Facultad de Arquitectura e Ingeniería
Universidad Internacional SEK

Trabajo de investigación de fin de carrera titulado:

Aplicación de técnicas de análisis de datos para obtener líneas de investigación específicas para el Ecuador. Caso de estudio: Computer Science en Scopus

Realizado por:

Ing. Geovanny Javier Páez García

Director del proyecto:

Ph.D. Diego Fernando Riofrío Luzcando

Como requisito para la obtención del título de:
Master en Tecnologías de la Información con Mención en Seguridad de Redes y Comunicaciones

Marzo, 2019

DECLARACIÓN JURAMENTADA

Yo, GEOVANNY JAVIER PÁEZ GARCÍA, con cédula de identidad 172264611-2, declaro bajo juramento que el trabajo aquí desarrollado es de mi autoría, que no ha sido previamente presentado para ningún grado a calificación profesional; y, que ha consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración, cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la UNIVERSIDAD INTERNACIONAL SEK, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normativa institucional vigente.

Ing. Geovanny Javier Páez García
Marzo, 2019

DECLARACIÓN DEL DIRECTOR DE TESIS

Declaro haber dirigido este trabajo a través de reuniones periódicas con el estudiante, orientando sus conocimientos y competencias para un eficiente desarrollo del tema escogido y dando cumplimiento a todas las disposiciones vigentes que regulan los Trabajos de Titulación.

PhD. Diego Fernando Riofrío Luzcando
Marzo, 2019

LOS PROFESORES INFORMANTES

Los Profesores informantes:

VERÓNICA ELIZABETH RODRÍGUEZ ARBOLEDA

CHRISTIAN DAVID PAZMIÑO FLORES

Después de revisar el trabajo presentado lo han calificado como apto para su defensa oral ante el tribunal examinador.

Ing. Verónica Rodríguez, MBA. Ing. Christian Pazmiño, Mg.

Marzo, 2019

DECLARACIÓN DE AUTORÍA DEL ESTUDIANTE

Declaro que este trabajo es original, de mi autoría, que se han citado las fuentes correspondientes y que en su ejecución se respetaron las disposiciones legales que protegen los derechos de autor vigentes.

Ing. Geovanny Javier Páez García
Marzo, 2019

Agradecimientos

En primer lugar, a Dios, por darme el regalo de la vida. Por su bendición y sabiduría que ha guiado todas mis decisiones y mis éxitos y logros personales.

A mi familia y a mis queridos padres, Norma y Manuel, que me dieron la vida. Por enseñarme el valor del sacrificio y esfuerzo y por los buenos valores que me inculcaron a lo largo de toda mi vida profesional y personal, sin ustedes esto no fuera posible. A mi hermano, a mis primos y a mis tíos y tías, que siempre me han apoyado en todo, y con los que he compartido mis éxitos y fracasos tanto profesionales como personales. Y en especial a Natasha que me ha acompañado día a día durante estos años, por ser un apoyo incondicional en todo este tiempo, por aguantarme los cambios de humor y por siempre ayudarme a salir adelante. Sólo puedo pedirle perdón y darle el mayor de mis agradecimientos.

A todos mis amigos que me apoyaron de algún modo a lo largo de mi carrera, y a lograr este objetivo.

A mi tutor de tesis Diego Riofrío, que con su paciencia, guía, consejos, conocimiento y tiempo, me ayudó a alcanzar un escalón más en mi vida profesional. Gracias por confiar en mí y por su ayuda, comprensión y por todo el tiempo para las tutorías y correcciones de la tesis.

Dedicatoria

Dedico esta tesis a mis padres que siempre estuvieron aquí para apoyarme y motivarme a ser mejor cada día.
A las personas que estuvieron conmigo que de alguna u otra manera me apoyaron a culminar este objetivo, gracias por todo.
Sé que todos mis seres más queridos estarán profundamente orgullosos de mi, y sólo por eso ha merecido la pena todo el esfuerzo y sacrificio de estos años.

Resumen

Las líneas de investigación deberían adaptarse de manera efectiva a todo el proceso de la investigación. Para que esto ocurra, es condición esencial que exista una metodología para obtención de líneas de investigación específicas para el Ecuador. Sin embargo, dicha metodología no existe y mucho menos las líneas específicas de importancia para el país.
En este trabajo se ha llevado a cabo un estudio que ha permitido comprender a detalle cómo proponer la obtención de líneas de investigación específicas para el Ecuador que ayuden al desarrollo del país, mediante técnicas de análisis de datos. El estudio se llevó a cabo únicamente con investigaciones en ciencias de la computación para asegurar que la variabilidad entre los distintos ámbitos de estudio no afectase a los resultados. Para llegar a este objetivo, se utilizó la metodología MIDANO con la cual se describió los procesos que participan en la creación de líneas de investigación.
En primer lugar, se realizó un diagnóstico de la situación actual de las líneas de investigación del Ecuador, a través de este análisis se hizo una asociación de las líneas de investigación propuestas por el MINTEL y las estandarizadas por la ACM. Esta asociación se utilizó para obtener las cadenas de búsqueda que permitieron conseguir los artículos de relevancia para el Ecuador.
A partir de esto se ha representado todos los conceptos y relaciones que emergieron del análisis de datos de los trabajos de investigación indexados en Scopus, obteniendo como resultados las líneas de investigación específicas para el país, las cuales se validaron mediante varios artículos de impacto. Para facilitar la comprensión de los resultados, estos fueron analizados, procesados y se ha creado una representación en tablas y gráficos del desarrollo. Con la finalidad de mostrar la utilidad de la solución propuesta, se presenta una aplicación práctica, que permite la extracción de líneas de investigación de artículos científicos.

Palabras Claves: analítica de datos, análisis de texto, creación de líneas de investigación, Scopus, MIDANO

Abstract

The lines of research should effectively adapt to the entire research process. For this to happen, it is essential to have a methodology in order to obtain specific lines of research for Ecuador. However, this methodology does not exist, much less specific lines of research of importance for the country.
In this thesis work, a research study has been carried out, it has allowed to understand in detail how to propose the obtaining of specific lines of research for Ecuador that help the development of the country, through data analysis techniques. The study was carried out only with research in Computer Science to ensure that the variability among the different study areas does not affect the results. To reach this goal, the MIDANO methodology was used to describe the processes participating in the creation of the lines of research.
First, a diagnosis of the current situation of the lines of research of Ecuador was carried out, through this analysis an association of the proposed lines of research by MINTEL and those standardized by the ACM was made. This association was used to obtain the search chains that allowed to obtain articles of relevance for Ecuador.
From this, all the concepts and relationships emerged from the analysis of data from the research works indexed in Scopus have been represented, obtaining as a result the specific lines of research for the country, they were validated through several impact articles. To facilitate the understanding of the results, they were analyzed, processed and a representation in tables and graphs of the development was created. In order to show the usefulness of the proposed solution, a practical application is presented, which allows the extraction of the lines of research of scientific articles.

Keywords: data analytics, text analysis, creation of research lines, Scopus, MIDANO

Tabla de Contenidos

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

1.1 Planteamiento del Problema

Las publicaciones científicas, más allá de ser un resultado académico de investigación, deben ser elementos definitorios para el logro del desarrollo de un país, es por ello que las mismas deben tener un alto grado de responsabilidad y pertinencia según el contexto propio de la investigación. A nivel mundial, a través de las universidades y los institutos de investigación, los países han buscado ser líderes en el ámbito de la investigación dentro de la región y a nivel mundial. Este afán ha llevado a un crecimiento acelerado de la cantidad de documentos científicos dentro de los indexadores como Scopus, lo que se ha hecho que estas se conviertan en un gran reservorio de información.

La información reduce la incertidumbre y, por lo tanto, permite tomar decisiones estratégicas para el país. Sin embargo, al tener una base de datos tan grande, por ejemplo, en Scopus se tiene alrededor de 23.343 revistas indexadas y activas en el 2017 (Elsevier, 2018), la capacidad de analizarlos o interpretar la información disminuye, por lo que es necesario el uso de herramientas que permitan la extracción de conocimiento útil a partir de grandes conjuntos de datos. Es aquí donde tiene cabida la analítica de datos, un área de investigación que pretende dar respuesta a la necesidad de procesar y analizar grandes volúmenes de datos, con el fin de encontrar y descubrir conocimiento útil contenido en los datos.

Durante los últimos años el Ecuador no ha sido ajeno a este proceso. A pesar de representar una pequeña parte de la producción científica mundial y regional, ha experimentado, como consecuencia de políticas públicas, un incremento sostenido de sus publicaciones y su presencia en Scopus (ver figura 1.1). Lamentablemente las publicaciones se han producido

sin ningún direccionamiento que permitan el desarrollo del país.

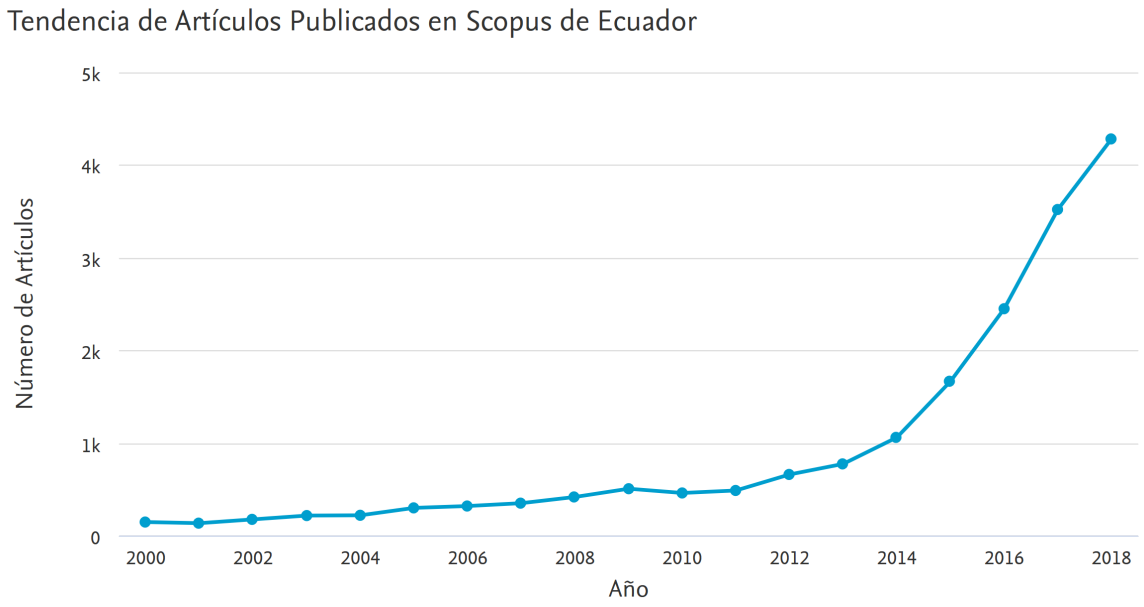

Fig. 1.1 Línea de Tendencia de las Publicaciones en Scopus de Ecuador (Elsevier, 2018).

Adicionalmente, organismos internacionales como la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO por sus siglas en inglés) y la Association for Computing Machinery (ACM), han definido una clasificación de líneas de investigación para modelos curriculares.

En el 2016, el Ministerio de Telecomunicaciones y de la Sociedad de la Información (MINTEL) del Ecuador, en un ecosistema de la academia, el gobierno y la industria, intentaron implementar un libro blanco de líneas de investigación (MINTEL, 2016). El mismo que no fue efectuado. Más tarde, en julio de 2018, el MINTEL publicó las primeras líneas de investigación prioritarias en la primera edición del "Libro Blanco de la Sociedad de la Información y del Conocimiento (LBSIC)", el cual es un instrumento de política pública dinámico que incluye estrategias y acciones para el desarrollo del país mediante la apropiación eficiente de las TICs y construir una Sociedad de la Información y del Conocimiento inclusiva. (MINTEL, 2018)

Estas líneas de investigación publicadas por el MINTEL son generales y no específicas, lo que genera una dificultad de los investigadores ecuatorianos para encontrar en que temas

concretos se debe investigar, los cuales sean pertinentes para el desarrollo del país.

La situación actual del Ecuador, en relación con la creación de las líneas de investigación de importancia para el país, se describe en la tabla 1.1, del mismo se percibe que se pueden implementar mejores metodologías para la creación de las líneas de investigación específicas para el Ecuador.

Tabla 1.1 Situación actual para la creación de líneas de investigación para el Ecuador

Resultados que se obtienen	**Actores Asociados**	**Variables Asociadas**	**Actividades que se realizan**
Libro Blanco de la Sociedad de la Información y del Conocimiento (LBSIC) en el que se detallan las líneas de investigación	Investigadores del Ecuador, MINTEL, educadores, profesores, estudiantes	Interés de las entidades públicas	El MINTEL junto con la Senescyt crearon y priorizaron las líneas de investigación según su interés

Fuente: Elaboración Propia

Para apoyar a la situación actual de la creación de las líneas de investigación en el Ecuador, la presente investigación busca proponer un nuevo escenario en el cual se mejore la creación de las líneas de investigación prioritarias del Ecuador en el área de conocimiento de ciencias de la computación, mediante técnicas de analítica de datos (AD) en bases de datos científicas de relevancia, con el fin de aportar líneas específicas o temáticas que contribuyan al desarrollo del investigador ecuatoriano a nivel mundial y también al desarrollo del país.

En la tabla 1.2 se muestra el nuevo escenario para la creación de líneas de investigación para el Ecuador y se observa que se pueden crear procesos para mejorar sustancialmente la investigación en el Ecuador y obtener mejores resultados.

Las líneas de investigación de importancia para el Ecuador son definidas mediante un análisis hipotético, es por ello que, el no poseer líneas de investigación específicas, genera una ambigüedad para la investigación del país, los investigadores ecuatorianos no conocen cuales son las necesidades del país para alcanzar su desarrollo. Adicionalmente, con el incremento sustancial en gastos en investigación y desarrollo según UNESCO (2018) de 0,129 en el 2006 a 0,441 en el 2014 se crearía un gasto innecesario que no beneficiaría al país, por lo que surge la importancia de tener líneas de investigación que sean indispensables

para el país.

Tabla 1.2 Nuevo Escenario para la creación de líneas de investigación para el Ecuador

Resultados que se desean obtener	**Actores Asociados**	**Variables Asociadas**	**Actividades de AD que se realizarían**	**Funcionalidades Nuevas**
Aplicativo de Analítica de datos que permitirá al investigador ecuatoriano conocer las líneas de investigación de importancia para el Ecuador en el área de las ciencias de la computación, lo que contribuirá a tener investigaciones de calidad para tener mejores investigadores y aportar al desarrollo del país.	Investigadores	Título del artículo científico, resumen del artículo científico, palabras claves del artículo científico, información de los docentes investigadores del Ecuador y del mundo, estudiantes investigadores	Obtención de información de la base de datos de Scopus, análisis estadístico descriptivo de la base de datos, agrupamiento de la búsqueda relacionada con las distintas áreas de las ciencias de la computación, minería de texto para obtener las líneas de investigación y predicción de tendencias	Obtener las líneas de investigación de importancia para el Ecuador, dar información de las líneas de investigación de importancia para el Ecuador

Fuente: Elaboración Propia

Obtener las líneas de investigación específicas para el Ecuador, le permite, por ejemplo, al país tomar decisiones nacionales en cuanto a lo que se está publicando y al investigador ecuatoriano que tema específico investigar dentro del libro blanco del MINTEL. Además, que aspectos de esas áreas a nivel mundial son nichos por cubrir, entre otras cosas. Decisiones que un país e investigadores inteligentes tomarían, para mejorar su posicionamiento científico y asociar las líneas de investigación para el beneficio del país.

Por lo tanto, el problema de investigación se centra en que, la gran cantidad de artículos científicos y la ambigüedad en la forma de análisis de los investigadores ecuatorianos para encontrar publicaciones de relevancia, incide en la dificultad de que un investigador pueda definir un tema específico en que investigar, el cual sea de importancia para el Ecuador.

1.2 Objetivo de la Investigación

1.2.1 Objetivo General

Aplicar técnicas de analítica de datos (AD) para la determinación de las lineas de investigación específicas para el Ecuador que contribuyan al desarrollo del País, mediante la utilización de las lineas del MINTEL y la ACM.

1.2.2 Objetivos Específicos

- Identificar las líneas de investigación de estándares internacionales, para alinearlas con las áreas de desarrollo del Ecuador y así obtener los querys de búsqueda de información mediante las palabras claves obtenidas de la asociación.
- Aplicar búsquedas automatizadas en bases de datos de literatura científica mediante un Script en Python para la obtención de los artículos científicos de importancia para el Ecuador.
- Proponer líneas de investigación alineadas al “Plan Nacional de Desarrollo 2017-2021 Toda una Vida” y al "Plan de la Sociedad de la Información y del Conocimiento - MINTEL", que le ayuden en el mejoramiento de la toma de decisiones tempranas para el desarrollo tecnológico del país mediante la aplicación de técnicas de analítica de datos.

1.3 Justificación

La implementación de técnicas de analítica de datos permite aprovechar el volumen creciente de datos de publicaciones científicas a nivel nacional y mundial, para extraer conocimientos que sirvan de apoyo a la toma de decisiones que estén alineadas al “Plan Nacional de Desarrollo 2017-2021 Toda una Vida" y que además puedan determinar las líneas de investigación específicas para el país. Estas líneas de investigación específicas serán de relevancia y de beneficio para el desarrollo del Ecuador a nivel regional y mundial.

Las líneas de investigación específicas o temáticas de investigación son importantes para los investigadores ecuatorianos, ya que reduce la dificultad que tienen los investigadores en encontrar temáticas específicas dentro de las áreas definidas por el MINTEL, lo que permitirá tener una investigación de calidad. Así como explican en sus artículos varios autores (Barros et al., 2018; Laudel, 2017; Muñoz-Écija et al., 2017), entre otros, las líneas de investigación

específicas son importantes para realizar nuevas investigaciones y con mejor calidad. Además, que permiten el desarrollo del país y saber en qué ámbito se debería investigar.

Las técnicas de analítica de datos permiten la representación correcta de documentos científicos para la obtención de patrones de investigación (Aguilar, 2013), por lo que en este contexto, la analítica documental es importante para determinar las líneas de investigación y su categorización, así como la tendencia actual de la investigación y cuales permiten el desarrollo del país. Junto a los estándares internacionales de líneas de investigación se puede determinar las áreas en las cuáles Ecuador puede tener un gran impacto para el desarrollo del país. Adicionalmente, este estudio de analítica de datos servirá a los investigadores para mejorar sustancialmente el estudio en documentos académicos.

En resumen, la presente investigación se justifica dado a que explorará los diferentes tipos de documentos científicos para obtener un panorama actual de la investigación en el campo de las ciencias de la computación, lo cual permitirá obtener una categorización y tendencia actual de las líneas de investigación, para así generar las líneas específicas para el Ecuador.

1.4 Marco Teórico

En esta sección se presenta el marco teórico necesario para la realización del presente trabajo de investigación dividido en tres secciones que se detallan a continuación:

En la primera sección se presenta un análisis de las líneas de investigación existentes a nivel internacional y en el Ecuador. Este análisis se utiliza para obtener conocimiento del enfoque y del estado actual de las líneas de investigación ya que estas se utilizarán como datos de entrada para conseguir los artículos científicos de Scopus.

En la segunda parte se presenta un estudio de la situación actual de las bases de datos científicas que serán la fuente principal de información para la obtención de datos de la presente tesis. Se realiza una breve reseña de cada base de datos más relevante para el país con sus principales características y el estado actual de las mismas.

Por último, se detalla un estudio de las diferentes técnicas de analítica de datos y cuales serán utilizadas en este proyecto. La analítica de datos se utilizará para obtener conocimiento de los artículos científicos, así como el logro de los objetivos planteados en este trabajo de

investigación.

1.4.1 Líneas de Investigación

Las líneas de investigación son categorías interdisciplinarias de objetos de investigación dentro de un campo de conocimiento, que permiten resolver problemas de la sociedad o descubrir algo. Estas líneas de investigación son utilizadas por uno o más grupos de investigación para impulsar el desarrollo del país. (Sunkel, 2006)

Sus alcances y desarrollos materiales de las prácticas y saberes involucrados son transversales a los proyectos de investigación. Las líneas de investigación son subsistemas estratégicos categorizados y organizados que sirven como una guía de acción para la resolución de problemas. (Sunkel, 2006)

1.4.1.1 Líneas de Investigación de Estándar Internacional

En la actualidad existen varios organismos internacionales que plantean líneas de investigación o áreas de conocimiento como estándares para investigadores, centros e institutos de investigación, entre otros.

A continuación se detallan las dos más relevantes para este trabajo de investigación cuyo caso de estudio son las ciencias de la computación: UNESCO y ACM

a) Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO por sus siglas en inglés)

La UNESCO crea un sistema de clasificación de líneas y áreas de investigación en todos los campos de la educación, esta se denominó Clasificación Internacional Normalizada de la Educación (CINE). La primera versión de CINE se lanzó en 1970 y fue revisada en 1997, en el 2011 se realizó la segunda revisión, y es la utilizada actualmente por los sistemas educativos a nivel mundial. Esta versión fue adoptada en la Conferencia General de la UNESCO en noviembre de 2011. (UNESCO, 2018)

CINE proporciona un marco integral para clasificar y organizar los programas educativos mediante la aplicación de definiciones uniformes y acordadas internacionalmente para man-

tener un estándar internacional de los sistemas educativos en todos los países.

La CINE abarca ramas, áreas y dominios de conocimiento de un amplio campo de estudio, basadas en programas de educación y clasificaciones relacionadas. Los campos de educación y formación son las líneas de investigación que adopta UNESCO, las cuales son referentes nacionales e internacionales para definir líneas de investigación del sector académico y productivo de un país.

b) Association for Computing Machinery (ACM)

La ACM (Association for Computing Machinery) es una Biblioteca digital que implementó un Sistema de Clasificación de Computación ACM (CCS). El CCS se creó en el año de 1998, teniendo su última versión en el año 2012, el mismo fue creado por un grupo de 120 voluntarios de la ACM, un tercio de ellos ACM Fellows, colaboraciones con el personal de ACM y con Semedica, una división de Silverchair. Semedica aportó su experiencia en la construcción de ontologías, mientras que los voluntarios de ACM Provideon la experiencia en el dominio. (ACM, 1998)

La primera versión fue lanzada en 1998, la misma fue actualizada y mejorada con nuevos términos y áreas de investigación en común. Adicionalmente, se agregaron taxonomías relevantes. La versión actual, 2012 del Sistema de Clasificación de Computación ACM (CCS) es un referente mundial para clasificar las áreas de las ciencias de la computación. Esta versión incorpora conceptos que sirven para indexadores y para la búsqueda en ellos. (ACM, 2012))

Esta clasificación es un beneficio para el investigador ya que, proporciona una referencia rápida del contenido del artículo, con lo cual facilita la búsqueda de artículos relacionados.

Las áreas de investigación de la ACM (ACM, 2012) son las siguientes:

- General y referencia
- Hardware
- Organización de sistemas informáticos
- Redes

- Software y su ingeniería
- Teoría de la computación
- Matemáticas de la informática
- Sistemas de información
- Seguridad y privacidad
- Computación centrada en el ser humano
- Metodologías de computación
- Computación aplicada
- Temas sociales y profesionales

Todas estas áreas de conocimientos abarcan totalmente a las ciencias de la computación, en más de mil líneas de investigación.

c) Ciclo de Sobre Expectación de Gartner para Tecnologías Emergentes 2018

Gartner es una empresa de investigación de tecnologías de información, la cual provee anualmente un estudio del mercado y sectores de las TICs. Uno de los resultados de estos estudios es el ciclo de sobre expectación de Gartner (Gartner, 2018), el cual menciona que las tendencias actuales para las tecnologías emergentes son por ejemplo los sistemas de recuperación automática, computación cuántica, biochips y asistentes virtuales. Las grandes tendencias futuras incluyen, inteligencia artificial, industria 4.0 y experiencias virtuales.

Este ciclo proporciona una representación gráfica de la madurez y la adopción de tecnologías y aplicaciones, una vista de cómo una tecnología o aplicación evolucionará con el tiempo. Este se muestra en la figura 1.2.

