Facultad de Ingeniería

INGENIERÍA INDUSTRIAL

Trabajo de Grado – Segundo Semestre 2020

[201022] Diseño de un algoritmo de minería de texto para realizar análisis semántico de publicaciones científicas sobre sistemas ciber Físicos.

Germán David González Morales[a,c] , Andrés Esteban Jácome Estupiñan[a,c] , Leidy Paola Pardo Cuellar[a,c], Daniel Esteban Salas Rodríguez[a,c]

Ana María Beltrán Cortés[b,c]

[a]*Estudiante de Ingeniería Industrial*
[b]*Profesor, Director del Proyecto de Grado, Departamento de Ingeniería Industrial*
[c]*Pontificia Universidad Javeriana, Bogotá, Colombia*

Abstract

The Sustainable Development Goals are the blueprint to achieve a better and more sustainable future for all. (*Take Action for the Sustainable Development Goals – United Nations Sustainable Development*, n.d.), and knowing their actual state, it is crucial to be aware of where each country stands and what they must do in order to accomplish the 2030 agenda. Nowadays fields of science, technology and innovation are key drivers of economic growth and development, making target 9.5 "Increase research and upgrade industrial technologies" important to reduce inequality gaps in the world. In Colombia, the National Planning Department (DNP) reports the monitoring figures for the SDGs in the country and in the case of target 9.5, it is measure based on the results of bibliometric analysis of scientific publications reported by the Colombian Observatory of Science and Technology (OCyT). The current investigation aimed to carry out a semantic and bibliometric analysis of academic publications regarding cyber physical systems in the Scopus database from the period 2017 to 2019, which purpose is to compare the results between the two analyses. A total of 1075 articles were analyzed with the bibliometric software Vantage Point and then was compared with an algorithm of text mining that was able to analyze semantic content executed in Python. The bibliometric results show how in recent years the number of publications has been growing exponentially specially for countries of the first world such as United States or China where the R&B spending is bigger unlike third World Countries as Colombia were there are only 8 publications of open access in the last three years. To complement the study using the text mining algorithm it was able to identify for the three most important topics: Engineering, Computer Science and Material Science what researchers are most focus on is the subtopics Network Security, Internet of things, embedded systems and Manufacture. It is important to clarify that SDG 9.5 is always rated with all the publications made since 2015, however this study was limited for publications of open access about industry 4.0 with a focus on cyber physical systems (CPS).

1. Justificación y planteamiento del problema

Naciones Unidas es una organización intergubernamental creada después de la Segunda Guerra Mundial con el objetivo de mantener la paz, seguridad y lograr la cooperación internacional en la solución de los problemas de carácter económico, social, cultural, humanitario y de desarrollo (Nations, n.d.). Para lograr su propósito en los años venideros, ha diseñado 17 objetivos de desarrollo sostenible (ODS) que buscan la prosperidad en el planeta y las personas. Para Colombia como país miembro de la ONU, "la Agenda 2030 es una oportunidad de generar transformaciones y dar impulso a temas de interés a nivel internacional, nacional y local" (La Agenda 2030 en Colombia - Objetivos de Desarrollo Sostenible, n.d.) haciendo explícita la importancia de aumentar la investigación y actualizar las tecnologías industriales (Objetivo 9.5) pues son motores fundamentales del crecimiento y desarrollo económico para el país.

En los últimos años con la llegada de la Industria 4.0 los sectores industriales y manufactureros están empezando a cambiar los esquemas de producción haciendo uso de alta tecnología en la integración del hombre con la máquina aumentando el valor agregado manufacturero (VAM) del 40,5% en el año 2000 al 44,7% en el 2016. Globalmente se observa que "En Asia oriental y sudoriental, así como en Europa y América del Norte, más del 47% del total del VAM procedía de sectores de alta tecnología; en cambio, en Oceanía (sin incluir Australia y Nueva Zelanda) y en África subsahariana, la proporción fue de solo 1,9% y 14,9%, respectivamente." (*Informe de los Objetivos de Desarrollo Sostenible 2019*, 2019).

Todos estos avances tecnológicos son posibles gracias a la investigación y desarrollo, es por eso que la ONU dentro de los ODS invita e incita a los países miembros a que aumenten la proporción del PIB que se invierte en investigación y desarrollo, siendo este el objetivo del indicador 9.5. En el último informe emitido por la ONU se observa cómo esta inversión "aumentó del 1,52% en el 2000 al 1,68% en 2016" alcanzando una inversión de 2 billones de dólares comparado con los 739.000 millones en el 2000. Esto representa una tasa media de crecimiento anual del 4,3%. Sin embargo, "Existen grandes disparidades entre las regiones, en 2016 en Europa y América del Norte, el 2,21% del PIB se gastó en I+D (Investigación y Desarrollo), en comparación con el 0,42% y el 0,83%, respectivamente, en el África subsahariana y Asia occidental. Estas disparidades indican la necesidad continua de un firme apoyo normativo para aumentar la financiación en I+D en las regiones en desarrollo" (*Informe de los Objetivos de Desarrollo Sostenible 2019*, 2019).

La Conferencia de las Naciones Unidas sobre Comercio y Desarrollo (UNCTAD), sugiere a organizaciones como el Banco Mundial, el Instituto de Estadística de la UNESCO y la División de Estadística de las Naciones Unidas, que la información recolectada sobre el estado de los ODS se debe clasificar utilizando bibliometría pues es una herramienta que ayuda a detectar palabras claves y definir si la información sí trata sobre ODS o prioridades nacionales (unctad, 2019). Adicional, la OCDE en el manual de Oslo sobre la medición de las actividades científicas y tecnológicas, resalta en los lineamientos sobre la recolección e interpretación de datos que la bibliometría proporciona información complementaria para medir el estado actual de la innovación en publicaciones científicas, revistas técnicas y comerciales (Wessel et al., 2019).

Desde esta perspectiva, medir el avance en el ODS 9.5 mediante la producción académica es un reto para todos los países. Los sistemas de producción industrial se están transformando a una producción inteligente, conectada y descentralizada (Kagermann, Wahlster y Helbig 2013; Hermann, Pentek y Otto 2016). La cuarta revolución industrial llega para generar un "impacto positivo en todas las dimensiones de la sostenibilidad de una manera integrada en el desarrollo de modelos de negocio sostenibles; sistemas de producción sostenibles y circulares; cadenas de suministro sostenibles; diseño de producto sostenible; y desarrollo de políticas para asegurar el logro de los objetivos sostenibles" (Machado et al., 2020).

La Industria 4.0 abarca gran cantidad de temáticas, sin embargo, Beier (2020) identificó que uno de los términos que más predomina dentro la investigación científica es el de "Hombre-máquina", el cual se relaciona con los conceptos de comunicación, sistemas ciber físicos, conexiones, colaboración, e interfaces de persona a máquina y de máquina a persona (Beier et al., 2020) pues estas conexiones minimizan errores humanos y aumentan la productividad (Abdrabou et al., 2018).

El artículo realizado por Lee (Lee et al., 2015) indica que los sistemas ciber físicos (CPS) proporcionan un marco para las conexiones cercanas entre dispositivos físicos y el mundo cibernético. Por esta razón el despliegue eficiente de un CPS es crucial. Posteriormente Trappey (Trappey et al., 2016) presentó una visión general de la literatura disponible de CPS y patentes de carteras de las 5 C (conexión, conversión, computación, cognición y configuración) de los sistemas de fabricación. La investigación benefició a las pequeñas y medianas empresas a integrar soluciones en cuanto a Industria 4.0 y a adoptar un entorno industrial global cambiante.

El concepto de Industria 4.0 está ganando cada vez más atención en todo el mundo, donde las tecnologías relacionadas con Internet de las cosas (IoT), Big Data (BD) y sistemas ciber físicos (CPS) se combinan para lograr un mayor nivel de eficiencia operativa, productividad y automatización (Tesch da Silva et al., 2020). La República (López, 2018) también expone que hay un beneficio de hasta el 30% en disminución de costos en cuanto a la implementación de sistemas ciber físicos. El uso de análisis bibliométrico y minería de texto en las publicaciones científicas en torno a la Industria 4.0, en particular sobre los sistemas ciber físicos, puede conducir a una medición del estado global en el avance del objetivo de desarrollo sostenible 9.5 Es por esto que este trabajo se enfoca en resolver la pregunta ¿el análisis del contenido semántico de publicaciones científicas sobre sistemas ciber físicos produce resultados diferentes en comparación con el análisis bibliométrico de las mismas? Además de establecer las diferencias entre los análisis se desea analizar si, en realidad, los resultados semánticos son un complemento para los resultados bibliométricos.

2. Antecedentes

2.1 Objetivos de desarrollo sostenible

Los objetivos de desarrollo sostenible (ODS) son un conjunto de políticas adoptadas por las Naciones Unidas, que buscan proteger al planeta y garantizar que todas las personas gocen de paz y prosperidad para el año 2030 (PNUD - Programa de las Naciones Unidas para el Desarrollo, n.d.).

Fernandes de Mesquita (2017) consideró un análisis bibliométrico en la base de datos ISI Web of Knowledge en donde se recolectaron 1157 artículos para hacer una revisión literaria consolidando el estado investigativo de los ODS en el ambiente académico, en donde encontró que "dentro de la industria minera las compañías líderes siempre se esfuerzan por ser innovadoras y desarrollar nuevas tecnologías que mejoren la eficiencia energética y reduzcan el consumo de agua" (Fernandes de Mesquita et al., 2017) pero "los investigadores clave y los principales artículos no han estado discutiendo estos temas" (Fernandes de Mesquita et al., 2017). El autor también sugiere que, aunque hay un creciente número de publicaciones en los últimos años, estas aún se centran en un foco medio ambiental de los indicadores de la ONU (Fernandes de Mesquita et al., 2017).

En otra investigación Beier (2018) comparó los resultados de una encuesta realizada a 109 empresas industriales chinas sobre las expectativas que tienen en cuanto a sostenibilidad industrial y concluyó que la gran mayoría están de acuerdo con que la Industria 4.0 podría traer grandes beneficios en los ODS, específicamente en el indicador 9 "Industria, innovación e infraestructura." (Beier et al.,2018).

Abdrabou (2018) en su investigación muestra que los conceptos transversales a la Industria 4.0 son sistemas ciber físicos, IoT, BigData, economía sostenible, computación en la nube y computación cognitiva en donde los investigadores y académicos se están interesando cada vez más por abordar las necesidades actuales y futuras de la humanidad con base en los ODS. La dependencia de las industrias manufactureras en el trabajo manual se volvió cada vez menos importante haciendo que la conexión entre humanos y máquinas minimice errores humanos y aumenten la productividad (Abdrabou et al., 2018).

Abba Adam (2019) realizó un análisis bibliométrico de la investigación publicada en la base de datos Scopus sobre los sistemas ciber físicos con el fin de comprender las tendencias y comparar las contribuciones de publicaciones entre diferentes regiones, instituciones, autores, áreas temáticas y otros dominios. En el estudio se encontró que de 2009 a 2011 el número de publicaciones anuales se mantuvo estable mientras que entre el 2012 y el 2018 hubo un crecimiento exponencial en la cantidad de artículos publicados, registrando el pico en el año 2018 con un total de 782 publicaciones. Además, se estableció que los países que lideran la investigación son China y Estados Unidos en los cuales, las áreas con mayor número de publicaciones son la ingeniería y ciencias de la computación. Los términos que más predominan dentro del análisis bibliométrico son sistemas ciber físicos, Industria 4.0 e internet de las cosas (Abba Adam, 2019).

