la base de datos que un usuario debe usar debe complementarse con criterios como: la posibilidad de acceder al texto completo o metadatos, que cuente con la información histórica necesaria, la existencia de métricas de evaluación entre revistas o artículos, la revisión por pares y que el tipo acceso se adecue a las necesidades del usuario al igual que la estructura de los datos [8].

Entre las bases de datos más populares se encuentran:

- Scopus: De tipo bibliográfica con resúmenes. Tiene disponible más de 23 mil revistas revisadas por pares y acceso a más de 44 millones de registros de patentes [28].
- Web of Science: Plataforma que recopila información de bases de datos bibliográficas. En su versión no comercial, llamada Core Collection, no ofrece textos completos ni patentes. En la versión comercial, ofrece más de 49 millones de registros de patentes y en total 182 millones de registros de todos los formatos [29].
- Google Scholar: Buscador gratuito de información científica que abarca únicamente desde 2016 hasta la actualidad. Recopila documentos científicos alojados en sitios web, por lo tanto, depende de ellos para tener disponible metadatos o textos completos. Además, no cuenta con métodos para controlar la validez y veracidad del contenido. No garantiza que el articulo esté publicado en una revista científica [30].

Una vez el usuario escoge la base de datos que se adapta a su necesidad, se necesita procesar la información y escalar los métodos de análisis. Para esto, se requieren herramientas computaciones que soporten la limpieza de los datos, su agregación y posterior disposición de los resultados. Estas tareas pueden ser delegadas a una solución comercial; por ejemplo, VantagePoint es un software diseñado para analizar literatura científica y patentes, el cual ofrece dentro de sus servicios, funcionalidades como limpieza de texto, mapas de colaboración, matriz de co-ocurrencia, tablero de monitoreo y reporteria enfocada en el resumen del material [31]. Sin embargo, el análisis se limita al procesamiento de un gran volumen de datos con conclusiones generales sin foco en un

área específica. Esta herramienta es reconocida por la amplia variedad de formatos de bases de datos bibliográficos que tiene la capacidad de importar.

Precisamente estas soluciones comerciales, unidas a las diversas bases de datos disponibles, la naturaleza no estructurada de los documentos y las herramientas computaciones necesarias para el escalamiento, hacen que la detección de temáticas emergentes y la priorización de temas no sean unas tareas sencillas, especialmente para los actores involucrados en los ecosistemas de innovación que deben tomar decisiones con respecto a la asignación y ejecución de recursos.

En este trabajo final de maestría, el problema a abordar es la detección y priorización de temas de investigación para equipos que desarrollan tareas de Investigación y Desarrollo en el área de Machine Learning. Desde el punto de vista de la analítica, este problema se puede plantear en primera instancia, mediante el uso de técnicas de minería de texto y procesamiento de lenguaje natural, enfocado en la detección de términos relevantes en sectores económicos en crecimiento. Una vez procesados los datos de salida, es posible usar métricas que permitan establecer el nivel de emergencia de cada uno de ellos.

## 1.4 Definiciones y conceptos básicos

En esta sección se presentan las principales definiciones y conceptos básicos asociados a la detección y priorización de temáticas emergentes.

- **Tech Mining:** Se define como la aplicación de técnicas de minería de texto sobre información de ciencia y tecnología proveniente de bases de datos especializadas; su objetivo principal es monitorear y pronosticar los cambios tecnológicos. Entre sus aplicaciones más importantes se pueden encontrar: pronóstico de tecnologías emergentes, creación de indicadores para el rastreo de las actividades de innovación, inteligencia competitiva de producto, entre otros [4].

- **Minería de texto:** Se refiere al campo del conocimiento dedicado a estudiar el procesamiento y análisis de los datos tipo texto en cualquier contexto. Su objetivo,

además de llevar el texto no estructurado a un formato estructurado, es reconocer patrones y descubrir nuevas perspectivas implícitas. Las técnicas para realizar este tipo de análisis van desde reglas heurísticas para la clasificación de palabras hasta técnicas de Machine Learning [4] [32].

- **Procesamiento de Lenguaje Natural (Natural Lenguage Processing, NLP):** Es el subarea de la inteligencia artificial (IA) encargada de explorar y modelar el entendimiento de las maquinas del lenguaje humano. Su propósito es detectar patrones lingüísticos y convertirlos en conocimiento replicable a través de herramientas tecnológicas que generen valor. Para ello, recopila información de áreas transversales como la psicología, ciencias de la computación, reconocimiento de patrones de voz, visión artificial, lingüística computacional, lingüística aplicada, etc [33] [8].

- **Base de datos indexada:** La indexación es un método tradicionalmente usado en las bases de datos que consiste en asignar un orden alfabético o numérico a los registros, para optimizar las búsquedas dentro de ellas. En el ámbito científico, las bases de datos indexadas son la forma de almacenamiento que permite a los usuarios realizar búsquedas sobre los metadatos de la información científica, especialmente relaciones que se dan en las citaciones entre autores. Dentro de la comunidad científica, la publicación de formatos, que cuenten con servicios indexados, genera credibilidad y reconocimiento, además de ser la principal herramienta de consulta actual [34].

- **Clustering:** Conjunto de técnicas de aprendizaje no supervisado de amplio uso en NLP, en donde se destaca el clustering basado en tópicos para extraer contenido y generar resumen de texto; esto implica agrupar las frases clave en temas [20]. Dentro de los problemas de tech minning, las técnicas de clustering son utilizadas para medir la similaridad entre los términos y la posibilidad de agruparlos como una temática individual [35].

## 1.5 Hipótesis

La hipótesis de este trabajo es la siguiente:

> Es posible detectar temáticas emergentes en el área de Machine Learning que permita establecer una prioridad en los temas a abordar dentro de las actividades de Investigación y Desarrollo, mediante la utilización de técnicas de minería de texto y el procesamiento de lenguaje natural.

## 1.6 Objetivos

### 1.6.1 Objetivo general

El objetivo general de este trabajo es el siguiente:

> Detectar las temáticas emergentes en el área de Machine Learning aplicando técnicas de minería de texto y procesamiento de lenguaje natural, generando un informe final con prioridades de temas abordar.

### 1.6.2 Objetivos específicos

1. Realizar el diseño del estudio para seleccionar los registros bibliográficos relevantes en el área de Machine Learning.
2. Aplicar el diseño del estudio anterior para seleccionar los registros bibliográficos de la muestra.
3. Preprocesar la muestra de registros.
4. Determinar los patrones emergentes a partir de la muestra preprocesada.
5. Validar los resultados obtenidos en el objetivo anterior.

## 1.7 Mapa del documento

El presente documento sigue la siguiente estructura: en el Capítulo 2 se realiza una discusión exhaustiva acerca de las metodologías y las técnicas disponibles para la detección de temáticas emergentes, además se expone detalladamente la metodología propuesta. Posteriormente, en el Capítulo 3, se lleva a cabo la aplicación de la metodología propuesta y se obtienen los resultados correspondientes. Por último, en el Capítulo 4, se analizan y discuten en profundidad los resultados obtenidos durante la implementación. Finalmente, en el Capítulo 5, se presentan las conclusiones y recomendaciones derivadas del trabajo realizado.

# 2. Discusión de técnicas y metodología propuesta para la detección de temáticas emergentes

Como se introduce en el capítulo anterior, existen múltiples técnicas para abordar cada paso sugerido en la identificación y priorización de temáticas emergentes. La combinación de diferentes técnicas de manera secuencial para obtener una lista de temas se denominará en el resto del documento como una metodología. En este capítulo se aborda el Objetivo 1 que consiste en estudiar y diseñar las técnicas para la detección de temáticas emergentes. Para ello se desarrolla una discusión entre las metodologías disponibles y se establecen las bases conceptuales necesarias en el entendimiento de las implementadas en los siguientes capítulos.

## 2.1 Mapas temáticos y Ontología de Ciencia de la Computación

### 2.1.1 Mapa temático

Cobo et al en el año 2011 [36] proponen un método de detección de temáticas en datos bibliográficos denominado mapas temáticos que consiste en ubicar los temas en un mapa de 4 cuadrantes basado en las métricas de relevancia y desarrollo. Cada cuadrante representa un momento en el ciclo de vida de un tema, iniciando como temas emergentes y finalizando como temas en declive o básicos en el contexto del área de investigación.

Los mapas temáticos se encuentran implementados en una librería de código abierto desarrollada para R llamada Bibliometrix cuyo objetivo es facilitar el análisis de información

bibliográfica. Entre sus funcionalidades más populares se encuentran la conexión directa con las API de Scopus, PubMed y Web of Science; la implementación de métodos de preprocesamiento de lenguaje natural, la representación datos de tipo texto, y la visualización de matrices de co-citación y colaboración científica bibliométricos [37] [38].

A continuación, se describen los pasos del algoritmo disponible en Bibliometrix para los mapas temáticos:

**Paso 1. Representación de los textos**

Dados los textos y posibles palabras presentes dentro de todos los documentos, cada texto es representado por un vector $d_i$:

$$d_i = \{w_{i,1}, \dots, w_{i,q}\}$$

Este vector está compuesto por el peso de cada palabra $w_{i,j}$ dentro del documento $i$. Existen diversos métodos para estimar el peso. En este caso, los pesos se estiman a partir de una distribución binaria de acuerdo con la aparición de la palabra en el texto; 0 para no ocurrencia y 1 para ocurrencia [39].

Al obtener todos los vectores $d_i$ como filas, se obtiene la matriz D y su representación de q-dimensiones, llamada matriz de co-ocurrencia, calculada así: $A = D^T D$. La diagonal principal de la matriz de co-ocurrencia A, representa el número de documentos en los que dos palabras que co-ocurren [39].

**Paso 2. Cálculo de métrica de similitud y representación como grafo**

Una vez obtenida la primera versión de la matriz A, compuesta por la ocurrencia y no ocurrencia de las palabras en el texto, se obtiene la métrica Association Strength (AS) que representa la similitud entre dos palabras. Un valor de 0 en la métrica significa que las dos palabras no co-ocurren en ningún documento y un valor de 1 que las dos palabras co-ocurren dentro de todos los documentos. A continuación, el método de cálculo:

Dadas dos palabras, $j\ y\ j'$, la co-ocurrencia se calcula como [39]:

$$AS_{jj'} = \frac{a_{jj'}}{\hat{a}_{jj}\hat{a}_{j'j'}}$$

Donde:

- $a_{jj'}$ representa al número de co-ocurrencias entre las palabras.
- que $\hat{a}_{jj}\hat{a}_{j'j'}$ es el número esperado de ocurrencias de las palabras asumiendo que son estadísticamente independientes.

La matriz A, se modifica y sus valores son ahora la métrica AS. A su vez, esta se puede representar como un grafo, en donde los nodos corresponden a cada palabra y las aristas a la métrica [39].

**Paso 3. Cálculo de la densidad y centralidad de Callon**

Dado el cálculo de AS y la representación de la matriz A como un grafo, se aplican algoritmos de detección de comunidades (métodos de agrupamiento para grafos) en los cuales se obtienen como salida subgrafos que mantiene coherencia entre sus nodos y aristas, cada subgrafo corresponde a un tema detectado. Este tipo de algoritmos pueden ser tipo de aglomerativo o divisivo, la diferencia reside en el método de inclusión de las aristas. La decisión de qué algoritmo de detección de comunidades a aplicar, puede influir en los resultados de la detección de temas y se recomienda que sea adecuando para el problema a abordar.

Luego de obtener los subgrafos del paso anterior, se proyectan los resultados obtenidos calculando las siguientes métricas para cada tema detectado: Centralidad de Callon ($\mathrm{CC_k}$) y Densidad de Callon ($\mathrm{CD_k}$) [39]:

$$CC_k = 10 \text{ x } \sum_{j\in k, h\in k'} AS_{jh}$$

$$CD_k = 100\ x \sum_{j,j'\in k} \frac{AS_{jj'}}{n_k}$$

La centralidad mide la relevancia de un tema en todo un dominio de investigación, es decir, el nivel de interacción de un subgrafo con otro y por tanto qué tan central es. La densidad, por su parte, mide el nivel de desarrollo de este representado por la densidad del subgrafo [40].

**Paso 4. Mapa temático**

Al obtener el número de temas (número de subgrafo o comunidades), y la relevancia y desarrollo de cada uno, se ubican gráficamente en un plano de 4 cuadrantes, como se muestra en la Fig. 1.