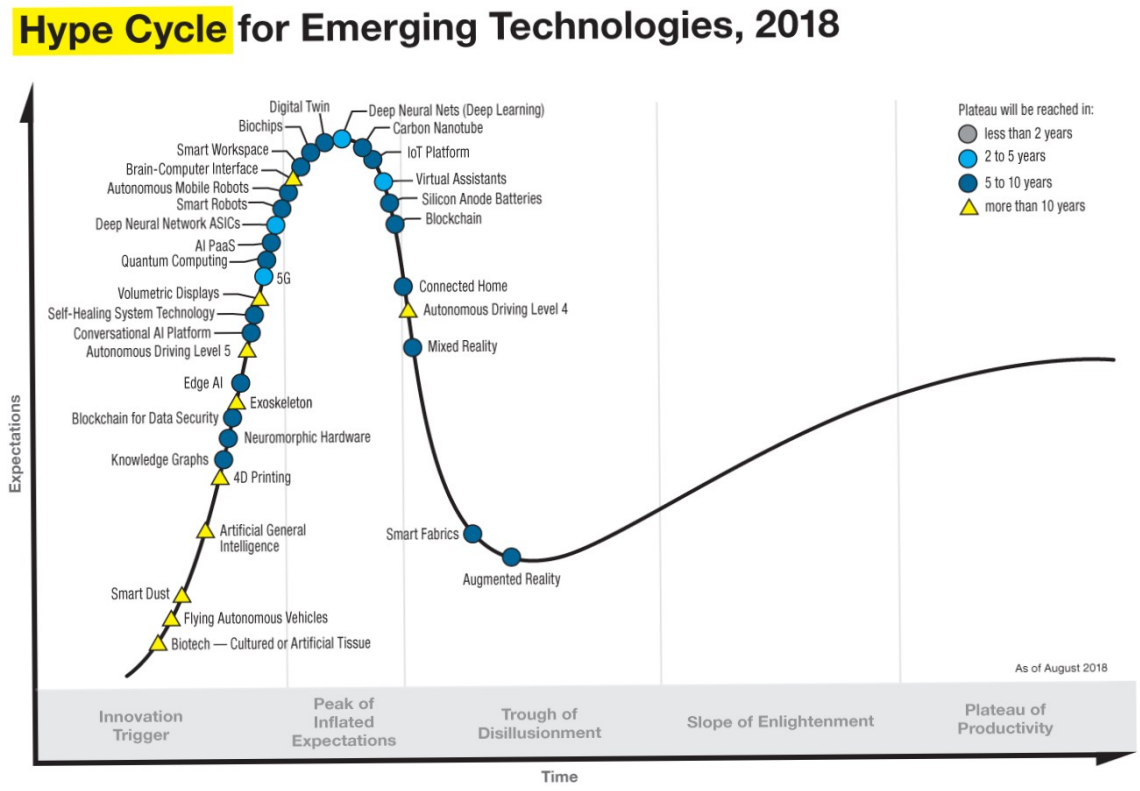

Fig. 1.2 Ciclo Hart de Gartner para tecnologías emergentes, 2018. Fuente: Gartner (2018)

1.4.1.2 Líneas de Investigación en el Ecuador

El MINTEL realizó un taller sobre fomento de la investigación a través del aprovechamiento de las TIC en el año de 2016, generando en conjunto con la academia las líneas de investigación para el país, estas líneas de investigación fueron: Internet de las Cosas, Espectro Radioeléctrico, Televisión Digital Terrestre, Big Data, Ciudades Inteligentes, Desarrollo de Software, Gobierno Electrónico, TIC y Cultura, TIC y Educación, TIC y Producción, TIC y Salud y TIC y Sociedad. (MINTEL, 2016)

En el gobierno de Lenin Moreno, presidente actual de la república del Ecuador, se ha creado el "Plan Toda una Vida", cuya misión es "garantizar el acceso progresivo de las personas a sus derechos en todo el ciclo de vida, a través de la generación de políticas públicas para el desarrollo social y humano de la población; y, proponer, coordinar y ejecutar de forma eficiente, eficaz y transparente el Plan Toda una Vida, dirigido a grupos con necesidades

básicas insatisfechas y en riesgo". (Secretaría-Técnica, 2017)

El mismo tiene como visión "construir La Secretaría Técnica del Plan Toda una Vida en el organismo estatal, que impulsará la generación de políticas públicas, a más de la coordinación y ejecución de los programas y misiones: Misión Ternura, Impulso Joven, Mis Mejores Años, Menos Pobreza Más Desarrollo, Casa Para Todos, Las Manuelas, Las Joaquinas y Plan Mujer, proyectos que promueven el acceso a la satisfacción de las necesidades básicas de los grupos de la población en condiciones de extrema pobreza y vulnerabilidad". (Secretaría-Técnica, 2017)

Dentro del Plan Toda una Vida, en el 2018, se ha logrado crear el Libro Blanco de la Sociedad de la Información y del Conocimiento (LBSIC) del MINTEL con el objetivo de dar a conocer las estrategias y líneas de investigación de la Sociedad de la Información y del Conocimiento en el Ecuador, con el fin de impulsar el desarrollo del país y mejorar la administración y ejes públicos.

En febrero del año 2019, se ha desarrollado una actualización al LBSIC y se ha creado el Libro Blanco de Líneas de Investigación, Desarrollo e Innovación y Transferencia del Conocimiento en TIC, donde se establecen 7 líneas de investigación priorizadas según MINTEL (2019).

a) Libro Blanco de Líneas de Investigación, Desarrollo e Innovación y Transferencia del Conocimiento en TIC

El objetivo general del Libro Blanco de Líneas de Investigación, Desarrollo e Innovación y Transferencia del Conocimiento en TIC es dar a conocer las estrategias que contribuirán al desarrollo del Ecuador y la sociedad ecuatoriana, en el ámbito de la información y del conocimiento, con la finalidad de impulsar la equidad, el crecimiento económico y la inclusión social. (MINTEL, 2018)

Entre los objetivos específicos se encuentran los siguientes:

- Dar a conocer el estado actual de la sociedad de la información y del conocimiento del Ecuador.

- Orientar la ejecución de los instrumentos de implementación de políticas públicas de información y telecomunicaciones.
- Impulsar la participación de los distintos actores y de toda la sociedad ecuatoriana sobre los ejes establecidos para la construcción de la Sociedad de la Información y del Conocimiento de forma articulada.
- Implementar las líneas de investigación, ejes y programas de acción. (MINTEL, 2018)

En el año 2018, el Ministerio de Telecomunicaciones y de la Sociedad de la Información (MINTEL, 2018), dentro de un encuentro con la industria de las telecomunicaciones, expuso los 4 ejes estratégicos que serán parte de la agenda digital para el desarrollo del país 2018-2021, los cuales son los siguientes:

1. **Infraestructura y conectividad**. Este eje constituye el crecimiento en la infraestructura de conectividad, al acceso y al servicio universal. Además, impulsará la ampliación del servicio a zonas aun no atendidas.
    - Los instrumentos de este eje son: el "Plan Nacional de Telecomunicaciones y Tecnologías de la Información", el "Plan de Servicio Universal" y el "Plan Maestro para la Migración a la Televisión Digital Terrestre".
2. **Gobierno electrónico**. Este eje establece incrementar la participación ciudadana por medios digitales, automatización y mayor agilidad en trámites y regulaciones. Además, mejorar los servicios del estado y fomentar la transparencia.
    - Los instrumentos de este eje son: el "Plan Nacional de Gobierno Electrónico", el "Registro Nacional Único de Trámites y Regulaciones" y el "Proyecto de Ley de Optimización y Simplificación de Trámites".
3. **Seguridad de la información y datos personales**. Este eje propone como política el fortalecer la disponibilidad, integridad y confidencialidad de la información crítica. Con esto, se garantiza los derechos de la ciudadanía y el desarrollo de la economía
    - Los instrumentos de este eje son: el "Pacto con Niñas, Niños y Adolescentes por un Internet Seguro", los "Protocolos de Atención a Casos de Violencia Digital", el "Proyecto de Ley de Protección a Datos Personales" y la "Estrategia Nacional de Ciberseguridad".

4. **Sociedad de la Información y del Conocimiento**. Este eje establece el fortalecimiento de la industria TIC y la transformación digital de las empresas mediante el uso adecuado y eficiente de las TICs. Además propone el desarrollo de emprendimientos y del comercio electrónico.

    - El instrumento de este eje es: el "Plan de la Sociedad de la Información y del Conocimiento".

En esta nueva versión del libro se establecieron 7 líneas de investigación priorizadas en los ámbitos de la Sociedad de la Información y del Conocimiento según MINTEL (2019), añadiendo 2 líneas a las definidas en el 2018. Las cuales son las siguientes:

- **TIC y Educación:** La incorporación de las tecnologías de la información y comunicación en la educación permite transformar los procesos de aprendizaje y enseñanza, e implementar herramientas y metodologías innovadoras que faciliten y potencien dichos procesos.
- **Redes e infraestructuras de telecomunicaciones:** Este eje permitirá la optimización del uso y transmisión de las ondas electromagnéticas y el control de los servicios de telecomunicaciones mediante el establecimiento de nuevos e innovadores mecanismos técnicos y regulatorios en el ámbito del sector. Esto mejorará la prestación de servicios de telecomunicaciones y su calidad.
- **Tecnologías de radiodifusión digital:** Las tecnologías de radiodifusión digital permitirá al país tener múltiples programas dentro de la señal de una misma estación multiprogramación: noticiero, clases en vivo, deportes, películas, entre otros. Esto tendrá como beneficio adicional, la recepción de alertas de emergencia, permitiendo así que la televisión se sume a los dispositivos que entregan a la población mensajes que ayudan a salvar vidas y acceder a contenidos interactivos en temas de salud, turismo, entretenimiento, entre otros.
- **Ciudades Inteligentes, sostenibles e inclusivas:** Este eje se platea para mejorar la calidad de vida y sostenibilidad en las poblaciones con el apoyo de las TIC en la eficiencia de los servicios hacia el ciudadano, como la prestación y el acceso a recursos hídricos, energía, transporte y movilidad, educación, medio ambiente, gestión de residuos, vivienda, entre otros. La dinámica de las ciudades inteligentes es aumentar la eficiencia y eficacia de los procesos y servicios.

- **Big Data:** Este eje busca mejorar y optimizar la toma de decisiones en áreas importantes para el país como: salud, empleo, productividad, seguridad, manejo de desastres naturales, entre otros. Con esto, se puede prevenir los riesgos de privacidad de datos, mejorar la equidad de la economía y optimizar los procesos para beneficio de la sociedad y la industria.

- **Seguridad de la Información:** Con este eje se busca mitigar las amenazas del ciberespacio para mantener la protección de los datos e información de importancia para el Ecuador. Con esto, permitirá a la sociedad ecuatoriana estar protegida de amenazas hacia su economía como a su bienestar. Además, fortalecer los planes de respuesta a incidentes y a incrementar la confianza al uso del internet.

- **TIC y Producción:** Con la implementación de las tecnologías de la información y comunicación en el sector productivo del país, se busca fomentar las capacidades de competitividad, emprendimiento y crecimiento sostenido de la producción, del empleo y de la productividad, incrementando la inversión en investigación, desarrollo tecnológico e innovación.

1.4.2 Bases de Datos Científicas Internacionales y Regionales

Bases de Datos Internacionales

A continuación se especifican las bases de datos internacionales más relevantes y más utilizadas:

a) Scopus

Scopus es la mayor base de datos científica creada por Elsevier, la misma consta de citas y resúmenes de literatura revisada por pares, como son revistas científicas, libros y actas o memorias de congresos. Además, cuenta con una herramienta web que permite realizar búsquedas inteligentes y permite rastrear, analizar y visualizar la investigación a nivel mundial. La base de datos en mención abarca la producción científica mundial en los campos de ingeniería, ciencias exactas, ciencias sociales, artes, humanidades y medicina. (Elsevier, 2018)

Scopus cuenta con más de 60 millones de registros que incluyen más de 23.343 [1] revistas activas con pares evaluadores de las cuales 4.200 son de acceso libre, más de 130.000 libros disponibles y 10.000 añadidos cada año. En esta era que la información crece exponencialmente, la investigación en las bases de datos científicas como Scopus se vuelve cada vez mayor, teniendo investigaciones interdisciplinarias y con grupos de investigación mucho más grandes (Macías, 2017). Por lo que es necesario no dejar pasar por alto investigaciones cruciales realizadas alrededor del mundo y generar líneas de investigación a través de esta información que permitan el desarrollo del país.

Publicar en revistas incluidas en la base de datos SCOPUS de Elsevier es considerado en los procesos de evaluación y acreditación un criterio de alta calidad (Macías, 2017).

A partir de las citas que reciben los artículos publicados en las revistas incluidas en Scopus el Grupo de Investigación Scimago elabora un ranking que ordena las revistas en función de un indicador que se denomina Scimago Journal Rank (SJR). Este, se calcula mediante el algoritmo Page Rank de Google y analiza las citas de la revista durante un período de tres años. A diferencia del Factor de Impacto, en el cálculo del Scimago Journal Rank (SJR) intervienen no sólo el número de citas sino también la influencia de la revista (gran cantidad de citas, sin auto citas) y su prestigio por lo que no todas las citas tienen el mismo valor. (Colledge et al., 2010)

Adicionalmente, Scopus cuenta con varias Interfaces de programación de aplicaciones (APIs por sus siglas en inglés) dedicadas para diferentes lenguajes de programación uno de ellos Python, que permite la extracción de la información de resúmenes y datos de citas de todas las revistas indexadas y que permite acceder a búsquedas de referencias bibliográficas completas, palabras claves, búsqueda de nueva investigación, etc. La consulta es gratuita, aunque la fuente de la que proceden los datos, no lo sea (Licencia Nacional Fecyt) (Yuen, 2018).

El acceso a ciertas áreas de la API no esta habilitado de forma predeterminada, ya que requiere un permiso especial de Elsevier. El suscriptor se distingue dependiendo de la IP desde el cual se hacen las consultas a la base de datos de Scopus. (Elsevier, 2018)

Las cuotas de la API se establecen cada 7 días y también cuenta con restricciones a 5000 registros por cada consulta que se realice, además, existen tiempos definidos para

[1] Fuente: Scopus, obtenido de: https://www.scopus.com/sources. Fecha de acceso: marzo 2019

la extracción de la información, dependiendo siempre del tipo de suscripción que se haya adquirido.

Para este proyecto se ha utilizado la suscripción de pago para obtener los datos más importantes de los artículos explicados en el capítulo 3. Además, al ser una suscripción de pago se tiene acceso a datos más completos de Scopus, por lo que para la extracción de estos artículos por cada consulta según ha establecido la API de Scopus, el tiempo promedio de consulta de 25 artículos es de 6 segundos. (Elsevier, 2018)

b) Web Of Science (WOS)

La Web Of Science (WOS) es una base de datos científica creada por el Instituto Científico de la Información (ISI), es una plataforma web que alberga las referencias de las principales publicaciones científicas a nivel mundial en el ámbito de todas las disciplinas del conocimiento. Fue creada en 1945, y es esencial para el apoyo de la investigación y de proyectos de investigación y para el reconocimiento mundial de investigadores por sus avances para la comunidad científica y tecnológica. (Clarivate, 2018)

Web of Science proporciona acceso a diferentes bases de datos como son:

- Book Citation Index– Social Sciences and Humanities (BKCI-SSH), desde el 2005
- Book Citation Index– Science (BKCI-S), desde el 2005
- Conference Proceedings Citation Index- Social Science and Humanities (CPCI-SSH), desde 1990
- Conference Proceedings Citation Index- Science (CPCI-S), desde 1990
- Arts and Humanities Citation Index (A and HCI), desde 1975
- Social Sciences Citation Index (SSCI), desde 1900
- Science Citation Index Expanded (SCI-EXPANDED), desde 1900
- Current Chemical Reactions (CCR-EXPANDED), desde 1986 hasta el 2009
- Index Chemicus (IC), desde 1993 hasta el 2009
- Emerging Sources Citation Index

Web of Science a través de sus estrictos procesos de evaluación para que una revista sea indexada, tiene la confiabilidad de que solo se incluye información de relevancia, creíble y muy influente para el mundo científico. Además, este sistema le permite un acceso a la investigación más confiable, integrada y multidisciplinaria, lo que le facilita descubrir las tendencias actuales de investigación. (Clarivate, 2018)

La plataforma de Web of Science contiene un proceso de búsqueda y descubrimiento a través de:

- Contenido principal multidisciplinario
- Tendencias emergentes
- Contenido específico del sujeto
- Contenido regional
- Datos de la investigación
- Herramientas de análisis

Journal Citation Reports (JCR) de Incites es una herramienta de análisis incluida en la Web of Science y se trata de un ranking de revistas ordenado a partir del recuento de las citas que reciben los artículos que en ellas se publican. Estas revistas son las incluidas en dos de los índices que forman la Colección Principal (Core Collection) de Web of Science: Science Citation Index Expanded (SCI Expanded) y Social Science Citation Index (SSCI). Se publican dos ediciones anuales de JCR: JCR Science Edition y JCR Social Sciences Edition. (Yuen, 2018)

Bases de Datos Regionales

Las bases de datos regionales más relevantes son:

a) Latindex

Latindex es una base de datos científica que alberga revistas de países de América Latina, el Caribe, España y Portugal. Fue creada por la Universidad Autónoma de México (UNAM) en 1995, y surgió como un producto para comparación en la red de universidades para reunir

y almacenar publicaciones científicas que son producidas en la región. (Cetto, 1998)

Latindex es un sistema de información que indexa revista científicas, técnico-profesionales y de divulgación científica y cultural. No contiene un sistema de búsqueda de publicaciones, solamente es una base de datos de revistas. El problema de Latindex radica en la dificultad de obtener artículos dentro de su base de datos, ya que solamente alberga la descripción de la revista.

Latindex no es un indexador por lo que no tiene métricas como Scopus, pero tiene cuatro tipos de bases de datos (directorio, catálogo, revistas en línea y portal de portales), y cada una de ellas tiene requisitos y criterios de calidad editorial para que las revistas sean incluidas en las bases de datos mencionadas anteriormente.

Scientific Electronic Library Online (SciELO)

Scientific Electronic Library Online (Biblioteca Científica Electrónica en Línea) es una base de datos de revistas en línea cooperativa en internet. Fue desarrollada por un conjunto de entidades de Brasil como BIREME, así como instituciones nacionales e internacionales con la comunicación científica y editores científicos. Es un proyecto que fue lanzado en 1998 con revistas brasileñas, teniendo una gran acogida a nivel mundial especialmente latinoamericano. A partir del 2012, con el apoyo del CNPq (Conselho Nacional de Desenvolvimiento Científico y Tecnológico) su operación se expandió a diferentes países. (Rosario et al., 2014)

SciELO en particular es un modelo para publicaciones electrónicas, eliminando así las impresiones innecesarias de varias revistas, esto además es una web que facilita la búsqueda de artículos científicos dentro de su base de datos, que cuenta adicionalmente con su ranking. El proyecto de SciELO fue creado para la cooperación de toda la comunidad científica en países en desarrollo de América Latina y el Caribe. SciELO asegura la visibilidad ya que es un modelo web que también es de acceso abierto a toda la literatura científica. Además, SciELO contiene métricas para medir el uso y el impacto de las revistas científicas indexadas en ella, pero lamentablemente en la actualidad no son muy utilizadas a comparación con las métricas de Scopus o WOS.

1.4.3 Analítica de Datos (AD)

En los últimos años, la analítica de datos gracias a que es una herramienta para generar conocimiento, predecir diferentes situaciones, entre otros, ha tenido un crecimiento importante en el sector académico e industrial de todo el mundo. Las mejores Universidades e industrias a nivel mundial utilizan la analítica de datos como una herramienta de ayuda para toma de decisiones.

La analítica de datos es la ciencia que permite la extracción, almacenamiento, análisis y/o transformación de datos para generar conocimiento. Esta técnica se puede aplicar a cualquier ámbito, con el fin de generar conocimiento y descubrir información específicas para una entidad, lo que ayuda a la toma de decisiones dentro y fuera de la entidad. En la actualidad los datos son el recurso más valioso para una institución, pero si los datos se convierten en información y en conocimiento, estos tendrán un mayor valor para la empresa. (Aguilar, 2013)

Los principales ámbitos de aplicación de la analítica de datos son:

- La industria para una toma de decisiones que permita el desarrollo industrial
- La ciencia e investigación para aprobar, rechazar y modificar modelos existentes

a) Tipos de Analítica de Datos

Carrascal y Jiménez (2018) definen 4 tipos de analítica de datos, las cuales se describen a continuación:

1. **Analítica Descriptiva:** Este tipo de análisis de datos proporciona información sobre el pasado de los datos, estos datos indican si algo esta bien o mal sin importar el por qué. Este tipo de analítica se utiliza simplemente como indicadores para tener un panorama del pasado.
2. **Analítica Diagnóstica:** Este tipo de analítica de datos es aún más completa que la anterior ya que los datos históricos se puede medir con otros datos para obtener resultados de el por qué sucedió algo. Esta analítica puede analizar dependencias y patrones en los datos.

3. **Analítica Predictiva:** Esta analítica es de tipo probabilista y mide la probabilidad de que suceda algún acontecimiento. Utiliza los dos tipos de analíticas anteriores (descriptiva y diagnostica) para determinar tendencias de los datos, agrupaciones, y excepciones. Esta es una herramienta muy valiosa de la analítica de datos para prevenir sucesos o datos. A pesar de lo importante y útil que resulta este tipo de analítica de datos para las empresas es relevante mencionar que solamente es una estimación de los datos siempre dependiendo del tratamiento de los datos y la calidad de los mismos.

4. **Analítica Prescriptiva:** Este tipo de analítica permite prescribir la acciones a tomar para problemas futuros o aprovechar las tendencias generadas por la analítica predictiva, este análisis ayuda a la toma de decisiones de las empresas u organismos para mejorar sustancialmente mejores tendencias y minimizar los riegos de los problemas que presente la empresa. La analítica Prescriptiva requiere además de los datos históricos, información externa debido a la naturaleza de los algoritmos estadísticos. Las herramientas y técnicas que utiliza este tipo de analítica son por ejemplo: machine learning, reglas de negocio y algoritmos. Este tipo es el más sofisticado, lo que hace que sea robusto pero también más complicado de realizar por las empresas ya que requiere un mayor esfuerzo y más recursos para lograr un valor agregado.

Dentro de la Analítica prescriptiva existe una categoría más amplia llamada **Inteligencia Artificial (IA)**, esta es una máquina con cualquier tipo de comportamiento inteligente. Una de las aplicaciones de la IA es el **Machine Learning** que es parte de la inteligencia artificial que aprende por sí misma. Y por último, dentro del Machine Learning se tiene al **Deep Learning**, que es parte del aprendizaje automático que utiliza redes neuronales. (Carrascal y Jiménez, 2018)

Además de estos tipos de analítica de datos existen varios metodologías para llegar a descubrir conocimiento como por ejemplo KDD (descubrir conocimiento en bases de datos) y MIDANO. KDD es un tipo de método que define un proceso no trivial para identificar patrones novedosos, útiles, y válidos para comprender los datos. (Carrascal y Jiménez, 2018)

En la sección 1.5 se detalla la metodología utilizada para el desarrollo de analítica de datos para este proyecto.

b) Técnicas de Analítica de Datos

Existen varias técnicas de analítica de datos, entre las pertinentes para la presente investigación se encuentran la minería de datos y de texto.

Minería de Datos

La minería de datos es una herramienta de la analítica de datos que permite la asociación, clasificación y agrupamiento de los datos para generar diferentes patrones de conocimiento y se dividen en las tres mencionadas a continuación: (Schab et al., 2018)

1. **Asociación:** Este tipo de técnica de minería de datos establece una relación entre un elemento que se organiza en un grupo de datos determinados. En minería de datos, se utiliza las reglas de asociación para analizar y predecir datos.
2. **Agrupamiento:** Este tipo de técnica realiza la división de un conjunto grande de datos en grupos homogéneos y significativos. La técnica de minería de datos más utilizada es el agrupamiento, ya que con esta se puede realizar diversos análisis predictivos y de toma de decisiones.
3. **Clasificación:** La clasificación es un tipo de técnica de minería de datos que permite predecir el comportamiento de los datos en el futuro mediante la clasificación de los datos predefinidos. Hay varios algoritmos de clasificación de minería de datos, los mas relevantes son: Naïve Bayes, Support vector machine, y Decision Tree. En la tabla 1.3 se realiza una comparativa de los diferentes tipos de técnicas de clasificación:

Support Vector Machine (SVM) se utiliza en este trabajo de investigación para generar modelos predictores, este es un algoritmo de aprendizaje de máquina supervisado. Básicamente, este algoritmo trata de encontrar un hiperplano para clasificar los conjuntos de datos. Existen dos tipos de clasificadores SVM:

- Clasificador Lineal SVM
- Clasificador no lineal SVM

Tabla 1.3 Comparación de algoritmos de clasificación

Algoritmo	Ventajas	Desventajas
Naïve Baye: Es un algoritmo probabilístico que calcula la frecuencia y las agrupaciones de valores en un registro determinado	Simple, muy rápido, predice el resultado correctamente y de buen rendimiento.	Requiere un número grande de datos para obtener buenos resultados. Se basa en instancias ya que almacena todas las muestras de entrenamiento.
Support vector machine: Es un algoritmo constante que permite la predicción de resultados mucho mejor	Predicción mucho mejor, estimación rápida del objetivo y utiliza menos parámetros.	Costoso Computacionalmente.
Decision Tree: Es un algoritmo basado en árboles de decisión. Este permite buscar el mejor camino para la clasificación lo que permite conseguir mejores resultados.	Resultados más exactos, costo computacional menor, menor tiempo de compilación del modelo, tiempo de búsqueda mejorado.	Exceso de ajuste, ramas vacías y no significantes.

Fuente: (Schab et al., 2018)

A continuación, en la figura 1.3, se muestra el flujograma con el que trabaja en forma general el SVM para la obtención de modelos predictivos.

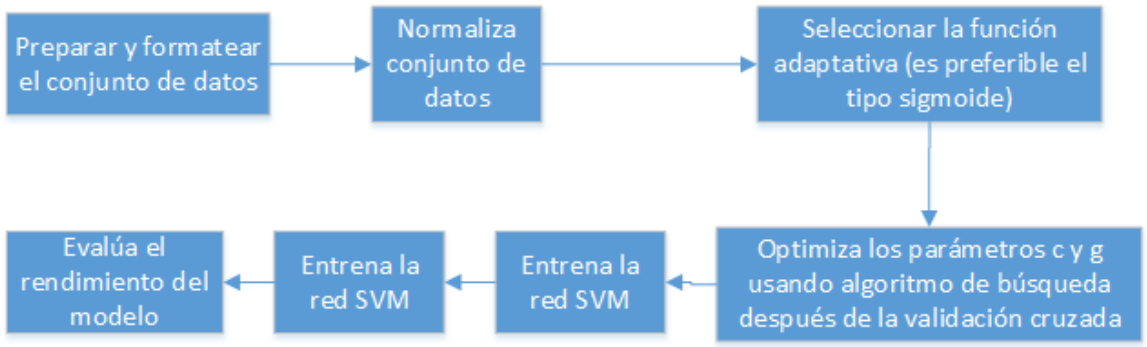

Fig. 1.3 Flujo de trabajo Support Vector Machine. Fuente: Schab et al. (2018)

Otra técnica de minería de datos es la minería de texto, esta se enfoca en el procesamiento del lenguaje natural. En la siguiente sección se especifica en detalle esta técnica, esta tiene una sección especial por ser la principal para la realización del presente proyecto.

Minería de Texto

Uno de los avances de las técnicas y modelos de la minería de datos es la minería de texto, esta es una técnica que permite obtener patrones de texto y una tendencia que tienen los datos. La minería de texto es necesaria para datos de texto complejos y que son difíciles de analizar o para volúmenes de datos demasiado grandes como documentos científicos o textos de organizaciones. (Hernández et al., 2004)

En este contexto, la minería de texto es el proceso encargado de obtener información y conocimiento que no exista en los documentos científicos, y que surge de relacionar a varias partes de ellos. Se establecen tres fases fundamentales para conseguir un resultado óptimo, la primera fase consiste en seleccionar los documentos científicos adecuados para realizar una analítica correcta. La segunda fase consiste en la extracción de la información de esos documentos, esto se realiza mediante el uso del procesamiento de lenguaje natural de los cuales se extrae características como palabras claves de los documentos, resumen, fecha de publicación, revista, etc. La tercer y última fase consiste en encontrar relaciones o asociaciones entre los datos claves que obtenemos de la segunda fase todo esto se realiza mediante la minería de datos. (Hernández et al., 2004)

La minería de texto se puede aplicar a:

- La clasificación de documentos científicos
- La elaboración de resúmenes
- La extracción de conocimiento
- El análisis de sentimientos
- La minería de opiniones
- La extracción de información

1.5 Metodología de Analítica de Datos

La siguiente sección específica la metodología de analítica de datos seleccionada para el desarrollo de la presente tesis con todas sus fases de desarrollo.