2.2 Análisis bibliométrico

La bibliometría fue propuesta por Alan Pritchard en 1969 y su contenido principal incluye análisis de coautores y análisis de co-palabras. El análisis de co-citas fue propuesto inicialmente por Small en 1973 y ha sido reconocido como una herramienta efectiva para detectar frentes de investigación, bases intelectuales y tendencias de desarrollo en la literatura científica (Li et al., 2020).

Mika (Vanhala et al., 2020) añade que el análisis bibliométrico es el análisis de citas y el análisis estadístico de publicaciones escritas y se basa en la construcción de un gráfico de citas y de una red o representación gráfica de citas entre documentos. Por su parte, el análisis de citas estudia la relación que hay entre los campos, autores, instituciones y países (McBurney & Novak, 2002). Según Jia (Huang et al., 2020) el análisis bibliométrico se realiza para identificar los autores, las instituciones y las áreas más influyentes, revelar los puntos críticos de investigación y dar una idea de la evolución del tema en este campo. Si bien es una gran herramienta para el análisis de información aún es difícil descubrir conexiones semánticas entre los temas de investigación (Glenisson et al., 2005).

La cantidad de publicaciones científicas referentes a la industria ha ido aumentando en los últimos años y se han realizado diferentes estudios mediante análisis bibliométrico para determinar cómo ha sido el comportamiento de la investigación en torno a la industria y cuáles son las principales áreas de crecimiento. Algunos de los estudios como los de Liane (Kipper et al., 2020) y Aidi (Ahmi et al., 2019) tuvieron como objetivo determinar cómo ha sido el crecimiento académico en temas relacionados con la Industria 4.0 durante los últimos años, analizando alrededor de 1000 documentos y extrayendo la información mediante la frecuencia de co-ocurrencia de las palabras clave. Dentro de la metodología se usaron softwares bibliométricos como SciMat y publish or perish para incorporar los datos obtenidos, posteriormente se utilizó VOSviewer para la visualización de datos, y SPSS y Microsoft Excel para el análisis de estos. Como resultado se obtuvo que la tasa de crecimiento de la literatura en la Industria 4.0 aumentó drásticamente año tras año desde 2012 y que los temas más representativos son los sistemas ciber físicos, Internet de las cosas y Big data.

Es importante resaltar que la mayoría de los autores hacen uso de indicadores de rendimiento y métricas para analizar cómo se ha desarrollado el tema de interés en los documentos de texto. Se utilizan herramientas descriptivas y relacionales ya que los idiomas y países pueden ayudar a explicar el contexto sociodemográfico de una mejor manera; por otro lado, la frecuencia con la que se publica sobre un tema cada año ayuda a entender la evolución de este. Para el caso de Pranab (Muhuri et al., 2019) el cual centró sus investigaciones en el análisis de crecimiento de la Industria 4.0 e Ignacio (Danvila-del-Valle et al., 2019) que investigó acerca del proceso de las capacitaciones en el área de recursos humanos se hizo uso de indicadores comunes como lo son el total de artículos, el total de citas recibidas por la publicación, las citas por artículo y el número de publicaciones realizadas en un año. Como resultado se obtuvo que las palabras clave más usadas por los autores en sus artículos hacen referencia a sistemas ciber físicos e Internet de las cosas, así como capital humano, rendimiento y desarrollo.

2.3 Construcción ecuaciones de búsqueda

Obtener la información adecuada para un análisis bibliométrico requiere la implementación de una ecuación de búsqueda, la cual es una expresión en un lenguaje interrogativo que está asociada a la consulta que se desee y que logra complacer la necesidad de obtener información relevante. Una ecuación de búsqueda se compone de palabras clave, de diferentes operadores de búsqueda que logren relacionar dichas palabras entre ellas y símbolos reservados (Codina, 2017). Existen también otros factores que pueden ocasionar que la búsqueda de resultados no sea lo esperado, por lo que es crucial definir bien los parámetros de búsqueda para no tener incongruencias en los resultados finales, dentro de los factores adicionales que puedan afectar el objetivo de la

búsqueda están: el idioma, el tipo de documento buscado, la necesidad de precisión o exhaustividad y, claramente, la finalidad de la búsqueda (Alfin EEES, 2019).

Los operadores de búsqueda fueron inicialmente presentados por Boole siglo XIX, esta secuencia de operadores daba pie a afirmar o negar la lógica de unas sentencias combinadas y en la actualidad existen tres operadores dominantes en todas las bases de datos, los operadores AND, OR y NOT. El operador XOR se usa en menor medida en las herramientas de búsqueda. Estos operadores booleanos representan en la ecuación de búsqueda funciones de intersección, inclusión y de omisión respectivamente (Codina, 2017).

2.4 Análisis bibliométrico y minería de texto

La minería de texto ha sido una herramienta utilizada en los últimos años para aproximarse a las investigaciones de bases de datos de texto. Fue mencionada por primera vez por Feldman y Dagan (Paass, 2015) y hace referencia a la precisa extracción de información de texto por medio del lenguaje de procesamiento natural. Para conseguir una buena extracción de datos es necesario realizar un preprocesamiento de texto. De acuerdo con Hotho, Nürnberger, y Paass (Paass, 2015) existen unos pasos básicos para realizar el preprocesamiento de texto. El etiquetado gramatical que clasifica las palabras en verbos, adjetivos y sujetos. La fragmentación del texto que agrupa las palabras para conformar oraciones. La desambiguación del texto en caso de que los datos no sean claros y por último el analizador sintáctico que genera un análisis de texto de las oraciones que se formaron en los pasos anteriores del procesamiento.

Según Sheela y Bharathi (Sheela & Bharathi, 2013) existen 8 técnicas reconocidas en la minería de texto. El problema radica en la diferencia de cada una de estas técnicas y si es necesario aplicarlas todas en orden de conseguir lo que se necesita. Un breve resumen de dichas técnicas se encuentra en la Tabla 1.

Tabla 1.Técnicas reconocidas de la minería de texto. Tomado de (Sheela & Bharathi, 2013)

Técnica	Características	Herramientas
Recuperación	Recupera información valiosa de texto no estructurado	Análisis de texto, minería inteligente
Extracción	Extrae información de bases de datos estructuradas	Buscador de tiempo, tala de texto
Resumen	Reduce la longitud del texto concentrándose en los puntos principales y el significado en general	Herramienta tópica de seguimiento, herramienta de extensión de texto
Categorización	Categorización del documento	Minería Inteligente
Grupos	Colección de documentos en grupo, agrupación, clasificación y análisis de documentos de texto	Incentivo, minería rápida

Yildiz (Yildiz, 2019) y Durmusoglu & ÇiFtçi (Unutmaz Durmusoglu & Kocabey ÇiFtçi, 2018) se aproximaron desde las bases de datos Scopus y Web of Knowledge respectivamente, usando palabras claves que permitieran filtrar de manera exitosa los artículos (ecuación de búsqueda). Por un lado, Yildiz realizó el análisis bibliométrico con ayuda de SciMat y VOSviewer, mientras que Durmusoglu & ÇiFtçi lo realizaron con herramientas para encontrar las frecuencias de las palabras repetidas, la teoría de “IDF” y el uso de “Stop words”. Sus resultados fueron similares al evidenciar que USA y China son países líderes en la publicación de artículos sobre Industria 4.0, al mismo tiempo se ve el énfasis en los temas que tuvieron auge como Internet de las cosas, Big data, integración, *Cyber-Physical*, nube, entre otros. Cabe aclarar que se hicieron análisis diferentes en los dos estudios, sin embargo, sus conclusiones se centran en que los países sin la tercera revolución industrial difícilmente podrán avanzar en la publicación de artículos en torno a Industria 4.0.

3. Objetivos

3.1 Objetivo general

Diseñar un algoritmo de minería de texto para efectuar un análisis semántico de las publicaciones académicas sobre sistemas ciber físicos en la base de datos Scopus, que estén escritas en inglés y hayan sido publicadas entre los años 2017 y 2019.

3.2 Objetivos específicos

1. Realizar el análisis bibliométrico de las publicaciones académicas sobre sistemas ciber físicos que se encuentran en la base de datos Scopus y que fueron publicadas entre 2017 y 2019.
2. Hacer el análisis semántico de las publicaciones académicas sobre sistemas ciber físicos mediante el algoritmo de minería de texto.
3. Medir el impacto del análisis semántico mediante la comparación, en términos cuantitativos y cualitativos, de sus resultados con los del análisis bibliométrico de las mismas publicaciones.

4. Metodología

En este proyecto se implementó la metodología de investigación CRISP-DM (Cross Industry Standard Process for Data Mining) que posee una estructura de seis fases: el entendimiento del negocio, la compresión de los datos, la preparación de los datos, el modelamiento a utilizar, evaluación de dicho modelo, y finalmente, la exposición de los resultados con base en los objetivos planteados del negocio (Carnerud, 2014). Esta metodología se aplicó a los artículos sobre sistemas ciber físicos, con el fin de realizar un análisis semántico mediante minería de texto y comparar sus resultados con los de un análisis bibliométrico. De esta forma encontrar beneficios de la técnica de minería de texto en la revisión de literatura científica sobre temas relacionados al ODS 9.5 “Aumentar la investigación y actualizar las tecnologías industriales”.

4.1 Objetivo específico 1

Los resultados del análisis bibliométrico se contrastaron con los presentados en el artículo de Abba Adam (2019) y por ello se empleó la misma ecuación de búsqueda referenciada en el estudio, a saber, “Cyber Physical System*” con filtros de limitación entre los años 2017 y 2019. Del mismo modo, la búsqueda se enfocó solo en publicaciones de tipo artículo de acceso abierto en idioma inglés. Esta consulta se realizó en la base de datos Scopus el día 18 de agosto de 2020, se descargó en formato RIS el cual es compatible con el software The Vantage Point utilizado para el análisis bibliométrico y como resultado se obtuvo un total de 1115 artículos. Es importante señalar que para cada artículo el archivo de consulta contiene campos como la información de la citación, bibliografía, resumen y palabras claves, afiliaciones, entre otros.

Cabe destacar que el software The Vantage Point se escogió para realizar el análisis bibliométrico sobre el archivo de descarga RIS obtenido desde Scopus, ya que es un software de análisis bibliométrico que ofrece un amplio repertorio de herramientas de refinamiento, análisis para información científica, técnicas de mercado y de patentes (Home - The Vantage Point, n.d.), el cual la Universidad tiene licenciado a través de la biblioteca. Asimismo, como complemento se usó el software VosViewer, una herramienta que permite construir y visualizar redes bibliométricas, la cual brinda la posibilidad de analizar publicaciones y artículos por medio de co-citaciones, citaciones y relaciones de coautoría. Adicionalmente, para la recopilación y estandarización de

las gráficas y datos encontrados se hizo uso de la herramienta Power BI, un software que permite conectar datos, modelarlos y visualizarlos con facilidad (Qué es Power BI | Microsoft Power BI, n.d.).

En primer lugar, al momento de cargar el archivo de los resultados de búsqueda en el software fue necesaria una preparación previa de los datos, por lo que en esta parte del proceso se hizo la depuración de registros duplicados, subconjuntos de palabras que no se agruparon automáticamente (palabras con el mismo significado pero con diferente grafía, por ejemplo: IoT, Internet of Things, etc). A su vez, se llevó a cabo la lectura óptima de estos datos obteniendo así una base lista para el análisis. En segundo lugar, se siguieron los pasos propuestos en el manual bibliométrico elaborado con anterioridad (Anexo 1) para su construcción se tomaron como referencia los manuales de Karolinska Institutet y el compendio de indicadores científicos ((CSIC), 2016). En este manual se definieron los indicadores que se esperaba obtener con el análisis bibliométrico, del mismo modo se tomaron indicadores de referencia presentados en Abba (2019) como el crecimiento en las publicaciones por año, el área de análisis en que fueron escritos los documentos, la productividad por autor, entre otros.