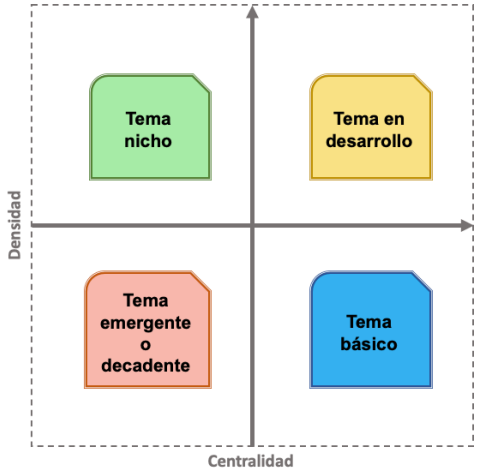


Fig. 1. Distribución de cuadrantes de ThematicMap ante valores de densidad y centralidad

Se les considera a los temas que cuentan con una alta centralidad y densidad como en desarrollo ya que denota que son importantes dentro de la comunidad y su vez los subtemas investigados guardan mucha relación entre ellos. De manera análoga, se consideran temas emergentes o decadentes cuando no demuestran ni densidad ni centralidad; dicha situación se presenta en dos momentos del ciclo de vida un tema de investigación: en su nacimiento cuando no existe una comunidad consolidada ni claridad sobre los temas centrales, o cuando la temática se desarrolló y se encuentra en el momento de convertirse en dos o más temas en desarrollo, por lo cual tanto su centralidad como densidad se degradan. Por su parte, un tema nicho se interpreta como un campo específico, que, si bien no es central dentro de la comunidad, es coherente y compacto en los subtemas que abordan. Finalmente, un tema básico es aquel que no tiene subtemas específicos en desarrollo, pero siguen siendo centrales para toda la comunidad [38].

### 2.1.2 Ontología de Ciencia de la Computación

El clasificador de Ontología de Ciencia de la Computación (CSO classifier, por sus siglas en inglés) es una herramienta especializada en la clasificación de artículos científicos en inglés según su área de investigación a partir de información como el resumen, título y palabras clave. Este servicio se basa en un modelo no supervisado que utiliza la taxonomía de la Ontología de Ciencia de la Computación. Dicha taxonomía ha recolectado más de 16 millones de publicaciones en el área de investigación de ciencias de la computación, lo que ha permitido la generación automática de una ontología que puede ser actualizada en cualquier momento. Esta ontología consta de aproximadamente 14.000 temas y 162.000 relaciones semánticas [41].

El clasificador consta de tres módulos que pueden ser usados independiente o secuencialmente, según el caso de uso, nivel de especificidad o escalabilidad requeridos por el usuario final para obtener temas relacionados con el texto. El primer módulo, el módulo sintáctico, es responsable de tomar los artículos sin preprocesar y buscar posibles cadenas de texto que coincidan explícitamente con algún término dentro de la ontología, a través de la construcción de n-gramas. A continuación, el módulo semántico busca categorías gramaticales y relaciones semánticas de la ontología, agregándole embeddings para inferir áreas de investigación implícitas. Finalmente, en el módulo de posprocesamiento se combinan los resultados anteriores para generar una lista de temas, que se revisa para eliminar posibles datos atípicos y complementarla agregando los super temas detectados dentro de la taxonomía [41] [42].

A continuación, se presentan las opciones disponibles para hacer uso del clasificador:

- **R-Classify:** es una herramienta desarrollada para R de acceso libre en línea que permite a los usuarios ingresar un resumen de su artículo en formato de texto o PDF. A través de su interfaz, la herramienta muestra los temas seleccionados y las anotaciones correspondientes dentro del documento para cada tema. Se debe tener en cuenta que esta herramienta funciona como una versión de prueba y no es posible clasificar múltiples textos ni modificar los parámetros del modelo

- **Librería cso-classifier:** es una librería de código abierto desarrollada en Python que permite la clasificación de textos individuales o múltiples en un proceso por lotes. Ella también permite la modificación de los parámetros del modelo, el uso independiente de los módulos y la paralelización del procesamiento. Es importante destacar que, a diferencia de otras implementaciones de modelos pre-entrenados que tienen una cuota límite para las solicitudes realizadas, esta librería no tiene restricciones en este sentido, ya que al obtener los archivos directamente, se descarga la última versión entrenada del modelo y no se depende de una conexión vía API para realizar las clasificaciones [42].

## 2.2 BERTopic

BERTopic es un método de aprendizaje profundo para el modelado de temas, desarrollado en 2020, que se basa en tres conceptos principales. El primer concepto es la representación del texto como vectores multidimensionales o embeddings, lo cual se logra mediante el uso de un modelo tipo BERT (Bidirectional Encoder Representations from Transformers, por sus siglas en inglés). El segundo concepto consiste en la agrupación de características utilizando algoritmos especializados en dicha tarea. Finalmente, el tercer concepto se relaciona con la medición de la relevancia de los temas detectados a través de un procedimiento de clasificación basado en la transformación TF-IDF (Term Frequency - Inverse Document Frequency, por sus siglas en inglés).

Es importante señalar que estos conceptos se utilizan en conjunto para lograr la detección efectiva de temas en un conjunto de textos. El uso de modelos tipo BERT ha demostrado ser una técnica muy efectiva para la representación de texto en varios campos de la inteligencia artificial. Además, los algoritmos de agrupamiento han sido ampliamente utilizados en la detección de temas en textos. Por último, la transformación TF-IDF es una herramienta comúnmente utilizada en la recuperación de información de los temas detectados y hacerlos interpretables para los usuarios.

El nombre de la técnica BERTopic se debe al uso de un modelo preentrenado para procesamiento de lenguaje natural llamado BERT. Este modelo preentrenado contiene

más de 345 millones de parámetros en su versión completa, y puede utilizarse como un modelo de clasificación, un sistema de preguntas y respuestas, o como un detector de patrones dentro de la estructura semántica de textos. Una de las ventajas más notables de BERT en comparación con otros modelos es que, gracias a su arquitectura bidireccional y su mecanismo de atención, es capaz de capturar el contexto completo de una frase al considerar las palabras que la rodean tanto antes como después. Modelos previos a BERT solo tomaban en cuenta la información en una sola dirección, ya sea de izquierda a derecha o de derecha a izquierda, lo que limitaba la capacidad del modelo para procesar información contextual completa.

A continuación, se describen los pasos que realiza BERTopic para obtener temas de un conjunto de documentos:

**Paso 1. Embedings de los documentos**

En la primera etapa del proceso, se recibe un conjunto de textos y se genera su representación numérica utilizando sentence-transformers, los cuales toman párrafos o frases como entrada y retornan una representación vectorial optimizada por similitud semántica entre sus componentes. Por defecto, BERTopic utiliza el modelo all-MiniLM-L6-v2 para textos en inglés, el cual se considera el modelo con mejor desempeño en tiempo de ejecución y el modelo paraphrase-multilingual-MiniLM-L12-v2 para textos en otros idiomas [43]. El tamaño del espacio vectorial es directamente proporcional al tamaño del diccionario de palabras o frases utilizado. En general, los datos de salida se consideran de alta dimensionalidad.

**Paso 2. Reducción de dimensionalidad**

Una vez obtenido el espacio vectorial compuesto por los embedings del Paso 1, es necesario reducir la dimensionalidad para facilitar la manipulación de los datos y su posterior agrupamiento. Para esta tarea, BERTopic utiliza el algoritmo UMAP (Uniform Manifold Approximation and Projection), el cual está especializado en conjuntos de datos de alta dimensionalidad y busca minimizar el tiempo necesario para su ejecución. De forma similar al método t-SNE (T-distributed Stochastic Neighbor Embedding), UMAP se basa en

la creación de un grafo integrado inicial y luego lo reduce manteniendo las relaciones más importantes entre sus nodos [44].

**Paso 3. Agrupamiento**

Para la detección de temas es necesario agrupar los términos que guardan una relación entre ellos, es decir, aquellos puntos del espacio vectorial de embedings reducidos en el Paso 2 que se encuentran cercanos entre sí. Esta tarea se lleva a cabo mediante algoritmos de agrupamiento, que pueden clasificarse en dos tipos importantes: basados en centroides y basados en densidad. Las técnicas basadas en centroides se fundamentan en la iteración sobre un punto inicial llamado centroide, mediante el cual se itera hasta converger de acuerdo con una métrica de referencia a optimizar. Sin embargo, estas técnicas son sensibles a los datos atípicos y a la elección del primer centroide. Por otro lado, las técnicas basadas en densidad agrupan secciones del espacio que se consideran densas, dejando de lado puntos atípicos [45].

BERTopic, por defecto, utiliza el algoritmo de agrupamiento basado en densidad HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) que garantiza la obtención de grupos compactos con cualquier forma y tamaño, así como la exclusión de datos atípicos [46].

**Paso 4. Lista de palabras**

Como resultado del Paso 3, se obtiene una serie de grupos detectados. Sin embargo, estos grupos no cuentan con un orden ni una visualización que represente un tema específico. Para abordar esta tarea, se utiliza la técnica de lista de palabras (conocida como "bag-of-words" en inglés) que implica combinar todos los documentos de un grupo y realizar un recuento de frecuencias en todo el texto. Esto permite que el usuario genere una descripción del tipo de grupo que se ha detectado guiándose por los más frecuentes [46].

**Paso 5. Técnica c-TF-IDF**

Por su parte, TF-IDF es una técnica estadística que busca cuantificar la importancia de una frase o una palabra en un documento y dentro de una colección documentos. Esta técnica es la base principal del modelo más popular para detección de temáticas llamada LDA (Latent Dirichlet Allocation, por sus siglas en inglés). Dentro de la comunidad,

generalmente se utiliza para calcular la importancia de las palabras entre documentos. En este caso, la unidad de texto se convierte en los grupos detectados en el Paso 4, por lo tanto se da paso a la técnica c-TF-IDF (class-based TF-IDF, por sus siglas en inglés), donde los documentos pasan a ser los grupos y por lo tanto se les llaman clases. El objetivo es determinar las palabras más importantes por cada grupo y así describirlo.

Estos cinco pasos representan el funcionamiento básico de BERTopic como metodología en su entrenamiento inicial. Una vez entrenado es posible hacer uso tanto de la detección de temáticas como de la aplicación de los patrones aprendidos para clasificar una nueva entrada al corpus dentro de los temas. Cabe resaltar que su implementación se encuentra en completo funcionamiento a través de una librería para Python que lleva el mismo. Entre sus funcionalidades más importantes se encuentra la visualización de los temas en el tiempo, una especie de Modelado Dinámico de Temas (Dynamic Topic Modeling en inglés), en donde se calcula el c-TF-IDF para cada periodo de tiempo y se obtienen las frecuencias.

Los cinco pasos anteriores describen el funcionamiento básico de BERTopic como metodología durante su fase de entrenamiento inicial. Una vez que el modelo ha sido entrenado, es posible utilizarlo tanto para detectar temas como para aplicar los patrones aprendidos y clasificar nuevas entradas en el corpus dentro de los temas existentes. Es importante destacar que BERTopic se encuentra plenamente implementado a través de una librería para Python que lleva el mismo nombre. Entre las funcionalidades más destacadas de esta herramienta se encuentra la visualización de los temas a lo largo del tiempo, que se asemeja a un Modelo Dinámico de Temas (Dynamic Topic Modeling en inglés). Este enfoque implica el cálculo del c-TF-IDF para cada periodo de tiempo y la obtención de las frecuencias correspondientes.

## 2.3 Metodología propuesta

Al combinar la detección de temas proporcionada por el clasificador de la Ontología de Ciencia de la Computación con las métricas calculadas a través del mapa temático de Bibliometrix, es posible desarrollar una metodología que tenga como entrada la metadata de los artículos científicos, es decir, el resumen, título y palabras clave, y como salida

proporcione una lista de temas emergentes o en desarrollo para un período específico de tiempo en cualquier subárea de las Ciencias de la Computación. Este enfoque es particularmente relevante en áreas como el Machine Learning ya que es posible rastrear tanto los temas técnicos como campos de aplicación para ellos.

En su estudio, Belfiore et al. [47] proponen un enfoque para caracterizar las áreas de investigación en el campo de la Inteligencia Artificial haciendo uso de una base de datos de aproximadamente 257.000 artículos publicados entre 1990 y 2022. Este enfoque combina el uso directo de la información de los artículos en el clasificador de la Ontología de Ciencia de la Computación para obtener los temas de investigación, y posteriormente los caracteriza en términos de centralidad y densidad derivados del análisis de grafos de co-ocurrencia con el mapa temático de Bibliometrix. A diferencia de los enfoques que utilizan estos métodos de manera independiente, los resultados de este estudio reflejan cambios estructurales dentro de los temas de investigación, especialmente la aparición de nuevas técnicas a partir del desarrollo de algunos temas.

La detección de temas a partir de las palabras clave obtenidas del clasificador de la Ontología de Ciencia de la Computación depende directamente de la taxonomía que, generalmente es una lista de entre una y seis palabras que describen el resumen y el título de un documento. La evolución de los temas y sus respectivas técnicas dependen de la aparición constante de las palabras clave a lo largo del tiempo, así como de la compacidad de la muestra de documentos de un área de investigación [47]. Por lo tanto, la metodología que combina el clasificador de la Ontología de Ciencia de la Computación con los mapas temáticos se considera útil para capturar las subáreas actuales en desarrollo y etapa emergente.