1.5.1 MIDANO

La metodología que se especifica a continuación fue creada por Aguilar (2013), con la finalidad de gestionar todo el proceso realizado en un proyecto de Analítica de datos (AD) y dar seguimiento del mismo en todas sus fases de desarrollo. Con esta metodología podemos obtener conocimiento sin vacío y con datos depurados y limpios. Para cualquier proyecto de Analítica de datos, MIDANO apoya en la eficiencia y eficacia del mismo. Aguilar (2013), en su metodología propone realizar un análisis de la situación actual de la entidad u organismo a ser estudiado para posteriormente realizar el proyecto de analítica de datos, esto garantiza que el proyecto tenga el éxito deseado y los resultados esperados.

La metodología MIDANO permite no solo identificar al ente de estudio si no también contextualizar la solución del problema desde la perspectiva del desarrollo de aplicaciones basadas en Analítica de datos (AD). Esta metodología está compuesta en tres fases:

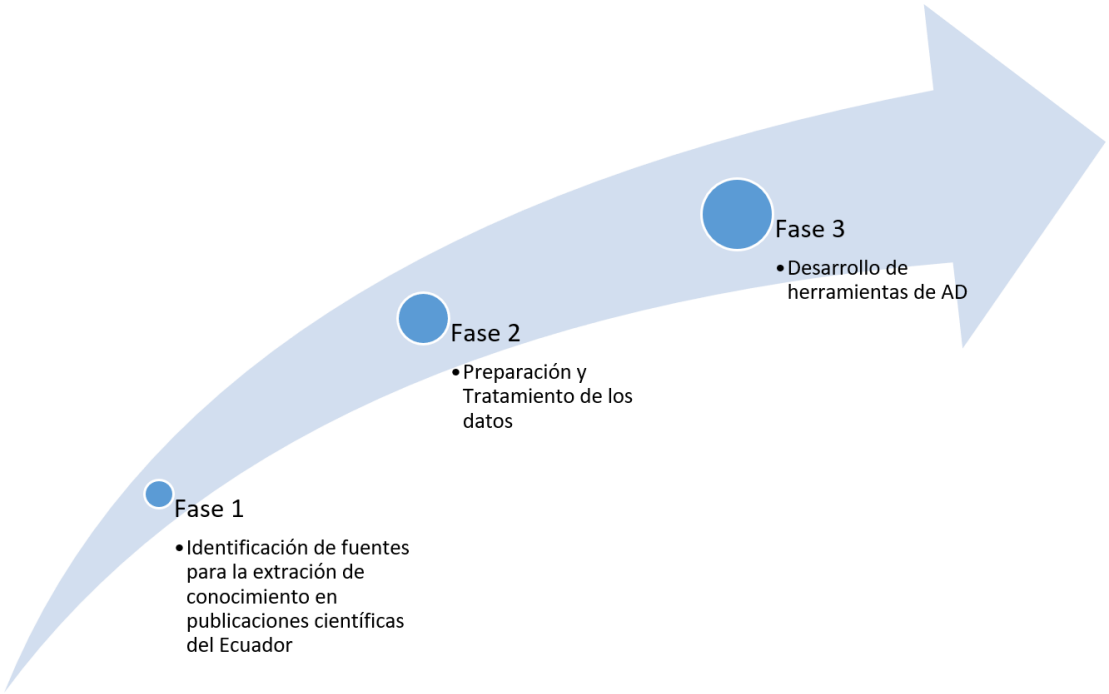

Fig. 1.4 Fases de desarrollo en MIDANO. (Aguilar, 2013)

1. Identificación de las fuentes de datos para la extracción de conocimiento en las publicaciones Científicas del Ecuador y de los problemas que se pueden resolver con ellas en el país.

2. Preparación y tratamiento de los Datos.

3. Desarrollo de herramientas de Analítica de datos.

En la Figura 1.4 se muestran las tres fases de un proyecto de analítica de datos, teniendo en cuenta que cada fase tiene su retroalimentación en las fases anteriores. Con esto se obtienen proyectos funcionales y eficaces. En algunos casos de proyectos no es necesaria la ejecución de todas las fases, pero en el proyecto actual se ejecutarán todas las fases para demostrar el potencial del mismo. Por lo tanto una característica de MIDANO es la flexibilidad. (Aguilar, 2013)

Cada una de las tres fases de la metodología MIDANO cuenta con etapas o procedimientos que estan descritas en la Figura 1.5 (Aguilar, 2013). En este caso es necesario recalcar que no todas las fases son de manera secuencial, en el presente proyecto se especifican dichas etapas en el Capítulo 4.

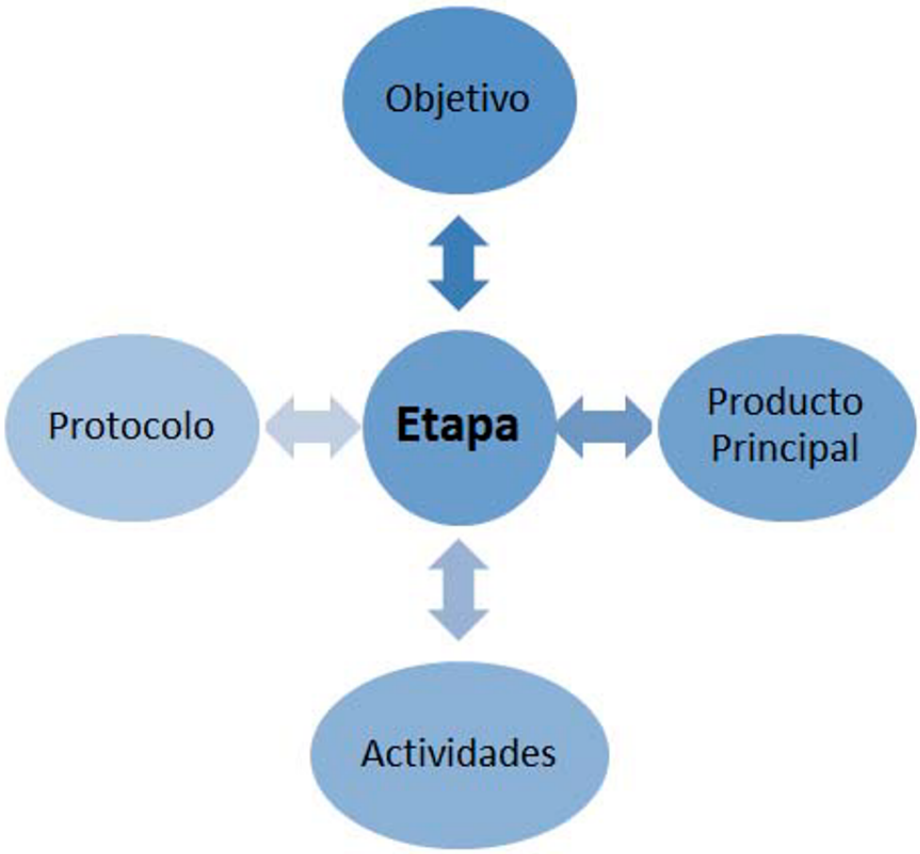

Fig. 1.5 Elementos de cada etapa de las fases de MIDANO. (Aguilar, 2013)

Como se expresa en la Figura 1.5, en cada etapa se especifican 4 procedimientos escenciales para que el proceso de ingeniería y extracción de conocimiento tenga éxito:

1. Objetivo: Especificar la meta que se requiere cumplir.
2. Producto principal: Resultado de la etapa.
3. Protocolo: Procedimientos, estudios o preguntas que se realizarán para conseguir la meta.
4. Actividades: Tareas necesarias para lograr el objetivo.

Estas 3 fases y actividades propuestas de cada etapa abarcan todo el dominio de conocimiento que puede encontrarse en la obtención de las líneas de investigación específicas para el Ecuador.

Fase 1: Identificación de las fuentes de datos para la extracción de conocimiento

La primera fase tiene como meta realizar un proceso de descripción de las fuentes de datos, es decir una ingeniería de conocimiento orientado a las bases de datos científicas y a las líneas de investigación que serán base para la extracción de los artículos científicos. Esta fase se enfoca en conceptualizar e identificar la solución del problema a investigarse, esto se logra desde la perspectiva del desarrollo de aplicaciones de Analítica de datos.

Las tareas de esta Fase son:

1. Conocimiento de las líneas de investigación nacionales e internacionales
2. Caracterización de las líneas de investigación para obtener las cadenas de búsqueda
3. Análisis y extracción de la información de las fuentes de datos
4. Obtención de los artículos científicos

Fase 2: Preparación de los Datos

La Fase 2 tiene como objetivo aplicar Analítica de datos sobre el problema descrito anteriormente, aquí es donde se realiza una limpieza de datos, un historial de datos asociados al problema en estudio y preparación de datos del problema. Es decir, esta fase se encarga de realizar varias operaciones con los datos, con el fin de obtener datos consistentes y aptos para desarrollar un modelo de Analítica de datos. En esta etapa se realiza un tratamiento, limpieza y una preparación de los datos que serán utilizados para el desarrollo de la herramienta de Analítica de datos.

Las tareas de esta Fase son:

1. Dominio de la Aplicación
2. Tratamiento de Datos

Fase 3: Desarrollo de herramientas de AD

La última fase busca desarrollar una herramienta de software que permita utilizar diferentes modelos de Analítica de datos con los datos anteriormente tratados para obtener conocimiento y en escenarios futuros identificar su potencial uso.

Las tareas de esta Fase son:

1. Especificación de los requerimientos de la herramienta computacional a utilizarse
2. Desarrollar el modelo de Analítica de datos
3. Implementación usando el modelo de Analítica de datos
4. Validación/Interpretación

En la siguiente sección se detalla la selección de las herramientas de analítica de datos que más se adapten para realizar las fases 2 y 3 de MIDANO.

1.6 Alcance de la Investigación

Los artículos científicos en las indexadoras como Scopus son documentos importantes de investigación, pero no sirve de nada si solo están almacenados, por lo cual, la analítica de datos es muy provechosa ya que con ella se puede obtener conocimiento de los datos como por ejemplo indicadores, modelos predictivos, etc. Es por ello que, el alcance de este proyecto de investigación es proponer líneas de investigación específicas que sirvan para el desarrollo del Ecuador.

Para el caso de Ecuador, el plan de desarrollo del MINTEL, se adapta a las necesidades de desarrollo del país. Este fue creado en el año 2014 y es el único que desarrolla un libro blanco de líneas de investigación que son de beneficio para el país, por lo que es el más idóneo para la presente investigación.

Diferentes entidades internacionales como UNESCO y ACM han definido estándares para clasificar las áreas y líneas de investigación, en el caso de UNESCO, se hace una clasificación de áreas en general. Para la presente investigación, ACM al ser un referente mundial en las ciencias de la computación, ha creado una clasificación de tópicos del área mencionada, es por ello que, esta se utuliza como referente para determinar las líneas de investigación específicas para el Ecuador.

Por otra parte, las bases de datos científicas más utilizadas a nivel mundial son Scopus y WOS, indexan las revistas por medio de la medición del impacto. Scopus al ser una base de datos mucho más grande, y poseer una variedad de artículos de calidad sirve para determinar los artículos científicos que pueden ser usados como base para determinar las líneas de investigación específicas para el Ecuador.

1.6.1 Python

Para muchas personas Python es un lenguaje de programación que enamora. Como explica McKinney (2012) en su libro, desde su creación en 1991, Python ha llegado a ser un lenguaje de programación dinámico y popular. Para la analítica de datos, computación exploratoria interactiva y visualización de datos, Python ha sido punto de comparación con otras herramientas especializadas en el área tanto de código abierto y comerciales como, por ejemplo: R, MATLAB, SAS, Stata y otros. En los últimos años, el soporte mejorado de las librerías de Python (especialmente de Pandas) lo ha convertido en una fuerte alternativa para tareas de

manipulación de datos. Es así como, junto con su fuerza en la programación de propósito general, es una excelente opción como lenguaje único para crear aplicaciones centradas en datos.

El lenguaje Python es fundamental en el contexto de la analítica de datos y Big Data, ya que dispone de las herramientas para casi todos los aspectos relacionados con la matemática y computación científica. Actualmente, es el lenguaje de preferencia para el análisis de datos no solo por su versatilidad, si no también por sus funcionalidades ya que incorpora herramientas o algoritmos para análisis y visualización de datos que se usan en Matlab y R y son mejorados por su eficiencia y eficacia. En los últimos años Python es de preferencia para la analítica de datos así como también para los análisis matemáticos. (Calixto, 2017)

Python es un lenguaje de programación con una sintaxis limpia y muy sencilla de leer y entender (McKinney, 2012). Además posee muchas librerías para realizar estadística de forma eficiente, y contiene una comunidad muy amplia para el soporte de cualquier inconveniente. La ventaja más grande de Python comparada con otros lenguajes o herramientas es que es posible realizar todo el ciclo de vida de la analítica de datos en el mismo lenguaje sin tener que usar otras herramientas, ya que contiene las librerías necesarias y suficientes para hacer el análisis completo y eficiente de los datos, un ejemplo de las herramientas y las más utilizadas de analítica de datos son: NumPy / SciPy, matplotlib, pandas y statmodels.

Capítulo 2

Estado del arte

La importancia de la analítica de datos en los últimos años, se ha incrementado radicalmente, debido a la necesidad de generar conocimiento e indicadores que permitan establecer estrategias en el campo de la academia, industria y para el desarrollo del país. En uno de los ámbitos que se podría aplicar esto son los documentos científicos, al existir registros tan numerosos de investigaciones, ya sean artículos o libros, en las bases de datos científicas es necesario transformar toda esa información en conocimiento que incremente e incentive el desarrollo del país. (Aggarwal y Zhai, 2012)

La analítica de datos ha tenido un auge en la era de la información digital, ya que cada día existen millones de registros que se procesan y almacenan cada segundo alrededor del mundo. (Russom et al., 2011)

En este capítulo se realiza un estudio de la literatura relacionada con la temática de la analítica de datos en artículos científicos, para lo cual se describen las investigaciones encontradas en 3 secciones.

La sección 2.1 describe varios trabajos sobre el proceso de búsqueda y el comportamiento de documentos científicos. Además, se realiza un estudio de la literatura de las diferentes bases de datos científicas e indexadoras y artículos de las ventajas y diferencias de cada una. La sección 2.2 refiere pesquisas que utilizan minería y analítica de datos en colecciones documentales y otras que utilizan la metodología MIDANO. Por último, la sección 2.3 describe las literaturas encontradas relacionadas con la creación de líneas de investigación, ninguna de estas fue creada con analítica de datos por lo que el presente proyecto es importante para futuros trabajos.

2.1 Proceso de Búsqueda y Comportamiento de Documentos Científicos

En la literatura se pueden encontrar varios autores que han realizado estudios acerca de los procesos de búsqueda y el comportamiento en documentos científicos en diferentes bases de datos de relevancia. Entre ellos Ellis (1997) analiza la diferencia entre el conocimiento técnico y experimental en la búsqueda de información y encontró que este estudio sitúa al investigador como productor de información y a los profesores como los consumidores de dicha información. Por otra parte, Meho y Tibbo (2003) confirma el modelo de Ellis y además, encontró una descripción más completa del proceso de búsqueda de información incluyendo cuatro nuevas características a las encontradas por Ellis, en vista de esto, el estudio desarrolla un nuevo modelo que agrupa estas características en etapas que son: búsqueda, acceso, procesamiento y finalización. Sin embargo, solamente Moral (2016), describe por completo el proceso, donde se detallan todos los aspectos que deben intervenir en él. De hecho, muchos de los trabajos más relevantes y libros que abordan el tema se centran en un aspecto muy básico para analizar los documentos científicos.

Los modelos existentes de búsqueda de información se describen en textos académicos como Meho y Tibbo (2003) explicado anteriormente en el que se detalla un modelo medianamente completo de la búsqueda de información. Foster (2004) ofrece un modelo no lineal que ilustra tres procesos centrales y tres niveles de interacción contextual, por otro lado Freund et al. (2006) describen el desarrollo de una taxonomía de género y experimentos en clasificación utilizando aprendizaje automático supervisado para la búsqueda de información y presentan una evaluación utilizando datos empresariales. Todos estos requieren que la información sea procesada y analizada para poder interpretarla. Esto restringe la usabilidad del modelo ya que se utiliza más esfuerzo y recursos para entenderlo completamente. En este trabajo, por el contrario, trata de especificar un modelo ágil mediante un script de obtención de información y análisis del mismo.

Adicionalmente, la búsqueda de información está estrechamente relacionada con las bases de datos científicas, de donde se extraerán los documentos y mediante técnicas de analítica de datos se puede obtener conocimiento. Actualmente, existen varias bases de datos y por lo tanto también se han analizado cada una de ellas y se ha recopilado los aportes relacionados con este trabajo de investigación. Los mismos se detallan a continuación. (Šubelj et al., 2015)

Morales y Aguado (2010) realizan un análisis de las investigaciones en diferentes bases de datos regionales obteniendo como resultado la importancia de la incorporación de las Tecnologías de Información y Comunicación en la sociedad. Así como Peralta et al. (2015), realiza una caracterización científica de las diferentes bases de datos obteniendo indicadores y características de las fortalezas y debilidades del impacto de las investigaciones y la visibilidad de la comunidad científica a nivel nacional y mundial. Estos trabajos se enfocan en la obtención de indicadores para generar conocimiento mediante ellos, pero existe mucha ambigüedad en la estadística de estos ya que no son modelos predictores de impacto. El presente trabajo de investigación va un poco más allá, desarrollando modelos que permitan obtener las líneas de investigación de importancia para el Ecuador, mediante técnicas de analítica de datos y no solo utilizando indicadores.

En la literatura (Falagas et al., 2008; Gavel y Iselid, 2008; Miguel et al., 2007; Torres-Salinas et al., 2012; Torres-Salinas y Jiménez-Contreras, 2012) se ha encontrado trabajos que especifican características de las medidas de impacto de las bases de datos, así como una comparativa entre ellas. Además, cada una de ellas realiza un análisis de los documentos científicos más importantes a nivel internacional. Por ejemplo, en el artículo de de Moya-Anegón et al. (2007) mediante una comparativa de Scopus y Ulrich, permite determinar qué tan homogéneo es el mundo académico, se determinan variables para obtener indicadores importantes que pueden ser utilizados en Scopus para saber la distribución por áreas geográficas de las investigaciones. Adicionalmente se realiza un estudio de recuperación de la información y del diseño de políticas para el uso de base de datos en la promoción científica. Así también en el artículo de Gavel y Iselid (2008), se presentan algunas características cualitativas de relevancia entre las bases de datos más importantes a nivel mundial y analiza las características principales de cada una de ellas.

Conjuntamente, en el artículo de Falagas et al. (2008), se realiza una comparativa mucho más amplia incluyendo a Google Scholar y a PubMed exponiendo las debilidades y fortalezas de cada una de ellas, así como una revisión sistemática de las páginas web oficiales y de los sistemas de búsqueda de información de cada una de las bases de datos. Este autor no solo se centra en determinar la mejor fuente de información, sino que también realiza un estudio del impacto de cada una y de su funcionalidad y utilidad. Con este análisis Falagas et al. (2008) obtuvo un mejor panorama de las bases de datos para determinar la cobertura de cada una de ellas y de la precisión de sus sistemas de búsqueda.

Por otro lado, los trabajos de Torres-Salinas y Jiménez-Contreras (2012) y Torres-Salinas et al. (2012) realizan un análisis de las bases de datos para medir el impacto de cada una y determinar estrategias para mejorar la investigación en varios sectores. En el primer artículo se propone nuevos modelos de bibliometría que no tiene las fuentes de información y que son muy utilizadas en el ámbito científico de Europa. Para lograr esto se basó en casos de estudios reales de las Universidades de Granada y Navarra obteniendo como resultado propuestas que se ajustan a dichas instituciones. En cambio, en su segundo trabajo se realiza un análisis de Google Scholar y discute el impacto que puede tener este con relación a las bases de datos científicas más grandes del mundo que son WOS y Scopus.

Además, existen varios autores (Castillo-Esparcia et al., 2011; Fernández-González et al., 2017) que desarrollan estudios de las bases de datos científicas para generar modelos o políticas públicas. En su investigación Fernández-González et al. (2017) realiza un análisis de datos abiertos en publicaciones y bases de datos científicas, mediante un proceso de captura, transportación, transformación y analítica, y finalmente permite visualizar información que apoya la toma de decisiones. Como resultado de su trabajo de investigación, el autor, genera metamodelos para la analítica de datos en datos abiertos con lo cual mejora el procedimiento de generación de conocimiento. Por otro lado, Castillo-Esparcia et al. (2011) realiza un análisis de las revistas y bases de datos científicas ya que estas desempeñan un papel muy importante en la difusión de resultados de investigaciones y del avance científico. Como resultado de esta investigación Castillo-Esparcia et al. (2011) realiza una inducción al modelo actual de publicaciones científicas, su tendencia a través del tiempo y cómo estas pueden evolucionar. Los trabajos descritos en esta sección son muy importantes para la elección correcta de las bases de datos que serán la fuente de información del proyecto y también para tener en cuenta todas las características de cada una de ellas.

2.2 Minería y Analítica de Datos en Colecciones Documentales

En el caso de analítica de datos se han encontrado algunos estudios en relación a análisis de datos en documentos científicos, pero con otros enfoques y objetivos. A continuación, se especifica cada artículo y su contribución.

Espejo y Apolo (2011), realizan una mejora a la calidad metodológica en el desarrollo del kinesiotaping, esto lo logran mediante técnicas de analítica de datos en artículos científicos de alto impacto referentes a dicho tema. Por otra parte, Michán et al. (2008), determina indicadores de crecimiento de la investigación y otras medidas estadísticas que permitan mejorar los resultados y la toma de decisiones de las investigaciones, esto lo realizan a través de un análisis de datos de los artículos científicos por países, áreas de estudio, grupos de investigación, resúmenes, tipos de artículos, idioma, entre otros. Michán et al. (2008) se centran en la analítica de datos para generar metodologías e indicadores para la toma de decisiones. En la presente investigación se desarrolla una metodología que permite obtener las mejores líneas de investigación para el país.

Vílchez-Román y Huamán-Delgado (2017) emplearon herramientas para analizar de forma cuantitativa las palabras clave y los autores de los estudios examinados, para obtener estrategias que permitan la estabilidad y cambio en la teoría organizacional. Flores (2017) utiliza algoritmos de minería de datos de los usuarios registrados y a documentos (tesis, artículos y proyectos de investigación) cargados en una página web universitaria, con el fin de realizar una retroalimentación tanto a docentes como a estudiantes y administradores del sistema para mejorar las estrategias de la docencia o la enseñanza. Al igual que los anteriores autores estas investigaciones plantean modelos diferentes a los utilizados en el presente proyecto es por ello que los modelos en esta tesis han derivado en un estudio inductivo principalmente porque no existen modelos anteriores que describan de manera holística y desde un punto de vista conceptual sobre técnicas de analítica de datos para obtener líneas de investigación.

Además, Godoy (2017) realiza una revisión de la literatura publicada en los últimos años referente a técnicas de aprendizaje de máquina empleadas para la minería de texto. Teniendo como principal característica a las técnicas más usadas en los artículos estudiados las cuales son support vector machine (SVM), k-means (K-M), k-nearest neighbors (K-NN), naive bayes (NB), self-organizing maps (SOM). Los pares que aparecen con mayor frecuencia son SVM/NB, SVM/K-NN, SVM/decission tree. Este artículo ha sido de vital importancia para determinar las herramientas y algoritmos adecuados para la obtención de resultados exitosos.

Por otra parte algunos autores ha utilizado técnicas de analítica de texto para obtener conceptos, entre estos Deirmengian et al. (2015) utiliza técnicas de minería de texto en resultados de laboratorio alfa-defensina del líquido sinovial para la identificación de los conceptos de especie o virulencia del organismo. Liberatore et al. (2018) describen el desarrollo de

analítica de datos en una colección documental, lo que permite la interpretación semántica de los documentos. Además, con estas características se obtiene una extracción de conceptos. Miñarro (2018) desarrolla una herramienta de análisis semántico para extracción de palabras claves mediante la utilización de herramientas de análisis de sentimientos. Velandia et al. (2017) mediante técnicas de analítica de datos y minería de texto logran obtener y extraer conceptos para conocer la opinión de usuarios en Twitter y sus intereses dentro de la plataforma. Estos artículos son importantes para este trabajo de investigación ya que se centran en la extracción de conceptos, pero en ámbitos diferentes, lo que ha permitido tener una idea de las diferentes técnicas que se han utilizado para este fin.

Por último, Moral (2016) realiza un análisis de los investigadores y sus trabajos para comprender como realizan la búsqueda de información del tema. Esta investigación se realizó mediante estudios cuantitativos y modelos holísticos conceptuales, en este trabajo además se presenta un análisis de las relaciones entre conceptos para la visualización de la información, también en esta investigación se realiza en aplicaciones prácticas un análisis de los sistemas de información y se efectúa una caracterización de los mismos.

En el análisis realizado para la obtención de literaturas relacionadas con la extracción de conceptos en documentos científicos se han encontrado muy pocos autores que realicen esto, y menos aún autores que mediante la extracción de conceptos y la analítica de datos generen líneas de investigación de importancia para un país en el ámbito de las ciencias de la computación. De hecho, solo hay dos modelos de búsqueda de información relacionados con el dominio de la informática (Almeida et al., 2014; Ebner et al., 2006), pero ninguno de ellos describe el proceso en un contexto de investigación. Por otro lado, Moral (2016) especifica el proceso desde un contexto de investigación, pero no se centra en las líneas de investigación, sino, en la búsqueda de información en las bases de datos científicas y un estudio del investigador, además, Moral (2016) realiza un análisis de las relaciones que existen entre los documentos, y crea clústeres para determinar características del investigador. Además, expone una técnica de categorizar a los documentos científicos por medio de clústeres lo que es importante para la presente tesis.

Entonces, para generar un modelo mucho más preciso y con resultados para beneficio de los investigadores ecuatorianos, se ha decidido mejorar estos procesos de extracción de conceptos para obtener resultado como las líneas de investigación de importancia para el país. Es por esto que se ha decidido utilizar a la analítica de datos como herramienta para cumplir este objetivo, específicamente la metodología MIDANO utiliza por varios autores

(Aguilar et al., 2017; Caldera et al., 2018; Lozada et al., 2017; VALENCIA, 2017). Esto indican que sus resultados fueron muy eficientes, por ello se ha optado por incluir esta metodología para el desarrollo de analítica de datos. Como conclusión de las investigaciones los autores Aguilar et al. (2017); Caldera et al. (2018); Lozada et al. (2017); VALENCIA (2017) describen a la metodología MIDANO como una ayuda sustancial para la identificación y priorización de los procesos de interés, y una toma de decisiones más eficiente. Las técnicas y la metodología utilizada para realizar la extracción de conocimiento es la adecuada ya que la analítica de datos es una poderosa herramienta para generar conocimiento. Además, mediante la utilización de diferentes técnicas de analítica de datos como el procesamiento del lenguaje natural se logra obtener buenos resultados.