De manera análoga, se hizo uso de la herramienta *Citation Overview Tool* de Scopus, ésta brinda diferentes ventajas a la hora de realizar una búsqueda en el sitio, por ejemplo, las tendencias de citas para un conjunto de documentos, en este caso particular se tomaron tanto los resultados del indicador de autores con más publicaciones en un tiempo determinado, como los temas predominantes en la búsqueda. Además, de los 1115 artículos encontrados se eliminaron 40 ya que estaban escritos en un idioma diferente al inglés, no eran de open access o tenían clave de acceso al momento de abrir el archivo descargado. A partir de lo anterior quedaron 1075 artículos para ser usados en los análisis bibliométrico y semántico.

4.2 Objetivo específico 2

4.2. 1 Preparación del corpus a analizar con la herramienta de minería de texto

Para esta fase se tomaron los 1075 artículos usados en el análisis bibliométrico, se descargaron en formato pdf y por medio de la página https://pdftotext.com/es/ se convirtieron en formato de texto (.txt), así se definió el corpus usado para el proceso de minería de texto. De esta manera, una vez leídos los artículos desde el algoritmo de Python (Anexo 2) se implementaron las siguientes actividades en torno a la preparación y limpieza previa a la ejecución de los algoritmos de minería de texto*: tokenización,* normalización a minúscula de los documentos y *lematización*. Por una parte, *la tokenización* divide el texto en entidades significativas las cuales pueden ser palabras u oraciones, además segmenta el texto en palabras llamadas "token" según los espacios en blanco. Para el código implementado se utilizó la función *Word_Tokenize* de la librería *NLTK*. Por otra parte, la normalización a minúscula permitió homogeneizar los caracteres cuyo significado fuera semejante y de esta manera reducir el vocabulario. La *lematización* separa las palabras en su lema (raíz) y esto puede usarse como representación de las diferentes conjugaciones y formas en que cada palabra puede flexionarse. Para este algoritmo se utilizó la función WordNetLemmatizer de la librería *NLTK.*

Es importante mencionar que también se realizó la eliminación de *Stop Words*, es decir, se removieron las palabras que se caracterizan por ser poco relevantes y no aportar información acerca de los contenidos de un texto, éstas suelen ser conectores y artículos. Para este paso se utilizó la lista de *Stop Words* encontrada en Google Code que cuenta con una lista predefinida de palabras en el idioma inglés que finalmente fueron eliminadas de los artículos. De igual forma, como parte del preprocesamiento de datos se realizó un análisis de parte del discurso (en inglés Part-of-Speech (pos)), de esta forma, las categorías lingüísticas de los "tokens" fueron reconocidas y, después, se filtraron las unidades léxicas con categorías como sustantivos, verbos y adjetivos. Las etiquetas nombradas fueron retenidas porque es usual asociarlas a términos informativos y útiles durante la identificación de temáticas (Ochoa et al., 2013).

El modelo del espacio vectorial (vector space model (VSM)) es frecuentemente empleado en el proceso de recuperación de información y análisis de temáticas (Gupta & Lehal, 2009) dada su simplicidad conceptual y

el atractivo de la metáfora subyacente del uso de la proximidad espacial para la proximidad semántica (Manning et al., 2002). En el esquema VSM, los documentos son representados en forma de vectores de características que están ubicados en un espacio euclidiano multidimensional que facilita las comparaciones entre términos en relación a distancias elucídeas, cosenos y otras medidas (H. Christopher D. Manning, 2008), así mismo, los significados de las palabras son representados a través de un esquema de pesos donde se buscan hallar los patrones estadísticos del uso de los términos con el fin de reconocer el significado con que las personas emplean estas unidades lingüísticas (Turney & Pantel, 2010).

Por lo anterior, el modelo VSM procesa los documentos al considerarlos como bolsas de palabras por lo que ignora el orden exacto de los términos del texto y se pondera el número de ocurrencias de cada término en relación a la frecuencia de aparición de la entidad en el documento (H. Christopher D. Manning, 2008). El esquema más utilizado para ponderar los términos dentro de los documentos es el TF-IDF que surge de la composición de dos métricas, la frecuencia del término (en inglés term frequency - TF) y la frecuencia inversa del documento (en inglés inverse document frequency - IDF). La primera métrica valora como relevante a un término que tiene alta frecuencia en los artículos, mientras que la segunda beneficia a los términos con baja frecuencia dentro del documento y se interpreta como un indicador de informatividad donde un término semánticamente importante dentro de una temática ocurrirá escasamente dentro del corpus (Manning et al., 2002). En consecuencia, el esquema TF-IDF expresa la ponderación relativa del término considerando que adquiere un gran peso cuando la unidad es frecuente en el documento (TF alto) y raro en otros textos del corpus (es decir IDF alto).

En este orden de ideas, este trabajo modeló los documentos preprocesados considerando el VSM y ponderó los términos dentro de los documentos mediante el esquema TF-IDF. Para esto se utilizó la función *TF-IDF vectoricer* de la librería de *Sklearn* de Python en el corpus. Es así como se construyó la matriz término – documento (MTD, en inglés term - document matrix) donde las columnas corresponden a los textos y las filas conciben términos únicos, además en la celda xij se establece la frecuencia del término i – ésimo en el documento j – ésimo (H. Christopher D. Manning, 2008).

Aun cuando la matriz término-documento permite reconocer lo que realmente se ha escrito dentro de cada texto, posee la falencia de ser una matriz dispersa por lo que establecer la temática del documento es un trabajo computacionalmente difícil (Manning et al., 2002). En este sentido se han planteado estrategias como la descomposición en valores singulares para establecer la representación semántica de los textos basándose en la idea que el significado de una palabra puede ser aprendido de un entorno lingüístico, al capturar de forma cuantitativa la estadística de ocurrencia de los términos y considerar que estos términos poseen una estructura latente enlazada a tópicos que expresan el sentido del texto (Turney & Pantel, 2010).

La idea principal detrás del valor de descomposición singular (Singular Value Descomposition (SVD)) es recopilar todos los contextos dentro de los cuales aparecen las palabras del vocabulario y establecer factores comunes que representen conceptos subyacentes (Deerwester et al., 1990), por ende, al implementar SVD se pueden determinar categorías que permitan describir el contenido de un corpus, lo que permite establecer una organización jerárquica de alta calidad de los conceptos a través de un conjunto de datos en diferentes niveles de granularidad (García-Morales et al., 2012). De esta forma (Evangelopoulos et al., 2012) señalan que los beneficios de emplear SVD radican en que: i) evita la subjetividad humana cuando las categorías son preexistentes y ii) destila nuevas categorías basadas en datos cuando no existen teorías bien establecidas que anticipen las categorías de codificación. En este sentido, mediante el SVD se facilita la construcción de un sistema que prediga qué términos están realmente implicados en una consulta o aplicados a un documento, al realizar análisis sobre la base de la muestra falible encontrada realmente (Bayer et al., 1990) como resultado obteniendo una matriz de dimensiones similares término – documento pero mucho más informativa.

4.2. 2 Definición de clústeres a ser estudiados bajo el análisis semántico

Para cumplir el segundo objetivo, se agruparon los artículos en clústeres usando como referencia los valores de la matriz TF-IDF resultante de aplicar SVD contemplando dos posibilidades: un algoritmo jerárquico aglomerativo calculado con el enlace de Ward y un algoritmo *k-means*, ambos con base en la distancia euclidiana. Se determinó el número de clústeres mediante la iteración con diferentes *k*, donde para el método jerárquico se obtuvo un resultado de un gráfico con la función *dendrogram* de la librería *scipy* que utiliza por

defecto la similaridad del coseno, y para *k-means* se utilizó un diagrama de codo que representa la suma de cuadrados de las distancias versus el *k*, ambos evaluados bajo los indicadores de Davies-Bouldin y Calinski-Harabasz.

De manera semejante, para que los métodos fueran comparables entre sí, además de corroborar si el número *k* era el adecuado, se calculó el indicador de Davies-Bouldin en donde entre más cercano el valor a 0, los clústeres son más compactos y sus centros están mejor separados unos de otros, y el de Calinski-Harabasz que refleja la similitud dentro de cada clúster y las diferencias entre diferentes clústeres en donde a mayor valor del indicador, mejor es el resultado (Maulik & Bandyopadhyay, 2002).

Después de obtenidos los clústeres y con el objetivo de determinar un tópico el cual pueda representar cada uno de estos, se utilizó la técnica Latent Dirichlet Allocation (LDA) (Blei et al., 2003), ésta es un modelo probabilístico que genera una distribución de probabilidad entre términos y documentos asociados a un tópico específico, y en la que cada documento se considera como una colección de tópicos y cada temática como una colección de palabras. Una vez obtenidas las temáticas de los clústeres, se utilizó la métrica de coherencia del tópico (en inglés *Topic Coherence*) (Röder et al., 2015), donde sin importar el resultado negativo o positivo, entre más cercano a 0 se garantiza que cada uno de estos tópicos contiene información que tiende a ocurrir en los mismos documentos, por ende, la temática representa que es coherente e interpretable por seres humanos.

4.2. 3 Análisis de palabras más relevantes por tema y por país

En este apartado, dado que el ODS 9.5 es validado por país y no se puede asegurar que los clústeres obtenidos anteriormente quedaron divididos de esa forma, se decidió analizar por separado los documentos de los dos países que mayor número de publicaciones científicas relacionadas a Cyber Physical Systems tienen: Estados Unidos (268) y China (210). Además, se separaron los artículos publicados en Colombia en el periodo considerado (8).

Con la herramienta *Citation Overview Tool de* Scopus se separaron los documentos que pertenecen a cada uno de los 3 temas más publicados de cada país, una vez identificadas las publicaciones de estos temas se hizo un análisis bibliométrico, encontrando la frecuencia en las palabras claves por país y por tema utilizando el programa The Vantage Point, con el fin de ser contrastadas con las palabras más relevantes obtenidas en el análisis semántico, el cual se realizó haciendo uso de la matriz TF-IDF para calcular las palabras de mayor relevancia y encontrar subtemas dentro de estos grupos.

4.2.4 Análisis de palabras para la conformación de posibles tópicos por país

Por una parte, para el análisis de palabras por país se utilizó la base de datos por tema y por país usada en el apartado 4.2.3 sin el filtro de los temas, y posteriormente se aplicó el proceso de Latent Dirichlet Allocation (LDA), enunciado en el numeral 4.2.2 para establecer los términos que con mayor probabilidad conforman las temáticas abordadas dentro de los documentos segregados por país.

4.2.5 Análisis de coocurrencias del corpus

Por otra parte, se emplearon para hacer una comparación los resultados de las coocurrencias identificadas bibliométricamente sobre las palabras clave de los artículos con el programa VOSviewer y las halladas mediante minería de texto donde se usó la función *countvectorizer* de la librería *Sklearn,* la cual asigna el número mínimo y máximo de grupos a formar entre palabras, para esto se hizo un análisis de colocaciones buscando desde bigramas hasta 5-gramas con ayuda de la misma función.

4.3 Objetivo específico 3

Los resultados del análisis bibliométrico y semántico fueron analizados a través de herramientas cuantitativas y cualitativas como: la frecuencia de palabras del *abstract* para los clústeres conformados y los países por tema con el uso de bibliometría. Por otro lado, palabras más probables a pertenecer a un tópico (LDA) y palabras más relevantes con ayuda del análisis semántico.