En el estudio realizado por Kenji Contreras et al. [48], se llevó a cabo una comparación entre el método LDA, considerado como el más tradicional, y BERTopic, con el objetivo de detectar temas en el tiempo. Los autores destacan que BERTopic presenta una capacidad superior para detectar temáticas diversas, específicas y coherentes, a pesar del costo computacional que conlleva. Es importante destacar que BERTopic cuenta con un preentrenamiento que enriquece la información a partir del texto de entrenamiento, proporcionando información adicional que no depende exclusivamente de la muestra utilizada. Según los autores, la elección entre el uso de modelos tradicionales o BERTopic

depende de la interpretabilidad semántica y la granularidad que se espera en la detección de los temas.

En este trabajo se propone la combinación de las tres técnicas presentadas, así:

- **Paso 1:** se obtendrán las palabras clave que describen el texto mediante el clasificador de la Ontología de Ciencia de la Computación, tal como se sugiere en [47] con el fin de inferir las áreas y subáreas de la taxonomía de Ciencias de la Computación que describen a las publicaciones.
- **Paso 2:** se rastrearán las subáreas temáticas obtenidas en el paso anterior usando mapas temáticos [47], mediante la implementación disponible en Bibliometrix [37] [47]. Esto permite agrupar los términos detectados de acuerdo con su momento en el ciclo de vida como temas en desarrollo, básicos, nicho o emergentes/declive.
- **Paso 3:** se obtendrán, a partir de los textos originales y usando BERTopic, las técnicas específicas que se relacionan con los temas en desarrollo y emergentes del mapa temático obtenido en el paso anterior a lo largo del tiempo, tal como se muestra en [49] [48].

La metodología propuesta se diferencia de las implementadas en los trabajos previamente referenciados al ofrecer una combinación de técnicas que permiten lograr una detección específica de temas emergentes tanto en el ámbito técnico como en el de aplicaciones dentro del subárea de Machine Learning. Los trabajos anteriores han centrado su atención en el campo de la Inteligencia Artificial en general o directamente en Ciencias de la Computación, y suelen optar por metodologías tradicionales para contrastar los resultados de una nueva propuesta que, en la mayoría de los casos, se basa en un solo algoritmo.

# 3. Aplicación de la metodología propuesta para la detección de temáticas emergentes

En este capítulo se abordan los Objetivos 2, 3 y 4 que abarcan la aplicación en general de la metodología propuesta desde la obtención de los datos, el preprocesamiento, el muestreo y el almacenamiento hasta la aplicación de mapas temáticos, la clasificación de la Ontología de Ciencia de Computación y finalmente el desglose de los temas detectados mediante BERTopic.

## 3.1 Paso 1: Clasificador de Ontología de Ciencia de la Computación

### 3.1.1 Consulta de los datos

Como se menciona en el capítulo introductorio, para obtener datos bibliográficos es necesario seleccionar una base de datos indexada que contenga documentos especializados en el área de interés. En este caso, dado que Machine Learning se considera una subárea de Inteligencia Artificial y esta, a su vez, es un campo asociado a las Ciencias de la Computación, se selecciona Scopus para el desarrollo de la metodología. Scopus cuenta con una amplia oferta de revistas que aportan al área de las Ciencias de la Computación, una plataforma robusta para la consulta de información que cuenta con más de 23.452 revistas con revisión por pares, lo que garantiza la calidad del contenido [28].

Para la descarga de las publicaciones, se ingresa al portal de búsqueda oficial de Scopus[1], específicamente a la sección de documentos, y se configura la consulta con los parámetros que se muestran en la Tabla 1. Al ingresar los parámetros, se obtiene una consulta en la sintaxis avanzada de la plataforma que es necesaria para replicar la búsqueda en caso de que sea necesario, como se muestra a continuación:

TITLE-ABS-KEY ( "machine learning" ) AND ( LIMIT-TO ( PUBYEAR , 2021 ) OR LIMIT-TO ( PUBYEAR , 2020 ) OR LIMIT-TO ( PUBYEAR , 2019 ) OR LIMIT-TO ( PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2017 ) OR LIMIT-TO ( PUBYEAR , 2016 ) OR LIMIT-TO ( PUBYEAR , 2015 ) OR LIMIT-TO ( PUBYEAR , 2014 ) OR LIMIT-TO ( PUBYEAR , 2013 ) OR LIMIT-TO ( PUBYEAR , 2012 ) ) AND ( LIMIT-TO ( PUBSTAGE , "final" ) ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) ) AND ( LIMIT-TO ( SRCTYPE , "j" ) ) AND ( LIMIT-TO ( SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA , "ENGI" ) OR LIMIT-TO ( SUBJAREA , "MATH" ) )

Tabla 1. Parámetro de consulta del portal oficial de Scopus

| Parámetro | Descripción | Valor |
|---|---|---|
| Cadena de texto | Cadena de texto que debe coincidir con el texto del título, resumen o palabras clave de la publicación | “machine learning” |
| Año de publicación | Intervalo de años en los que se buscará la información | De 2012 a 2021 |
| Tipo de documento | Tipos de documento disponibles en Scopus: Artículo, artículo de prensa, libro, capítulo de libro, artículo de revista, artículo de datos, editoriales, erratum, carta, nota, artículo de retracción, crítica y encuesta corta [28] | Artículo |

[1] https://www-scopus-com.ezproxy.unal.edu.co/search/form.uri?display=basic#basic

| | | |
|---|---|---|
| Tipo de fuente | Tipos de fuentes disponibles en Scopus: Revistas, publicaciones comerciales, series de libros y material de conferencias [28] | Revistas |
| Etapa de publicación | Cada publicación inicia el ciclo de vida en Scopus como un artículo en prensa y finaliza con la versión final | Final |
| Idioma | Scopus ofrece publicaciones en el idioma de publicación de cada revista | Inglés |
| Área | Las publicaciones están asociadas a una o varias áreas de investigación | Ciencias de la Computación, Ingeniería y Matemáticas |

### 3.1.2 Obtención de los datos

Dada la consulta replicable obtenida en el paso anterior, se procede a descargar los datos desde la plataforma Scopus a un almacenamiento local o remoto destinado para este propósito. Scopus ofrece diversos formatos de descarga, incluyendo archivos CSV, RIS, BibTeX, Plain text, Mendeley, ProQuest RefWorks, EndNote, Zotero y SciVal. Para acceder a la descarga de estos formatos, es necesario cumplir con las cuotas límite de la plataforma. Por ejemplo, para el formato CSV, existen dos opciones de descarga: la primera permite descargar hasta 2.000 publicaciones mediante una descarga directa desde el portal, incluyendo toda la información disponible, mientras que la segunda opción permite descargar hasta 20.000 publicaciones mediante un enlace externo que solo proporciona información sobre los documentos y sus citaciones [50].

En la ejecución de la metodología, uno de los objetivos es utilizar herramientas de software libre. Por lo tanto, se selecciona el formato CSV como el más adecuado para este propósito, y se opta por la descarga por enlace para todos los años. En la Fig. 2 se presentan los campos seleccionados para la descarga de información en el formato CSV. La información se descarga en archivos separados por año, generando un total de 10 archivos CSV con la información original. Cada año cuenta con menos de 20.000

publicaciones, excepto en 2021, donde se obtuvieron un total de 24.542 publicaciones. Debido a las limitaciones de descarga, se priorizó la obtención de los primeros 20.000 artículos de 2021 con la métrica CiteScore más alta, asegurando así que se seleccionaran las publicaciones más relevantes. La distribución de publicaciones por año se muestra en la Tabla 2.

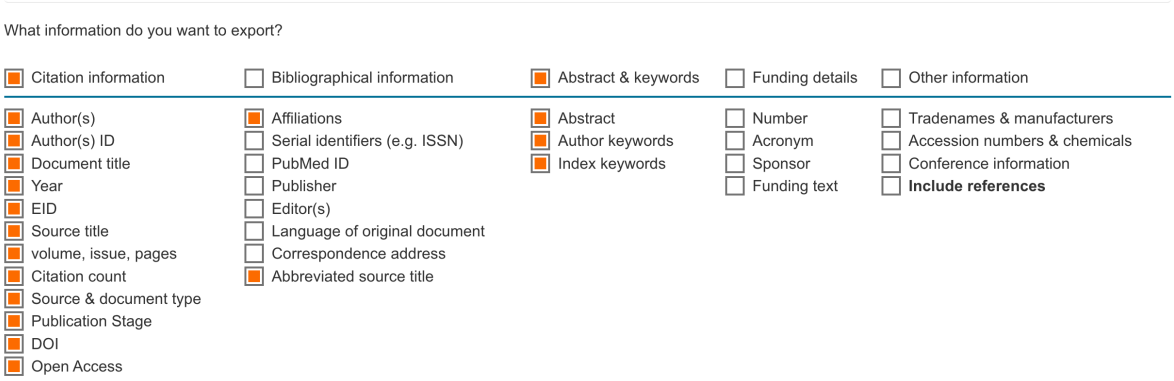


Fig. 2. Información escogida para la descarga de publicaciones en el portal oficial de Scopus

Tabla 2. Distribución de publicaciones por año originales y luego de preprocesamiento

| Año | Número de publicaciones originales | Número de publicaciones luego preprocesamiento |
|---|---|---|
| 2012 | 1.579 | 1.557 |
| 2013 | 2.007 | 1.983 |
| 2014 | 2.405 | 2.364 |
| 2015 | 3.302 | 3.240 |
| 2016 | 3.956 | 3.898 |
| 2017 | 5.077 | 4.986 |
| 2018 | 8.032 | 7.925 |
| 2019 | 13.951 | 13.732 |
| 2020 | 18.787 | 18.598 |
| 2021 | 20.000 | 19.766 |
| Total | 79.096 | 78.049 |

### 3.1.3 Preprocesamiento y almacenamiento

Una vez descargados los archivos por año en formato CSV, se creó un archivo único concatenando estos. En el resto de este documento, se referirá a este archivo como almacenamiento de las publicaciones originales.

En términos generales, la aplicación de la metodología requiere al menos 4 campos de metadatos por publicación: año, resumen, título y palabras clave indexadas, a los que se agrega el EID como identificador único de cada publicación en la base de datos. Cada uno de estos campos tiene un preprocesamiento diferencial, como se puede observar en la Fig. 3 Se consideran completamente duplicadas las publicaciones con resúmenes o palabras clave indexadas idénticas y que difieren en el resto de los campos, por lo que se eliminan de la base de datos. Además, el resumen suele tener una marca de Copyright que no aporta información semántica ni sintáctica al análisis, la cual se elimina en cada publicación donde se detecta. Por último, se eliminan las filas con campos vacíos, se convierten todos los textos a minúsculas y se almacenan en formato JSON (el cual es requerido para la librería cso-classifier) con el EID como llave y los 4 campos mencionados como valores.

Al finalizar el preprocesamiento, se pasa de un total de 79.096 publicaciones a 78.049 publicaciones. La distribución de publicaciones resultantes por año se puede observar en la Tabla 2.

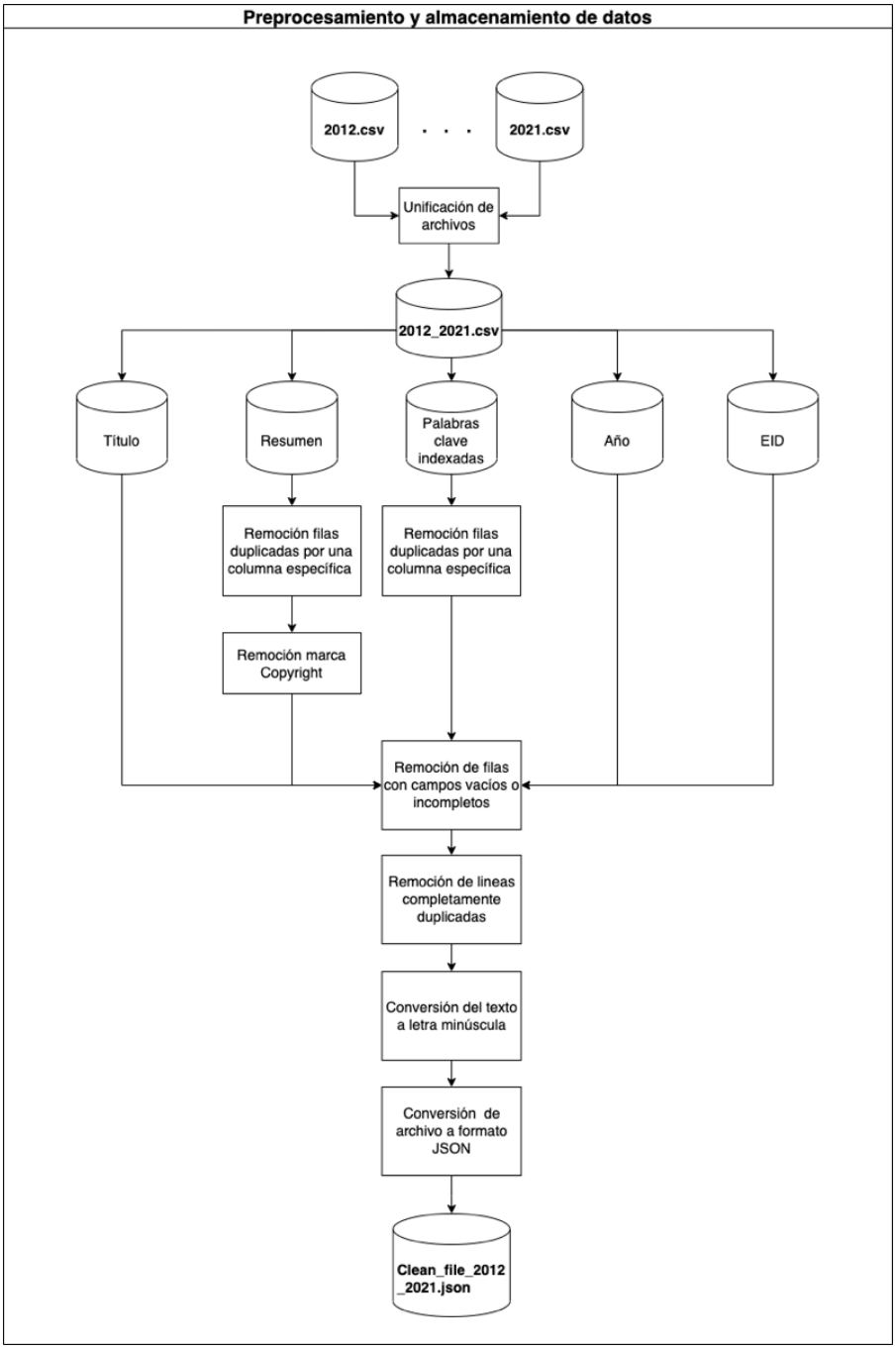


Fig. 3. Flujo del proceso de preprocesamiento y almacenamiento de datos original

### 3.1.4 Aplicación

El primer paso, desde los datos, en la ejecución de la metodología consiste en extraer por cada publicación contenida en el archivo tipo JSON, una lista de temas y super temas