2.3 Creación de Líneas de Investigación

Para la creación de líneas de investigación en el ámbito de las ciencias de la computación no se ha encontrado autores que realicen esto ya que existen estándares internacionales como la ACM que define las áreas de investigación. Pero esto genera un problema en países en desarrollo como el Ecuador, en donde la investigación actual no se realiza correctamente y no tiene un beneficio para el desarrollo del país (Larrea, 2006). Se han encontrado varios autores en tres documentos que realizan la creación de líneas de investigación, pero en ámbitos diferentes y no adaptables a un país.

Es el caso de Mora-Riapira et al. (2015) desarrollan la definición de las líneas de investigación que sirven para empresas de cualquier tamaño en Colombia, esto lo hacen mediante una investigación descriptivo-exploratoria documental de artículos científicos empresariales obteniendo indicadores que les permitieron obtener las áreas de estudio. Así es que en el resultado de esta investigación, los autores señalan la necesidad de fortalecer investigaciones que desarrollen líneas de investigación de importancia para el Ecuador.

Por otro lado Acosta y Medina (1997) desarrollan la creación de las líneas de investigación de enfermería pero lo hacen mediante el análisis de la evolución de la investigación, por medio de indicadores y basándose en un modelo conceptual de enfermería.

Por último, Bacino et al. (2018) hacen hincapié en la necesidad de realizar actividades de gestión, docencia e investigación para ser un docente universitario que aporte al país. Con esto, realizan un instrumento de medición para evaluar el grado de conocimiento del

uso de las bases de datos científicas por un profesor, con el fin de proponer líneas de trabajo de interés para el catedrático, es decir, líneas de investigación para el educador en específico. Como resultado de este estudio, dan a conocer que existe una brecha en el conocimiento de las bases o indexadores de colecciones científicas por parte de los docentes, lo que demuestra que las investigaciones realizadas por los mismos no son tan eficientes como deberían serlo, lo que implica una deficiencia en el desarrollo de un país tercermundista.

Bacino et al. (2018) analiza los trabajos de investigación de los docentes universitarios y su impacto en la sociedad y el país, y encontró que no son tan eficientes como deberían serlo, lo que implica que la investigación es necesaria para el desarrollo del país. Es por ello que la presente tesis se centra en crear líneas de investigación que permitan generar proyectos que sean de beneficio para un país. Estas son importantes para obtener resultados de calidad y que sean de alto impacto para el desarrollo del país.

Capítulo 3

Solución adoptada

En este capítulo se presenta la solución adoptada para el problema mediante el desarrollo de analítica de datos para la obtención de líneas de investigación específicas para el Ecuador.

Además, se desarrolla la metodología MIDANO (Aguilar, 2013) (más detalle de la metodología en el capítulo 1 de la presente tesis), especificando todas sus fases y tareas para la obtención de los resultados.

El primer paso es conocer la organización objeto de estudio. En este caso el objeto de estudio es los objetivos del "Libro Blanco de la Sociedad de la Información y del Conocimiento" (LBSIC) del MINTEL y el Plan Toda una Vida de la República del Ecuador, ya que estos objetivos se alinean a las ciencias de la computación. El interés del país es tener líneas de investigación adecuadas para un correcto desarrollo del país. El detalle de estos objetivos y del libro blanco se especifican en el capítulo 1.

En el caso del MINTEL y de la presente investigación se utilizarán sus líneas prioritarias de investigación para ser analizadas mediante MIDANO, siendo la construcción de nuevas líneas de investigación específicas para el Ecuador, el proceso de analítica de datos prioritario.(Aguilar, 2013)

Como se explica en el Capítulo 1 en el planteamiento del problema existen escenarios actuales y futuros del Ecuador en el ámbito de las líneas de investigación que se deben realizar para cumplir los objetivos, por lo expuesto, se utiliza estos escenarios para obtener los procesos y subprocesos de MIDANO para el desarrollo de la analítica de datos.

Especificación del flujograma del proceso autonómico

El flujograma del proceso autonómico se realiza para el escenario futuro, es decir, los procesos determinados en el flujograma están basados en lo que se requiere obtener gracias a la analítica de datos. Para generar el flujograma se ha planteado lo siguiente:

- Diseño del flujograma
- Descripción de las tareas y relaciones

Dentro del diseño del flujograma del proceso autonómico de MIDANO se tienen las siguientes etapas:

- Fase 1: Monitoreo
- Fase 1: Análisis
- Fase 2: Planificación
- Fase 3: Ejecución

El flujograma del proceso autonómico describe las tareas específicas que tendrá que seguir el proceso de analítica de datos para obtener el resultado esperado y que responden a los requerimientos de cada escenario.

El resultado esperado de la realización del flujograma seró la implementación de un aplicativo de analítica de datos que permitirá al investigador ecuatoriano o la Sociedad informática conocer las líneas de investigación específicas para el Ecuador en el área de las ciencias de la computación, esto permitirá tener investigaciones de calidad y que contribuyan al desarrollo del país y tener mejores investigadores.

En la figura 3.1 se presenta el flujograma del proceso autonómico de Analítica de datos adaptada para la realización del proyecto y del Desarrollo de Técnicas de analítica de datos para determinar las líneas de investigación específicas para el Ecuador.

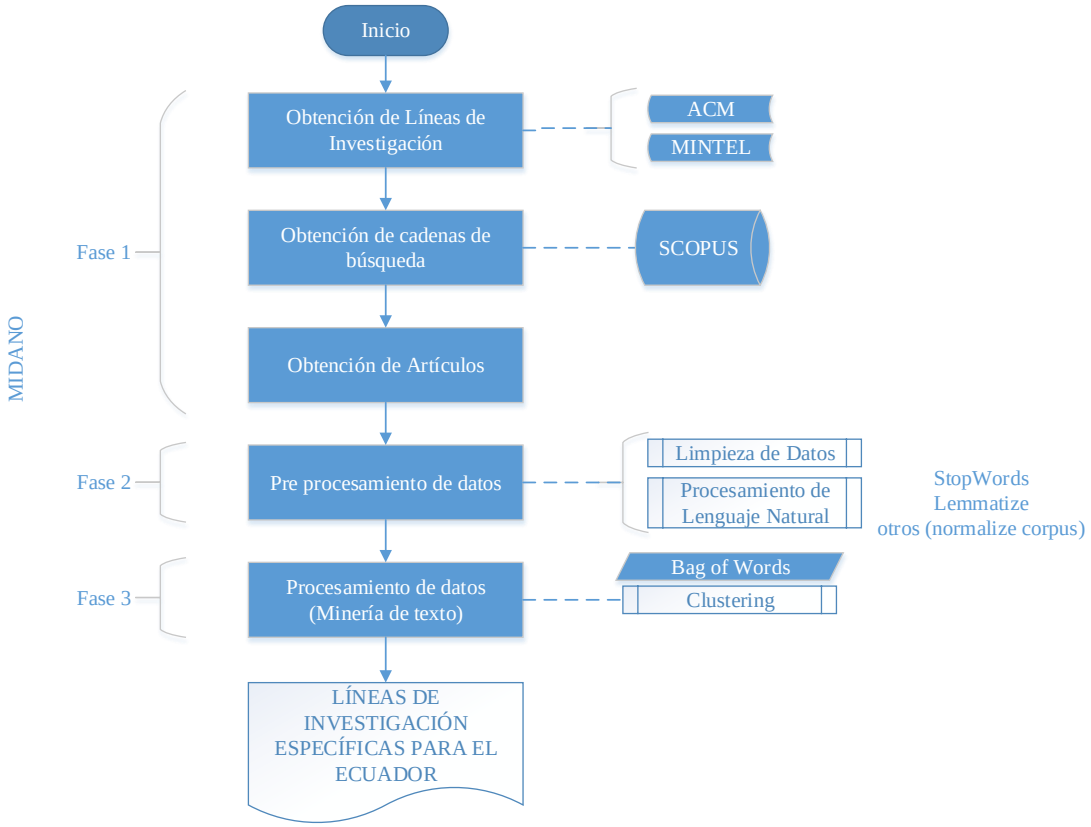

Fig. 3.1 flujograma del proceso autonómico de Analítica de Datos.

Como se puede observar en el flujograma de procesos autonómicos definido, se realizarán todos los procesos descritos para obtener como resultado final el listado de las líneas de investigación específicas para el Ecuador.

Para definir las tareas se tomó como referencia las fases y los procesos del flujograma. Estas tareas son definidas como una acción para iniciar y finalizar un proceso dentro del desarrollo de la analítica de datos.

Previamente a detallar las tareas de cada proceso, en la tabla 3.1 se presenta la descripción de los procesos y las relaciones que existen para cada etapa del flujograma del proceso autonómico.

Esta tabla explica de manera general los procesos a seguir de cada fase de MIDANO en el flujograma, así como las relaciones que existen entre procesos.

Tabla 3.1 Descripción de los procesos del flujograma y relaciones para el proceso autonómico

Etapa	**Tarea flujograma del proceso autonómico**	**Descripción**
Monitoreo	Obtención de las Líneas de Investigación	Se realiza un análisis de asociación entre las fuentes de datos de la ACM y el MINTEL para obtener las palabras claves que servirán como entrada del siguiente proceso.
y	Obtención de las cadenas de búsqueda	Mediante las líneas de investigación obtenidas, se genera la cadena de búsqueda que servirá como entrada del siguiente proceso.
Análisis	Obtención de artículos	A través de la cadena de búsqueda, se realiza la extracción de los artículos científicos de relevancia para el Ecuador de la base de datos Scopus.
Planificación	Pre procesamiento de datos	Se realiza una limpieza de datos y por medio del procesamiento del lenguaje natural se obtienen los datos que serán procesados para obtención de conocimiento
Ejecución	Analítica de texto	Mediante técnicas de analítica de datos se obtienen las líneas de investigación específicas para el Ecuador

Identificar los procesos de analítica de datos permite mejorar las tareas y obtener mejores resultados del análisis de datos. Además, aporta una visión más amplificada de lo que se debe realizar para conseguir los resultados esperados, haciendo que las tareas sean más eficientes y se adapten a las necesidades del país.

Como se observa en la tabla 3.1 además de identificar los procesos se realiza un estudio de lo que se debería obtener de ellos y también las relaciones de dependencia que existen entre ellos, es decir, la salida de un proceso es la entrada de otro.

Una vez definidos los procesos y las relaciones existentes entre los procesos detallados en la tabla anterior, a continuación, en la tabla 3.2 se presenta el detalle de las fuentes de datos para el desarrollo de los procesos del flujograma que servirán para obtener los resultados descritos en la tabla 3.1. Estas fuentes de datos están descritas en el capítulo 1, y se analizan en secciones siguientes del presente capítulo.

Tabla 3.2 Fuente de datos de los procesos del flujograma

Tarea	**Fuentes de datos requeridas**	**Efectos esperados**
Obtención de las Líneas de Investigación	ACM y MINTEL	Análisis de las líneas de investigación de la ACM y del MINTEL y de la relación que existe entre ambas.
Obtención de las cadenas de búsqueda	ACM y MINTEL	Mediante un análisis se establecen las palabras claves que servirán de base para obtener la cadena de búsqueda a ser utilizada por Scopus.
Obtención de artículos	API Scopus	Se realiza la extracción de los documentos científicos mediante la cadena de búsqueda predefinida obteniendo como resultado los artículos de relevancia para el Ecuador
Pre procesamiento de datos	Base de datos de Artículos científicos de Scopus	Mediante varios algoritmos se realiza la limpieza de datos y el análisis de lenguaje natural para tener datos de calidad para el procesamiento
Analítica de texto	BDD de artículos científicos de Scopus pre procesados	Mediante técnicas de analítica de datos (clustering) obtener las líneas de investigación específicas para el Ecuador

En las secciones siguientes, se presenta con detalle cada uno de los procesos con sus respectivas tareas y el resultado en cada etapa de MIDANO. Es así como, el proceso a seguir para la obtención de los resultados de esta investigación se ilustra en la figura 3.2 es un gráfico a detalle del flujograma presentado en la figura 3.1.

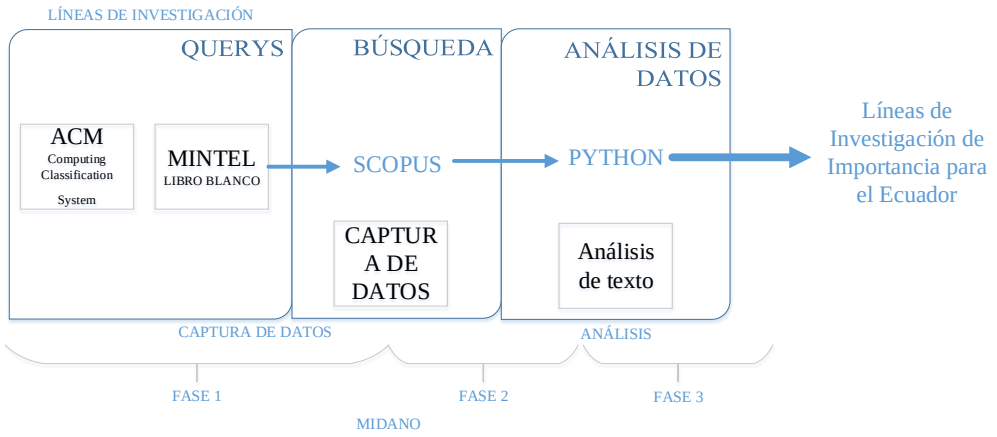

Fig. 3.2 Proceso del desarrollo de analítica de datos.

3.1 Obtención de Líneas de Investigación

En esta sección se presenta el análisis realizado de las líneas de investigación tanto de la ACM como el MINTEL. Además, se detallan las tareas realizadas para la obtención de las palabras claves para la generación de la cadena de búsqueda que se detalla en la siguiente sección.

A continuación, en la tabla 3.3 se definen todas las tareas del proceso de "Obtención de líneas de investigación" incluido en la etapa de monitoreo y análisis, para obtener como resultado las palabras claves para la generación de las cadenas de búsqueda que serán utilizadas en Scopus.

Tabla 3.3 Tareas para el primer proceso del flujograma del proceso autonómico en la etapa de monitoreo y análisis

Nombre de las tareas	Analizar las líneas de investigación de la ACM y el MINTEL, analizar las asociaciones entre ellas y obtener las palabras claves para generar la cadena de búsqueda
Descripción	Se realiza un análisis de las líneas de investigación de cada entidad para luego realizar una asociación que permita definir las palabras claves. Estas tareas se realizarán lógicamente.
Fuente de datos	ACM y MINTEL
Tipo de tarea de analítica de datos	Descubrimiento
Técnica de Analítica de Datos	No supervisado, descriptivo
Con que otro proceso se relaciona	Obtención de las cadenas de búsqueda
Etapa del flujograma del proceso autonómico	Monitoreo y Análisis

Para realizar las tareas detalladas anteriormente, se efectuó una extracción de todas las líneas de investigación de la ACM para ejecutar un cotejamiento con las líneas de investigación del MINTEL que como se especifica en el Capítulo 1 son apenas 7.

Al ser un tipo de análisis lógico como se ilustra en la figura 3.3, no necesita una herramienta de analítica de datos para obtener los resultados. Las herramientas utilizadas para esta sección han sido la ACM Computing Classification System (CCS) y el Libro Blanco del MINTEL (2019).

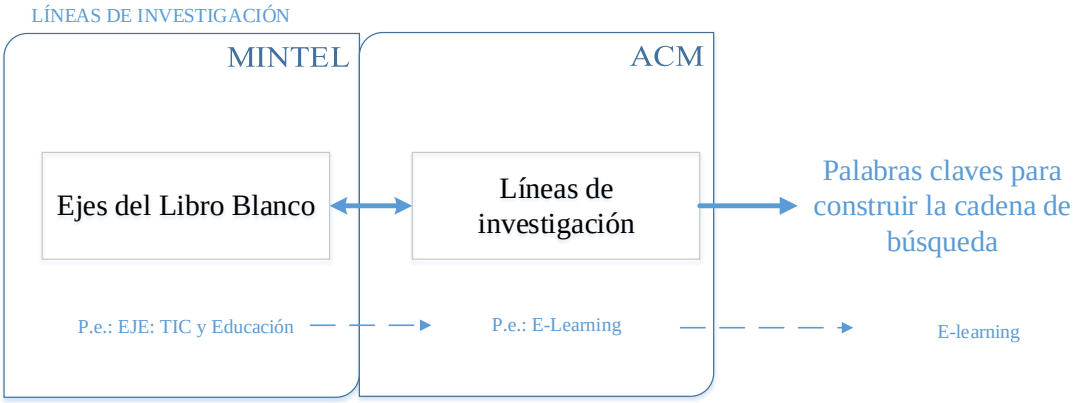

Fig. 3.3 Proceso de Obtención de líneas de investigación.

A continuación, se presenta con detalle cada una de ella.

Recopilación de las Líneas de Investigación de la ACM

La ACM cuenta con más de 1500 líneas de investigación agrupadas en 13 áreas principales que se listan a continuación:

1. General and reference
2. Hardware
3. Computer systems organization
4. Networks
5. Software and its engineering
6. Theory of computation
7. Mathematics of computing
8. Information systems
9. Security and privacy
10. Human-centered computing
11. Computing methodologies

12. Applied computing
13. Social and professional topics

De las 13 áreas principales, se derivan líneas de investigación que se detallan a continuación:

1. nivel 0 = 13 líneas de investigación
2. nivel 1 = 84 líneas de investigación
3. nivel 2 = 543 líneas de investigación
4. nivel 3 = 1087 líneas de investigación
5. nivel 4 = 353 líneas de investigación
6. nivel 5 = 33 líneas de investigación

Este proyecto se basa principalmente en esta clasificación internacional (ACM) para obtener las cadenas de búsqueda que permitirán extraer los artículos científicos de Scopus.

Líneas de investigación del Libro Blanco de la Sociedad de la Información y del Conocimiento

En el año 2018, el Ministerio de Telecomunicaciones y de la Sociedad de la Información desarrolló y trabajó con una única visión sustentada en cinco ejes que se describen a continuación:

1. Infraestructura y Conectividad: por mandato constitucional, esta línea de investigación tiene relevancia para los objetivos del Plan toda una vida. En este eje se impulsa el despliegue de la infraestructura de telecomunicaciones en todo el país.
2. Gobierno Electrónico: Este eje se refiere al uso de las TIC por parte del gobierno y sus instituciones para aumentar la eficacia y eficiencia en la gestión pública. En el Ecuador se ha implementado tres temáticas principales las mismas que son: Gobierno Abierto (participación de la ciudadanía por internet), el Gobierno Cercano (Facilidades de acceso a la información y acceso a trámites en línea para la sociedad) y el Gobierno Eficaz y Eficiente (Intercambio e interconexión de servicios de entidades del gobierno para reducir los requisitos en línea y presenciales en trámites de la ciudadanía)

3. Inclusión y Habilidades Digitales: Este eje tiene como objetivo fortalecer la inclusión y habilidades digitales de los ciudadanos para mejorar las oportunidades laborales de los ecuatorianos.

4. Seguridad de la Información y Protección de Datos Personales: Establece las directrices de seguridad de la información en los sectores financieros, bienes de consumo, energía y recursos renovables, tecnología y salud. Esta línea de investigación o eje es muy importante ya que según el MINTEL en Ecuador existen varías brechas de seguridad de la información.

5. Economía Digital y Tecnologías Emergentes: Establece la transformación digital de las empresas, la evolución del Comercio Electrónico, el impulso de la innovación y emprendimientos de base tecnológica, la dinamización de la industria TIC y del aprovechamiento de las Tecnologías Emergentes.

Actualmente, el Ministerio de Telecomunicaciones y de la Sociedad de la Información actualizó las líneas de investigación prioritarias para el país, estas se describen a continuación:

1. TIC y Educación: Este eje hace referencia a la apropiación y participación de la Educación con las teorías del campo de las tecnologías, es decir, es una relación que existe entre la cognición y la tecnología. Este eje es muy importante para el Ecuador, debido a que la educación ecuatoriana debería hacer uso de la ciencia y tecnología para mejorar su desarrollo e independencia.

2. Redes e infraestructuras de telecomunicaciones: como se especifica anteriormente este eje se mantiene siendo una de las mayores prioridades para el país y para los objetivos del Plan toda una vida. En este eje se impulsa el despliegue de la infraestructura de telecomunicaciones en todo el país.

3. Tecnologías de radiodifusión digital: En Ecuador se ha venido desarrollando un crecimiento en las tecnologías de radiodifusión en frecuencia modulada. Es así como este eje hace referencia al mejoramiento de estas tecnologías digitales.

4. Ciudades Inteligentes, sostenibles e inclusivas: Este eje es trascendental para las ciudades más representativas del Ecuador, puesto que la sostenibilidad debe ser un mecanismo vital para las ciudades inteligentes con la cooperación de los ciudadanos en la democratización de las decisiones adoptadas sobre el futuro. Conjuntamente,

este eje hace referencia a cómo la tecnología puede ser utilizada por la ciudadanía para que estos obtengan más beneficios tanto personales como de trámites. Este eje intenta conseguir una mejor ciudad capaz de ofrecer mejores servicios a los ecuatorianos y que esta pueda utilizar eficientemente los recursos de las tecnologías de la información y comunicación. Tiene como objetivo gozar de un mejor control ciudadano, conservación de energías, mejor prestación de servicios y progreso en la calidad de vida.

5. Big Data: Este eje describe el tratamiento de los grandes volúmenes de datos que se generan a diario en el país. Esta línea de investigación es muy importante para un mejoramiento en la toma de decisiones en base a lo que se pueda hacer con los datos.

6. Seguridad de la Información: Este eje se mantiene con los definidos precedentemente, establece las directrices de seguridad de la información en todos los sectores del Ecuador.

7. TIC y Producción: Este eje hace referencia a la mejora sustancial de la industria TIC, mediante varias temáticas y varios elementos de las líneas de investigación, como el comercio electrónico.

3.1.1 Asociación de las fuentes de datos

Para la asociación de las fuentes de datos se ha realizado un levantamiento de la información tanto de la ACM como del MINTEL. Como se especifica en las tareas de MIDANO del presente proyecto en la tabla 3.3 se ha desarrollado lo siguiente:

La obtención de las palabras claves que tengan importancia para el Ecuador en el área de las ciencias de la computación. Con estas, se realizarán la construcción de las cadenas de búsqueda que se explica en la siguiente sección.

A continuación, se detalla cada eje del MINTEL y su asociación con la ACM o con sinónimos que serán útiles para la generación de las cadenas de búsqueda.

1. TIC y Educación: Este eje hace referencia a la apropiación y participación de la Educación con las teorías del campo de las tecnologías, por lo que las líneas de investigación de la ACM y sus sinónimos que están más relacionadas con este eje son:

- Computer-assisted instruction

- Interactive learning environments
- Collaborative learning
- Learning management systems
- Distance learning
- E-learning
- Special capabilities
- technology-enhanced learning
- computer-based instruction
- computer-based training
- computer-aided instruction
- web-based training
- online education
- virtual education

Como detalla anteriormente, las líneas de investigación se asemejan con el eje TIC y Educación porque este hace referencia a optimizar las capacidades de los ecuatorianos para mejorar las habilidades digitales, en estas se encuentran por ejemplo E-learning para aprendizaje electrónico, políticas tecnológicas, aprendizaje a distancia, aprendizaje colaborativo, entre otros.

2. Eje de Redes e infraestructuras de telecomunicaciones: Este eje definido por el MINTEL es muy importante para el beneficio del país ya que la infraestructura es vital para el desarrollo del Ecuador, por lo que las líneas de investigación de la ACM y sus sinónimos que están más relacionadas con este eje son:

- Network
- BIG Networks
- Network types

- internet usage
- internet use
- broad networks
- extensive network
- wide-ranging network
- wide network
- Wide area networks
- WAN
- widespread network

Como detalla anteriormente, las líneas de investigación se asemejan con el eje Redes e infraestructuras de telecomunicaciones por que este abarca a todas las intercomunicaciones que existen en el país, así como las redes y como poder mejorar la comunicación entre el Gobierno y la sociedad.

3. Tecnologías de radiodifusión digital: Este eje hace referencia al mejoramiento de estas tecnologías digitales y de radiodifusión, por lo que las líneas de investigación de la ACM y sus sinónimos que están más relacionadas con este eje son:

- Radio frequency and wireless interconnect
- Cognitive radios
- 5G

Como detalla anteriormente, las líneas de investigación se asemejan con el eje Tecnologías de radiodifusión digital por que este abarca todos los tipos radiodifusión, así como el mejoramiento de la comunicación digital como es la tecnología 5G.

4. Ciudades Inteligentes, sostenibles e inclusivas: Este eje intenta conseguir una mejor ciudad capaz de ofrecer mejores servicios a los ecuatorianos y que esta pueda utilizar eficientemente los recursos de las tecnologías de la información y comunicación, por lo que las líneas de investigación de la ACM y sus sinónimos que están más relacionadas con este eje son:

- Smart Workspace
- Biochips
- Digital Twin
- Carbon Nanotube
- IoT Platform
- Virtual Assistants
- Silicon Anode Batteries
- Smart Robots
- Autonomous Mobile Robots
- AI PaaS
- Neuromorphic Hardware
- Exoskeleton
- Biotech
- Flying Autonomous Vehicles
- Smart Dust
- Artificial General Intelligence

Como detalla anteriormente, las líneas de investigación se asemejan con el eje Ciudades Inteligentes, sostenibles e inclusivas por que este abarca la sostenibilidad de las ciudades inteligentes, así como los métodos o herramientas para obtener una ciudad inclusiva con la tecnología.

5. Big Data: Este eje hace referencia al tratamiento de los grandes volúmenes de datos que se generan a diario en el país, por lo que las líneas de investigación de la ACM y sus sinónimos que están más relacionadas con este eje son:

- Deep Neural Nets
- Deep Learning

- Big Data
- Deep Neural Network ASICs
- Blockchain for Data Security

Como detalla anteriormente, las líneas de investigación se asemejan con el eje Big Data ya que este abarca todos los tipos de analítica de datos que pueden tener impacto en la sociedad ecuatoriana y todo lo relacionado al tratamiento y procesamiento de datos.