5. Resultados

A continuación se presentan los resultados divididos en los dos métodos, a) análisis bibliométrico y b) análisis semántico, dando así cumplimiento a los objetivos 1 y 2 de la investigación. Finalmente se hace una comparación cuantitativa y cualitativa de los resultados de ambos métodos a lo largo del documento para responder con la pregunta de investigación planteada en la justificación en torno al ODS 9.5 y cómo varían estos resultados de países potencias en comparación a países subdesarrollados como Colombia.

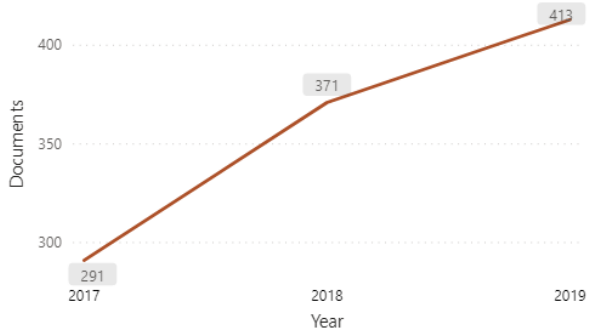
5.1 Objetivo específico número 1

El análisis bibliométrico "permite valorar la actividad científica en un campo específico, en determinados períodos y su impacto" (Arbeláez & Onrubia, 2014) a través de los indicadores bibliométricos los cuales "aportan información sobre variables cuantitativas y cualitativas de las publicaciones científicas, tales como el número y distribución de publicaciones, la productividad por autores, el número de autores firmantes, el número y distribución de referencias bibliográficas, el número de citas recibidas por un trabajo, entre otras" (Arbeláez & Onrubia, 2014). A partir de la anterior definición se analizaron los siguientes indicadores:

5.1.1 Frecuencia de publicaciones por año

Sobre el total de artículos se halló la frecuencia respecto al año de publicación y se obtuvo como resultado una tendencia creciente (Gráfica 1), en donde se evidencia un aumento del 11.3% para el año 2019 con respecto al 2018 con 413 artículos. Se puede afirmar que al igual que en el documento de Abba Adam (2019), el número de publicaciones sobre sistemas ciber físicos ha aumentado a través de los años, siendo un tema cada vez más relevante en la investigación científica.

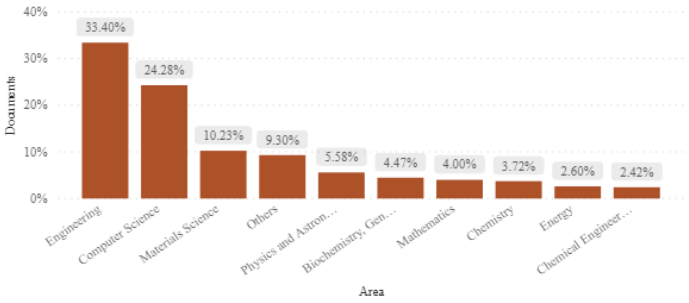
Gráfica 1. Número de documentos publicados por año



5.1.2 Análisis de área de publicación

En las 10 primeras áreas de publicación, en comparación con el estudio de Abba (2019) se evidencia un aumento en el porcentaje de *Materials Science* el cual era un 6% y ahora representa un 10.2% de la totalidad de los artículos, además, como tema principal sigue liderando *Engineering* con un 33.4%, seguido del área de *Computer Science* con un 24.3% (Gráfica 2).

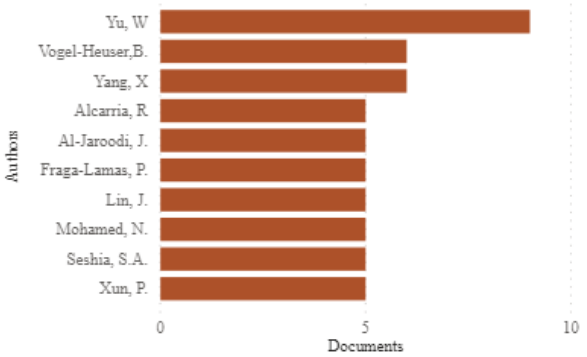
Gráfica 2.Porcentaje de áreas de enfoque de los documentos



5.1.3 Productividad de los autores

Los resultados mostraron que, con un total de 9 publicaciones, el autor más productivo durante los tres años seleccionados fue Yu, W. (Gráfica 3), afiliado al College of Computer National University of Defense Technology, Shanghai, China. Este dato se tomó directamente desde la herramienta *Citation Overview Tool* de Scopus ya que con The Vantage Point no fue posible realizar la limpieza y estandarización de los autores. Según Aliaga, Francisco M. & Correa, Ana D. (2011) la inexperiencia de muchos autores al manejar sus nombres en bases de datos, así como las diferencias culturales, sobre todo con la redacción de nombres chinos, dificulta la estandarización y lectura de estos, esto explica porqué cada autor tiene un código diferente que lo identifica dentro de las bases de datos.

Gráfica 3. *Número de documentos publicados por autor en los tres años*



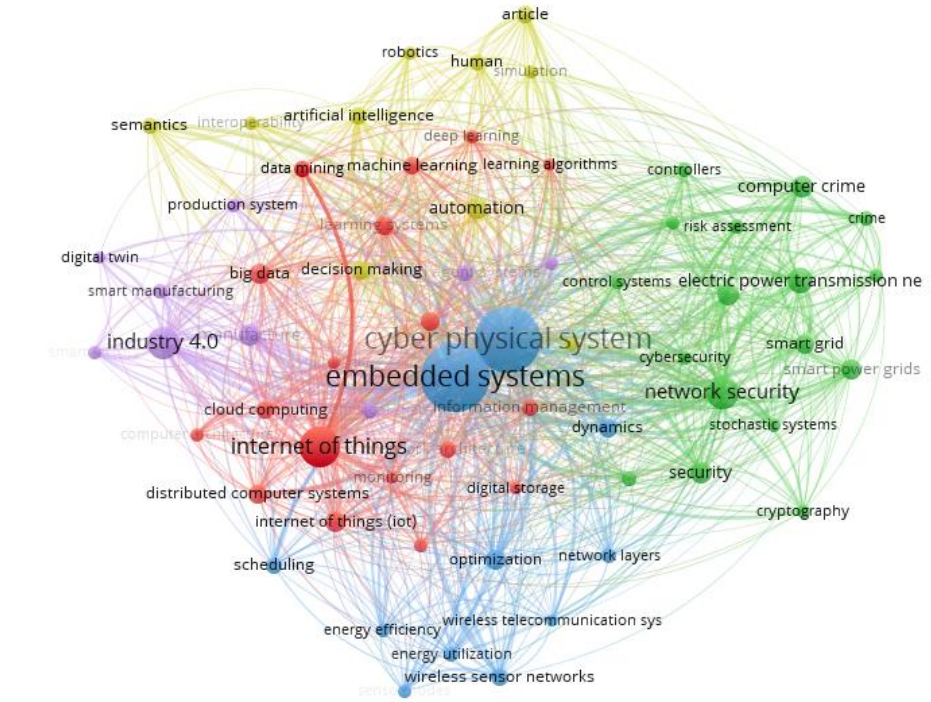
5.1.4 Análisis de las palabras claves

El mapa de palabras claves (Gráfica 4), fue realizado en VOSviewer, cada color representa un clúster que contiene palabras agrupadas con base en la coocurrencia de estas. En términos generales el clúster azul hace referencia a *Cyber Physical Systems*, el verde a *Security Networks* e *Informatic*, el rojo representa *Big Data* e *Industry 4.0*, el amarillo representa la noción de *Automation* entre hombre y máquina, y el morado agrupa los términos en torno a *Manufacture*. A su vez, en esta gráfica se pueden detallar las palabras pertenecientes a cada clúster.

Conviene enfatizar que se decidió manejar una frecuencia mayor a 20 veces en coocurrencia ya que permitía una mejor visualización de las correlaciones de las palabras más importantes. En este caso, existen dos tipos de pesos de atributos que representan la magnitud con la que dos palabras están correlacionadas: la fuerza del enlace y la fuerza total del enlace. La primera hace referencia al número de publicaciones en que estas dos palabras aparecen juntas, y la segunda hace referencia a la fuerza de una palabra con todas las que está conectada, es decir, al número de veces que aparece esta palabra con todas las que está conectada (Jan van Eck

& Waltman, n.d.), cabe aclarar que por método predeterminado esta fuerza está normalizada por una constante multiplicativa dada por el programa, por eso las fuerzas se obtienen con valores decimales. Es así como se puede observar que *Embedded Systems* y *Cyber Physical System* tienen una fuerza de enlace de 201.53 y *Embedded Systems* se encuentra enlazada con una fuerza total de 524 a 33 palabras.

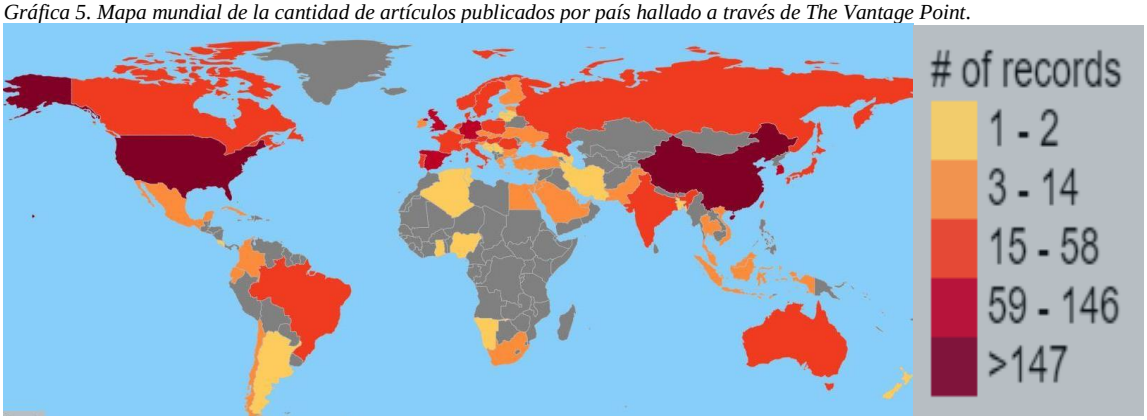
Gráfica 4.Correlación de palabras claves en el mapa



5.1.5 Análisis por países

Teniendo en cuenta que el ODS 9.5 se mide por países, se procedió a analizar individualmente los más importantes a partir del total de publicaciones realizadas. La Gráfica 5 representa la densidad de publicaciones por país en donde el color más oscuro hace referencia a una mayor cantidad de publicaciones. Se observó que los países con más publicaciones sobre sistemas ciber físicos son potencias mundiales como Estados Unidos, China y Rusia, y de acuerdo con Dilmus D. James, James H. Street, Allen D. Jedlicka and Sibila Seibert (1981) los países con menor investigación y publicaciones científicas suelen ser los países tercermundistas.

Gráfica 5. Mapa mundial de la cantidad de artículos publicados por país hallado a través de The Vantage Point.



Por otro lado, en cuanto a la cantidad de citaciones en los documentos por país, la Tabla 2 muestra el top 10 de los países con más citaciones en sus artículos, encabezado por Estados Unidos el cual tiene un total de 294 artículos que han sido citados 4503 veces y seguido por China con 4416 citaciones de sus 283 artículos. Se estima que el número de citaciones es directamente proporcional a la cantidad de publicaciones por país, indicando que a mayor cantidad de artículos publicados, mayor oportunidad de incrementar el índice de citaciones.