referentes a la Ontología de Ciencia de la Computación que caractericen al resumen de la publicación. Para esta tarea se usa la librería cso-classifier con los parámetros descritos en la Tabla 3. Dado que los tiempos de procesamiento de la clasificación son significativamente altos por documento con los parámetros configurados, se decide tomar una muestra aleatoria estratificada con respecto al año de 20.000 publicaciones. La distribución de publicaciones resultantes por año se puede observar en la Tabla 4.

Al obtener la lista de temas y super temas, se lleva a cabo una depuración basada en el conocimiento experto, como se muestra en la Figura 4. Esta depuración consta de dos pasos: en primer lugar, se eliminan los términos que se consideren genéricos, es decir, que no se centren en una técnica o método específico de interés en Machine Learning o cuya frecuencia en el texto sea desproporcionada con respecto al resto de términos. En segundo lugar, se unifican los temas que se consideran sinónimos. El objetivo de esta depuración es descartar y unificar temas que puedan interferir con el funcionamiento del algoritmo de mapas temáticos de Bibliometrix, que se basa en dar igual peso a todos los términos y su frecuencia. Si este paso no se lleva a cabo, existe el riesgo de que los términos sin depurar dominen el agrupamiento de las temáticas emergentes o se creen grupos genéricos que no sean interpretables ni permitan discriminar técnicas y/o métodos.

### 3.1.5 Resultados

Finalmente, los resultados del paso anterior se exportan a un archivo en formato CSV, que fue elegido el medio para ingresar los datos a Bibliometrix. Esta elección se basa en la compatibilidad de dicho formato con la herramienta. Además, el archivo CSV permite almacenar de manera estructurada la información obtenida, lo que facilita su posterior análisis y procesamiento. Así, se asegura la integridad y la transferencia eficiente de los resultados hacia el siguiente paso.

En la Fig. 5, se presenta una comparativa visual entre las nubes de palabras generadas a partir del texto original sin remoción de términos genéricos (esquina superior izquierda), el texto con términos genéricos eliminados (esquina superior derecha) y el texto final con términos genéricos y sinónimos removidos (esquina inferior izquierda). Este análisis visual revela que al eliminar términos como "Machine Learning" o "Artificial Intelligence", que

representan la frecuencia dominante en la muestra pero no aportan un mayor conocimiento semántico para entender las temáticas, surgen términos específicos de interés como "classifiers" o "support vector machine". Este enfoque permite destacar los términos apropiados, lo que resulta en una representación más precisa y enriquecedora de los temas abordados. Además, se observa la aparición de la técnica "SVM", que hace referencia a la unificación de sinónimos de término "support vector machine". Esta identificación precisa y consolidada de términos relacionados contribuye a una comprensión más completa de los patrones y tendencias presentes en el corpus, enriqueciendo así el análisis de los resultados obtenidos.

Tabla 3. Parámetros configurados para el uso de cso-classifier

| Parámetro | Descripción | Valor |
|---|---|---|
| Workers | Número de hilos que se usarán para la ejecución de una clasificación [42] | 4 |
| Modules | El módulo por usar: sintáctico (*syntactic*), semántico (*semantic*) o ambos (*both*) [42] | both |
| Enhancement | Controla la inferencia de los súper temas: primer súper tema (*first*), todos los súper temas detectados (*all*) o ninguno (*no*) [42] | first |
| Explanation | Retorna o no una explicación de la clasificación | False |
| Delete_outliers | Usa o no usa el módulo de detección de datos anómalos del postprocesamiento [42] | True |
| Fast_classification | Usa o no una versión más simple del módulo semántico. El uso de la versión simple es 15 veces más rápido en tiempo de procesamiento que el modelo completo [42] | True |
| Silent | Retorna o no el progreso de la clasificación en pantalla durante la ejecución [42] | False |

Tabla 4. Distribución de publicaciones a las que se aplica CSOClassifier

| Año | Número de publicaciones de la muestra |
|---|---|
| 2012 | 405 |
| 2013 | 492 |
| 2014 | 626 |
| 2015 | 823 |
| 2016 | 946 |
| 2017 | 1.283 |
| 2018 | 1.994 |
| 2019 | 3.480 |
| 2020 | 4.841 |
| 2021 | 5.110 |
| Total | 20.000 |

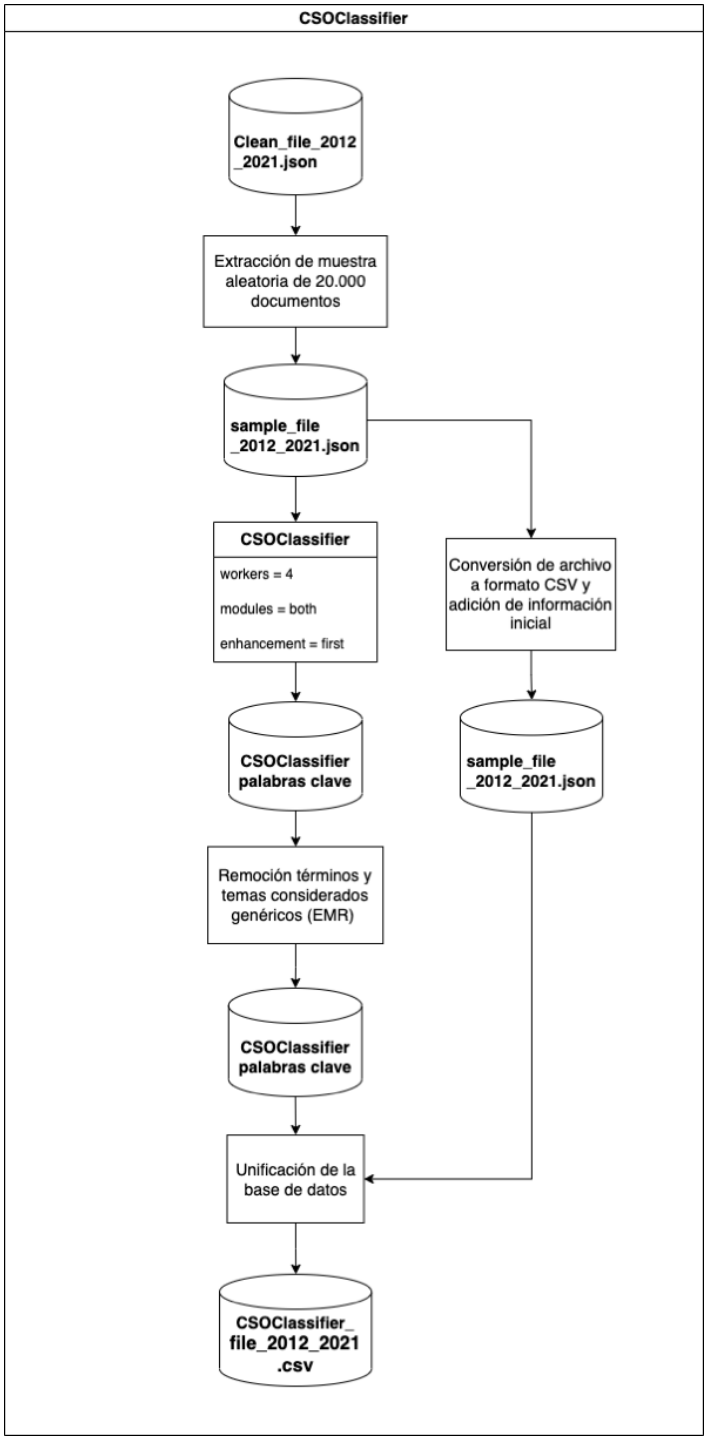


Fig. 4. Flujo de CSOClassifier para un modelo

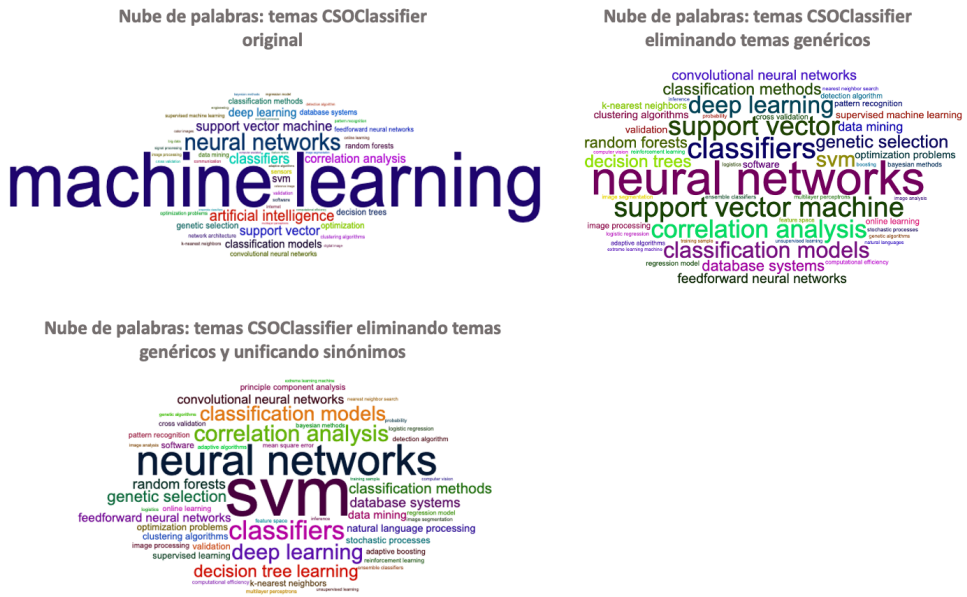


Fig. 5. Nubes de palabras para cada tipo de depuración de conocimiento experto

## 3.2 Paso 2: Mapa temático de Bibliometrix

### 3.2.1 Aplicación

El segundo paso dentro de la metodología consiste en usar las funciones thematicMap y thematicEvolution de Bibliometrix para detectar temas en desarrollo y emergentes, y a su vez analizar su evolución en el tiempo. Los parámetros usados se encuentran descritos en la Tabla 5.

Para analizar la evolución de los temas en el área de investigación, es necesario establecer intervalos de tiempo que agrupen el comportamiento de la tendencia con respecto al número de publicaciones por año. Como se puede apreciar en la Fig. 6., se han presentado cambios en la tendencia del número total de publicaciones desde el año 2012 hasta el 2021, lo que indica cambios estructurales dentro del área de investigación. Por lo tanto, se han configurado los intervalos de tiempo de la siguiente manera: el primer periodo comprende los años 2012 hasta 2014, el segundo periodo abarca los años 2015 a 2018, el tercer periodo solamente incluye el año 2019, al igual que el cuarto que se enfoca en el

año 2020, y finalmente, el quinto periodo corresponde al año 2021, que es donde se identificarán los temas en desarrollo y emergentes.

Al analizar la evolución de los temas a través de los intervalos de tiempo se detectan cambios estructurales dentro del área de Machine Learning y la granularidad de las técnicas que lo componen. Como observa en la Fig. 7, se evidencian cambios como el crecimiento de las implementaciones de las redes neuronales entre 2015 y 2018 que se derivan en su posterior uso, en 2019, en campos como la biométrica, robótica y el procesamiento de lenguaje natural, además de la expansión de la infraestructura necesaria en computación en la nube. Entre 2019 y 2020, se observa una evolución en las aplicaciones de las redes neuronales como el aprendizaje profundo, los vehículos autónomos y la introducción al análisis de sentimientos como un subtema derivado de la disponibilidad de información en redes sociales. Entre 2020 y 2021, se consolidan aplicaciones vigentes de las redes neuronales como los sistemas de recomendación, las redes convolucionales y el aprendizaje por refuerzo.