6. Seguridad de la Información: Este eje al hacer referencia a políticas de seguridad de la información y a la privacidad y protección de la información del Ecuador, por lo que las líneas de investigación de la ACM y sus sinónimos que están más relacionadas con este eje son:

- Cryptography
- Formal methods and theory of security
- Security services
- Intrusion/anomaly detection and malware mitigation
- Security in hardware
- Systems security
- Network security
- Database and storage security
- Software and application security
- Human and societal aspects of security and privacy

Como detalla anteriormente, las líneas de investigación se asemejan con el eje Seguridad de la Información ya que este abarca todo lo referente a seguridad y privacidad de la información, además, estás líneas de investigación tienen áreas como Seguridad en servicios que se pueden utilizar en los sistemas estatales, seguridad de redes, seguridad del Hardware, seguridad de bases de datos, seguridad y privacidad de la sociedad, entre otros.

7. TIC y Producción: Este eje hace referencia a la transformación digital y evolución del comercio electrónico que permitirá el desarrollo del país mediante la nueva economía, además, se ha incluido las tendencias de las tecnologías emergentes definidas en Gartner del año 2018 (se tomarán en cuenta las que están en crecimiento y las que están en la cima del ciclo Gartner) [1] y las líneas de investigación de la ACM que están más relacionadas con este eje son:

- 4D Printing
- Knowledge Graphs
- Edge AI
- Autonomous Driving Level 5
- Conversational AI Platform
- Self-Healing System Technology
- Volumetric Displays
- Quantum Computing
- Brain-Computer Interface
- Blockchain
- Digital economy
- Internet Economy
- Web Economy
- New Economy

Como detalla anteriormente, las líneas de investigación se asemejan con el eje TIC y Producción ya que estas líneas tienen como objetivo mejorar la tecnología del país mediante la nueva economía, los nuevos avances científicos en la industria, la industria digital 4.0, las nuevas tendencias de producción, entre otros.

[1] Fuente: Gartner, obtenido de: https://www.gartner.com/doc/3885468/hype-cycle-emerging-technologies-

Como resultado de este proceso, se obtuvieron 74 palabras claves que son: "Computer-assisted instruction", "Interactive learning environments", "Collaborative learning", "Learning management systems", "Distance learning", "E-learning", "Special capabilities", "technology-enhanced learning", "computer-based instruction", "computer-based training", "computer-aided instruction", "web-based training", "online education", "virtual education", "Network", "BIG Networks", "Network types", "internet usage", "internet use", "broad networks", "extensive network", "wide-ranging network", "wide network", "Wide area networks", "WAN", "widespread network", "Radio frequency and wireless interconnect", "Cognitive radios", "5G ", "Smart Workspace", "Biochips", "Digital Twin", "Carbon Nanotube", "IoT Platform", "Virtual Assistants", "Silicon Anode Batteries", "Smart Robots", "Autonomous Mobile Robots", "AI PaaS", "Neuromorphic Hardware ", "Exoskeleton ", "Biotech ", "Flying Autonomous Vehicles", "Smart Dust ", "Artificial General Intelligence ", "Deep Neural Nets", "Deep Learning", "Big Data", "Deep Neural Network ASICs", "Blockchain for Data Security ", "Cryptography", "Formal methods and theory of security", "Security services", "Intrusion/anomaly detection and malware mitigation", "Security in hardware", "Systems security", "Network security", "Database and storage security", "Software and application security", "Human and societal aspects of security and privacy", "4D Printing ", "Knowledge Graphs ", "Edge AI ", "Autonomous Driving Level 5 ", "Conversational AI Platform ", "Self-Healing System Technology ", "Volumetric Displays ", "Quantum Computing ", "Brain-Computer Interface", "Blockchain", "Digital economy", "Internet Economy", "Web Economy", y "New Economy".

Estas palabras claves serán utilizadas para creación de las cadenas de búsqueda para Scopus.

3.2 Obtención de cadenas de búsqueda

En esta sección se presenta el análisis realizado de las palabras claves obtenidas en la sección previa. Además, se detallan las tareas realizadas para la obtención de la cadena de búsqueda que será utilizada en Scopus para obtención de artículos científicos.

A continuación, en la tabla 3.4 se definen todas las tareas del proceso de "Obtención de cadenas de búsqueda" incluido en la etapa de monitoreo y análisis, para obtener como resultado la cadena de búsqueda que servirá para realizar la consulta en Scopus.

Tabla 3.4 Tareas para el segundo proceso del flujograma del proceso autonómico en la etapa de monitoreo y análisis

Nombre de las tareas	Analizar las palabras claves, construir las cadenas de búsqueda y obtener la cadena de búsqueda unificada
Descripción	Se realiza un análisis de las palabras claves obtenidas en el proceso anterior para construir la cadena de búsqueda que servirá para obtener los artículos científicos desde Scopus.
Fuente de datos	ACM y MINTEL
Tipo de tarea de analítica de datos	Descubrimiento
Técnica de Analítica de Datos	No supervisado, descriptivo
Con que otro proceso se relaciona	Obtención de Artículos
Etapa del flujograma del proceso autonómico	Monitoreo y Análisis

El procedimiento para generar la cadena de búsqueda es muy simple, mediante las palabras claves obtenidas anteriormente, se las encierra entre comillas y dentro de la palabra reservada ALL. Como se detalla en la figura 3.4.

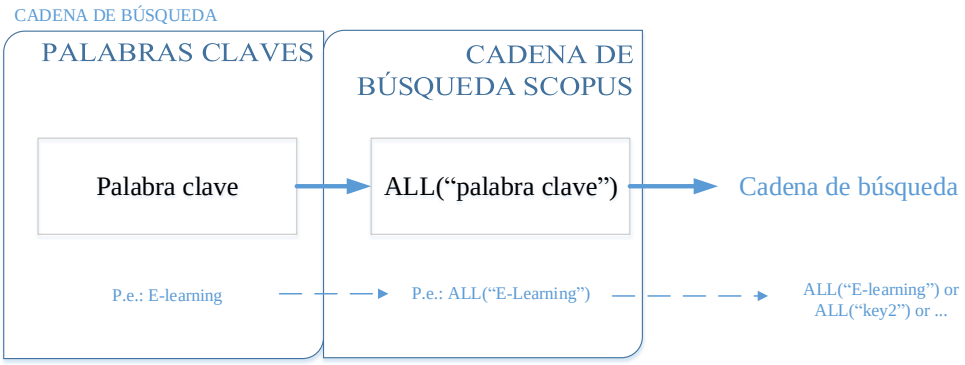

Fig. 3.4 Proceso de Obtención de cadena de búsqueda.

Por ejemplo:

- En el Eje TIC y Educación, una de las palabras claves obtenidas es E-learning, entonces la cadena de búsqueda para esa palabra clave es: ALL("E-learning") or
- En el Eje Redes e infraestructuras de telecomunicaciones, una de las palabras claves obtenidas es BIG Networks, entonces la cadena de búsqueda para esa palabra clave es: ALL("BIG Networks") or

- En el Eje Tecnologías de radiodifusión digital, una de las palabras claves obtenidas es Cognitive radios, entonces la cadena de búsqueda para esa palabra clave es: ALL("Cognitive radios") or
- En el Eje Ciudades Inteligentes, sostenibles e inclusivas, una de las palabras claves obtenidas es Autonomous Mobile Robots, entonces la cadena de búsqueda para esa palabra clave es: ALL("Autonomous Mobile Robots") or
- En el Eje Big Data, una de las palabras claves obtenidas es Big Data, entonces la cadena de búsqueda para esa palabra clave es: ALL("Big Data") or
- En el Eje Seguridad de la Información, una de las palabras claves obtenidas es Network security, entonces la cadena de búsqueda para esa palabra clave es: ALL("Network security") or
- En el Eje TIC y Producción, una de las palabras claves obtenidas es Autonomous Driving Level 5, entonces la cadena de búsqueda para esa palabra clave es: ALL("Autonomous Driving Level 5") or

Entonces la cadena de búsqueda unificada se debe unir todas las cadenas de búsqueda dentro de las ciencias de la computación, es decir:

SUBJAREA(COMP) AND DOCTYPE(ar) and (ALL("Autonomous Driving Level 5") or ALL("Network security") or ALL("Big Data") or ALL("Autonomous Mobile Robots") or ALL("Cognitive radios") or ALL("BIG Networks") or ALL("E-learning")) AND NOT ACCESSTYPE(OA))

En donde, las siguientes palabras reservadas en Scopus significan:

- SUBJAREA = Área de estudio de Scopus (COMP=computer science)
- DOCTYPE = Tipo de documento en Scopus (ar=artículos)
- ALL = Busca la palabra en campos como el título, resumen, keywords, entre otras
- ACCESSTYPE = Tipo de acceso de los artículos científicos

Tomando en cuenta todas las palabras claves el resultado de la cadena de búsqueda unificada es:

```
doc_srch = ElsSearch2('','scopus?count=25&view=COMPLETE&content=all
&&facets=language&query=SUBJAREA(COMP) AND DOCTYPE(ar) and
(ALL("BIG Networks") or ALL("Network types") or ALL("internet usage")
or ALL("internet use") or ALL("broad networks") or ALL("extensive network")
 or ALL("wide-ranging network") or ALL("wide network")
 or ALL("Wide area networks") or ALL("WAN") or ALL("widespread network")
or ALL("Document types") or ALL("Cross-computing tools and techniques")
 or ALL("e-governance") or ALL("e-government") or ALL("egovernment") or
 ALL("public policies") or ALL("public policy") or ALL("public plans") or
 ALL("public strategy") or ALL("information system")
 or ALL("technology policy") or ALL("Computing policy")
or ALL("Computer-assisted instruction") or
 ALL("Interactive learning environments") or ALL("Collaborative learning")
 or ALL("Learning management systems") or ALL("Distance learning") or
 ALL("E-learning") or ALL("Special capabilities") or ALL("LMS") or
 ALL("technology-enhanced learning") or ALL("computer-based instruction")
 or ALL("computer-based training") or ALL("computer-aided instruction")
 or ALL("web-based training") or ALL("online education") or
 ALL("virtual education") or ALL("Cryptography") or
 ALL("Formal methods and theory of security") or ALL("Security services")
 or ALL("Intrusion/anomaly detection and malware mitigation") or
 ALL("Security in hardware") or ALL("Systems security") or
 ALL("Network security") or ALL("Database and storage security")
 or ALL("Software and application security") or
 ALL("Human and societal aspects of security and privacy")
 or ALL("Biotech") or ALL("Flying Autonomous Vehicles") or
 ALL("Smart Dust") or ALL("Artificial General Intelligence") or
 ALL("4D Printing") or ALL("Knowledge Graphs") or
 ALL("Neuromorphic Hardware") or
 ALL("Blockchain for Data Security") or
 ALL("Exoskeleton") or ALL("Edge AI") or
 ALL("Autonomous Driving Level 5") or
 ALL("Conversational AI Platform") or
 ALL("Self-Healing System Technology") or ALL("Volumetric Displays")
 or ALL("5G") or ALL("Quantum Computing") or ALL("AI PaaS")
 or ALL("Deep Neural Network ASICs") or ALL("Smart Robots") or
```

```
 ALL("Autonomous Mobile Robots") or ALL("Brain-Computer Interface")
 or ALL("Smart Workspace") or ALL("Biochips") or ALL("Digital Twin")
 or ALL("Deep Neural Nets") or ALL("Deep Learning") or
 ALL("Carbon Nanotube") or ALL("IoT Platform") or
 ALL("Virtual Assistants") or ALL("Silicon Anode Batteries") or
 ALL("Blockchain") or ALL("Digital economy") or ALL("Internet Economy")
 or ALL("Web Economy") or ALL("New Economy") or ALL("Big Data")
) AND NOT ACCESSTYPE(OA))
```

3.3 Obtención de Artículos

En esta sección se presenta el análisis realizado de la cadena de búsqueda obtenida en la sección previa. Además, se detallan las tareas para la obtención de artículos de Scopus para obtención de artículos científicos.

A continuación, en la tabla 3.5 se definen todas las tareas del proceso de "Obtención de Artículos" incluido en la etapa de monitoreo y análisis, para obtener como resultado los artículos científicos de relevancia para el país.

Tabla 3.5 Tareas para el tercer proceso del flujograma del proceso autonómico en la etapa de monitoreo y análisis

Nombre de las tareas	Analizar la cadena de búsqueda, construir un script para extracción de artículos automáticamente y obtener los datos de los artículos mediante el API de Scopus
Descripción	Se realiza un análisis de la cadena de búsqueda obtenidas en el proceso anterior para desarrollar un script automatizado con la funcionalidad de extracción de información relevante de artículos científicos.
Fuente de datos	Scopus
Tipo de tarea de analítica de datos	Descubrimiento
Técnica de Analítica de Datos	No supervisado, descriptivo
Con que otro proceso se relaciona	Pre procesamiento de datos
Etapa del flujograma del proceso autonómico	Monitoreo y Análisis

Para obtener los artículos científicos se ha utilizado el API de Scopus, este provee algunas funcionalidades y restricciones explicadas en el Capítulo 1. Además, se ha desarrollado un

script en Python para extraer los datos de Scopus en el área de Ciencias de la Computación de manera automatizada. Se modificó el API de Scopus para poder obtener los "facets", la misma es una cabecera que devuelve Scopus en la cual se pueden definir varios parámetros como, por ejemplo: las tres principales palabras claves utilizadas en todos los artículos de la búsqueda, al igual que los países que más participaron en el desarrollo del artículo, entre otros.

La modificación del API de Scopus ha sido necesaria para conseguir todos los datos de los artículos científicos. Además, se ha desarrollado un script para automatizar búsquedas que permite obtener los datos más importantes de un artículo científico.

El código fuente de se encuentra en Data-Scopus/*.py.

Base de datos científica Scopus

La base de datos Scopus se utiliza en este proyecto como fuente principal de información de los artículos científicos que más le interesa al Ecuador para ser posteriormente analizados.

Se ha decidido utilizar las siguientes variables de Scopus para la extracción de la información de los artículos científicos:

- title: Título del documento científico
- abs: Resumen del documento científico
- coverDate: Fecha de publicación del documento científico
- doi: Identificador digital del documento científico
- identificadorArt: Electronic ID del artículo
- keywords: Palabras claves del artículo científico
- links: URL de Scopus del artículo
- pais: País de publicación del artículo
- revista: Revista en la cual se publicó el artículo
- revistaCod: ISSN o ISBN de la Revista o Congreso en la que se publicó el artículo

- tipoDoc: Tipo de documento publicado (artículo, review, artículo en prensa, etc.)

Para un mejor manejo de los datos se ha decidido almacenar la información recopilada de los artículos científicos en un archivo de Excel. Es así, que se ha creado un modelo de datos que se explica a continuación:

Diseño del modelo de datos

En la tabla 3.6 se especifica la Vista Minable Operativa (VMO) generada para el desarrollo del proyecto y de la base de datos utilizada para la carga de la información de los artículos científicos de relevancia para el Ecuador.

Tabla 3.6 Vista Minable Operativa

Variable	Descripción	Fuente
titulo	Titulo del artículo científico	Scopus
abs	Resumen del artículo científico	Scopus
keywords	Palabras claves del artículo científico	Scopus
identificadorArt	Identificador de Scopus del artículo científico	Scopus
doi	Identificador DOI del artículo científico	Scopus
links	URL de Scopus del artículo	Scopus
pais	País en el que se publicó el artículo científico	Scopus
tipoDoc	Tipo de documento del artículo científico	Scopus
revista	Revista en el que publicó el artículo científico	Scopus
revistaCod	Código ISSN en el que se publicó el artículo	Scopus
coverDate	Fecha de publicación del artículo	Scopus

Cabe indicar que se realizó la extracción de los datos más importantes de un artículo para tener una base de datos más completa. Adicionalmente, esta base de datos puede servir para otro tipo de investigaciones.

Para el tratamiento de datos se define lo siguiente:

- **Extracción**: Líneas de investigación ACM y Libro Blanco MINTEL para generar la búsqueda de artículos de relevancia mediante el API Scopus
- **Transformación**: Obtención de datos, limpieza y análisis de lenguaje natural
- **Carga**: Carga a tabla de hechos a un archivo csv

El proceso de extracción de la información de artículos científicos de relevancia para el Ecuador se especifica en la figura 3.5.

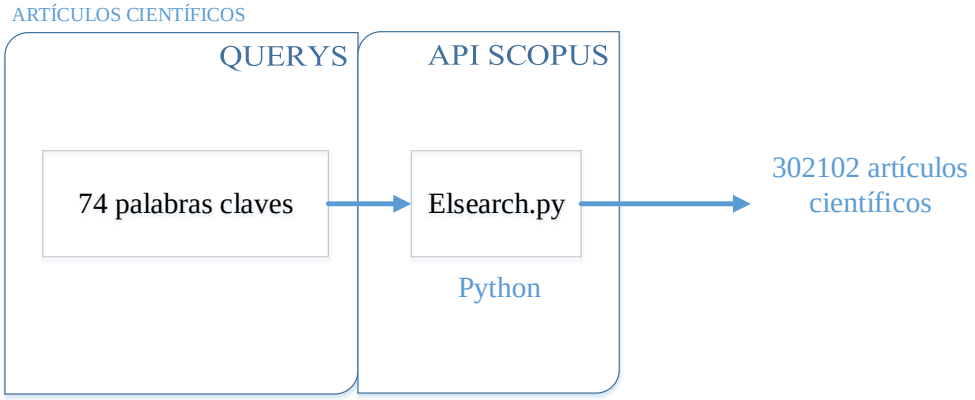

Fig. 3.5 Proceso de Obtención de artículos.

Con la cadena de búsqueda unificada construida en la sección anterior, se ha realizado la extracción de la información de los artículos científicos mediante el API de Scopus y Python de manera automática.

Los resultados de la consulta anterior se realizaron con corte de 2019/02/19 y estos fueron:

```
La cantidad total de artículos obtenidos para ese corte fueron de
302,102 resultados.
```

A pesar de que Scopus tiene restricciones de datos y no permite la descarga de la información de más de 5000 artículos con el API por cada consulta, se ha optimizado esta consulta dividiéndola por años, es decir, se tuvo varias consultas, además, se ha realizado una segunda división por cada palabra clave, al estar unidas simplemente con un OR estas cadenas se pueden separar para obtener consultas más pequeñas. Adicionalmente, se ha realizado la consulta mediante hilos de procesamiento lo que ha permitido obtener el 100% de la información.

Como se detalla en la tabla 3.7, los artículos científicos se han concentrado en su mayoría en los últimos 15 años. Lo que es lógico ya que la extracción de información se realiza para obtener líneas de investigación específicas de relevancia para el país.

Tabla 3.7 Extracción de artículos científicos por año

Año	**Nro. Artículos**
2019	9563
2018	35572
2017	30085
2016	26643
2015	24462
2014	25061
2013	21405
2012	20633
2011	17881
2010	15315
2009	14106
2008	11733
otros	67643
total	302102

Con lo explicado anteriormente la consulta a Scopus se ha realizado correctamente en un tiempo transcurrido para la extracción de la información de: 20 horas 14 minutos aproximadamente.

Este tiempo depende exclusivamente de la plataforma de Scopus ya que como lo han definido se realiza una consulta de 25 documentos por cada 6 segundos lo que equivale a un total de 1209 minutos que son más de 20 horas de consulta. El tiempo se disminuyó ya que se realizó procesamiento paralelo para las consultas.

Se han extraído todas las variables detalladas anteriormente: abs, coverDate, doi, identificadorArt, keywords, links, pais, revista, revistaCod, tipoDoc y titulo

De los más de 300 mil datos recolectados, las columnas más importantes son las del título de la publicación, el resumen y las palabras claves. Pero se decidió extraer más datos para tener una mayor cantidad de indicadores que pueden ser útiles para otros trabajos de investigación.

3.4 Pre procesamiento de datos

En esta sección se presenta el análisis realizado a los artículos científicos obtenidos en la sección anterior. Además, se especifican las tareas realizadas para la limpieza y pre procesamiento de datos. El pre procesamiento de datos es muy importante para eliminar registros impuros que pueden conducir a la extracción de patrones o reglas poco útiles.

A continuación, en la tabla 3.8 se definen todas las tareas del proceso de "Pre procesamiento de datos" incluido en la etapa de Planificación, para obtener como resultado datos limpios y listos para ser procesados por técnicas de analítica de texto.

Tabla 3.8 Tareas para el cuarto proceso del flujograma del proceso autonómico en la etapa de Planificación

Nombre de las tareas	Analizar los artículos científicos, realizar la limpieza de datos y realizar un análisis de lenguaje natural de los mismo para obtener datos con calidad
Descripción	Se realiza un análisis de los artículos científicos obtenidos en el proceso anterior para desarrollar un script en Python que permita la limpieza y el análisis de lenguaje natural de los datos de artículos científicos.
Fuente de datos	Base de datos artículos científicos obtenidos de Scopus
Tipo de tarea de analítica de datos	Pre procesamiento, limpieza y análisis de lenguaje natural
Técnica de Analítica de Datos	Lenguaje Natural
Con que otro proceso se relaciona	Procesamiento de datos (Minería de texto)
Etapa del flujograma del proceso autonómico	Planificación

Este proceso al ser más complejo que los anteriores, se ha desarrollado el macro algoritmo del mismo, este especifica los algoritmos utilizados para la obtención de los resultados, en este caso los datos limpios y de calidad. En la tabla 3.9 se especifica el desarrollo de los algoritmos utilizados para este proceso.

Tabla 3.9 Macro algoritmo para el cuarto proceso autonómico en Etapa de planificación

Algoritmo	**Tipo de analítica de datos**	**Herramientas**
Limpieza de datos de las publicaciones científicas	Minería de texto	Eliminación de datos incompletos o con errores
Obtención de los datos con calidad para ser procesados	análisis de lenguaje natural	Minería de datos Pyhton, minería de texto, herramientas de análisis de lenguaje natural como Stop Words, Lematizer, Steamizer, entre otros

Para la preparación y tratamiento de los datos se ha realizado una limpieza correcta de la información (como lo especifica MIDANO). Además, se ha eliminado datos vacíos de texto y caracteres con ruido. Posteriormente se realizó un análisis de lenguaje natural para obtener datos de calidad, como por ejemplo, las palabras conectoras como "the" o "and" que solo son conectores y no ayudan al procesamiento de los datos.

Tal como se especifica en la figura 3.6, el pre procesamiento es necesario para obtener datos con calidad para ser utilizados con técnicas de minería de texto.

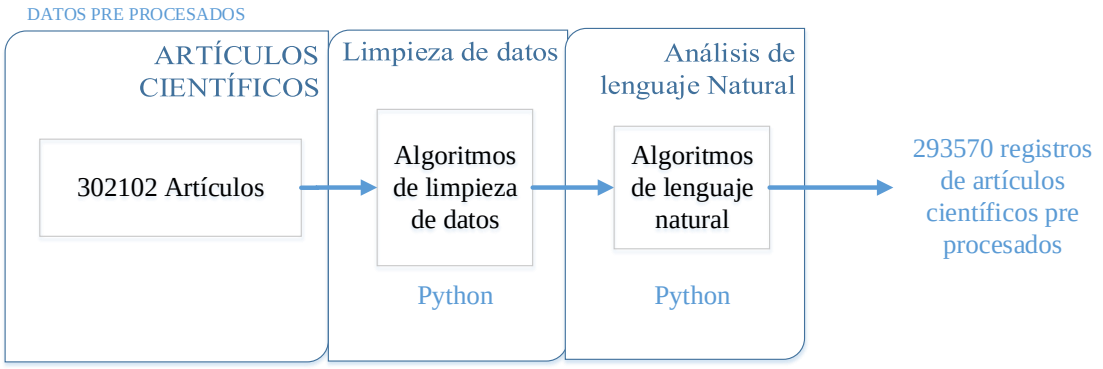

Fig. 3.6 Proceso de Pre procesamiento de datos.

Una de las principales formas de preprocesamiento es filtrar datos inútiles y datos con error o datos incompletos. Además, se ha realizado un procesamiento del lenguaje natural, empezando por la eliminación de palabras inútiles o StopWords como por ejemplo artículos o conjunciones. Se denominan palabras inútiles debido a que no aportan información relevante para generar conocimiento, por ejemplo las imprecisiones y las ambigüedades.

En la figura 3.7 se especifica el flujo de los algoritmos para la obtención de datos limpios y con calidad, además, más adelante se detallan los algoritmos utilizados para el análisis de lenguaje natural:

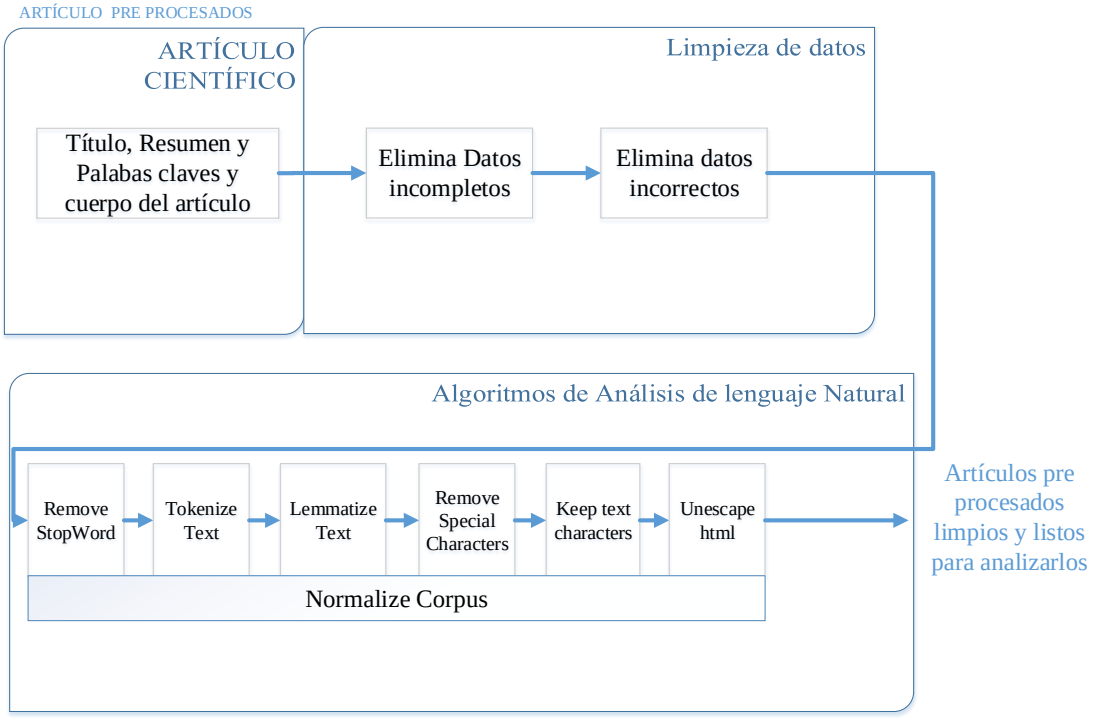

Fig. 3.7 Algoritmos del Pre procesamiento de datos.