Tabla 2.Número de citas de artículos por país

Country	Number of citations per article	Number of documents
Estados Unidos	4503	294
China	4416	283
Reino Unido	1818	76
Alemania	1521	75
Italia	703	58
España	622	67
Australia	469	35
Korea del Sur	427	68
Francia	192	44
India	99	44
Total	**14770**	**1044**

5.1.6 Análisis de afiliaciones por países

Por último, se realizó el análisis de afiliaciones por países (Tabla 3), donde se encontró que el centro de investigación con mayor cantidad de publicaciones (6) está ubicado en China. Este resultado concuerda con lo presentado en el documento de Abba (2019). Además, se destaca que la segunda institución en el ranking tuvo la misma cantidad de artículos y se sitúa en Estados Unidos.

Tabla 3.Top 10 de instituciones-afiliaciones en la búsqueda

Institutions	Documents
College of Computer National University of Defense Technology Changsha 410073, China	6
Department of Computer and Information Sciences Towson University, Towson, MD 21252. United States	6
Department of Electronic Information and Electrical Engineering Changsha University, Changsha, 410022, China	5
College of Electrical Engineering Zhejiang University Hangzhou, 310027, China	4
School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, 100083, China	4
School of Computing Science and Engineering, VIT University, Chennai, Tamil Nadu, India	4
School of information Science and Engineering, Central South University, Changsha, 410083, China	4
School of Software, Tsinghua University;Beijing, 100084, China	4
Beijin Key Laboratory of Knowledge Engineering for Materials Science,Bejing, 100083, China	3
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China	3
Total	**43**

5.2 Objetivo específico número 2

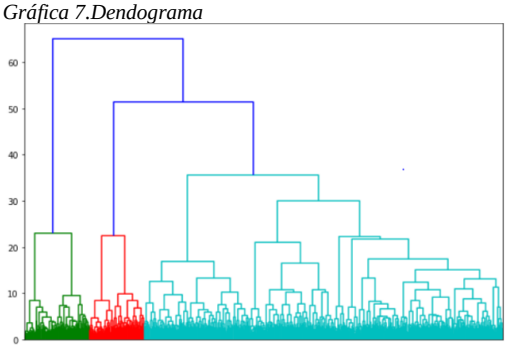
El corpus inicial estuvo compuesto por 683 oraciones y 16440 “tokens” (antes de preprocesamiento), además al aplicar las técnicas mencionadas en el apartado 4.2.1 se obtuvieron 2353 “tokens” como *exploration, method, ability*, entre otras y se pueden visualizar en el Anexo 3. En particular los términos más relevantes considerando la métrica TF-IDF se señalan en la Gráfica 6.

*Gráfica 6.*Nube de palabras matriz TF-IDF.

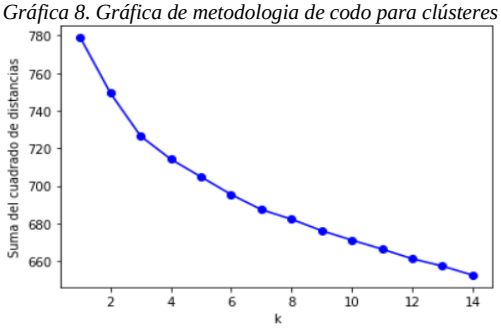
5.2.1 Definición de clústeres

Como se mencionó en la metodología del objetivo específico 2 y en pro de realizar un análisis semántico adecuado donde los artículos se agrupen con base en la similitud de sus características (*clustering*), se emplearon las metodologías aglomerativa y *k-means.* Por una parte, para el método aglomerativo se obtuvo un dendograma con una visualización de *k*=3 clústeres (Gráfica 7). Por otra parte, conforme al método de *k-means*, se obtuvo la gráfica de codo (Gráfica 8) que representa la suma de cuadrados de las distancias versus el *k*.

Gráfica 7.Dendograma



Gráfica 8. Gráfica de metodologia de codo para clústeres óptimos.



Siguiendo la literatura, para el método de *k-means* se debe realizar también una evaluación visual de la gráfica de codo, en donde el valor del *k* óptimo es representado por el punto de inflexión o pico. Como no se encuentra ninguno de estos dos puntos, no hay una razón para preferir un número de clústeres frente a otro. Por esto, se decide variar el valor de k para diferentes iteraciones evaluadas bajo los indicadores de Davies-Bouldi y Calinski-Harabasz obteniendo los siguientes resultados (Tabla 4).

Tabla 4. Número de citas de artículos por país

	k-means	Aglomerativo	k-means	Aglomerativo	k-means	Aglomerativo	k-means	Aglomerativo	k-means	Aglomerativo
Número de Clústeres	3		4		5		6		7	
Davies Bouldi	4,21	4,1	4,49	4,33	4,97	5,61	4,48	5,91	4,48	5,41
Calinski Harabasz	31,15	27,91	25,97	23,65	22,53	20,24	20,53	17,14	20,53	17,14

Dado que las dos metodologías tuvieron resultados similares con respecto al indicador de Davies y Calinski se decidió agrupar los clústeres bajo la metodología jerárquica aglomerativa pues el dendograma muestra claramente tres clústeres y el valor de sus indicadores es mejor que los otros *k* evaluados.

5.2.2 Análisis de los clústeres

A partir del análisis bibliométrico y en el momento de realizar una búsqueda, se toman las palabras claves que funcionan como etiquetas de los artículos científicos pues describen, en términos generales, los temas tratados por los documentos y ayudan a clasificar los textos de manera eficaz. No obstante, con la finalidad de encontrar subtemas o términos específicos dentro de dichas etiquetas se decidió utilizar la técnica LDA del análisis semántico para cada clúster, en donde se encuentra información acerca de qué tan bien describen las palabras claves los tópicos tratados en los documentos.

Como parámetros iniciales para realizar el modelamiento LDA se tomaron a consideración los valores *update_every* = 5, *chunksize* = 10000, y *passes* =100 de la librería *gensim* bajo la función *models.Ldamodel*. El primer parámetro garantiza un análisis profundo para cada documento, logrando categorizar cada uno de estos en tópicos. El segunda usa un valor más grande que el del corpus a trabajar con el fin de que todos los documentos se incluyan dentro de cada iteración. El último hace referencia a la precisión del modelo donde su valor no suele ser alto buscando la optimización computacional del algoritmo.

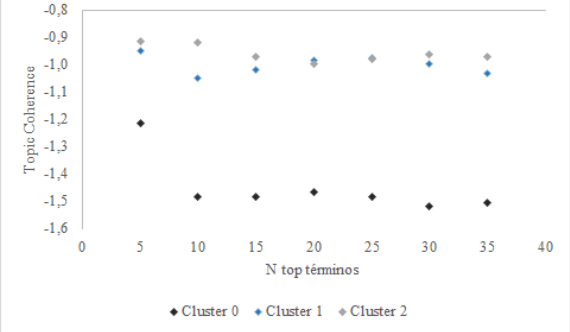
Para corroborar que las palabras estaban correctamente relacionadas a los temas se utilizó el indicador de coherencia del tópico, con el cual también se determinó la cantidad de tópicos a trabajar. Este indicador representa la pureza de los datos y entre más cerca de 0 su resultado es más coherente dentro del grupo de palabras. Con el fin de determinar el número de tópicos a trabajar se iteró sobre la cantidad de tópicos hasta estabilizar el indicador de coherencia y la probabilidad que brinda cada término sobre el tópico. Los resultados se pueden ver en la Tabla 5 donde se decidió trabajar con 50 tópicos para el clúster 0 y con 35 tópicos para el clúster 1 y 2.

*Tabla 5.*Resultados del indicador de coherencia del tópico.

Cluster	0	1	2
Indice de Coherencia	-1.24	-1.03	0.93

Una vez obtenidos el número de tópicos a trabajar se calculó el número *n* de palabras que describen al mismo. En la Gráfica 9 se pueden observar los resultados de las iteraciones donde se obtuvo que el mejor índice de coherencia que describe un tópico se logra con cinco términos.

Gráfica 9. Resultados del indicador de coherencia del tópico para el top de términos.



En efecto, el clúster 0 está conformado por 741 documentos, en el clúster 1 hay 179, mientras que en el 2 hay 155. Las palabras claves obtenidas en el análisis bibliométrico de cada uno se evidencian en la parte superior de la Tabla 6 y en la parte inferior sus respectivos grupos de términos asociados a los tópicos obtenidos con el LDA.

Tabla 6.Palabras más probables para conformar un tópico mediante LDA

Cluster 0			Cluster 1			Cluster 2		
Cyber physical production systems	Cyber physical systems	Smart Manufacturing	Network security	Cyber Physical System	Embedded Systems	Network security	Cyber Physical Systems	Internet of things
Internet of things		Industry 4.0	Industry 4.0	Internet of things		Smart Grid		Wireless sensor networks
1. LDA	**2. LDA**	**3. LDA**	**10. LDA**	**11. LDA**	**12. LDA**	**18. LDA**	**19. LDA**	**20. LDA**
Motion	Component	Energy	Game	Estimation	Learning	Performance	Digital	Module
Force	Event	Power	Strategy	Measurement	Feature	Simulation	Intelligent	Simulation
Constraint	Interface	Sensor	Defender	Error	Machine	Variable	Twin	Net
Frequency	Modeling	Consumption	Cost	Matrix	Dataset	Plant	Smart	Virtual
Motor			Equlibrium	Estimator	Training			
			Player					
4. LDA	**5. LDA**	**6. LDA**	**13. LDA**	**14. LDA**	**15. LDA**	**21. LDA**	**22. LDA**	**23. LDA**
Record	Device	User	Vehicle	Signal	Emergency	Economy	Economy	Disturbance
Event	Mobile	People	Traffic	Output	Safety	Platform	Economic	Agent
Level	User	Participant	Road	Multiplicative	Response	Digital	Technological	Machining
	Sensor		Scheme	Stealthy	Mechanism	Artificial	form	Cutting
						Employment	Organization	
7. LDA	**8. LDA**	**9. LDA**	**16. LDA**	**17. LDA**		**24. LDA**		
Node	Service	Production	Experiment	Device		Maturity		
Failure	Smart	Manufacturing	Fault	Vulnerability		Criterion		
Cyber	Technology	Industrial	Point	Smart		Workshop		
Layer	Device	Machine	Technique	Grid		Indicator		
Link		Product						
Coupling		Architecture						

Los grupos "LDA" hacen referencia a palabras que por probabilidad pueden pertenecer a un tema tratado en el clúster. Así, el color rojo en los grupos de palabras representa para qué términos no encontró una conexión semántica o contexto dentro de su grupo, esto mediante una inspección manual.

Como resultado se encontró que múltiples LDA tienen una relación semántica y están coherentemente relacionados, adicionalmente de estar formados por términos que pertenecen a las temáticas descritas por las palabras claves. Por ejemplo, en el clúster 0 el LDA 8, arrojó las palabras como *Service, Smart, Technology, y Device,* todas pertenecientes a un mismo contexto, esto validado manualmente con los documentos de los clústeres resultantes y contrastados con la literatura especializada como se puede evidenciar en el artículo "An IoT Based Architecture for Enhancing the Effectiveness of Prototype Medical Instruments Applied to Neurodegenerative Disease Diagnosis" (Depari et al., 2019). Así mismo pasa con el LDA 2 y el 9 donde los términos constituyen un mismo contexto y hablan de *Cyber Physical Production Sytems* como se pudo verificar en el título "Cyber-Physical production systems architecure based on multi-agent's design pattern- comparison of selected approaches mapping four agent pattern*"* (Cruz Salazar et al., 2019) y Smart Manufacturing en el título "Industrial multi-energy and production management scheme in cyber-physical environments: a case study in a battery manufacturing plant" (Pei et al., 2019).