La detección de temáticas básicas y nicho se realiza específicamente para el periodo 2021, y los grupos de temas corresponden a técnicas estándar en la comunidad y subtemas nuevos que inician su desarrollo. Como se muestra en la Fig. 8, el hecho de que los cuatro cuadrantes del mapa cuenten con al menos tres grupos indica una granularidad de términos y una agrupación entre técnicas y aplicaciones. En el caso de los temas básicos, se agrupan técnicas y aplicaciones que se consideran clásicas e indispensables en cualquier equipo de Machine Learning conformado en la actualidad, como los algoritmos de clustering, la diferencia entre el aprendizaje supervisado y no supervisado, y técnicas como las máquinas de soporte vectorial y la minería de datos. Por otro lado, los temas nicho se refieren a términos de comunidades pequeñas pero que han surgido a partir de algunos temas ya desarrollados, como las expresiones emocionales derivadas del procesamiento de lenguaje natural o el análisis de imágenes fáciles derivado del reconocimiento de patrones y de la visión por computadora.

Tabla 5. Parámetros configurados para el uso de thematicMap

| Parámetro | Descripción | Valor |
|---|---|---|
| n | Número de términos a incluir | 1.000 |
| minfreq | Frecuencia mínima de un grupo | 5 |
| stemming | Usa o no usa técnica de *stemming* con el algoritmo *Porter* | False |
| n.labels | Número de etiquetas por cada grupo | 1 |
| community.repulsion | Indica la repulsión entre los elementos de la red | 0.1 |
| cluster | Algoritmo de detección de comunidades a usar | Walktrap |

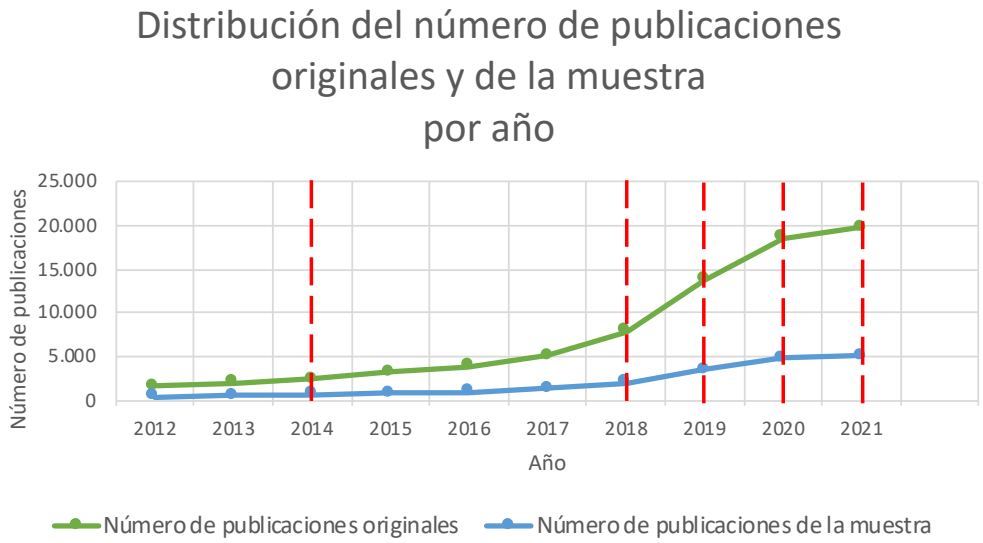


Fig. 6. Tendencia de la muestra de publicaciones en el tiempo y definición de intervalos de análisis

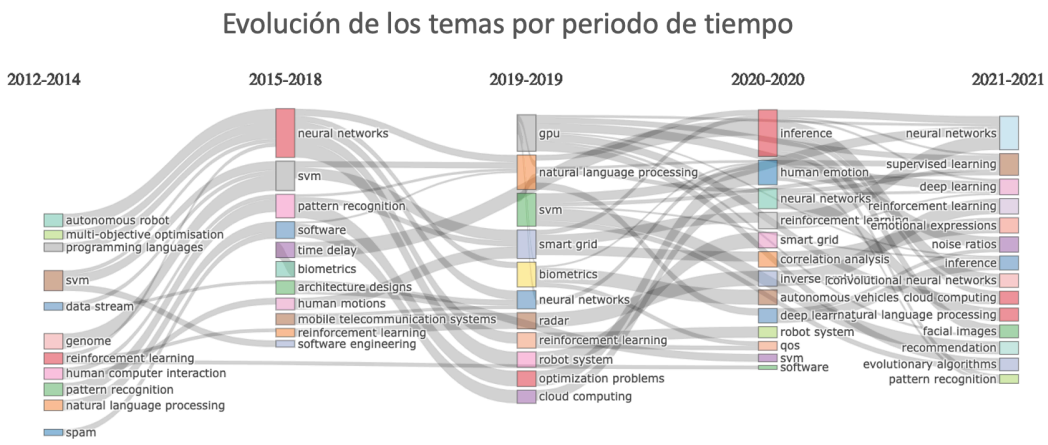


Fig. 7. Evolución de temas por intervalos de tiempo

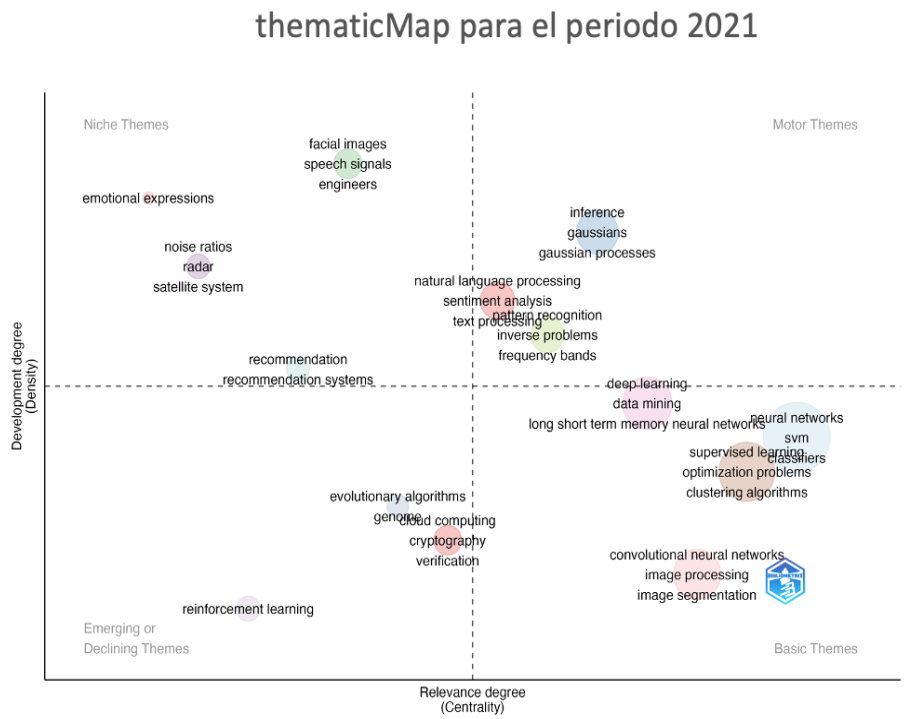


Fig. 8. thematicMap para el periodo 2021

## 3.2.2 Resultados

La detección de temáticas en desarrollo y emergentes se llevó a cabo específicamente para el periodo de 2021, donde se observaron cambios estructurales en la evolución en

los intervalos, y se evidenció una mayor granularidad en los métodos y técnicas en cada grupo. En el caso de los temas en desarrollo, se destacan el procesamiento de lenguaje natural, la inferencia asociada a procesos gaussianos y el reconocimiento de patrones enfocado al procesamiento de señales. Por su parte, dentro de los temas considerados emergentes se encuentran el aprendizaje por refuerzo, la criptografía aplicada a la computación en la nube y los sistemas de recomendación. Aunque estos últimos no se encuentran en el cuadrante inferior izquierdo, su nivel de centralidad y densidad los ubica como temas emergentes debido al crecimiento de sus aplicaciones en el mundo de la publicidad y el e-commerce.

En la Tabla 6 se observan los términos más frecuentes para cada grupo identificado como en desarrollo y emergente.

Tabla 6. Temas en desarrollo y emergentes con sus términos más frecuentes resultado de thematicMap

| Tipo de tema | Nombre del tema | Términos más frecuentes |
|---|---|---|
| En desarrollo | Inferencia | Procesos Gausianos, métodos numéricos, distribuciones de probabilidad, redes neuronales fuzzy, inferencia probabilística |
| En desarrollo | Procesamiento de Lenguaje Natural | Análisis de sentimientos, procesamiento de texto, clasificación, minería de texto, *embeddings* de palabras |
| En desarrollo | Reconocimiento de patrones | Detección de anomalías, datos dimensionales, detección de señales, detección de desempeño |
| Emergentes | Computación en la nube y criptografía | Sistemas en tiempo real, verificación, ciberseguridad |
| Emergentes | Aprendizaje por refuerzo | Aprendizaje por refuerzo |
| Emergentes | Sistemas de recomendación | Sistemas de recomendación |

## 3.3 Paso 3: BERTopic

### 3.3.1 Preprocesamiento y almacenamiento

Como primer paso para el entrenamiento de BERTopic, es necesario realizar la preprocesamiento de los resúmenes de las publicaciones que se utilizarán como entrada para la detección de temas. Para ello, se utiliza el archivo resultante del flujo descrito en la Fig. 3 y se le añaden pasos adicionales al proceso.

En primer lugar, se realiza la eliminación de términos considerados genéricos dentro de los resúmenes. A continuación, se lleva a cabo un proceso de detección de partes del lenguaje, conocido como Part-Of-Speech (POS), con el fin de eliminar preposiciones y pronombres que no contribuyen al entendimiento del texto.

Finalmente, se convierte el archivo de salida en formato JSON, incluyendo el campo EID como identificador y el año de publicación, que se utilizará para analizar la evolución de los temas en el tiempo. El flujo modificado presentado en la Fig. 9.

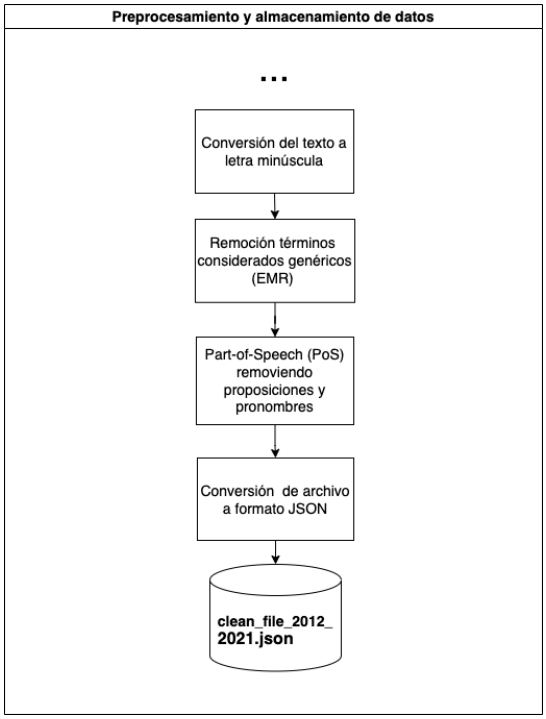

Fig. 9. Modificación al flujo de preprocesamieno y almacenamiento para BERTopic

### 3.3.2 Entrenamiento

Para el entrenamiento se utilizan 78.051 resúmenes de documentos correspondientes al periodo de 2012 a 2021 usando los parámetros de entrenamiento se establecen para cada uno de los componentes, tal y como se muestra en la Tabla 7. Ese proceso se divide en dos partes complementarias. En primer lugar, se lleva a cabo la creación de los embeddings de un modelo base automático que se puede almacenar para futuras modificaciones debido al alto costo computacional. En segundo lugar, se realiza la explotación del primer modelo automático mediante fine-tuning para mejorar los resultados.

En la primera etapa del proceso, que se llevó a cabo en un ordenador tipo MacBook Pro Corei5 de 4 núcleos del año 2019, se emplearon aproximadamente 2 horas y se obtuvieron un total de 778 temas: 776 temas densos y 2 temas que agrupan los datos considerados atípicos o genéricos.

Los temas se encuentran organizados de manera ascendente según su frecuencia dentro del conjunto de documentos. En la Fig. 10 se muestra el top 20 de los temas más frecuentes, de los cuales aproximadamente el 70% corresponden a temas de aplicación y solo el 30% a técnicas y métodos propios de Machine Learning. Además, se observaron términos duplicados en varios temas, cuyas palabras complementarias no ofrecen interpretabilidad. Por ejemplo, el tema 2 y el tema 7 se refieren al internet de las cosas (IoT por sus siglas en inglés) como su término principal, pero la frecuencia de las palabras complementarias no es concluyente sobre el tipo de aplicación o técnicas a las que se refieren. Esto indica que el modelo automático generado por el entrenamiento no agrupa correctamente las técnicas y casos de aplicación para que sean interpretables por el usuario.