El algoritmo de preprocesamiento definido en la figura 3.7 comienza desde los datos obtenidos desde Scopus (título, resumen, palabras claves, cuerpo del artículo), de los cuales se eliminan los datos incompletos, mediante la librería Pandas de Python. Posteriormente, se eliminan los datos incorrectos, haciendo uso de la misma librería. Datos incompletos son aquellos casos en los que no se pudo extraer la información de algún metadato del artículo como por ejemplo el resumen ya sea por las restricciones o por errores del API. Datos incorrectos son datos que no aportan nada a la analítica de datos, por ejemplo que existan muchos caracteres especiales de los cuales no se pueda extraer ningún concepto.

Por ejemplo, si se inicia con dos artículos:
Artículo. 1. "A machine learning approach to predict perceptual decisions: an insight into face http://20..." **Artículo. 2**: Contiene datos incompletos y erróneos.

De estos datos se eliminaría el artículo 2 del proceso de limpieza. Seguidamente, se ejecutan los algoritmos de análisis de lenguaje natural y obtener datos adecuados para ser procesados.

En la tabla 3.10 a continuación, se puede observar el resultado de la ejecución de cada algoritmo de análisis de lenguaje natural para el caso de ejemplo.

Tabla 3.10 Resultado de la ejecución de cada algoritmo de análisis de lenguaje natural para un ejemplo

	Algoritmo	Resultado
1	Remove StopWords	machine learning approach perceptual decisions face http://20...
2	Tokenize Text	[machine, learning, perceptual, decisions, face, http://20...]".
3	Lemmatize Text	[machine, learn, perceptual, decisions, face, http://20...]
4	Remove special characters	[machine, learn, perceptual, decisions, face, http20...]
5	Keep text characters	[machine, learn, perceptual, decisions, face, http...]
6	Unescape html	[machine, learn, perceptual, decisions, face...]

Continuando con el mismo ejemplo, luego de ser preprocesado el artículo quedaría:

Artículo. 1. "[machine learn perceptual decision face pareidolia...]".

A continuación, se detalla cada uno de los algoritmos utilizados para el análisis de lenguaje natural:

1. Remove Stopwords: Es un algoritmo que filtra datos inútiles y los elimina.
Este algoritmo se especifica en el programa y se extendió la funcionalidad a ciertas palabras inútiles utilizadas generalmente en el resumen del artículo como "mr", "mrs", "also", "ask", "make", entre otras. (Perkins, 2014)

2. Tokenize Text: Es un algoritmo que realiza un escaneo léxico y devuelve las palabras como tokens para podernos analizarlos por separado, lo que hace útil para usarlos en la obtención de las líneas de investigación. (Perkins, 2014)
Este algoritmo se especifica en el programa, a continuación, se presenta la función.

3. Lemmatize Text: Es un algoritmo que realiza el proceso de agrupar las diferentes formas de inflexión de una palabra para que puedan analizarse como un solo elemento. Este

proceso se realiza con texto en inglés con mayor eficiencia. (Perkins, 2014)
Por ejemplo, la palabra "going" se reduce a "go"

4. Remove special characters: Es un algoritmo que elimina caracteres innecesarios y especiales como la puntuación. (Perkins, 2014)

5. Keep text characters: Es un algoritmo que conserva los textos útiles para el análisis de datos. (Perkins, 2014)

6. Unescape html: Es un algoritmo que elimina las etiquetas HTML ya que estas no aportan nada a la analítica de datos. (Perkins, 2014)

7. Normalize corpus: Todas las funciones anteriores se implementaron en esta función para la normalización del corpus para poder manejar el texto correctamente. (Perkins, 2014)

Además, a las funciones mencionadas se creó una función que permite convertir el texto en formato ASCII ya que existen algunos caracteres de difícil interpretación. Esta función se definió como "parse document".

Todos estos algoritmos fueron necesarios para tener los artículos científicos de relevancia para el Ecuador con datos limpios y de calidad, listos para ser analizados mediante técnicas de analítica de datos. En la siguiente sección se desarrolla el análisis de los datos para obtener las líneas de investigación específicas para el Ecuador.

3.5 Procesamiento de datos - Minería de texto

En esta sección se presenta el análisis realizado a los datos de los artículos científicos pre procesados obtenidos en la sección anterior. Además, se detallan las tareas realizadas para la obtención de las líneas de investigación específicas para el Ecuador.

A continuación, en la tabla 3.11 se definen todas las tareas del proceso de "Minería de datos" incluido en la etapa de Ejecución, para obtener como resultado las líneas de investigación específicas de relevancia para el Ecuador.

Tabla 3.11 Tareas para el quinto proceso del flujograma del proceso autonómico en la etapa de Planificación

Nombre de las tareas	Analizar los datos de los artículos científicos mediante la clusterización y obtener las líneas de investigación específicas para el Ecuador
Descripción	Se realiza un análisis de los datos de los artículos científicos pre procesados obtenidos en el proceso anterior para desarrollar un modelo de agrupamiento que permita la obtención de líneas de investigación específicas de relevancia para el Ecuador en el ámbito de las ciencias de la computación.
Fuente de datos	Base de datos Artículos Científicos pre procesados
Tipo de tarea de analítica de datos	Minería de texto y análisis de lenguaje natural
Técnica de Analítica de Datos	Agrupamiento/Clustering
Con que otro proceso se relaciona	N/A
Etapa del flujograma del proceso autonómico	Ejecución

La minería de texto al ser un proceso más complejo que los anteriores, se ha desarrollado el macro algoritmo del mismo, este especifica los algoritmos utilizados para la obtención de los resultados. En la tabla 3.12 se especifica el desarrollo de los algoritmos utilizados para este proceso.

Tabla 3.12 Macro algoritmo para el quinto proceso autonómico en Etapa de Ejecución

Algoritmo	**Tipo de analítica de datos**	**Herramientas**
Obtener las líneas de investigación específicas para el Ecuador	Minería de texto	Clustering con Affinity Propagation en Python

Para desarrollar esta tarea se ilustra el proceso mediante la figura 3.8. En la cual se puede observar que se obtiene las líneas de investigación de cada artículo. Además, esta figura puede ser representada como la arquitectura madre del proyecto de analítica de datos para obtener las líneas de investigación específicas mediante la minería de texto.

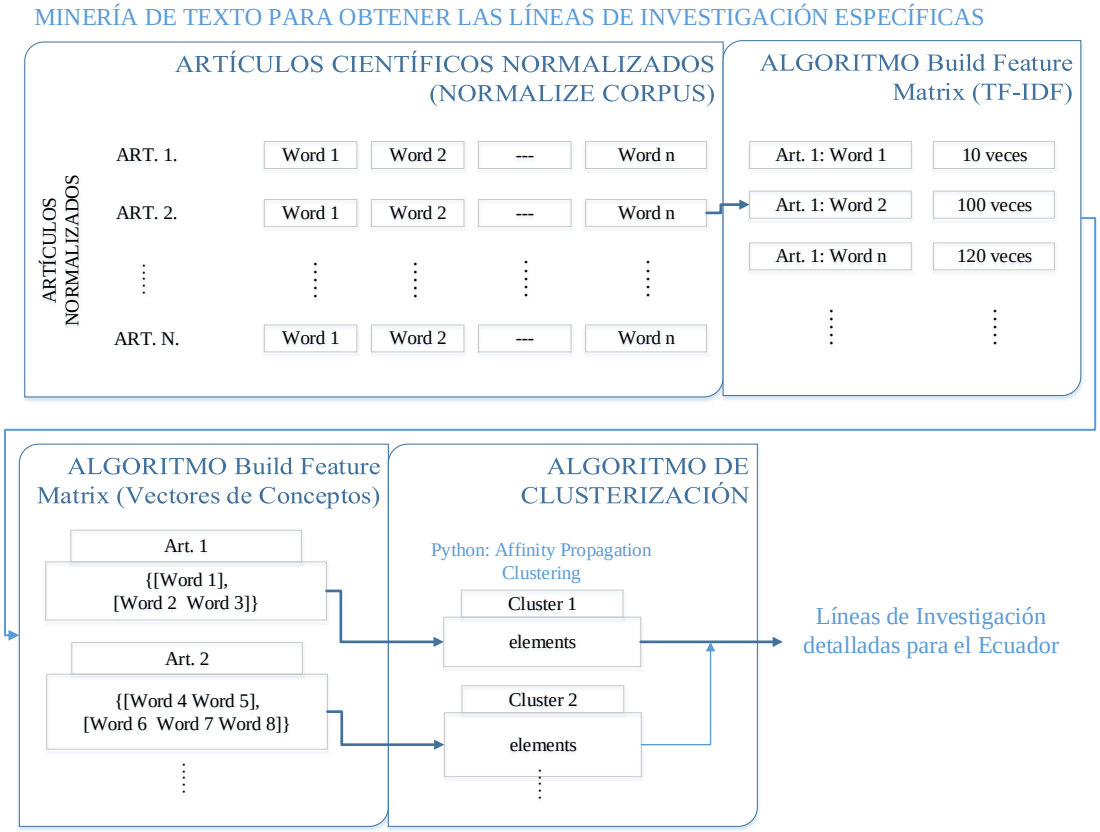

Fig. 3.8 Flujo del Procesamiento de datos (minería de texto).

Como se puede observar en la figura 3.8, para la obtención de las líneas de investigación específicas se aplican varios algoritmos. De los resultados del preprocesamiento, mediante la normalización del corpus, se construye una matriz de características, la cual despliega la relación entre los artículos normalizados y las palabras claves de estos. Esta matriz genera los ejemplares para la construcción de las líneas de investigación.

Una vez obtenida la matriz de características, se realiza el agrupamiento de las palabras claves más utilizadas lo que da como resultado las líneas de investigación específicas para el Ecuador. Todos estos algoritmos se explican en detalle más adelante.

Para obtener las líneas de investigación específicas para el Ecuador, se ha generado una aplicación web en Jupyter de Pyhton, la cual es una herramienta web de código abierto que permite compartir código en vivo. Los usos de esta herramienta son muy variados y se ajustan perfecto para el rendimiento del programa desarrollado para obtener las líneas de

investigación específicas para el Ecuador

A continuación, se presenta un detalle de las herramientas y técnicas de analítica de datos utilizadas para cumplir el objetivo que fue el de obtener las líneas de investigación específicas para el Ecuador.

1. Build feature matrix, luego con los datos limpios se realiza la extracción de las características tf-idf (frecuencia de término – frecuencia inversa de documento) es decir, la frecuencia de ocurrencia del término en los datos de los artículos científicos. Este algoritmo nos ayuda a determinar el vector que contiene todas las palabras con más ocurrencias dentro del abstract, palabras claves y los títulos de los documentos. Además, este provee una matriz de características que se utiliza para el modelo de entrenamiento.

La matriz de características contiene puntos de datos de los artículos científicos normalizados representados como vectores en el plano de longitud fija. Esta matriz también es conocida como document-term matrix, y se utiliza para representar vectores de palabras. En la figura 3.9 se ilustra un ejemplo de vectores de las palabras del bag of words en el espacio de características.

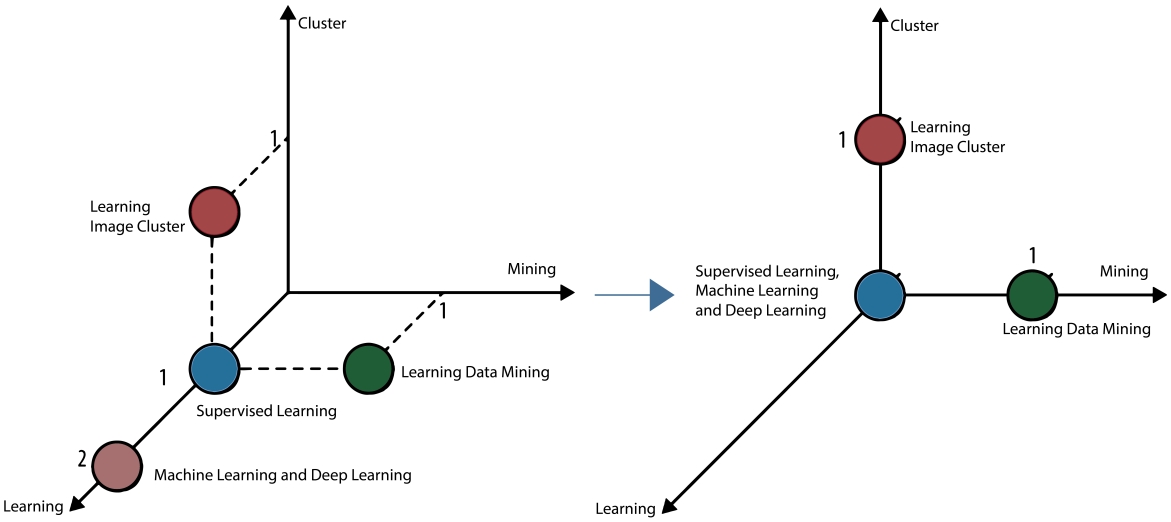

Fig. 3.9 Ejemplo de vectores en el espacio de características.

En la figura 3.9 se pueden observar que cada circulo representa vectores de palabras de cada artículo, cada número representa la cantidad de ocurrencias de la palabra en el documento. Por ejemplo, en la ilustración se puede observar 4 vectores de palabras, el "Learning Data mining", se representa en las coordenadas (1,0,1), ya que este se relaciona con los vectores que contienen la palabra "Learning" y con "Mining". Así se construye las

relaciones existentes entre vectores de palabras dentro del documento científico.

En la tabla 3.13 se muestra un pequeño ejemplo de una matriz de términos de documentos científicos, esta matriz se crea simplemente con la comparación de los vectores del bag of words en el espacio de características, y las columnas representan todas las posibles palabras en el bag of words.

Tabla 3.13 Ejemplo de matriz de términos de documentos científicos

	machine	learning	and	Deep	Supervised	Image	Clúster	Data	Mining	Analitycs
Machine learning and Deep Learning	1	2	1	1	0	0	0	0	0	0
Supervised Learning	0	1	0	0	1	0	0	0	0	0
Learning Image Clúster	0	1	0	0	0	1	1	0	0	0
Learning Data Mining	0	1	0	0	0	0	0	1	1	0
Data Analitycs	0	0	0	0	0	0	0	1	0	1

Como se puede observar la tabla 3.13 es una matriz dispersa, y está estrechamente relacionada con los métodos de filtrado basados en frecuencia de las palabras, es decir este algoritmo hace que las palabras raras sean más prominentes e ignore las palabras comunes, por lo que se crean las líneas de investigación de cada artículo.

3. Algoritmo De Clusterización

Luego de obtener los vectores de palabras más representativos (ejemplares) de cada artículo analizado con el algoritmo anterior, se realiza la clusterización de las mismas, obteniendo como resultado final las líneas específicas de investigación de las cuales se obtienen también los elementos que conforman el clúster.

Como describen Guan et al. (2011) y Vasantha y Majji (2013), la mejor técnica para agrupar textos es affinity propagation, además Guan realiza una comparativa de varias técnicas una de ellas y una de las más conocidas es Kmeans. Demostrando mediante varios experi-

mentos de documentos de texto que el Algoritmo affinity propagation es el más eficiente para realizar clústers en textos. En estos experimentos se midieron tanto la rapidez, rendimiento, convergencia, entropía y tiempo de ejecución.

Guan et al. (2011) y Vasantha y Majji (2013) en sus trabajos científicos explican que en comparación con el algoritmo clásico de agrupación en clúster k-means, Affinity Propagation (AP) no solo reduce la complejidad computacional de la agrupación de textos y mejora la precisión, sino que también evita la inicialización aleatoria y la captura en el mínimo local. AP también es más robusto y menos sensible a la distribución de datos que k-means. En otras palabras, hace una mejora importante en las tareas de agrupamiento de texto.

En base a los resultados experimentales de los artículos de Guan et al. (2011) y Vasantha y Majji (2013) en donde se especifica la efectividad del algoritmo Affinity Propagation para realizar agrupamiento en texto o lenguaje natural, este algoritmo ha sido utilizado en la presente tesis para obtención de mejores resultados, AP es basado en el concepto de paso de mensajes, por lo que se adaptó correctamente a la necesidad de creación de líneas de investigación específicas para el Ecuador.

Una de las ventajas del algoritmo AP es que no requiere que el número de clústers sea determinado o estimado antes de ejecutar el algoritmo. Este algoritmo encuentra similitudes en las líneas de investigación y las agrupa en clústers que tiene elementos. Además, es un algoritmo de agrupamiento en clústeres, ejemplos y valores atípicos, de bajo error, alta velocidad, flexible y fácil de codificar.

Affinity Propagation es un algoritmo de agrupación en clúster que identifica un conjunto de "ejemplares" que representa el conjunto de datos. La entrada del Affinity Propagation es la similitud de pares entre cada par de puntos de datos, s [i, j] (i, j = 1, 2, ..., N), es decir se basa en la matriz de características descrita anteriormente. El proceso de Affinity Propagation se puede ver como un proceso de paso de mensajes con dos tipos de mensajes intercambiados entre los puntos de datos: responsabilidad y disponibilidad. La responsabilidad, r [i, j], es un mensaje del punto de datos i a j que refleja la evidencia acumulada de qué tan adecuado es el punto de datos j para servir como ejemplo para el punto de datos i. La disponibilidad, un [i, j], es un mensaje del punto de datos j a i que refleja la evidencia acumulada de cuán apropiado sería para el punto de datos i elegir el punto de datos j como ejemplo. Es así como realiza la clusterización en este caso de las líneas de investigación. (Vasantha y Majji, 2013)

Luego de obtener los ejemplares de cada artículo científico, se realiza la agrupación de estos mediante el algoritmo Affinity Propagation, para finalmente, obtener los clústeres que representan las líneas de investigación de importancia para el país. El algoritmo genera la agrupación de los ejemplares mediante la afinidad o cercanía que existe entre ellos, y el centroide o centro del grupo se establece como el nombre del clúster. Por ejemplo, para un artículo "The so-called ... of manufacturing industries has been conceived as the fourth industrial revolution or Industry 4.0, a paradigm shift propelled by the upsurge and progressive maturity of new Information and Communication Technologies (ICT) applied to industrial processes and products...", su ejemplar sería: "Data fusion machine learning industrial prognosis Trends perspectives towards Industry" y su clúster: "Industrial processes using machine learning".

3.6 Resultados

A continuación, se detalla los resultados obtenidos de la minería de texto. En esta etapa de MIDANO se presentan los resultados obtenidos del Desarrollo de técnicas de analítica de datos para la obtención de líneas de investigación específicas para el Ecuador.

Del proceso antes detallado se obtuvieron 347 clústeres, de los cuales se presentan algunos en la figura 3.10 en la que se puede obvervar las agrupaciones de los artículos y sus elementos. Estas agrupaciones son las líneas de investigación específicas para el Ecuador.

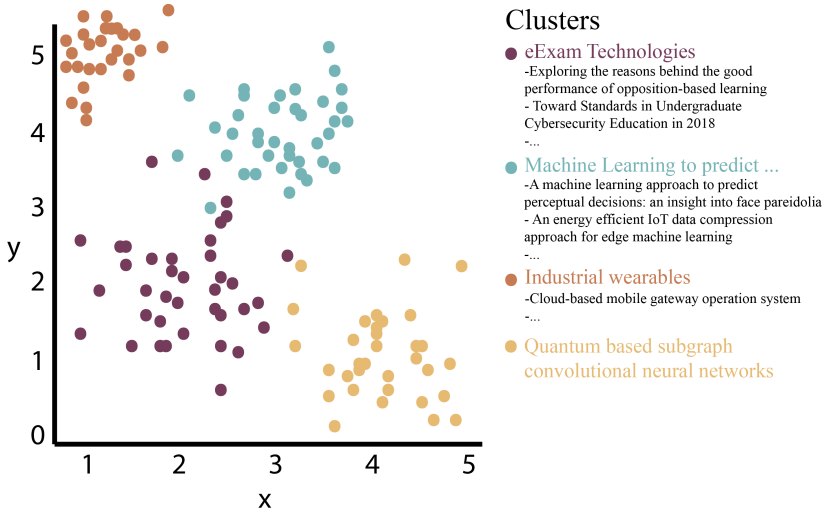

Fig. 3.10 Ejemplo de clústeres de las líneas de investigación específicas.

Luego de realizar el análisis a los documentos científicos se ha conseguido un total de 347 líneas de investigación específicas. La tabla 3.14 presenta algunos de los resultados.

En el Anexo A están detalladas las 347 líneas de investigación específicas en el ámbito de las ciencias de la computación para el Ecuador obtenidas como resultado de la tesis.

Tabla 3.14 Resultados de la aplicación de AD para obtención de líneas de investigación específicas

Artículo Original	**Resultados obtenidos**
City-view image location identification by multiple geo-social media and graph-based image cluster refinement... Recently, landmark image identification has shown great promise for the addressed problems, where most previous approaches are either visual-based or location-based. However, regarding city-view image location identification, there could be a number of buildings in a close proximity. Moreover, it is common that photos were taken indoors. The conditions may degrade the performance of previous approaches. To remedy the deficiencies, this paper unifies visual features... image location identification. Besides, this paper shows an effective and memory-efficient implementation based on sparse coding, where a new dictionary selection approach is presented. Further, this paper proposes a location-aware graph-based regrouping approach, leveraging spanning graph construction, on clusters of photos to refine clustering results. Experimental results show the improvement over the baselines (location-based, visual-based, etc.)....	Categorize scene images in multi views

Continúa en la siguiente página

Tabla 3.14 – *Continúa en la página previa*

Artículo Original	**Resultados obtenidos**
Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services... Facial expression recognition is an active research area for which the research community has presented a number of approaches due to its diverse applicability in different real-world situations such as real-time suspicious activity recognition for smart security, monitoring, marketing, and group sentiment analysis. However, developing a robust application with high accuracy is still a challenging task mainly due to the inherent problems related to human emotions, lack of sufficient data, and computational complexity. In this paper, we propose a novel, cost-effective, and energy-efficient framework designed for suspicious activity recognition based on facial expression analysis for smart security in law-enforcement services. The Raspberry Pi camera captures the video stream and detects faces using the Viola Jones algorithm. The face region is pre-processed using Gabor filter and median filter prior to feature extraction. Oriented FAST and Rotated BRIEF (ORB) features are then extracted and the support vector machine (SVM) classifier is trained, which predicts the known emotions (Angry, disgust, fear, happy, neutral, sad, and surprise). Based on the collective emotions of the faces, we predict the sentiment behind the scene. Using this approach, we predict if a certain situation is hostile and can prevent it prior to its occurrence. The system is tested on three publically available datasets: Cohen Kande (CK+), MMI, and JAFEE. A detailed comparative analysis based on SURF, SIFT, and ORB is also presented. Experimental results verify the efficiency and effectiveness of the proposed system in accurate recognition of suspicious activity compared to state-of-the-art methods and validate its superiority for enhancing security in law enforcement services...	Facial expression recognition

Continúa en la siguiente página

Tabla 3.14 – *Continúa en la página previa*

Artículo Original	**Resultados obtenidos**
Hard negative generation for identity-disentangled facial expression recognition... Various factors such as identity-specific attributes, pose, illumination and expression affect the appearance of face images. Disentangling the identity-specific factors is potentially beneficial for facial expression recognition (FER). Existing image-based FER systems either use hand-crafted or learned features to represent a single face image. In this paper, we propose a novel FER framework, named identity-disentangled facial expression recognition machine (IDFERM), in which we untangle the identity from a query sample by exploiting its difference from its references (e.g., its mined or generated frontal and neutral normalized faces). We demonstrate a possible recognition via generation scheme which consists of a novel hard negative generation (HNG) network and a generalized radial metric learning (RML) network. For FER, generated normalized faces are used as hard negative samples for metric learning. The difficulty of threshold validation and anchor selection are alleviated in RML and its distance comparisons are fewer than those of traditional deep metric learning methods. The expression representations of RML achieve superior performance on the CK +, MMI and Oulu-CASIA datasets, given a single query image for testing...	Facial expression recognition

Continúa en la siguiente página

Tabla 3.14 – *Continúa en la página previa*

Artículo Original	**Resultados obtenidos**
Design and evaluation of 3D CAPTCHAs... Most current 2D CAPTCHAs are vulnerable to automated character recognition attacks and the latest attacks can successfully break the 2D text CAPTCHAs at a rate of more than 90%. In this work, we present two novel 3D CAPTCHAs, which are more secure than current 2D text CAPTCHAs against automated character recognition attacks. Our approach is to display CAPTCHA characters on 3D objects. We exploit the difficulty that machines have in rotating 3D objects to find the correct viewpoint and in further recognizing characters in 3D, while we believe humans can easily perform these tasks. Using an offline automated character recognition attack, we find that 82% of new text reCAPTCHAs are broken, while approximately 60% of our 3D CAPTCHAs are broken and only if characters are focused on and zoomed in from a direct viewpoint. When CAPTCHAs are presented in slightly different views, the attack success rate is rapidly diminished to 0%. In addition, we use commercial Deep Neural Networks-based text and object detection classifiers to attack our systems, and demonstrate that our approach is extremely difficult to break with these classifiers, even if CAPTCHA characters are presented in direct, 2D view. With emulated relay attacks, fewer than 16% of our CAPTCHAs are accurately solved by human solvers, while more than 90% of current 2D text-based CAPTCHAs are solved. Also, we performed an IRB-approved user study to evaluate the usability of our approach. Participants agreed that our approach was usable in spite of the extra time required for 3D model rotation...	Design 3D CAPTCHAs

Continúa en la siguiente página

Tabla 3.14 – *Continúa en la página previa*

Artículo Original	**Resultados obtenidos**
Exploring the reasons behind the good performance of opposition-based learning... Cognitive research suggests that understanding the semantics, or the meaning, of representations involves both ascension from concrete concepts denoting specific observations (that is, extension) to abstract concepts denoting a number of observations (that is, intension), and vice versa. Consonantly, extant conceptual schemas can encode the semantics of a domain intensionally (e.g., ER diagram, UML class diagram) or extensionally (e.g., set diagram, UML object diagram). However, prior IS research has exclusively focused on intensional representations and the role they play in aiding domain understanding. In this research, we compare the interpretational fidelity of two types of representational encoding of cardinality constraints, an intensional schema using an ER diagram and its extensional analog using a set diagram. We employ cognitive science research to conceptualize that extensional representations will enable enhanced understanding as compared with intensional representations. Further, given that prior research suggests that the semantics of cardinality constraints remain challenging to understand, we focus on mandatory and optional cardinality constraints associated with relationships in these representations. Based on our laboratory experiments, we find that understanding with an extensional representation was (1) at least as good as that with an intensional representation for mandatory cardinality constraints and (2) significantly better for optional cardinality constraints. We also conducted an applicability check of our results via focus groups and found support for the perceived significance of extensional representations in practice. Overall, this research suggests that the tradition in IS research of exclusively focusing on intensional encoding of domain semantics should be reexamined...	eExam technologies

A continuación se realiza el análisis de los resultados obtenidos y como esto puede influir en el desarrollo del país.