Es preciso señalar que en el caso del clúster 1, las palabras de los LDA 12 y 15 crean un contexto y tiene relación entre sí para pertenecer a temas como *Embedded Systems* y *Network Security* evidenciado en títulos como "Machine Learning for Security and the Internet of Things The Good, the Bad, and the Ugly" (Liang et al., 2019) y "Data Integrity Attacks Against Dynamic Route Guidance in Transportation-Based Cyber-Physical

Systems Modeling, Analysis, and Defense" (Lin et al., 2018) respectivamente. Para el clúster 2 no se encontraron grupos de LDA completamente uniformes, sin embargo, en el caso del LDA 21 se pueden evidenciar términos como *Plataform, Digital, Artifitial,* y *Employment,* refiriéndose a un mismo tema como se puede ver en los artículos "A CPS-Based Simulation Platform for Long Production Factories" (Iannino et al., 2019) y "A cyber-physical systems approach to cognitive enterprise".

Asimismo, que la técnica de LDA puede resaltar información de temas que no fueron tenidos en cuenta por las etiquetas bibliométricas, como sucede con el caso del LDA 13, en el que palabras como Vehicle, Traffic, y Road, son coherentes entre sí, y hablan sobre un contexto de transporte, una temática que no había aparecido durante esta investigación pero sí mencionada en títulos como "DISASTER Dedicated Intelligent Security Attacks on Sensor-Triggered Emergency Responses"(Mosenia et al., 2017). Bajo este análisis se encontró una relación semántica entre los temas arrojados por la técnica LDA y las palabras claves que describen los documentos.

5.2.3 Análisis de palabras más frecuentes (bibliométrico) y relevantes (semántico) por tema y por país.

Para este análisis se tomaron los países con mayor número de publicaciones científicas además de Colombia por ser un país de interés para la investigación, se escogen las tres primeras temáticas obtenidas a partir del análisis bibliométrico donde se contrastan sus resultados con los obtenidos en el análisis semántico con la matriz TF-IDF, con el fin de encontrar subtemas en las temáticas populares para cada país como lo muestran las Tabla 7, Tabla 8 y Tabla 9.

Tabla 7.Palabras más frecuentes y relevantes para la temática de Engineering

Engineering			
Análisis	United States	China	Colombia
Bibliométrico	Cyber-Physical systems	Cyber-Physical systems	Cyber-Physical systems
	Smart grid	Cyber-Physical social systems	Additive Manufacturing
	internet of things	Cyber-security	Industry 4.0
	Cloud Computing	Internet of things	RAMI 4.0
	Security	Smart grid	Distributed control system
Semántico	Atack	Attack	Manufacturing
	Node	Node	Node
	Power	Power	Controller
	CPS	Manufacturing	Power
	Security	CPS	Circuit

Tal como se muestra en la Tabla 7, se evidenció por medio del análisis bibliométrico que para estos tres países en la temática de *Engineering* el término más frecuente es *Cyber Physical Systems*, esto debido a que la ecuación de búsqueda está centrada en este tópico. Sin embargo, para Estados Unidos y China *Internet of things, Smart grid* y *Security* son los términos más recurrentes; caso contrario para Colombia en donde *Additive manufacturing, RAMI 4.0* y *Distributed systems* son los más mencionados. Así, al contrastar estos resultados con el análisis semántico se encuentran nuevas palabras como *Attack, Node, Manufacturing, Controller, Power*, entre otras que resultan ser temáticas actuales y pertenecientes a lo que se conoce como Industria 4.0. Por consiguiente, combinando estos dos análisis se puede concluir que para las potencias de Estados Unidos y China la seguridad cibernética es definitivamente un tema popular entre sus autores, en contraste con Colombia donde los temas están enfocados principalmente en la manufactura en industria inteligente.

Tabla 8. Palabras más frecuentes y relevantes para la temática Computer Science

Computer Science			
Análisis	United States	China	Colombia
Bibliométrico	Cloud Computing	Smart manufacturing	Industrial control Systems
	Cyber-Physical Systems	Cyber-physical systems	CPPS
	Internet of Things	Security	DC_AC power converters
	Security	Internet of things	Distributed Control systems
	Smart Grid	Smart grids	Design patterns
Semántico	Attack	Attack	Attack
	Node	Node	Img
	Power	Manufacturing	Voltage
	CPS	CPS	Idg
	Device	Power	Primary

Por otra parte, el segundo tema con mayor cantidad de artículos publicados fue *Computer Science* el cual a partir de lo obtenido en el análisis bibliométrico y en el análisis semántico tiene una gran relación con los temas de ingeniería, de lo que se infiere que la mayoría de los artículos publicados de ingeniería se centra en temas computacionales con principal interés de integrar los CPS a la manufactura inteligente, la cual está enfocada a la seguridad cibernética y la computación en la nube. El caso de Colombia difiere un poco de los resultados obtenidos para los demás países ya que palabras como *Attack* en el análisis semántico hacen referencia a temas de seguridad, también se pueden encontrar términos como *voltage* y siglas como *Idg* y *Img* que hacen referencia a *industrial control systems.*

Tabla 9.Palabras más frecuentes y relevantes para la temática Material Science/Biochemistry.

Material Science/Biochemistry			
Análisis	United States (Material Science)	China (Material Science)	Colombia (Biochemistry)
Bibliométrico	Cloud Computing	Smart manufacturing	Industrial control Systems
	Cyber-physical systems	Cyber-physical systems	CPPS
	Internet of Things	Smart grids	DC_AC power converters
	Security	Internet of things	Distributed Control systems
	Smart manufacturing	Cyber security	Design patterns
Semántico	Energy	Sensor	Attack
	Optimization	Energy	Database
	Water	Product	Voltage
	Evironment	Optimization	Experiments
	Cost	Attack	Primary

Por último, el tercer tema más publicado es *Material Science* en el que coinciden Estados Unidos y China, pues según el análisis bibliométrico, las palabras siguen siendo similares a los temas anteriores como es *Cloud computing*, *Internet of things* o *Cyber security.* No obstante, al hacer el análisis semántico se encuentran términos nuevos como: *Energy, Optimization, Water* y *Product,* dando mayor información de posibles subtemas asociados a la aplicación de esta ciencia. Al contrario de las grandes potencias, para Colombia el tercer tema se focaliza en *Biochemistry* y los términos más frecuentes son: *Industrial control Systems, DC_AC Power converts o Design Patterns.*

5.2.4 Análisis de frecuencia de palabras

En el caso del análisis bibliométrico (Tabla 10), se evidencia que el término más frecuente en los tres países es *Cyber physical Systems*, seguido de *Electric power transmission networks* y *Embedded systems*. Además, es

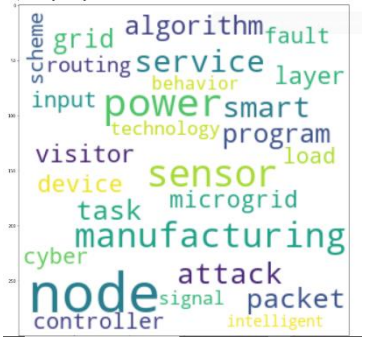
importante mostrar que países potencias como Estados Unidos y China se mueven bajo las mismas temáticas que por el contrario, en Colombia el enfoque es diferente. A su vez es interesante mencionar que la inversión en investigación y desarrollo en Colombia es baja y la cantidad de artículos publicados en los años evaluados solo llega a 8 y afecta directamente en un -6% por debajo de la meta que se tiene trazada para el 2030. (La Agenda 2030 en Colombia - Objetivos de Desarrollo Sostenible, n.d.). De lo anterior es posible inferir que la cantidad de artículos es proporcional al dinero invertido para investigación, también se observa que actualmente países subdesarrollados abordan temáticas que países potencias han trabajado en años anteriores.

Tabla 10. Frecuencia de palabras claves por país

Palabras Claves		
United States	China	Colombia
Cyber-Phisical System	Cyber-Phisical System	Cyber-Phisical System
Electric Power transmission networks	Electric Power transmission networks	Additive Manufacturing
Embeded systems	Embeded systems	Distributed Control Systems
Internet of things (IoT)	Internet of things	RAMI 4.0
Network Security	Network Security	Industry 4.0

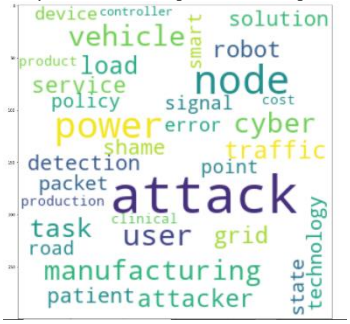
Por un lado, en el caso de Estados Unidos el término más frecuente es *Cyber physical System* seguido de *Electric power transmission networks*, *Embedded systems*, I*nternet of things* y *Network security*, se presentan los resultados del análisis semántico (Gráfica 10) producto de una técnica LDA en donde se evidencia que palabras como *Node*, *Power*, S*ensor*, *Smart* logran ser coherentes juntas y crear un contexto. Es así como estas palabras pueden establecer una relación con los términos más frecuentes en el análisis bibliométrico. Sin embargo, la nube de palabras expone un panorama más amplio de las temáticas que se están publicando en Estados Unidos con términos como *Microgrid*, *signal device*, *algorithm*, *controller*, entre otros.

Gráfica 10. Nube de palabras más probables encontradas con LDA para Estados Unidos



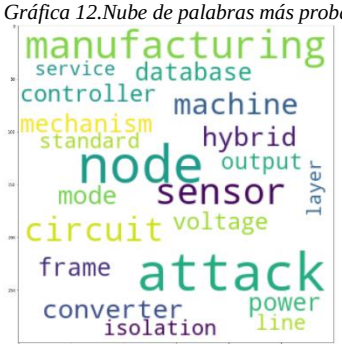
Por otro lado, los resultados de China (Tabla 10) indican que al igual que en Estados Unidos, el término más frecuente continúa siendo *Cyber Physical systems*, seguido de términos como: *Electric power transmission networks*, *Embedded systems, Internet of things* y *Network Security*. En cuanto al análisis semántico, tienen especial relevancia palabras como: *Attack, power* y *node*. Adicionalmente, se encuentran palabras como: *Robot, smart* y *technology*, términos muy cercanos a Industria 4.0 y el internet de las cosas pero que hasta ahora no habían sido mencionados en el análisis bibliométrico.

Gráfica 11.Nube de palabras más probables encontradas con LDA para China



Con referencia a los términos más frecuentes de Colombia (Tabla 10), se encontraron *Cyber Physical Systems*, *additive manufacturing* y *Distributed control* systems. De forma similar, según el análisis semántico las palabras más relevantes siguen siendo *Attack*, *node* y *sensor*. No obstante, nuevas palabras aparecen como es el caso de *circuit, mechanism, controllers* y *database*, entre otras; mostrando así un enfoque mucho más claro acerca de los subtemas que se están tratando en los artículos y que en este caso estarían relacionados a circuitos, energía y voltaje.

Gráfica 12.Nube de palabras más probables encontradas con LDA para Colombia



5.2.5 Análisis de coocurrencias

Se realizó un análisis de coocurrencias para todos los artículos del corpus (Tabla 11), con la finalidad de encontrar y medir la frecuencia de apariciones conjuntas de dos o más términos, para así posteriormente comparar la coocurrencia de palabras dentro del corpus y los resultados hallados mediante el análisis bibliométrico (Gráfica 4).

Tabla 11.Coocurrencia de palabras en el corpus del texto.