Dados los resultados, se pasa al segundo punto del entrenamiento en donde se utiliza la función *reduce_topics* para reducir el número de temas a 50 y generar mejor agrupamiento de los términos. En la Fig. 11, se muestra el top 20 de los temas más frecuentes dado el

fine-tuning, y se puede observar que tanto los temas de aplicación como las técnicas y métodos presentan términos diferenciales y son interpretables.

Tabla 7. Parámetros usados para el entrenamiento de BERTopic

| Componente | Parámetros usados |
|---|---|
| Calculate_probabilties | True |
| ctfidf_model | ClassTfidfTransformer() |
| embedding_model | <bertopic.backend._sentencetransformers.SentenceTransformerBackend at 0x1402189d0> |
| hdbscan_model | HDBSCAN(min_cluster_size=10, prediction_data=True) |
| language | english |
| low_memory | False |
| min_topic_size | 10 |
| n_gram_range | (1, 1) |
| nr_topics | None |
| representation_model | None |
| seed_topic_list | None |
| top_n_words | None |
| umap_model | UMAP(angular_rp_forest=True, low_memory=False, metric='cosine', min_dist=0.0, n_components=5, tqdm_kwds={'bar_format': '{desc}: {percentage:3.0f}%\| {bar} {n_fmt}/{total_fmt} [{elapsed}]', 'desc': 'Epochs completed', 'disable': True}) |
| vectorizer_model | CountVectorizer() |
| verbose | True |

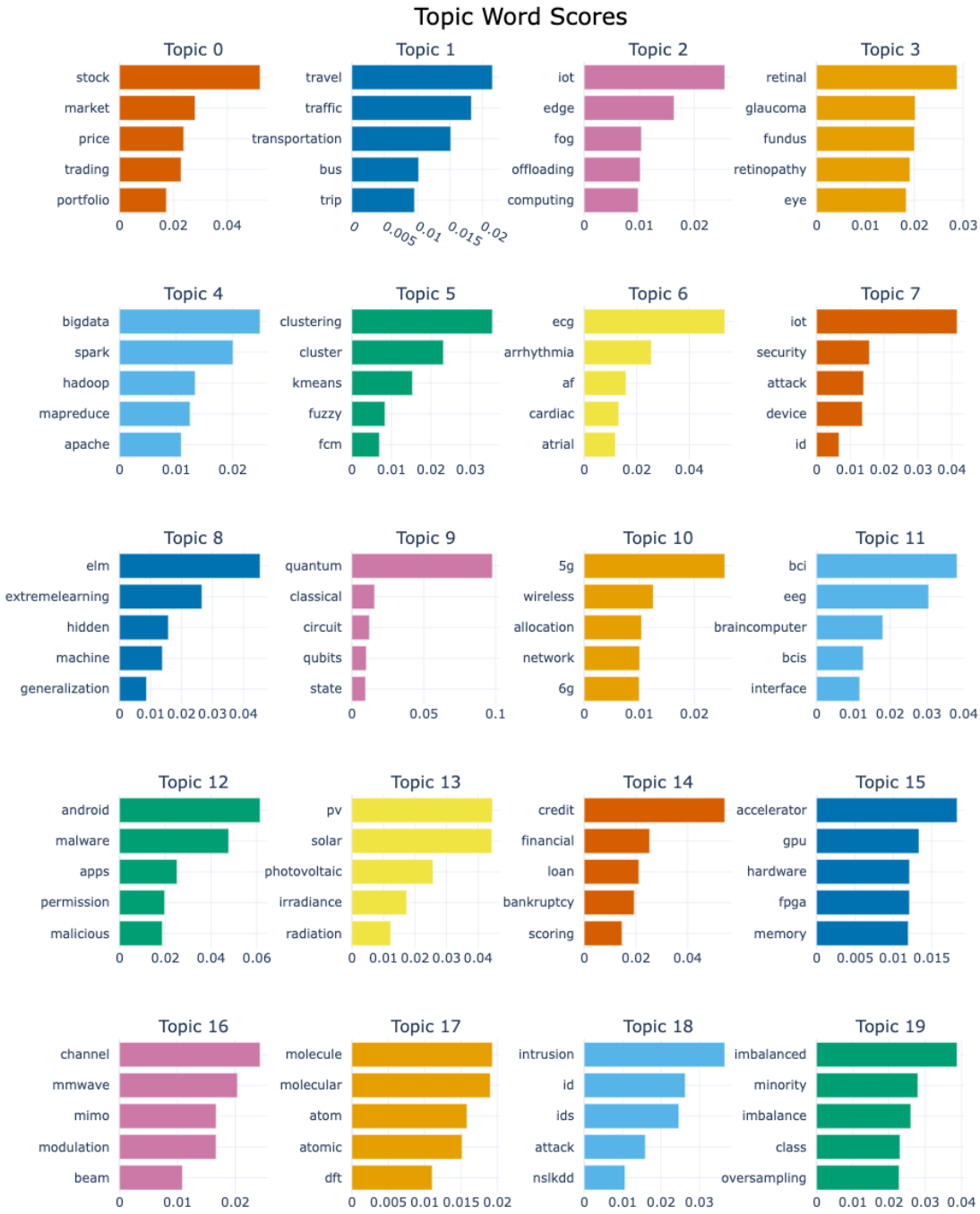


Fig. 10. Top 20 de temas detectados por BERTopic en su entrenamiento inicial

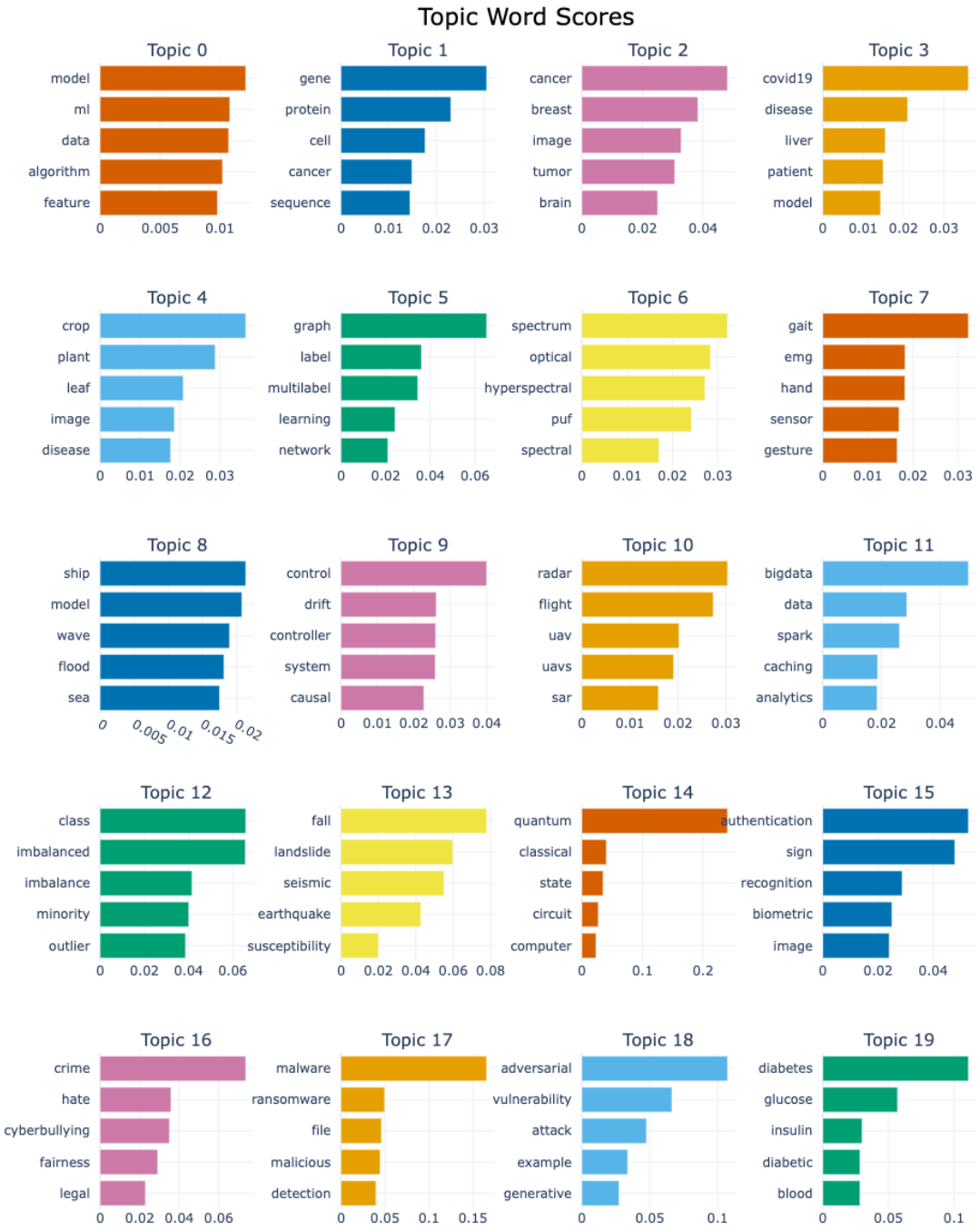


Fig. 11. Top 20 de temas detectados por BERTopic para reducción a 50 temas

## 3.4 Priorización

Para concluir el análisis de los temas identificados por BERTopic, es necesario realizar su interpretación y evaluar su nivel de emergencia a lo largo del tiempo. Para lograr esto, los temas son clasificados manualmente por un experto en tres categorías distintas. El primer tipo abarca los campos de aplicación en los que se evidencia un área de aplicación, pero no se detectan palabras clave relacionadas con métodos de Machine Learning. El segundo grupo se define como categoría técnica, en la cual se identifican exclusivamente términos técnicos en desarrollo e inherentes a Machine Learning. Por último, se consideran los híbridos que consisten en temas que presentan términos técnicos y, además, identifican dentro del mismo tema su correspondiente campo de aplicación. Estos últimos se consideran los más importantes dentro del análisis ya que vinculan técnicas directamente con sus aplicaciones y esta especificidad hacia donde enfocar los esfuerzos es beneficiosa para los grupos involucrados en la investigación de Machine Learning.

Una vez se han identificado los temas granulares e interpretables, resulta imperativo analizar su comportamiento a lo largo del tiempo con el fin de determinar su nivel de emergencia durante los períodos más recientes de la muestra. La evolución de los temas a lo largo del tiempo se evalúa mediante su frecuencia dentro del corpus. En BERTopic, esta frecuencia se define como la cantidad de documentos en los que un tema está presente. En otras palabras, para cada periodo t, un documento solo puede ser asignado a un tema.

Como se observa en la Fig. 12, la frecuencia normalizada de los 20 temas top detectados presentan, a nivel general, una tendencia positiva con estacionalidades diferentes. Cabe resaltar que durante el periodo de 2019 a 2020, la pendiente de la frecuencia aumenta significativamente para algunos temas mientras que otros inician el descenso. El aumento significativo y sostenido del número de documentos que contiene un tema denota un signo de emergencia.

Para determinar el nivel de emergencia de los 50 temas granulares detectados, se usa la métrica de cambio porcentual entre la frecuencia de un tema dentro de un año específico y su año inmediatamente anterior. Los temas considerados emergentes serán los que

tengan un cambio porcentual relativo mayor al promedio del periodo, para cualquiera de los dos periodos. A continuación, los pasos necesarios para realizar la priorización:

1. Se calcula el cambio porcentual de la frecuencia entre temas para los periodos 2018 a 2019 (periodo 1) y 2019 a 2020 (periodo 2).
2. Se calcula el promedio del cambio porcentual por periodo.
3. Se extraen los temas cuyo cambio porcentual supere el promedio del periodo, para cualquiera dos periodos.

Los temas emergentes granulares técnicos e híbridos se relacionan con las subáreas obtenidas en el paso 2 de la metodología mediante los mapas temáticos de Bibliometrix. Esto genera una lista de subáreas emergentes y temas granulares que las describen.