3.7 Discusión de Resultados

Antes de discutir los resultados, se debe tener en cuenta que aún después de analizar a fondo toda la información obtenida de Scopus, estos deben ser contextualizados e interpretados correctamente. Los resultados están directamente vinculados con las publicaciones de Scopus en el área de investigación de Computer Science. Para generar las líneas de investigación este proyecto se ha centrado únicamente en el área de las ciencias de la computación, pero el modelo propuesto y la metodología pueden ser utilizadas como punto de partida para modelar procesos de búsqueda de información en otros campos, y así generar las líneas de investigación específicas para el Ecuador en todos sus ámbitos.

Adicionalmente, el estudio realizado y los modelos son basados en una metodología bien estructurada por lo que se obtuvieron resultados completos y detallados. Además, se ha interpretado la información recopilada, mediante la comprensión personal de los conceptos que se detallan en el modelo, para tener resultados coherentes. Por lo tanto, existe un riesgo limitado de incluir los conocimientos y creencias del autor.

Para generar el modelo resultante se utilizaron únicamente artículos científicos de revistas, ya que son los que más impacto tienen a nivel mundial, pero pueden existir más conceptos relacionados con la investigación en diferentes contextos, como por ejemplo, el documento puede ser de tipo artículos de conferencia, documentos de tesis de maestría, tesis de doctorado, compendio de documentos, libros de texto, informes, páginas web, publicación de foros, posters, entre otros. Todos estos tipos de documentos puede agregarse al modelo mediante una nueva relación de generalización sin afectar la consistencia y validez.

Además, incluso si el estudio realizado ha sido muy fructífero, el tamaño de la muestra siendo únicamente de una base de datos (Scopus) utilizada para obtener el modelo, pueden ser una amenaza para la validez de los resultados, aunque se haya logrado la saturación teórica. Sin embargo, los estudios del modelo realizado son un excelente punto de partida para un estudio de mayores dimensiones.

También se consideró si sería necesario tener un modelo más detallado como por ejemplo incluyendo todos los tipos de documentos posibles u otra alternativa. Pero los resultados obtenidos muestran una coherencia de las líneas de investigación específicas para el Ecuador.

Utilizando los artículos científicos de Scopus alineados con el MINTEL, se utilizó un

enfoque para elaborar la teoría, en lugar de intentar corroborar una hipótesis inicial. También se debe tener en cuenta que los modelos y las líneas de investigación resultantes dependerán de la época en la que se realice la captura de datos, además de la situación actual del país y de la experiencia de los investigadores. Así es que se puede prever que algunos de los conceptos incluidos en las líneas de investigación o los resultados no serán de relevancia en el futuro, pero de acuerdo con el criterio de inclusión del ciclo de gartner se puede determinar que las líneas de investigación serán de importancia en los próximos años.

Con respecto a la adaptación de la visualización e interacción de la información, la solución propuesta es una prueba de concepto que propone una nueva forma de calcular o predecir las líneas de investigación de relevancia para el Ecuador y para cualquier país, dependiendo de su plan de gobierno y su plan de desarrollo. Esta solución es altamente adaptable a cualquier país según las reglas definidas.

Capítulo 4

Validación de Resultados

En este capítulo, se realiza la validación de los resultados obtenidos de las líneas de investigación específicas propuestas. Para realizar esta validación primero se crea un modelo predictivo mediante una máquina de vectores de soporte (SVM) y utilizando como datos para este modelo las líneas de investigación obtenidas en el capítulo anterior. Posteriormente, se obtiene artículos científicos de la base de datos Scielo, y, por último, se realiza los resultados de la validación. Este proceso se muestra en la figura 4.1.

Esta validación se realiza para comprobar la efectividad del modelo y de las líneas de investigación específicas propuestas.

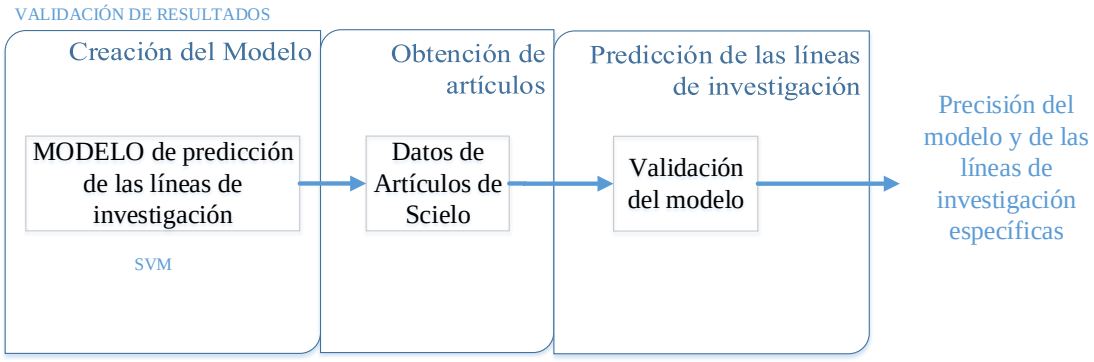

Fig. 4.1 Flujo de Proceso para la validación de resultados.

Se ha seleccionado el desarrollo del modelo mediante algoritmos SVM ya que este es de rápida implementación y es el mejor clasificador de documentos de texto en según Kabir et al. (2015). Como detallan los resultados de Kabir et al. (2015), su método propuesto

proporciona una mayor precisión en comparación con los métodos basados el clasificador Bayesian (NB) y otros clasificadores.

4.1 Creación del Modelo de Predicción de Líneas de Investigación

Para generar un modelo que permita la predicción de las líneas de investigación se ha utilizado el algoritmo Support Vector Machine (SVM). Este algoritmo trata de buscar la clasificación de títulos de los artículos en líneas de investigación definidas en el capítulo anterior.

El proceso para la creación del modelo que permita la predicción de las líneas de investigación es el siguiente:

1. Selección del tipo de clasificador SVM.
2. Definición de los sets de datos de prueba y entrenamiento.
3. Entrenamiento del modelo.
4. Validación del modelo.
5. Pruebas del modelo.

Para la predicción de las líneas de investigación se ha decidido definir el modelo con el clasificador lineal, y una implementación en Python que se ajusta a la Support Vector Machine denominada SGDClassifier. Además, este tipo de clasificador se adapta mucho mejor a documentos de texto según Guan et al. (2011), quien mediante varios experimentos demostró que este algoritmo reduce el tiempo de ejecución y tiene una entropía más baja. Esta funciona con datos representados como matrices densas o dispersas para la obtención de características que es lo realizado en la creación de las líneas de investigación específicas.

Posteriormente a la selección del clasificador, se procedió a establecer los parámetros definidos por el algoritmo SVM en Python para el entrenamiento del modelo:

- **Loss:** define la función utilizada para el modelo, en este caso se definió "hinge" que da un modelo SVM lineal
- **penalty:** es el término de regularización del modelo. En este caso se seleccionó "12", que es la regularización estándar para modelos lineales.
- **n-iter:** define el número de pasadas al set de entrenamiento, es conocido también como épocas. Para este modelo se seleccionó "12" iteraciones.
- **random-state:** define a la semilla del generador de números pseudoaleatorios para ser usados cuando se mezclan los datos en el modelo. En este caso se seleccionó "42" semillas.

Para la selección del mejor conjunto de entrenamiento y pruebas, se ha realizado los experimentos mostrados en la tabla 4.1. Se entrenó el modelo con los porcentajes de datos de la tabla y se seleccionó el que mejor resultado de precisión obtenía.

Tabla 4.1 Experimentos para elegir el set de entrenamiento y pruebas

Set de Entrenamiento	Set de Pruebas	Precisión del modelo
57%	43%	0.7001515465113469
67%	33%	0.8123548948534333
70%	30%	**0.8565721272235711**
85%	15%	0.7445886865886499

Como se puede observar en la tabla 4.1 el set de pruebas y entrenamiento que mejor resultados obtuvo fue el modelo entrenado con los valores de 70% para entrenamiento y el 30% para pruebas de todo el conjunto de datos. Estos valores de precisión del modelo se obtuvieron realizando la ejecución de este con el set de pruebas definidos en cada experimento.

Tomando en consideración que la precisión de la SVM es mayor a un 85% se puede considerar un modelo altamente confiable, por lo tanto, se ha decidido probarlo con fuentes de datos externas como son los artículos científicos indexados en Scielo. En la siguiente sección se especifica las fuentes de datos y la información obtenida de Scielo.

4.2 Obtención de artículos de Scielo

Scielo es una plataforma creada por Brasil para publicar artículos científicos a nivel regional. Esta plataforma no tiene un API como Scopus, por lo que se realizó la extracción de la información de manera manual, mediante la consulta a la plataforma web [1].

Esta consulta se realizó mediante los siguientes parámetros:

- Artículos en inglés: debido a que el modelo se realizó únicamente con artículos en inglés.
- Filtros de búsqueda: Año de publicación mayores a 2018, ya que estos artículos son los últimos generados en Scielo y de los 700 artículos indexados en la plataforma, es una muestra representativa de la totalidad de artículos presentes en esta base de datos.
- Área de conocimiento: Ciencias de la Computación.

Tomando en cuenta los parámetros de búsqueda y la extracción de datos de Scielo, se obtuvieron un total de 109 artículos relacionados con el ámbito de las ciencias de la computación. Estos datos se utilizaron como entrada al modelo de predicción creado anteriormente, en el que se obtuvieron los resultados presentados en la siguiente sección.

4.3 Predicción de las líneas de investigación

Finalmente, se ejecutó el modelo de predicción con los 109 artículos obtenidos de Scielo, para obtener las líneas de investigación de cada uno de ellos, y se realizó una comprobación manual de las líneas de investigación obteniendo como resultado que 93 de las 109 áreas están correctamente predichas.

La figura 4.2 muestra los resultados de la validación, como se puede visualizar de los 109 artículos, el 93 líneas fueron predichas correctamente y solo 16 se predijeron incorrectamente.

[1] Fuente: Scielo, obtenido de: https://scielo.org

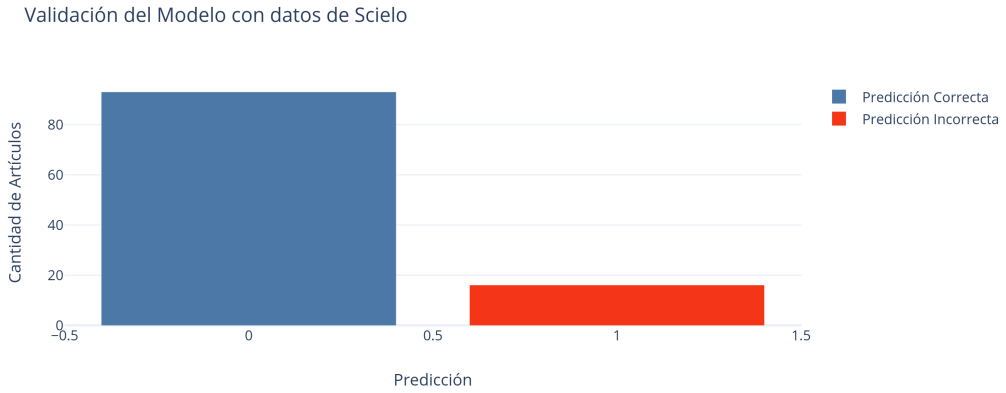

Fig. 4.2 Gráfico de la validación del modelo con datos externos (Scielo).

En la tabla 4.2 se muestran algunos de los ejemplos ejecutados con el modelo de predicción. Para ver el detalle completo de la validación de los datos de los 109 artículos de Scielo y los valores predichos y esperados dirigirse al Anexo B.

Tabla 4.2 Resultados de la validación del modelo SVM

Nombre del Artículo Scielo	**Línea de investigación Esperada**	**Línea de investigación Predicha**
A Social Commerce Intention Model for Traditional E-Commerce Sites	Social commerce media	Social commerce media
A South African universityes readiness assessment for bringing your own device for teaching and learning	A model of technology adoption in the theory of university career	A model of technology adoption in the theory of university career
Influencing Factors Analysis for a Social Network Web Based Payment Service in China	Focus to identify user preferences based on the analysis of social networks	Focus to identify user preferences based on the analysis of social networks
Perceptions of scratch programming among secondary school students in KwaZulu-Natal, South Africa	Algorithm programming applications of key chaos	Algorithms of programming of distributed links located in the IoT
Information Security Management Practices: Study of the Influencing Factors in a Brazilian AIR FORCE Institution	Cyber security in communications solutions	Cyber security in communications solutions
Some password users are more equal than others: Towards customisation of online security initiatives	Security of the information	Security of the information

Capítulo 5

Conclusiones

En la actualidad, con el gran avance que ha tenido la tecnología se han creado nuevas plataformas que permiten obtener documentos científicos de una manera más rápida, fácil y en formato digital. Así como la creación de herramientas con un acceso mucho más rápido y eficiente a estos documentos desde cualquier lugar del mundo. Sin embargo, esto también ha llevado a la existencia de una cantidad enorme de información difícil de procesar y descifrar. Es así que en esta tesis, se propone un enfoque innovador cuyo objetivo es brindar un apoyo a la exploración de estos documentos para que se conviertan en conocimiento y sean de fácil entendimiento. Así como también con el planteamiento de las líneas de investigación que sean de mucho interés para el país lo que contribuye al desarrollo de este con proyectos de investigación de relevancia.

En este capítulo, se presentan los aportes originales que se han derivado de este trabajo. Así como algunas recomendaciones.

5.1 Conclusiones

El comportamiento de búsqueda de información ha sido ampliamente estudiado por otros autores, pero todos ellos han realizado sus investigaciones en aspectos psicológicos del proceso, lo que resulta en modelos que se centran solamente en algunas variables y no en todas las necesarias para construir conocimiento. La solución planteada en este proyecto cubrió y definió todos los aspectos y variables necesarios para la construcción de unas líneas de investigación que realmente sean necesarias para el Ecuador en el área de las ciencias de la computación.

Debido a la complejidad del proceso de búsqueda de información, se ha decidido realizar un estudio de las líneas de investigación incluidas en el Libro Blanco del MINTEL, para asociarlas con las líneas de investigación de la ACM y generar las querys o cadenas de búsquedas necesarias para el propósito de la investigación que se realizarían en Scopus.

Adicionalmente, el estudio se centró en tener una búsqueda de documentos científicos en el área de las ciencias de la computación. Este análisis dio como resultado una cadena de búsqueda compleja creada a partir de 74 palabras claves que representan los principales conceptos y relaciones que existen entre las líneas de investigación del MINTEL, la ACM y los conceptos de tendecia en tecnologías emergentes segun Gartner. Los detalles completos de la realización de este estudio y de su procedimiento para obtener las querys necesarias están descritas en el capítulo 3.

Para procesar las cadenas de búsqueda se utilizó la API de Scopus, la cual facilitó la obtención de los documentos científicos importantes para el Ecuador. Se desarrolló además una mejora a su API para añadir más agilidad y eficiencia en la obtención de datos de los artículos. Así es que junto con Python se ha creado un script que permite obtener los datos más importantes de los artículos científicos mediante cualquier cadena de búsqueda.

Además, se ejecutó una limpieza de los datos obtenidos mediante diferentes técnicas como stopwords o lematizer para que en la información no existan datos basura. Con esto se consiguió un modelo completo de las tareas de búsqueda de información que pueden ser utilizadas también por investigadores ecuatorianos para obtener artículos científicos afines a sus líneas de interés.

La automatización del proceso de búsqueda utilizando Python como lenguaje de programación fue ventajosa en relación con la búsqueda de la información manual, ya que por la rapidez de procesamiento del lenguaje de programación con la librería Pandas, se obtiene como resultado una búsqueda más rápida comparada con las herramientas manuales que provee Scopus, cuyos resultados además, tienen restricciones. Adicionalmente, Scopus provee un API exclusivamente para aplicativos desarrollados en dicho lenguaje de programación y permite la extracción de mucha más información y datos de los artículos científicos.

Para la obtención de las líneas de investigación específicas para el Ecuador se tomó como base principal los documentos extraídos de Scopus y así obtener los datos para ser analizados

mediante técnicas de analítica de datos. En esta investigación, se han señalado cuáles son los aspectos a los que debe adaptarse el sistema, como el título del artículo, su resumen, sus palabras claves, con estos definir tareas de análisis de datos y obtener los resultados esperados.

Sin embargo, en el ámbito de la investigación, uno de los principales aspectos que encabeza todas las tareas, y especialmente las actividades de búsqueda de información, es la calidad de la información. Es por ello que lo principal fue definir las técnicas de minería de texto que permitieron realizar una limpieza total de la información obtenida.

Adicionalmente, en este trabajo se toma en cuenta las políticas del estado ecuatoriano, los estándares internacionales y las tendencias de la investigación. Por lo tanto, se ha desarrollado un sistema de minería de datos que permitió analizar, y encontrar patrones para clasificar las líneas de investigación específicas para el Ecuador. Para este objetivo se utilizó la clusterización para clasificar el conjunto de datos de los artículos científicos, el algoritmo de agrupación utilizado fue AfinityPropagation ya que este reduce la complejidad computacional para la agrupación de textos, además, mejora la precisión y evita la inicialización aleatoria y la captura del mínimo local. Es por ello que este algoritmo permite calcular las funciones de pertenencia de los datos, en este caso las líneas de investigación dentro de cada grupo.

Para validar estos resultados se realizó un análisis en las líneas de investigación con artículos que se han desarrollado regionalmente en la base de datos de Scielo. Tomando en cuenta que la precisión de la validación fue mayor a un 85% se puede considerar un modelo altamente confiable.

Finalmente, la descripción completa de todos los aspectos que intervienen en el proceso de búsqueda de información ha permitido identificar carencias y problemas que debería abordarse en trabajos futuros para mejorar el diseño de un sistema de información para la obtención de líneas de investigación de mayor impacto.

5.2 Trabajos Futuros

A continuación, se presentan varios trabajos futuros que se pueden abordar a corto, mediano y largo plazo como consecuencia de los resultados derivados de este trabajo.

- Evaluar el modelo propuesto por medio de varios investigadores de las ciencias de la computación dentro del Ecuador, esto implica determinar en qué medida los investigadores consideran que el modelo puede describir los conceptos y relaciones existentes entre las líneas de investigación y el plan de desarrollo del país por medio de entrevistas.

- El proceso para la obtención de líneas de investigación específicas presentado en este trabajo puede ser evaluado o analizado en diferentes áreas del conocimiento a parte de las ciencias de la computación, para así determinar si el proceso puede adaptarse y mantener su validez en diferentes áreas de estudio. También sería interesante establecer en qué medida el proceso es válido en diferentes dominios o campos de investigación y para los diferentes tipos de documentos científicos que existen.

- Crear un marco de referencia para elaborar diferentes proyectos de analítica de datos para obtención de líneas de investigación desde bases de datos de artículos científicos existentes.

- Además el modelo puede ser utilizado como una base para establecer un sistema de búsqueda de información que permita obtener los intereses del investigador y con ello crear redes que pueden ser utilizadas para conectarse con personas a nivel internacional y así, mejorar sustancialmente los resultados de la investigación.

- Explorar como se podría usar el sistema de agrupamiento utilizado para diseñar un sistema que proporcione sugerencias a los investigadores en función de sus intereses.

- Se puede mejorar y acelerar la etapa del preprocesamiento de los documentos científicos mediante la utilización de más técnicas y herramientas de analítica de texto como por ejemplo la integración con diferentes fuentes de información mediante cubos de agregación, reducciones dimensionales, así también con algoritmos de transformación de datos como Smmothing, generalización y agregación.

Referencias

ACM (1998). The acm computing classification system. [online] https://www.acm.org/publications/computing-classification-system/1998/ccs98.

ACM (2012). The acm computing classification system 2012. [online] https://www.acm.org/publications/class-2012.

Acosta, G. y Medina, E. (1997). Líneas de investigación en enfermería. *Revista Cubana de Enfermería*, 13(2):103–106.

Aggarwal, C. C. y Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Aguilar (2013). *The MIDANO*. ULA.

Aguilar, J., Aguilar, K., Jerez, M., y Jiménez, C. (2017). Implementación de tareas de analítica de datos para mejorar la calidad de servicios en las redes de comunicaciones. *Publicaciones En Ciencias Y Tecnología*, 11(2):63–77.

Almeida, H. V., Liu, Y., Cunniffe, G. M., Mulhall, K. J., Matsiko, A., Buckley, C. T., O'Brien, F. J., y Kelly, D. J. (2014). Controlled release of transforming growth factor-β3 from cartilage-extra-cellular-matrix-derived scaffolds to promote chondrogenesis of human-joint-tissue-derived stem cells. *Acta biomaterialia*, 10(10):4400–4409.

Bacino, G., Moro, L. E., Massa, S. M., Pirro, A., y Hinojal, H. (2018). Ambientes de aprendizaje enriquecidos con tecnología. In *XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste).*

Barros, M. V., Salvador, R., Piekarski, C. M., y de Francisco, A. C. (2018). Mapping of main research lines concerning life cycle studies on packaging systems in brazil and in the world. *International Journal of Life Cycle Assessment*. Article in Press.

Caldera, Y. M., Castro, J. L. A., y Hidrobo, F. J. (2018). Análisis de los problemas de rendimiento en un eva (entorno virtual de aprendizaje) a través de la extracción de conocimiento. *Ingeniería al Día*, 4(1):3–24.

Calixto, S. (2017). Estudio comparativo de herramientas para redes neuronales artificiales (rna): Weka, matlab y neurosolutions.

Carrascal, A. I. O. y Jiménez, G. A. (2018). Estudio sobre estilos de aprendizaje mediante minería de datos como apoyo a la gestión académica en instituciones educativas. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*, (29):1–13.

Castillo-Esparcia, A. et al. (2011). El rol de las publicaciones científicas en comunicación en el eees: indexación e impacto.

Cetto, A. M. (1998). Ciencia y producción científica en américa latina. el proyecto latindex. *International Microbiology*, 1(3):181–182.

Clarivate (2018). Web of science. [online] https://clarivate.com/products/web-of-science/.

Colledge, L., de Moya-Anegón, F., Guerrero-Bote, V. P., López-Illescas, C., Moed, H. F., et al. (2010). Sjr and snip: two new journal metrics in elsevier's scopus. *Insights*, 23(3):215.

de Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A., y Herrero-Solana, V. (2007). Coverage analysis of scopus: A journal metric approach. *Scientometrics*, 73(1):53–78.

Deirmengian, C., Kardos, K., Kilmartin, P., Gulati, S., Citrano, P., y Booth, R. E. (2015). The alpha-defensin test for periprosthetic joint infection responds to a wide spectrum of organisms. *Clinical Orthopaedics and Related Research®*, 473(7):2229–2235.

Ebner, N. C., Freund, A. M., y Baltes, P. B. (2006). Developmental changes in personal goal orientation from young to late adulthood: from striving for gains to maintenance and prevention of losses. *Psychology and aging*, 21(4):664.

Ellis, R. (1997). *SLA Research and Language Teaching.* ERIC.

Elsevier (2018). Scopus. [online] https://www.elsevier.com/solutions/scopus Fecha de acceso: marzo 2019.

Espejo, L. y Apolo, M. (2011). Revisión bibliográfica de la efectividad del kinesiotaping. *Rehabilitación*, 45(2):148–158.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., y Pappas, G. (2008). Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342.

Fernández-González, J., Grindlay, A., Serrano-Bernardo, F., Rodríguez-Rojas, M., y Zamorano, M. (2017). Economic and environmental review of waste-to-energy systems for municipal solid waste management in medium and small municipalities. *Waste Management*, 67:360–374.

Flores, N. (2017). Extracción de patrones semánticamente distintos a partir de los datos almacenados en la plataforma paideia.

Foster, A. (2004). A nonlinear model of information-seeking behavior. *Journal of the American society for information science and technology*, 55(3):228–237.

Freund, L., Clarke, C. L., y Toms, E. G. (2006). Towards genre classification for ir in the workplace. In *Proceedings of the 1st international conference on Information interaction in context*, pages 30–36. ACM.

Gartner (2018). 5 trends emerge in the gartner hype cycle for emerging technologies. *gartner.com*, Available at: https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/ (Accedido: 20 March 2019).

Gavel, Y. y Iselid, L. (2008). Web of science and scopus: a journal title overlap study. *Online information review*, 32(1):8–21.

Godoy, P. E. (2017). Un primer enfoque para el reconocimiento de lenguaje de señas basado en un guante inteligente que utiliza técnicas de machine learning. Master's thesis, Universidad de las Fuerzas Armadas ESPE. Maestría en Gestión de Sistemas de ....

Guan, R., Shi, X., Marchese, M., Yang, C., y Liang, Y. (2011). Text clustering with seeds affinity propagation. *IEEE Transactions on Knowledge and Data Engineering*, 23(4):627–637.

Hernández, J. O., FERRI RAMIREZ, C., y RAMIREZ QUINTANA, M. J. (2004). Introducción a la minería de datos.

Kabir, F., Siddique, S., Kotwal, M. R. A., y Huda, M. N. (2015). Bangla text document categorization using stochastic gradient descent (sgd) classifier. In *2015 International Conference on Cognitive Computing and Information Processing (CCIP)*, pages 1–4. IEEE.

Larrea, C. (2006). Universidad, investigación científica y desarrollo en américa latina y el ecuador. In *Ponencia presentada ante el Congreso "Universidad y Cooperación para el Desarrollo" en la Universidad Complutense de Madrid*, pages 26–28.

Laudel, G. (2017). How do national career systems promote or hinder the emergence of new research lines? *Minerva*, 55(3):341–369. Cited By :2.

Liberatore, G., Vuotto, A., y Fernández, G. (2018). Desarrollo de una herramienta para el análisis y representación semántica de colecciones documentales a través del factor tf-idf.

Lozada, H., Camarago, E., y Aguilar, J. L. (2017). Análisis del comportamiento de los precios del petróleo (ac2p). *Ingeniería al Día*, 3(2):20–45.