Grupos de palabras continuas	Co-ocurrencias
Cyber Physical System	1067
Wireless sensor networks	1023
Smart grid	595
Data mining	336
Embedded systems	113
Distributed sysems	104

A partir de lo anterior, se puede deducir que mediante el procedimiento de coocurrencias realizado sobre el corpus de análisis semántico se logró extraer temas relevantes y transversales a todo el documento, esto gracias a la formación de bigramas frecuentes, por ejemplo, *Embedded systems* con una frecuencia de 113 apariciones en los documentos, por lo cual se entiende que es un grupo de palabras relevantes. De igual manera *Embedded systems* es el segundo conjunto de palabras más relevantes en la Gráfica 4. Además cabe resaltar que, en el caso de los resultados del análisis semántico, existen otros términos con un mayor valor, por lo que se infiere que a

pesar de no coincidir los valores de las coocurrencias, sí logran coincidir la mayoría de los términos obtenidos mediante los dos procesos.

Al intentar realizar la comparación de los dos análisis, las coocurrencias basadas en el corpus ofrecen resultados similares a los de la Gráfica 4 como la jerarquía de los términos obtenidos; sin embargo, se observan diferencias en las magnitudes ya que los métodos usan métricas de cálculo diferentes y esto dificulta su comparación. Lo anterior permite inferir que las coocurrencias encontradas mediante el contenido semántico no se repiten lo suficiente dentro del corpus de los documentos, dando a entender que los términos giran en torno a distintos subtemas con tanta variabilidad que no son suficientes para crear conexiones relevantes, pero que técnicas como la clusterización y el LDA siguen siendo las mejores para profundizar en los documentos.

6. Limitaciones, conclusiones y recomendaciones

6.1 Limitaciones

Al momento de abordar el estudio se hallaron limitaciones que dificultaron el inicio del desarrollo del documento. Por una parte, buscar herramientas con el fin de correr el análisis bibliométrico fue complicado ya que se encontraron pocos softwares a disposición, pues muchos de estos eran de pago anticipado y los gratuitos no manejaban un sistema que permitiera encontrar los indicadores buscados.

Por otra parte, respecto a la base de datos es importante mencionar que la limpieza y tratamiento de ésta es un paso esencial para el éxito del estudio. De acuerdo con esto, Scopus divide los artículos en diferentes campos de trabajo por medio de etiquetas que identifican los temas de los que trata un artículo, una de las limitaciones se presentó al tratar de indagar sobre las temáticas de cada documento, pues la información obtenida de sus etiquetas a partir del análisis de la base de datos limitaba el manejo de estos. Asimismo, la información de las etiquetas no se encontraba contenida allí cuando se descargaba la base de datos y únicamente se podía visualizar en el buscador, por lo que evidentemente se perdía la información acerca de la agrupación por temas de los artículos, lo cual dificultó la comparación con los resultados obtenidos de la minería de texto, y obligó a realizar esta revisión de manera manual con la posibilidad de cometer errores. En cuanto a la revisión de fuentes bibliográficas, el método que más se implementó fue el análisis bibliométrico. Todo lo anterior explica cómo los recursos e indicadores disponibles son limitados para la óptima comparación de los resultados entre los dos enfoques propuestos en el trabajo.

6.2 Conclusiones

Este artículo condujo una investigación desde los análisis bibliométrico y semántico como métodos complementarios desde sus diferentes enfoques que brindaron información sobre un conjunto de artículos evaluados, desde sus aspectos más generales como palabras claves, autores, país de publicación, entre otros, y el análisis del corpus en su totalidad con el fin de agrupar documentos similares (clústeres) que ayudaran a identificar los temas más significativos dentro de la literatura sobre *Cyber physical Systems* o sistemas ciber físicos.

Fue así que a partir de la bibliometría se logró evidenciar cómo *Cyber Physical Systems* es una temática que desde el año 2009 hasta el año 2019 ha tenido una tendencia creciente, particularmente enfocada en las temáticas de *Engineering, Computer Science y Materials Science*, esto debido en gran medida a que con la llegada de la cuarta revolución industrial "las tecnologías de fabricación y de la información se han integrado para crear innovadores sistemas de manufactura" (Dra Carmen Berenice Ynzunza-Cortés et al., 2017), los cuales incorporan al hombre con la máquina logrando mayor optimización de los procesos. Además se constata cómo Estados Unidos, China y Alemania, cuentan con el mayor número de publicaciones científicas (69%) y

mayor número de citaciones por otros autores. En contraposición a Colombia que cuenta con sólo 8 artículos publicados del tema en los tres años evaluados.

Con el fin de entender el contexto de las palabras contenidas en el corpus se logró diseñar un algoritmo de minería de texto con el que se encontró que los 1075 artículos están clasificados en tres grandes clústeres, para el primero las temáticas se relacionan con modelos, manufactura, industria, energía y máquinas inteligentes, el segundo clúster está relacionado con temas de transporte, seguridad, mecanismos inteligentes y estrategias y el último clúster contiene temas enfocados a la implementación de los *Cyber Physical Systems*, redes e internet de las cosas en diferentes subtemas como lo son el digital, económico, simulación, entre otros.

Finalmente, con esta investigación se concluye que sí hay diferencia entre los resultados obtenidos para las dos metodologías, pues a partir de la elaboración y proceso del análisis semántico se obtuvo la información necesaria para describir y profundizar sobre los conceptos que engloban las temáticas identificadas en el análisis bibliométrico. Esto, gracias al contexto creado entre palabras contenidas en el corpus. Lo anterior, permite englobar análisis cualitativos de texto sobre la literatura científica que pueden ser usados para establecer el estado actual del ODS 9.5.

6.3 Recomendaciones

Desde la metodología planteada y para futuros estudios se recomienda tomar publicaciones de diferentes bases de datos que puedan complementar la información y no solo de tipo open Access, esto con la finalidad de englobar la mayor cantidad de textos y que no se delimiten los datos a un tema específico, y así poder visibilizar en términos generales y de manera transversal el estado actual de la investigación científica.

Con el objetivo de encontrar relación y contexto entre los términos como complemento a esta investigación, se recomienda no ingresar una matriz término-documento a los clústeres, por el contrario, se sugiere utilizar una matriz término-término que haya sido sometida a un proceso de reducción de dimensionalidad LDA.

Finalmente, para trabajos futuros, escenarios experimentales con otros esquemas de ponderación serán considerados para analizar la influencia que la medida aquí utilizada tuvo en los resultados señalados. Investigaciones futuras pueden utilizar patrones léxicos si estos se construyen a través de herramientas no supervisadas o aplicando un enfoque semi-supervisado.

Respecto a los Anexos o Apéndices

La Tabla 12 describe el nombre y contenido de cada anexo presentando en esta investigación.

Tabla 12. Anexos utilizados para el desarrollo de la investigación.

Anexos	Descripción
Anexo 1	Manual Bibliométrico con los indicadores escogidos a tratar por los autores
Anexo 2	Algoritmo en Python para mineria de texto
Anexo 3	Listado de Tokens obtenido después de la preparación del corpus

Referencias

(CSIC), O. and Sci. R. G. (2016). Compendium of bibliometric science indicators. In OECD, Paris. http://oe.cd/scientometrics

Abba Adam, N. B. Z. H. G. M. S. I. D. K. A. (2019). Comprehensive Bibliometric Analysis of Published Research in Cyber Physical System from 2009 to 2018. International Journal of Recent Technology and Engineering (IJRTE), 8(3S2). https://doi.org/10.35940/ijrte.C1237.1083S219

Abdrabou, Y., Mohsen, F., El Mougy, A., Bahron, H., & Alam, S. (2018). IOP Conference Series: Earth and Environmental Science Related content CHEOPS: Cultural Heritage Enhancement Over Cyber-Physical Systems Research Frontiers and Way Forward. IOP Conf. Ser.: Earth Environ. Sci, 117, 12049. https://doi.org/10.1088/1755-1315/117/1/012049

Ahmi, A., Elbardan, H., & Raja Mohd Ali, R. H. (2019, May 3). Bibliometric analysis of published literature on industry 4.0. ICEIC 2019 - International Conference on Electronics, Information, and Communication. https://doi.org/10.23919/ELINFOCOM.2019.8706445

Alfin EEES. (n.d.). Retrieved May 12, 2020, from http://www.mariapinto.es/alfineees/buscar/como.htm#inicio

Arbeláez, M. C., & Onrubia, J. (2014). Análisis bibliométrico y de contenido. Dos metodologías complementarias para el análisis de la revista colombiana Educación y Cultura. Revista de Investigaciones UCM, 14(23), 14–31. https://bit.ly/31A2xbH

Bayer, A. E., Smart, J. C., & McLaughlin, G. W. (1990). Mapping intellectual structure of a scientific subfield through author cocitations. Journal of the American Society for Information Science, 41(6), 444–452. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<444::AID-ASI12>3.0.CO;2-J

Beier, G., Niehoff, S., & Xue, B. (2018). More sustainability in industry through Industrial Internet of Things? Applied Sciences (Switzerland), 8(2). https://doi.org/10.3390/app8020219

Beier, G., Ullrich, A., Niehoff, S., Reißig, M., & Habich, M. (2020). Industry 4.0: How it is defined from a sociotechnical perspective and how much sustainability it includes – A literature review. Journal of Cleaner Production, 259. https://doi.org/10.1016/j.jclepro.2020.120856

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In Journal of Machine Learning Research (Vol. 3).

Broadus, R. N. (1987). Toward a definition of "bibliometrics." Scientometrics, 12(5–6), 373–379. https://doi.org/10.1007/BF02016680

Carnerud, D. (n.d.). Exploration of text mining methodology through investigation of QMOD-ICQSS proceedings.

Chinchilla Rodríguez Director, Z., & de Moya Anegón DEPARTAMENTO BIBLIOTECONOMÍA Y DOCUMENTACIÓN FACULTAD DE BIBLIOTECONOMÍA Y DOCUMENTACIÓN, F. DE. (n.d.). TESIS DOCTORAL.

Codina, L. (2017, October 28). Ecuaciones de búsqueda: qué son y cómo se utilizan en bases de datos académicas · 1 - Operadores booleanos. https://www.lluiscodina.com/ecuaciones-de-busqueda-bases-datos-operadores-booleanos/

Cruz Salazar, L. A., Ryashentseva, D., Lüder, A., & Vogel-Heuser, B. (2019). Cyber-physical production systems architecture based on multi-agent's design pattern—comparison of selected approaches mapping four agent patterns. International Journal of Advanced Manufacturing Technology, 105(9), 4005–4034. https://doi.org/10.1007/s00170-019-03800-4

Danvila-del-Valle, I., Estévez-Mendoza, C., & Lara, F. J. (2019). Human resources training: A bibliometric analysis. Journal of Business Research, 101, 627–636. https://doi.org/10.1016/j.jbusres.2019.02.026

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

Depari, A., Fernandes Carvalho, D., Bellagente, P., Ferrari, P., Sisinni, E., Flammini, A., & Padovani, A. (2019). An IoT Based Architecture for Enhancing the Effectiveness of Prototype Medical Instruments Applied to Neurodegenerative Disease Diagnosis. Sensors, 19(7), 1564. https://doi.org/10.3390/s19071564

Dra Carmen Berenice Ynzunza-Cortés, I., Manuel Izar-Landeta, J., Jacqueline Guadalupe Bocarando-Chacón, D., Aguilar-Pereyra, F., & en Martín Larios-Osorio, M. I. (2017). El Entorno de la Industria 4.0: Implicaciones y Perspectivas Futuras Implications and Perspectives of Industry 4.0. In ConCiencia Tecnológica, ISSN-e 1405-5597, No. 54 (julio-diciembre), 2017, págs. 33-45 (Issue 8). Departamento de Desarrollo Académico. https://dialnet.unirioja.es/servlet/articulo?codigo=6405835&info=resumen&idioma=ENG