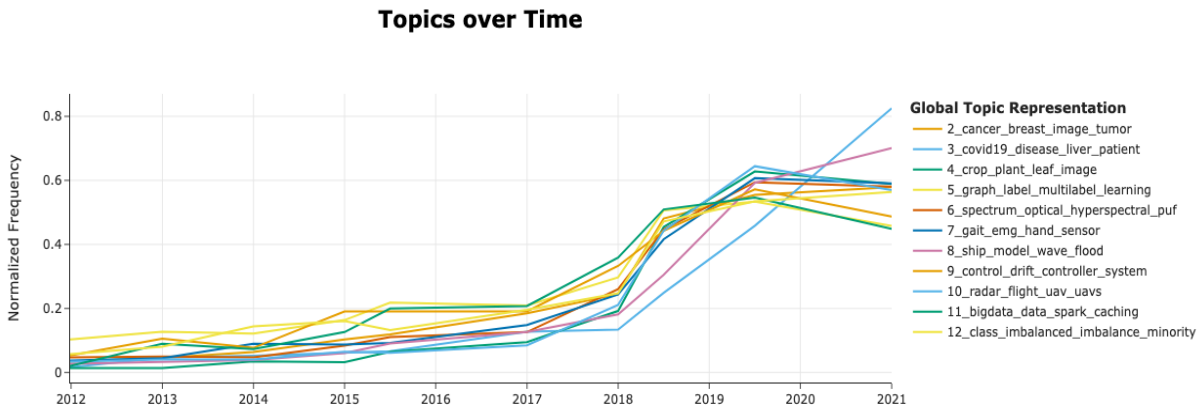


Fig. 12. Evolución de los temas detectados por BERTopic

## *3.5* Resultados

Como resultado de la clasificación manual de los temas, se identificaron 2 temas considerados generales o atípicos, 22 temas de aplicación, 14 temas técnicos y 12 temas hibridos. Entre los temas más frecuentes dentro del grupo de aplicaciones se encuentran el uso de modelos de Machine Learning para la investigación de receptores biológicos en

el tratamiento del cáncer, el uso de imágenes médicas con el mismo propósito y las investigaciones relacionadas con la pandemia de COVID-19. En cuanto a los temas técnicos más frecuentes, se destacan el aprendizaje basado en grafos, la detección de espectros electromagnéticos y los algoritmos de Big Data como Spark. Por último, en cuanto a los temas híbridos, se incluyen las imágenes agrícolas, los sensores electromagnéticos y los sistemas de control para vehículos autónomos.

En la Tabla 8 se identificaron un total de 26 temas que superaron la frecuencia promedio anual en al menos uno de los 50 temas evaluados. De estos, 13 son temas de aplicación, 7 son temas técnicos y 6 son híbridos. Además, de los 26 temas, 9 superaron el promedio durante el periodo 1, 15 durante el periodo 2 y solo 2 mantuvieron un crecimiento por encima del promedio en ambos periodos. Estos últimos temas están relacionados con la detección de terahertz en radiación electromagnética y el uso de modelos para combatir la pandemia de COVID-19.

En la Tabla 9, se muestran los temas técnicos que superan el promedio de los periodos asociados a la subárea identificada en el mapa temático que guarda relación con dicho tema. Sin embargo, existe un tema específico que no puede ubicarse dentro de ninguna de las subáreas y se refiere al desarrollo y escalado de técnicas de aprendizaje basado en grafos. Este tipo de algoritmos experimentó un aumento significativo entre los años 2021 y 2022, especialmente con el escalado de algoritmos de agrupamiento.

Tabla 8. Temas detectados por BERTopic que superan el promedio de frecuencia por periodo

| Nombre del tema | Descripción | Clasificación | Periodo donde supera el promedio |
|---|---|---|---|
| 48_superconductors_superconductor_material_superconducting | Uso de técnicas de Machine Learning para detectar nuevos materiales superconductores | Aplicación | Periodo 1 |
| 44_dental_caries_teeth_tooth | Uso de técnicas de Machine Learning y Deep Learning para analizar imágenes | Aplicación | Periodo 1 |

| | | | |
|---|---|---|---|
| | odontológicas para diagnóstico temprano | | |
| 41_sarcasm_sarcastic_irony_figurative | Desarrollo de técnicas y metodologías para detectar el sarcasmo dentro del procesamiento de lenguaje natural | Aplicación-técnico | Periodo 1 |
| 38_patent_technology_technological_apple | Inscripción de patentes de métodos de Machine Learning en la industria tecnológica | Aplicación | Periodo 1 |
| 37_saliency_salient_map_visual | Desarrollo de técnicas de visión por computador para resaltar áreas de concentración en una imágen | Técnico | Periodo 1 |
| 34_multiview_view_clustering_learning | Desarrollo y escalamiento de técnicas de aprendizaje basado en grafos, especialmente agrupamiento usando grafos | Técnico | Periodo 1 |
| 28_missing_imputation_value_data | Desarrollo de técnicas para imputar datos faltantes o defectuosos | Técnico | Periodo 1 |
| 20_seizure_eeg_epilepsy_epileptic | Uso de técnicas de Machine Learning y Deep Learning para analizar resonancias magnéticas y electroencefalogramas para diagnosticar crisis de epilepsia | Aplicación | Periodo 1 |
| 8_ship_model_wave_flood | Uso de modelos de Machine Learning para la predicción y clasificación de eventos de inundación | Aplicación | Periodo 1 |

| | | | |
|---|---|---|---|
| 46_thz_terahertz_spectroscopy_contraband | Uso de modelos de Machine Learning para la detección de terahertz en radiación electromagnética | Aplicación | Periodo 1 y 2 |
| 3_covid19_disease_liver_patient | Uso de modelos de Machine Learning y Deep Learning para combatir la pandemia por Covid-19 | Aplicación | Periodo 1 y 2 |
| 47_discretization_continuous_attribute_interval | Desarrollo de algoritmos de discretización de datos | Técnico | Periodo 2 |
| 45_granule_granular_fuzzy_computing | Desarrollo de técnicas para solucionar incertidumbre o información completa | Técnico | Periodo 2 |
| 43_occupancy_building_occupant_energy | Uso de técnicas de Machine Learning para la predicción de la ocupación de infraestructura aplicado al cálculo de energía requerida | Aplicación | Periodo 2 |
| 40_corrosion_steel_stray_corrosioninduced | Uso de técnicas de Machine Learning para la predicción de la corrosión | Aplicación | Periodo 2 |
| 39_tourism_tourist_hotel_customer | Desarrollo de sistemas de recomendación para turismo | Aplicación-técnico | Periodo 2 |
| 33_compiler_automl_ml_compilation | Desarrollo de aplicaciones y técnicas para Automatic Machine Learning | Técnico | Periodo 2 |
| 31_biosensors_concentration_nanoparticles_nanomaterials | Uso de técnicas de Machine Learning para la detección de bioreceptores para el desarrollo de biosensores | Aplicación | Periodo 2 |

| | | | |
|---|---|---|---|
| 30_antenna_electromagnetic_frequency_rfid | Aplicación de técnicas de detección de anomalias para señales electromagnéticas | Aplicación-técnico | Periodo 2 |
| 26_docking_scoring_tunnel_islanding | Uso de técnicas de optimización estocástica para análisis molecular | Aplicación-técnico | Periodo 2 |
| 25_fetal_sperm_pregnancy_birth | Uso de técnicas de Machine Learning para pronosticar la supervivencia de un embrión | Aplicación | Periodo 2 |
| 24_rainfall_weather_drought_forecast | Uso de modelos de Machine Learning y Deep Learning para la predicción de sequías | Aplicación | Periodo 2 |
| 23_equation_pdes_differential_numerical | Desarrollo de algoritmos para la implementación de redes neuronales | Técnico | Periodo 2 |
| 16_crime_hate_cyberbullying_fairness | Aplicación de técnicas de procesamiento lenguaje natural para la detección y contención de speech cyberbulling | Aplicación-técnico | Periodo 2 |
| 14_quantum_classical_state_circuit | Desarrollo de técnicas de Machine Learning en computación cuántica | Aplicación-técnico | Periodo 2 |
| 13_fall_landslide_seismic_earthquake | Uso de modelos de Machine Learning y Deep Learning para detección de deslizamientos de tierra | Aplicación | Periodo 2 |

Tabla 9. Resultados de temas detectados por mapas temáticos y su relación con temas detectados por BERTopic

| Tipo de tema | Nombre del tema | Términos más frecuentes | Temas técnicos BERTopic asociados |
|---|---|---|---|

| | | | |
|---|---|---|---|
| En desarrollo | Inferencia | Procesos Gausianos, métodos numéricos, distribuciones de probabilidad, redes neuronales fuzzy, inferencia probabilística | 47_discretization_continuous_attribute_interval, 45_granule_granular_fuzzy_computing, 26_docking_scoring_tunnel_islanding |
| En desarrollo | Procesamiento de Lenguaje Natural | Análisis de sentimientos, procesamiento de texto, clasificación, minería de texto, *embeddings* de palabras | 41_sarcasm_sarcastic_irony_figurative, 16_crime_hate_cyber bullying_fairness |
| En desarrollo | Reconocimiento de patrones | Detección de anomalías, datos dimensionales, detección de señales, detección de desempeño | 30_antenna_electromagnetic_frequency_rfid |
| Emergentes | Computación en la nube y criptografía | Sistemas en tiempo real, verificación, ciberseguridad | 33_compiler_automl_ml_compilation, 14_quantum_classical_state_circuit |
| Emergentes | Aprendizaje por refuerzo | Aprendizaje por refuerzo | 37_saliency_salient_map_visual |
| Emergentes | Sistemas de recomendación | Sistemas de recomendación | 39_tourism_tourist_hotel_customer |

# 4. Discusión de los resultados

Al analizar los resultados de ambas técnicas de forma individual, se evidencia la capacidad superior de BERTopic para detectar una amplia variedad de temas de mayor complejidad en comparación con los detectados por el Clasificador de Ontología de Ciencias de la Computación. Esto se debe a la estructura de ambos modelos: BERTopic se especializa en analizar el contexto sintáctico y semántico de texto no estructurado para obtener temas densos sin supervisión, mientras que el Clasificador de Ontología de Ciencias de la Computación depende de la actualización de la taxonomía para obtener y clasificar los temas nuevos.

Una diferencia notable entre ambas detecciones es la presencia de datos atípicos. BERTopic excluye estos datos mediante el uso del algoritmo de agrupamiento HDBSCAN, mientras que el Clasificador de Ontología de Ciencias de la Computación es sensible a la aparición de términos genéricos que no describen el contexto, por lo que se hace necesario aplicar tareas de preprocesamiento experto en su pipeline.

Por otra parte, la integración de ambas técnicas ofrece una visión completa y actualizada de los temas detectados. Mientras que los mapas temáticos en conjunto con el Clasificador de Ontología de Ciencias de la Computación rastrean las subáreas oficiales existentes, BERTopic identifica temas latentes e incluso nuevos términos dentro de las comunidades científicas que no se aparecen dentro de la taxonomía del clasificador. Esta combinación proporciona una perspectiva completa para los grupos de investigación vinculados al Machine Learning.

Lo anterior, se evidencia en los resultados obtenidos en donde se observan cambios estructurales dentro de la comunidad científica. Por ejemplo, en los últimos años, se ha dado una amplia implementación de las redes neuronales, especialmente en el modelado de datos no estructurados como imágenes, audio y texto. Ambas técnicas de detección de temáticas capturan eficazmente el procesamiento del lenguaje natural y en el análisis de imágenes.

Otro caso que resalta dentro de los resultados es el aprendizaje por refuerzo ya que anteriormente el área Machine Learning se dividía en modelos supervisados y no supervisados, sin embargo, en los últimos años, los algoritmos de aprendizaje por refuerzo surgieron como un tercer tipo, especialmente a través de redes neuronales mediante las cuales se han extendido a todos los campos de aplicación. Ambas técnicas apuntan emergencia hacia este tema.

Cabe resaltar que, para lograr la extensión de temas nuevos, como el aprendizaje por refuerzo, es necesario el desarrollo de herramientas que soporten dichas implementaciones, tanto en la industria como en la comunidad científica. Este es el caso de la computación en la nube, la computación cuántica y el Machine Learning automático que se observan dentro de las subáreas de los mapas temáticos y dentro de los temas granulares extraídos con BERTopic.

Otro de los episodios que han marcado a la comunidad de investigadores, y que se refleja dentro de los temas detectados debido a su frecuencia elevada, ha sido la pandemia por COVID-19 que ha generado nuevos temas de investigación, especialmente con el auge del comercio electrónico a nivel mundial durante este periodo. Esto ha impulsado la aparición de sistemas de recomendación como algoritmos indispensables para cualquier negocio que involucre ventas o captación de usuarios de forma digital. Además, esta coyuntura reforzó el crecimiento acelerado del uso de Machine Learning e Inteligencia Artificial en los campos de la salud. Como evidencia de esto, 7 de los 21 temas detectados por BERTopic y clasificados como aplicaciones, se refieren directamente a nuevas modelos usados en el diagnóstico de enfermedades crónicas a través de imágenes o predicción de estado de salud de un paciente.

# 5. Conclusiones y recomendaciones

## 5.1 Objetivos

### 5.1.1 Objetivo general

El objetivo general de este trabajo fue definido como:

> Detectar las temáticas emergentes en el área de Machine Learning aplicando técnicas de minería de texto y procesamiento de lenguaje natural, generando un informe final con prioridades de temas abordar.