Erkens, M., Bodemer, D., & Hoppe, H. U. (2016). Improving collaborative learning in the classroom: Text mining based grouping and representing. International Journal of Computer-Supported Collaborative Learning, 11(4), 387–415. https://doi.org/10.1007/s11412-016-9243-5

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: Five methodological recommendations. European Journal of Information Systems, 21(1), 70–86. https://doi.org/10.1057/ejis.2010.61

Fernandes de Mesquita, R., Xavier, A., Klein, B., & Regina Ney Matos, F. (2017). Mining and the Sustainable Development Goals: A Systematic Literature Review. Geo-Resources Environment and Engineering, 2(January). https://doi.org/10.15273/gree.2017.02.006

García-Morales, V. J., Jiménez-Barrionuevo, M. M., & Gutiérrez-Gutiérrez, L. (2012). Transformational leadership influence on organizational performance through organizational learning and innovation. Journal of Business Research, 65(7), 1040–1050. https://doi.org/10.1016/j.jbusres.2011.03.005

Glenisson, P., Glänzel, W., & Persson, O. (2005). Combining full-text analysis and bibliometric indicators. A pilot study. Scientometrics, 63(1), 163–180. https://doi.org/10.1007/s11192-005-0208-0

Goal 9: Industrial innovation and infrastructure | UNDP. (n.d.). Retrieved April 9, 2020, from https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-9-industry-innovation-and-infrastructure.html

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. In Journal of Emerging Technologies in Web Intelligence (Vol. 1, Issue 1, pp. 60–76). Academy Publisher. https://doi.org/10.4304/jetwi.1.1.60-76

H. Christopher D. Manning, P. R. (2008). Introduction to Information Retrieval | BibSonomy. https://www.bibsonomy.org/bibtex/6b779b7c4b6e85b614707e21f4f86a48

Henning, Kagermann (National Academy of Science and Engineering). Wolfgang, Wahlster (German Research Center for Artificial Intelligence). Johannes, H. (Deutsche P. A. (2013). Germany - INDUSTRIE 4.0. Final Report of the Industrie 4.0 WG, April, 82. https://doi.org/10.13140/RG.2.1.1205.8966

Home - The VantagePoint. (n.d.). Retrieved October 18, 2020, from https://www.thevantagepoint.com/

Huang, J., You, J. X., Liu, H. C., & Song, M. S. (2020). Failure mode and effect analysis improvement: A systematic literature review and future research agenda. In Reliability Engineering and System Safety (Vol. 199, p. 106885). Elsevier Ltd. https://doi.org/10.1016/j.ress.2020.106885

Iannino, V., Colla, V., Denker, J., & Göttsche, M. (2019). A CPS-Based Simulation Platform for Long Production Factories. Metals, 9(10), 1025. https://doi.org/10.3390/met9101025

Indicadores de sostenibilidad para la industria minera extractiva en Uige, Angola. (n.d.). Retrieved May 12, 2020, from http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1993-80122019000200233&lng=es&nrm=iso&tlng=es

Informe de los Objetivos de Desarrollo Sostenible 2019. (2019).

Informe del Grupo Interinstitucional y de Expertos sobre los Indicadores de los Objetivos de Desarrollo Sostenible Nota del Secretario General. (n.d.).

Kipper, L. M., Furstenau, L. B., Hoppe, D., Frozza, R., & Iepsen, S. (2020). Scopus scientific mapping production in industry 4.0 (2011–2018): a bibliometric analysis. International Journal of Production Research, 58(6), 1605–1627. https://doi.org/10.1080/00207543.2019.1671625

La Agenda 2030 en Colombia - Objetivos de Desarrollo Sostenible. (n.d.). Retrieved May 2, 2020, from https://www.ods.gov.co/es/faq

Lee, C. H., Chen, C. H., Lin, C., Li, F., & Zhao, X. (2019). Developing a quick response product configuration system under industry 4.0 based on customer requirement modelling and optimization method. Applied Sciences (Switzerland), 9(23). https://doi.org/10.3390/app9235004

Lee, J., Bagheri, B., & Kao, H. A. (2015). A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters, 3, 18–23. https://doi.org/10.1016/j.mfglet.2014.12.001

Lee, J., Lapira, E., Bagheri, B., & Kao, H. an. (2013). Recent advances and trends in predictive manufacturing systems in big data environment. Manufacturing Letters, 1(1), 38–41. https://doi.org/10.1016/j.mfglet.2013.09.005

Li, X., Yuan, J., Shi, Y., Sun, Z., & Ruan, J. (2020). Emerging Trends and Innovation Modes of Internet Finance—Results from Co-Word and Co-Citation Networks. Future Internet, 12(3), 52. https://doi.org/10.3390/fi12030052

Liang, F., Hatcher, W. G., Liao, W., Gao, W., & Yu, W. (2019). Machine Learning for Security and the Internet of Things: The Good, the Bad, and the Ugly. IEEE Access, 7, 158126–158147. https://doi.org/10.1109/ACCESS.2019.2948912

Lin, J., Yu, W., Zhang, N., Yang, X., & Ge, L. (2018). Data Integrity Attacks Against Dynamic Route Guidance in Transportation-Based Cyber-Physical Systems: Modeling, Analysis, and Defense. IEEE Transactions on Vehicular Technology, 67(9), 8738–8753. https://doi.org/10.1109/TVT.2018.2845744

Liu, B. (n.d.). [(Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data)] [Author: Bing Liu] [Dec-2011].

López, N. (2018, February 17). ¿Está cerca la Industria 4.0 en Colombia? https://www.larepublica.co/internet-economy/esta-cerca-la-industria-40-en-colombia-2600242

Machado, C. G., Winroth, M. P., & Ribeiro da Silva, E. H. D. (2020). Sustainable manufacturing in Industry 4.0: an emerging research agenda. International Journal of Production Research, 58(5), 1462–1484. https://doi.org/10.1080/00207543.2019.1652777

Manning, C. D., Schütze, H., & Weikurn, G. (2002). Foundations of Statistical Natural Language Processing. SIGMOD Record, 31(3), 37–38. https://doi.org/10.1145/601858.601867

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12), 1650–1654. https://doi.org/10.1109/TPAMI.2002.1114856

Mayo, M. (n.d.). A General Approach to Preprocessing Text Data. Retrieved May 13, 2020, from https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html

McBurney, M. K., & Novak, P. L. (2002). What is bibliometrics and why should you care? IEEE International Professional Communication Conference, 108–114. https://doi.org/10.1109/ipcc.2002.1049094

Miner, G. D., Elder, J., & Nisbet, R. A. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. In Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Elsevier Inc. https://doi.org/10.1016/C2010-0-66188-8

Moisescu, M. A., Sacala, I. S., Dumitrache, I., Caramihai, S. I., Barbulescu, B., & Danciuc, M. (2019). A cyber-physical systems approach to cognitive enterprise. Periodicals of Engineering and Natural Sciences, 7(1), 337–342. https://doi.org/10.21533/pen.v7i1.378

Mosenia, A., Sur-Kolay, S., Raghunathan, A., & Jha, N. K. (2017). DISASTER: Dedicated Intelligent Security Attacks on Sensor-Triggered Emergency Responses. IEEE Transactions on Multi-Scale Computing Systems, 3(4), 255–268. https://doi.org/10.1109/TMSCS.2017.2720660

Mongi, C. E., Langi, Y. A. R., Montolalu, C. E. J. C., & Nainggolan, N. (2019). Comparison of hierarchical clustering methods (case study: Data on poverty influence in North Sulawesi). IOP Conference Series: Materials Science and Engineering, 567(1). https://doi.org/10.1088/1757-899X/567/1/012048

Muhuri, P. K., Shukla, A. K., & Abraham, A. (2019). Industry 4.0: A bibliometric analysis and detailed overview. Engineering Applications of Artificial Intelligence, 78, 218–235. https://doi.org/10.1016/j.engappai.2018.11.007

*Nations, U. (n.d.). Naciones Unidas | Paz, dignidad e igualdad
en un planeta sano.*

Ochoa, J. L., Valencia-García, R., Perez-Soltero, A., & Barceló-Valenzuela, M. (2013). A semantic role labelling-based framework for learning ontologies from Spanish documents. Expert Systems with Applications, 40(6), 2058–2068. https://doi.org/10.1016/j.eswa.2012.10.017

Pei, W., Ma, X., Deng, W., Chen, X., Sun, H., & Li, D. (2019). Industrial multi-energy and production management scheme in cyber-physical environments: A case study in a battery manufacturing plant. IET Cyber-Physical Systems: Theory and Applications, 4(1), 13–21. https://doi.org/10.1049/iet-cps.2018.5029

PNUD - Programa de las Naciones Unidas para el Desarrollo. (n.d.). Retrieved April 9, 2020, from https://www.undp.org/content/undp/en/home.html

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining, 399–408. https://doi.org/10.1145/2684822.2685324

Sheela, S., & Bharathi, T. (2013). Analyzing Different Approaches of Text Mining Techniques and Applications. International Journal of Computer Science Trends and Technology, 6. www.ijcstjournal.org

Sniderman, B. (n.d.). Industry 4.0 and manufacturing ecosystems Exploring the world of connected enterprises.

Tesch da Silva, F. S., da Costa, C. A., Paredes Crovato, C. D., & da Rosa Righi, R. (2020). Looking at energy through the lens of Industry 4.0: A systematic literature review of concerns and challenges. In Computers and Industrial Engineering (Vol. 143, p. 106426). Elsevier Ltd. https://doi.org/10.1016/j.cie.2020.106426

Trappey, A. J. C., Trappey, C. V., Govindarajan, U. H., Sun, J. J., & Chuang, A. C. (2016). A Review of Technology Standards and Patent Portfolios for Enabling Cyber-Physical Systems in Advanced Manufacturing. In IEEE Access (Vol. 4, pp. 7356–7382). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2016.2619360

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37, 141–188. https://doi.org/10.1613/jair.2934

unctad. (2019). A Framework for Science, Technology and Innovation Policy Reviews: Harnessing Innovation for Sustainable Development.

Unutmaz Durmusoglu, Z. D., & Kocabey ÇiFtçi, P. (2018). The evolution of the industry 4.0: A retrospective analysis using text mining. ACM International Conference Proceeding Series, 1–5. https://doi.org/10.1145/3234698.3234757

Vanhala, M., Lu, C., Peltonen, J., Sundqvist, S., Nummenmaa, J., & Järvelin, K. (2020). The usage of large data sets in online consumer behaviour: A bibliometric and computational text-mining–driven analysis of previous research. Journal of Business Research, 106, 46–59. https://doi.org/10.1016/j.jbusres.2019.09.009

VOSviewer - Visualizing scientific landscapes. (n.d.). Retrieved November 7, 2020, from https://www.vosviewer.com/

Wessel, R., Odermatt, J., & Vlastou-Dimopoulou, F. (2019). Organisation for Economic Co-operation and Development (OECD). In Research Handbook on the European Union and International Organizations. https://doi.org/10.4337/9781786438935.00024

Yildiz, T. (2019). Examining the Concept of Industry 4.0 Studies Using Text Mining and Scientific Mapping Method. Procedia Computer Science, 158, 498–507. https://doi.org/10.1016/j.procs.2019.09.081Qué es

Zahariev, R., Valchkova, N., & Wagatsuma, H. (2020). Service Robots for Special Education of Children with Disabilities: Robotized Systems for Social Applications. ACM International Conference Proceeding Series, 300–306. https://doi.org/10.1145/3407982.3408023