Este objetivo se cumplió de la siguiente manera: se compararon las metodologías y técnicas disponibles para la detección de temáticas emergentes, proponiendo así una metodología para el desarrollo del trabajo en el Capítulo 2. Posteriormente, se detectaron las temáticas emergentes en el área de Machine Learning mediante el uso de tres técnicas ensambladas de procesamiento de lenguaje natural: el Clasificador de Ontología de Ciencias de la Computación, los mapas temáticos y BERTopic, junto con varias técnicas de preprocesamiento, como se describe en el Capítulo 3. Además, se generó una lista final de temas emergentes detectados, que se recomienda como prioritaria para los equipos de Machine Learning que desarrollan actividades de I+D en el área abordada, tal como se menciona en el Capítulo 3. Finalmente, se lleva a cabo una discusión de resultados que contrasta la lista final de temas emergentes detectados con los cambios dentro del área de investigación, abordados en el Capítulo 4.

### 5.1.2 Objetivos específicos

En la etapa inicial de este trabajo, se llevó a cabo una detallada discusión acerca de las distintas técnicas disponibles para la detección de temáticas emergentes. Esta discusión abarcó desde el diseño de la muestra hasta algoritmos especializados con un componente temporal. Como resultado de este análisis, se propuso una metodología específica que se detalla y explora en profundidad en los Capítulos 1 y 2 de este trabajo de investigación.

Posteriormente, se procedió a aplicar la metodología propuesta, comenzando con una serie de tareas de preprocesamiento esenciales. Estas tareas incluyeron la obtención de los datos necesarios, la depuración de la muestra, la lematización de los textos, el análisis de Part-of-Speech y la eliminación de duplicados, entre otros aspectos fundamentales. Todos estos pasos se describen detalladamente en el Capítulo 3 del presente trabajo.

Además de abordar las tareas de preprocesamiento, en el Capítulo 3 también se exploran las etapas relacionadas con la detección de temáticas en el área de Machine Learning y la posterior priorización de las mismas. Como resultado de este análisis, se generó una lista de prioridades basada en los patrones emergentes identificados durante el estudio.

Finalmente, se llevó a cabo una discusión y validación de los resultados obtenidos, contrastándolos con la evolución de los temas en la comunidad científica y los cambios estructurales observados en el área de investigación estudiada. Esta discusión se desarrolla a lo largo del Capítulo 4, proporcionando una perspectiva crítica y una visión contextualizada de los hallazgos del estudio.

### 5.1.3 Evaluación de la hipótesis

En este trabajo se plantea la hipótesis de si es posible detectar temáticas emergentes en el área de Machine Learning que permita establecer una prioridad en los temas a abordar dentro de las actividades de Investigación y Desarrollo, mediante la utilización de técnicas de minería de texto y el procesamiento de lenguaje natural. La respuesta es que sí es posible; en este trabajo final se realizó la priorización de temáticas emergentes en el área de Machine Learning generando un informe de temas generales y específicos que permite a los equipos priorizar sus recursos escasos. El uso de esta priorización puede orientar a diversas instituciones en decisiones como temas de investigación a abordar, tecnologías a implementar o áreas de potencial inversión en investigación.

### 5.1.4 Recomendaciones

Para enriquecer la lista de prioridades se recomienda tomar información de registro de patentes e información comercial complementaria que genere mayor contexto a los temas emergentes detectados, lo cual permitirá que adicionalmente, se puedan tomar decisiones de inversión o de impacto comercial y no únicamente decisiones enfocadas a las actividades de Investigación y Desarrollo. Adicionalmente, se recomienda explorar la optimización de hiperparametros del modelo BERTopic, además de la exploración de experimentos con los componentes del mismo.

# 6. Bibliografía


[1] D. H. R. M. Daniele Rotolo, «What is an emerging technology?,» *Research Policy,* vol. 44, nº 10, pp. 1827-1843, 2015.

[2] Q. Wang, «A Bibliometric Model for Identifying Emerging Research Topics,» *Journal of the Association for Information Science and Technology,* vol. 69, nº 2, pp. 290-304, 2018.

[3] L. H. A. P. L. Shuo Xu, «Review on emerging research topics with key-route main path analysis,» *Scientometrics,* vol. 122, pp. 607-624, 2020.

[4] S. W. C. Alan L. Porter, «How Tech Mining Works,» de *Tech Mining*, New Jersey, John Wiley & Sons, INC, 2005, pp. 17-20.

[5] J. G. F. C. C. N. Alan L. Porter, «Emergence scoring to identify frontier R&D topics and key players,» *Technological Forecasting & Social Change,* vol. 146, pp. 628-643, 2019.

[6] E. M. Rogers, «Four Main Elements in the Diffusion of Innovations,» de *Diffusion of Innovations*, New York, NY, The Free Press, 1962, pp. 10-35.

[7] L. H. G. Y. K. L. X. A. Shuo Xu, «A topic models based framework for detecting and forecasting emerging technologies,» *Technological Forecasting & Social Change,* vol. 162, pp. 120-366, 2021.

[8] S. W. C. Alan L. Porter, «Finding the Right Sources,» de *Tech Mining*, New Jersey, John Wiley & Sons, INC, 2005, pp. 69-94.

[9] J. S. L. P. Y. Ying Huang, «A systematic method to create search strategies for emerging technologies based on the Web of Science: illustrated for ‘Big Data’,» *Scientometrics,* vol. 105, pp. 2005-2022, 2015.

[10] J. M. L. B. L. Zhentao Liang, «Combining deep neural network and bibliometric indicator for emerging research topic prediction,» *Information Processing and Management,* vol. 58, pp. 102-611, 2021.

[11] M. G. Christian Mühlroth, «Artificial Intelligence in Innovation: How to Spot Emerging Trends and Technologies,» *IEEE Transactions on Engineering Management ,* vol. 69, nº 2, 2022.

[12] D. C. N. C. N. Alan L. Porter, «Measuring tech emergence: A contest,» *Technological Forecasting & Social Change,* vol. 159, pp. 120-176, 2020.

[13] L. H. X. A. G. Y. F. W. Shuo Xu, «Emerging research topics detection with multiple machine learning models,» *Journal of Informetrics,* vol. 13, pp. 100-983, 2019.

[14] S. W. C. Alan L. Porter, «Technological Innovation and the Need for Tech Mining,» de *Tech Mining: Exploiting New Technologies for Competitive Advantage*, New Jersey, John Wiley & Sons, INC, 2005, pp. 3-8.

[15] D. J. Jackson, «What is an Innovation Ecosystem,» *National Science Foundation,* vol. 1, nº 2, pp. 1-13, 2022.

[16] J. Z. Manzoor Ahmad, «The Cyclical and Nonlinear Impact of R&D and Innovation Activities on Economic Growth in OECD Economies: a New Perspective,» *Journal of the Knowledge Economy,* 2022.

[17] OECD, «Gross domestic spending on R&D,» 2022. [En línea]. Available: https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm. [Último acceso: 18 04 2022].

[18] OECD, «Embrasing Innovation in Government: Global Trends,» OECD. Observatory of Public Sector Innovation, 2017.

[19] OECD, «OECD. Innovation Scoreboard. HERD as percentage,» OECD, 2022. [En línea]. Available: https://www.oecd.org/innovation/scoreboard.htm. [Último acceso: 18 04 2022].

[20] G. Medda, «External R&D, product and process innovation in European manufacturing companies,» *The Journal of Technology Transfer,* vol. 45, pp. 339-369, 2018.

[21] National Science Board (NSB), « U.S. and Global Science and Technology Capabilities,» National Science Foundation (NSF), 2022.

[22] National Center for Science and Engineering Statistics (NCSES), «New Data on U.S. R&D: Summary Statistics from the 2019–20,» *National Patterns of R&D Resources,* pp. 22-314, 2021.

[23] K. White, «Publication Output by Country, Region, or Economy and Scientific Field,» 2021. [En línea]. Available: https://ncses.nsf.gov/pubs/nsb20214/publication-output-by-country-region-or-economy-and-scientific-field. [Último acceso: 18 04 2022].

[24] N. M. E. B. J. E. T. L. J. M. H. N. J. C. N. M. S. E. S. Y. S. J. C. a. R. P. Daniel Zhang, «The AI Index 2022 Annual Report,» AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, 2022.

[25] World Intellectual Property Organization (WIPO), «IP Statistics Data Center,» Febrero 2022. [En línea]. Available: https://www.wipo.int/edocs/infogdocs/en/ipfactsandfigures/. [Último acceso: 18 04 2022].

[26] University of Groningen, «Information Literacy History: Types of databases,» 3 Marzo 2022. [En línea]. Available: https://libguides.rug.nl/c.php?g=470628&p=3283312. [Último acceso: 18 04 2022].

[27] «The History of the Scientific Journal,» 2022. [En línea]. Available: https://arts.st-andrews.ac.uk/philosophicaltransactions/brief-history-of-phil-trans/.

[28] Scopus, «Scopus Content Coverage Guide,» Elsevier, 2020.

[29] Clarivate, «Web of Science Coverage Details,» Clarivate, Agosto 2021. [En línea]. Available: https://clarivate.libguides.com/librarianresources/coverage. [Último acceso: 18 04 2022].

[30] Google Schoolar, «Inclusion Guidelines for Webmasters,» 2022. [En línea]. Available: https://scholar.google.com/intl/es/scholar/inclusion.html. [Último acceso: 18 04 2022].

[31] Vantage Point , «Vantage Points Products,» 2022. [En línea]. Available: https://www.thevantagepoint.com/products.html. [Último acceso: 18 04 2022].

[32] C. C. Aggarwal, de *Machine Learning for Text*, New York, Springer, 2018, p. 2.

[33] E. D. Liddy, «Natural Language Processing. In Encyclopedia of Library and Information Science 2nd Ed,» Marcel Decker Inc, New York, 2001.

[34] R. U. H. Ish Kumar Dhammi, «What is indexing,» *Indian Journal of Orthopaedics,* pp. 115-116, 2016.

[35] S. W. C. Alan L. Porter, «What Tech Mining Can Do for You,» de *Tech Mining*, New Jersey, John Wiley & Sons, INC, 2005, pp. 33-40.

[36] A. L.-H. E. H.-V. F. H. M.J. Cobo, «An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field,» *Journal of Informetrics,* vol. 5, nº 1, pp. 146-166, 2011.

[37] K.-S. Sr, «Bibliometrix,» 2023. [En línea]. Available: https://www.bibliometrix.org/home/index.php/layout/bibliometrix. [Último acceso: 02 04 2023].

[38] C. C. Massimo Aria, «bibliometrix: An R-tool for comprehensive science mapping analysis,» *Journal of Informetrics,* vol. 11, pp. 959-975, 2017.

[39] M. Aria, C. Cuccurullo, L. D'Aniello, M. Misuraca y M. Spano, «Thematic Analysis as a New Culturomic Tool: The Social Media Coverage on COVID-19,» *Sustainability ,* vol. 14, 2022.

[40] A. L.-H. E. H.-V. F. H. M.J. Cobo, «An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field,» *Journal of Informetrics,* vol. 5, pp. 146-166, 2011.

[41] T. T. A. M. A. B. F. O. E. M. Angelo A. Salatino, «The Computer Science Ontology: A Comprehensive Automatically-Generated Taxonomy of Research Areas,» *Data Intelligence,* vol. 2, nº 3, 2020.

[42] T. T. A. M. Angelo A. Salatino, «CSO Classifier Github Page,» Github, 01 04 2023. [En línea]. Available: https://github.com/angelosalatino/cso-classifier#about. [Último acceso: 05 04 2023].

[43] SBERT, «Pretrained Models,» 07 09 2022. [En línea]. Available: https://www.sbert.net/docs/pretrained_models.html. [Último acceso: 06 05 2023].

[44] A. P. Andy Coenen, «A deeper dive into UMAP theory,» Google, [En línea]. Available: https://pair-code.github.io/understanding-umap/supplement.html. [Último acceso: 06 05 2023].

[45] Google, «Algoritmos de agrupamiento,» Google, [En línea]. Available: https://developers.google.com/machine-learning/clustering/clustering-algorithms?hl=es-419. [Último acceso: 06 05 2023].

[46] M. Grootendorst, «BERTopic,» [En línea]. Available: https://maartengr.github.io/BERTopic/algorithm/algorithm.html. [Último acceso: 06 05 2023].

[47] A. S. F. O. Alessandra Belfiore, «Characterising Research Areas in the field of AI,» de *SIS2022 51a Reunión Científica de la Sociedad Estadística Italiana*, Caserta, 2022.

[48] G. V. S. E. S.-G. Kenji Contreras, «Using Topic Modelling for Analyzing Panamanian Parliamentary Proceedings with Neural and Statistical Methods,» de *2022 IEEE 40th Central America and Panama Convention (CONCAPAN)*, Panamá, 2022.

[49] M. Grootendorst, «BERTopic: Neural topic modeling with a class-based TF-IDF procedure,» de *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*, Netherlands, 2022.

[50] Scopus, «Scopus: Access and use Support Center,» Scopus, [En línea]. Available: https://service-elsevier-com.ezproxy.unal.edu.co/app/answers/detail/a_id/11234/supporthub/scopus/#anchor. [Último acceso: 06 04 2023].