Macías, H. A. (2017). El sentido de publicar en revistas scopus: el caso de los autores colombianos de las áreas negocios, administración y contabilidad. *Science of Human Action*, 2(1):10–27.

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".

Meho, L. I. y Tibbo, H. R. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American society for Information Science and Technology*, 54(6):570–587.

Michán, L., Russell, J. M., Sánchez Pereyra, A., Llorens Cruset, A., y López Beltrán, C. (2008). Análisis de la sistemática actual en latinoamérica. *Interciencia*, 33(10):754–761.

Miguel, S., Moya-Anegón, F., y Herrero-Solana, V. (2007). El análisis de co-citas como método de investigación en bibliotecología y ciencia de la información. *Investigación bibliotecológica*, 21(43):139–155.

MINTEL (2016). Libro blanco 2016. [online] https://www.telecomunicaciones.gob.ec/wp-content/uploads/downloads/2017/05/Presentacion-Rendicion-de-cuentas-2016.pdf.

MINTEL (2018). Libro blanco 2018. [online] https://www.telecomunicaciones.gob.ec/wp-content/uploads/2018/07/Libro-Blanco-de-la-Sociedad-del-Informaci%C3%B3n-y-del-Conocimento.pdf.

MINTEL (2019). Libro blanco 2019. [online] https://www.telecomunicaciones.gob.ec/mintel-y-senescyt-presentaron-el-libro-blanco-de-lineas-de-investigacion-desarrollo-e-innovacion-y-transfe

Mora-Riapira, E. H., Vera-Colina, M. A., y Melgarejo-Molina, Z. A. (2015). Planificación estratégica y niveles de competitividad de las mipymes del sector comercio en bogotá. *Estudios Gerenciales*, 31(134):79–87.

Moral, C. (2016). Modeling the visualization and exploration of document collections with user and purpose-based adaptation.

Morales, K. A. y Aguado, E. (2010). La legitimación de la ciencia social en las bases de datos científicas más importantes para américa latina. *Latinoamérica. Revista de Estudios Latinoamericanos*, (51):159–188.

Muñoz-Écija, T., Vargas-Quesada, B., y Chinchilla-Rodríguez, Z. (2017). Identification and visualization of the intellectual structure and the main research lines in nanoscience and nanotechnology at the worldwide level. *Journal of Nanoparticle Research*, 19(2). Cited By :2.

Peralta, M. J., Frías, M., y Gregorio, O. (2015). Criterios, clasificaciones y tendencias de los indicadores bibliométricos en la evaluación de la ciencia. *Revista cubana de información en ciencias de la salud*, 26(3):290–309.

Perkins, J. (2014). *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd.

Rosario, P., Pereira, A. S., Högemann, J., Nunez, A., Figueiredo, M., Núñez, J. C., Fuentes, S., y Gaeta, M. L. (2014). Autorregulación del aprendizaje: una revision sistemática en revistas de la base scielo. *Universitas Psychologica*, 13(2):781–798.

Russom, P. et al. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, 19(4):1–34.

Schab, E., Rivera, R., Bracco, L., Coto, F., Cristaldo, P., Ramos, L., Rapesta, N., Nuñez, J. P., Retamar, S., Casanova, C., et al. (2018). Minería de datos y visualización de información. In *XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste).*

Secretaría-Técnica (2017). Plan toda una vida. [online] https://bit.ly/2TVlBP7.

Šubelj, L., Bajec, M., Boshkoska, B. M., Kastrin, A., y Levnajić, Z. (2015). Quantifying the consistency of scientific databases. *PloS one*, 10(5):e0127390.

Sunkel, G. (2006). *El consumo cultural en América Latina: construcción teórica y líneas de investigación*. Convenio Andrés Bello.

Torres-Salinas, D., Cabezas-Clavijo, Á., et al. (2012). Herramientas para la evaluación de la ciencia en universidades y centros i+ d: descripción y usos.

Torres-Salinas, D. y Jiménez-Contreras, E. (2012). Hacia las unidades de bibliometría en las universidades: modelo y funciones.

UNESCO (2018). Clasificación internacional normalizada de la educación, cine 2011. [online] https://unesdoc.unesco.org/ark:/48223/pf0000220782.

VALENCIA, P. U. A. (2017). ImplementaciÓn de analÍtica de datos sobre datos geoespaciales en una aplicaciÓnde micro-localizaciÓn que sirva para generar un software de guÍadentro de la universidad politÉcnica salesiana campus sur. Master's thesis, UNIVERSIDAD POLITÉCNICA SALESIANA.

Vasantha, K. y Majji, J. (2013). An efficient text clustering algorithm using affinity propagation.

Velandia, C., Serrano, F. J., Martínez, M. J., et al. (2017). La investigación formativa en ambientes ubicuos y virtuales en educación superior.

Vílchez-Román, C. y Huamán-Delgado, F. (2017). Estabilidad y cambio en la teoría organizacional: un estudio basado en la minería de texto y el análisis de co-citación. In *II Congreso Internacional de Ciencias de la Gestión, Lima, Peru*.

Yuen, J. (2018). Comparison of impact factor, eigenfactor metrics, and scimago journal rank indicator and h-index for neurosurgical and spinal surgical journals. *World neurosurgery*, 119:e328–e337.

Anexo A

Líneas de Investigación Específicas para el Ecuador

1. The use of machine learning to predict student indicators
2. Patterns of fusion of knowledge
3. Hybrid Android malware detection systems
4. Selection sets and feature analysis.
5. Local mining of high data set
6. Convulsive and conventional neural networks for bag comparison and fruit recognition
7. Crowd learning variations in Gaussian processes.
8. Convolutional neuronal network based on short-term memory attention, mood disorders, or short-term detection based on speech responses.
9. Shared navigation between men and women
10. Use of frogs jump algorithms for image processing.
11. Internet of things: security and privacy.
12. Data fusion Cybernetic social systems.
13. DDoS network traffic merge analysis: flooding attacks
14. Deep learning architectures to synthesize visual discourse

15. Algebraic information systems, new equivalent DOM domain category
16. Machine learning integration of data in biology and medicine
17. Understanding the mechanism of training and creation of high quality knowledge
18. L approach fuzzifying approximation fuzzy approximation operators
19. Continuity of virtual work environments.
20. Impact of corporate social responsibility in social networks
21. Calculation of the analysis of the trust domain.
22. Systems of decision-making services based on analysis
23. Strategic dependency diagrams of non-technical cases.
24. Business intelligence analysis for value creation in companies
25. Deep Visual Saliency Stereoscopic Images
26. Relative kinematics in large networks
27. An improved gray group decision making approach.
28. RSSI based ranging outdoor wireless sensor networks
29. Multibiometric merger strategy.
30. Prolongation of the correlation based on deep learning for the trend of financial time series
31. Detection of local metric blur
32. Manufacturer data driven by retail collaboration competition
33. Learning of attack mechanisms in wireless sensor networks using the Markov decision processes
34. The excessive use of smartphones among students.
35. Information fusion Theory of sets.
36. Maximization of influence in social networks.

37. Chaos optimization algorithm based on manufacturers
38. Web use mining
39. How social networks contribute to functional conflict.
40. PPP reductions
41. Cyberbullying, Perpetration Parents, and moderators of moral disengagement
42. EHealth system
43. New security model based on the cloud
44. Online courses, career choice based on technology.
45. A model of technology adoption in the theory of university career.
46. Architectural tactics: Big Data, Cybersecurity, Analytics Systems.
47. Mobile advertising. Affective cognitive perspectives.
48. Microfinance: An information management approach
49. Mobile gateway operation system based on the cloud.
50. Industrial wearables.
51. Aspect based Kano categorization
52. Learning of deep and compact similarity for facial kinship simulation
53. Recognition of songs based on music fingerprints
54. Spontaneous speech gestures using adverse generative networks.
55. A download method that uses edge computing with decentralized P2P
56. Flexible SDN: ad hoc network control tactics.
57. Possible hesitant fuzzy sets: A novel model group decision making
58. Combining different metadata views for better recommendation accuracy.
59. Incorporation of URL groupings that detect web anomalies
60. Correlation based on semantics that detect APT hybrid low level

61. Key phrase extraction algorithm
62. Conditional discriminatory mining of sequential patterns.
63. Disentangled Variational Auto Encoder semi-supervised learning
64. Participation in online games, young users of social networks.
65. A visualization approach discovering colocation patterns.
66. Computer applications with graphene
67. Autonomous vehicles using learning by deep reverse reinforcement
68. A dynamic and efficient energy data center.
69. Deep learning based on the prediction of tensile strength
70. Reactive model based on the neighborhood consensus of continuous optimization.
71. Applications of nested genetic algorithm
72. Relational social recommendation academic domain
73. Prediction models in medicine using automatic learning algorithms.
74. Shallow neural networks in highly demanding data networks
75. Convolutional networks for image classification
76. Neural network applications
77. A new pdf method of justified texts.
78. Fusion Mobile artificial intelligence and energy efficiency: Traffic control, Wireless Networks
79. Live virtual machines
80. UAV applications for grouping of programming tasks
81. Deep metric learning through diffuse grouping
82. quantile classifier for unbalanced data
83. XML filtering techniques for cloud configurations

84. Influence of individual values of internet use
85. The excessive use of social networks: A perspective of using the dimension.
86. Cloud computing based on FHE.
87. Kinematic model that simulates the phenomena of waves of vocal folds.
88. Online shopping: offline stores
89. Cybersecurity in intelligent networks based on IoT
90. SDBPR Social distance aware Bayesian personalized ranking recommendation
91. Intelligent transport systems.
92. IoT FBAC Access control scheme based on IoT identity encryption
93. Internet of Things Applications
94. Blockchain based searchable encryption electronic health record sharing
95. Privacy preserving publication of transactional data.
96. PhD students online who write an academic publication
97. Intelligent public bicycle services
98. Tuning of scalable residential energy consumption data.
99. Path data compressed based on grammar
100. Semi-supervised dimensionality reduction
101. The adoption of blockchain
102. Motivational feedback increases the benefits for the user.
103. Location-based service application connectivity
104. The uncanny valley No need judgments avatar looks eerie
105. A robust confidence inference algorithm
106. Planning the trajectory using differential evolution
107. A new recommendation for a matrix factorization model based on semantic LOD

108. Meta-circuit machine for information networks
109. Learning natural language models
110. Access control systems for privileged accounts in Business Networks
111. Predicting service qualification through user product reviews
112. Architecture of cybernetic incidents.
113. Study of EKG signal sensors.
114. Processes of learning virtual reality of desk
115. Face-to-face and digital learning
116. 3D Design CAPTCHAs
117. VR vehicle identification through deep unsupervised architecture
118. Processing of hyperspectral document images.
119. Estimation of human intention based on the Markov model
120. Robotic welding using deep learning
121. MGPV A new and efficient scheme for secure data exchange between mobile users of the public cloud
122. Policies that ensure the exchange of information in organizational environments
123. Architecture security that allow IoT infrastructures
124. Identification authorship code through convolutional neural networks.
125. Block chain based on reliable and decentralized mobile crowdsourcing
126. Medical data security system, portable health systems.
127. Duplication correction codes
128. Spatial data infrastructure management
129. Optimization of the artificial bee colony
130. Cloud environment for multiple task scheduling

131. Recognition of facial expression
132. Categorization of images in multiple views
133. Dispersed non-negative collaborative representation for pattern classification
134. Unique infrared image through a deep convolutional neuronal network
135. Feature of matrix enriched with extraction factorization.
136. Recurrence based on the identification of diseases such as Parkinson's
137. Adaptive time explicit goal-oriented
138. Particle swarm optimization damping factor cooperative mechanism
139. Periodic Semantic Hierarchical Mining with GPS Patterns
140. Knowledge-based system
141. Industrial processes using machine learning.
142. Frequency control isolated induction motor using multi-level three-phase inverter
143. Cloud of Big Data processing.
144. Search optimization of big data pre-processing
145. A forensic framework of mobile networks and ISP networks.
146. High speed scientific networks.
147. Cognitive statistics that combines the recovery of 3D objects
148. Machine learning based on vehicular communications infrastructure.
149. Intelligent communication network based on cognitive radio and QoS
150. Reverse gate networks dense gate
151. Integrated algorithms, multiagent, fault tolerant and based on MOEA programming.
152. Multi-copy coupling with routing speed tolerant to network delay
153. A predictor of perceptual distinguishability in images contaminated by noise
154. Virtual machine containers

155. Quaternion Grassmann
156. Perceptions of students against the use of technology
157. Inertial neural networks
158. Automatic extraction of clauses, international construction contracts that use NLP based on rules
159. Industry preparation 4.0
160. Multi-agent fixed-time consensus systems
161. Flow learning model green energy experience
162. Heterogeneity of the GPGPU programmability.
163. Multipath DenseNet: A tightly supervised ensemble architecture connected convolutional networks
164. Predictors of electronic words
165. An approach to find good ontologies.
166. Detection of outgoing objects
167. Representation of knowledge, learning of entity descriptions
168. Classification of multiclass BCI methods based on Dempster tasks
169. Analysis of the feeling of memory in the short and long term based on attention
170. Semantic approach through an association based on the academy
171. Virtual coordinators
172. Stacking models for phishing web pages detection
173. Automatic processing based on workflow
174. Internet floating things with collective collaboration data
175. Deep integration of extraction features of learning sets
176. Genetic algorithms based on evolution of adaptive genotypes

177. Data dissemination scheme code.
178. Intelligent search for augmented keywords, and spatial entities.
179. Cultural values of technology
180. EExam Technologies.
181. Performance of quality cost analysis of CSOC alert
182. Automatic translation modeling
183. Mobile malware detection systems
184. Assessment of technological education.
185. Detection of anomalies in network communication
186. Influence of impact time using a graphics-based model
187. On circulant involutory MDS matrices
188. On line Elastic Similarity Measures time series
189. Mobile Video Marketing
190. Machine learning based on intelligent IoT devices
191. The exploitation of syntactic neighborhood attributes
192. Inherited IA systems
193. Privacy of crowd detection systems
194. Heuristic search of multiple objectives.
195. Flexibility adapted to the process driven by events
196. Deep line video stabilization: Multi Grid Warping
197. Direct conversion transceivers with automatic cancellation
198. Linearly separable limited minimum beam variance trainers
199. Multi-proportional neural networks
200. Preserves the distributed networks of the safe origin scheme.

201. Smart cities infrastructure architecture based on IoT
202. Watermark for the protection and privacy of multimedia content
203. Channel attack controlled by SGX LEGO
204. Energy efficient retailing using femtolet based fog network
205. Generation of network traffic based on flow using generative advertising networks.
206. Vulnerabilities and massive contamination attacks of MIMO systems.
207. A secure image encryption scheme based on cellular automata
208. Revocable encryption based on attributes and encrypted text
209. Non-interactive privacy preserving the prediction of the neural network
210. SCADA machine learning
211. Recovery of surface meshes of color patterns using the edgeLBP descriptors
212. Diagnosis of faults based on discrete neuronal network convolution
213. Localized detection of anomalies without historical data of reference density estimation
214. Information management
215. Cloud Computing.
216. Detection of fuel consumption of vehicles through smartphones and recurrent neural networks
217. Minimum redundancy
218. Content prediction
219. Data analysis platform for security and emergency management decision making
220. Convolutional prediction using LSTM neural network
221. Multi objective gray wolf optimizer-based decomposition
222. Deep learning by transferring MR images
223. Deep learning in IoT system

224. Virtual network cards
225. 5G networks
226. Propagation of labels based on prediction algorithm
227. Decision making in financial technologies
228. Hashing of unsupervised deep video
229. Classification of adaptive tweets through deep learning
230. Safety nets in aerial vehicles based on the approach inspired by the principles of blockchain
231. Efficient DDoS
232. Cascade learning of synthetic images
233. Autocoder node
234. Social tagging tools
235. Quick templates for 3D point estimation
236. Surveillance videos in smart cities
237. Set approximate skyline queries
238. Efficient medical diagnosis through intelligent systems
239. Prediction based on fuzzy numbers and neural network systems
240. Learning artificial neural networks, forecasting the performance curve.
241. Feature space distance metric learning
242. Method of decision support based on the consumption of products
243. High voltage lines for the exchange of instant messages
244. Social commerce media
245. Dysfunctional mechanisms of the use of social networks in adolescents
246. Recognition of actions using dynamic multiple view images

247. Efficient colony optimization algorithm to block the relocation problem
248. Theory of quantum security games.
249. Public key encryption without an Oracle random upgrade
250. Learning from peer recommendation.
251. Weighted archetypal analysis dependent on the summary query of multiple videos
252. SAR images of detection of ships grouped in groups of convolutional neuronal networks.
253. Massive detection MIMO
254. Digital company
255. Multilevel learning based on user link modeling
256. Account identification algorithm for adults using behavioral traces
257. Focus to identify user preferences based on the analysis of social networks.
258. Secure and reliable routing protocols for the Internet of Things
259. Application clustering-based decision tree approach SQL query error database
260. Identifying key players in social networks using multi objective artificial bee colony optimization approach
261. Cloud-based data resources
262. Distributed learning in the cloud
263. Efficient data request vehicle answering Ad hoc networks based fog nodes filters
264. Semi supervised learning grouping unknown expressions
265. High-speed intrusion detection networks based on reliable anomalies in real time
266. Quantum based subgraph convolutional neural networks
267. Evidence-based search for evidence
268. Software security reverses engineering attacks.

269. Smart phone technology focused on the citizen in real time, scalable and monitoring
270. Hybrid remote monitoring in real time.
271. Multiple view learning application
272. Non-monetary benefits of mobile commerce
273. Cybersecurity policy and behavior
274. Smart storage
275. User preferences using neural networks.
276. Access control based on emergency roles
277. An integrated neutrosophical method
278. Algorithm programming applications of key chaos.
279. Individual performance and participation in learning in virtual environments.
280. Custom recommendation matrix of weather information
281. Integrated learning approaches based on cloud computing
282. Electronic learning
283. An asynchronous Algorithm algorithm AES CCM
284. Automated access control for IoT and Smart Home using the cumulative Keyed hash string
285. Recognition of multivariate anomalies with a spatially restricted approach
286. Online services for security behavior.
287. Elderly mobile users adopt the ubiquitous mobile social service.
288. Electronic commerce applications of wireless data sensors
289. Towards a reliable and outsourced computing
290. Privacy in the self-adaptive access control system for intelligent IoT storage
291. Chatbot learning

292. Quantification of the aesthetic website using deep learning.
293. Privacy in distributed services.
294. Avatars of technology use
295. A statistical approach to selection of participants based on social networks
296. Information search behavior
297. Hierarchical group code agreement protocol in the cloud.
298. Hybrid decision method of cloud services
299. Development of intercultural knowledge
300. Technology Evaluate
301. IoT decentralized industrial infrastructures
302. Lightweight authentication protocol for remote users
303. 5G multiple server networks with self-certified public key cryptography
304. Bayesian analysis model for medicine
305. A fully leveled, highly impenetrable and efficient homomorphic signature scheme.
306. SPIDER data fusion technique with Peer 2 Peer, smart city security improvements
307. Dynamic network traffic classification system
308. Encryption of public key of security receivers.
309. Algorithm in the cloud to group large data
310. Cloud storage
311. Convolutional network
312. Hybrid grouping deep convolutional neuronal networks
313. Deep matrix factorization for large scale recommendation of scattered data
314. Robots for smart vehicles
315. Extraction of multilevel features of people with memory leaks

316. Dynamic encryption techniques based on protected dual authentication
317. Joint joint of users
318. Methodology of recognition of open sets.
319. Detection of fully automated brain tumors using a classification based on superpixels
320. A new graphic kernel method stock price trend prediction based financial news semantic structural similarity
321. Applying mutual information discretization support discovery rare unusual association rule cerebrovascular examination dataset
322. Algorithms of programming of distributed links located in the IoT
323. Query processing for industrial IoT applications
324. Cyber security in communications solutions
325. Efficient cloud of classification of neural networks of Elman in IoT
326. Internet of things
327. Optimization of hyperactive Bayesian parameters
328. Application of intelligent decision machine learning models
329. Neural networks in multi obstacle environment robot systems
330. Prediction model in social media.
331. Multiple objective wrap method
332. Boat detection
333. Quaternion based weighted nuclear norm minimization color image denoising
334. User characteristics based on mobile usage behavior
335. Restricted network clusters
336. Digital text book
337. Deep Bayesian learning methodology

338. Social management of the internet of things

339. Security of the information.

340. Markov for compression detection algorithms

341. Progressive approaches to computing

342. Data fusion multiple classification systems

343. Detection of human activity, health monitoring

344. Houses with internet of things

345. Facebook useful to learn

346. Risk management in the company: the role of information technology.

347. Identify personal traits in an adaptive learning environment

Anexo B

Datos de prueba para validación del modelo de predicción de líneas de investigación

Tabla B.1 Tabla de datos de prueba para validación del modelo de predicción de líneas de investigación

Título	**Cluster Esperado**	**Cluster Predicho**	**Error**
ENHANCE KNOWLEDGE COMMUNICATION AND LEARNING: A SURPRISE PARADOX	114	114	0
RESEARCH METHODOLOGY FOR NOVELTY TECHNOLOGY	179	179	0
Higher Order Markov Chain Model for Synthetic Generation of Daily Streamflows	14	14	0
A Convergence Indicator for Multi-Objective Optimisation Algorithms	221	221	0
A Trajectory Planning Model for the Manipulation of Particles in Microfluidics	106	106	0
Wavelet Cross-correlation in Bivariate Time-Series Analysis	188	188	0
Applications of Nachbin's Theorem concerning Dense Subalgebras of Differentiable Functions	106	245	1
Multiple Solutions for an Equation of Kirchhoff Type: Theoretical and Numerical Aspects	131	131	0

Continúa en la siguiente página

Tabla B.1 – *Continúa en la página previa*

Título	**Cluster Esperado**	**Cluster Predicho**	**Error**
Fusion of Online Assessment Methods for Gynecological Examination Training: a Feasibility Study	140	140	0
A Reduced Semidefinite Programming Formulation for HA Assignment Problems in Sport Scheduling	199	199	0
Truncated-fractional Taylor's Formula with Applications	253	253	0
Bandwidth efficiency improvement for differential Alamouti space-time block codes using M-QAM	99	99	0
The economic reality of home PV systems: Matching consumption to generation	105	105	0
Instantaneous bit error rate based ASM scheme for MPSK spatial modulation	104	104	0
Subtropical rain attenuation statistics on 12.6 GHz ku-band satellite link using Synthetic Storm Technique	336	336	0
Time series analysis of impulsive noise in power line communication (PLC) networks	20	21	1
A study of single transmit antenna selection with modulation	128	128	0
Information Security Management Practices: Study of the Influencing Factors in a Brazilian AIR FORCE Institution	220	220	0
A Mapping Study of Scientific Merit of Papers, which Subject are Web Applications Test Techniques, Considering their Validity Threats	247	247	0
Towards Classifying Sociocultural Aspects in Global Software Development	134	134	0
Parameter analysis of the Jensen-Shannon divergence for shot boundary detection in streaming media applications	91	91	0
Low-complexity near-ML detection algorithms for NR-STAR-MQAM spatial modulation	242	242	0
Soft-output decision-based detection of SSK, Bi-SSK AND QSM	148	148	0
Analysis of bounded distance decoding for Reed Solomon codes	64	64	0
Development of a plant health and risk index for distribution power transformers in South Africa	301	301	0

Continúa en la siguiente página

Tabla B.1 – *Continúa en la página previa*

Título	**Cluster Esperado**	**Cluster Predicho**	**Error**
The Influence of Velocity Field Approximationsin Tracer Injection Processes	267	267	0
A Comparison Among Simple Algorithms for Linear Programming	82	82	0
Sequences of Primitive and Non-primitive BCH Codes	317	317	0
Apnea Recognition with Wavelet Neural Networks	248	248	0
An Error Bound for Low Order Approximation of Dynamical Systems Subjected to Initial Conditions	214	214	0
Numerical Solution of Heat Equation with Singular Robin Boundary Condition	322	322	0
Generalized Exponential Bidirectional Fuzzy Associative Memory with Fuzzy Cardinality-Based Similarity Measures Applied to Face Recognition	214	214	0
A New Scheme for Fault Detection and Classification Applied to DC Motor	245	245	0
Ordinal Sums of De Morgan Triples	67	67	0
Artificial Intelligence: way forward for India	284	284	0
A Digital Forensic Readiness Architecture for Online Examinations	219	219	0
Design Requirements for a Teledermatology Scale-up Framework	260	260	0
On More or Less Appropriate Notions of 'Computation'	205	205	0
In-lecture Media Use and Academic Performance: Investigating Demographic and Intentional Moderators	44	45	1
An integrative modelling technique bridging the gap between business and information systems development	131	131	0
Fit for Review	11	11	0
Assessing South African ICT4D research outputs: A journal review	214	214	0
Semi-automated Usability Analysis through Eye Tracking	14	14	0
Comic-based instruction and vocabulary learning among 11th and 12th graders in two Chilean schools	15	15	0

Continúa en la siguiente página

Tabla B.1 – *Continúa en la página previa*

Título	**Cluster Esperado**	**Cluster Predicho**	**Error**
Finite state machine for the social engineering attack detection model: SEADM	167	167	0
Guidelines for ethical nudging in password authentication	295	295	0
NoSQL databases: forensic attribution implications	298	298	0
Developing an electromagnetic noise generator to protect a Raspberry Pi from side channel analysis	143	143	0
Methodology for The Development of an Ontology Network on The Brazilian National System for the Evaluation of higher Education (OntoSINAES)	225	225	0
Subjectivity of Diamond Prices in Online Retail: Insights from a Data Mining Study	177	177	0
Investigating the Post-Adoption Attitude of the Web Based Content Management System within Organization	193	193	0
A Social Commerce Intention Model for Traditional E-Commerce Sites	13	13	0
The New Face of Internet User Typology: The Case of Thailand	247	247	0
Editorial: Digital Transformation & Digital Business Strategy in Electronic Commerce - The Role of Organizational Capabilities	146	200	1
Effects of Product Smartness on Satisfaction: Focused on the Perceived Characteristics of Smartphones	145	145	0
Assessing the Buyer Trust and Satisfaction Factors in the E-Marketplace	0	0	0
SAM: a meta-heuristic algorithm for single machine scheduling problems	250	250	0
Neural network fault diagnosis system for a diesel-electric locomotive's closed loop excitation control system	54	54	0
Detection of GSM And GSSK signals with soft-output demodulators	284	284	0
Enhanced congestion management for minimizing network performance degradation in OBS networks	89	89	0
Thermal instability analysis of a synchronous generator rotor using direct mapping	39	39	0

Continúa en la siguiente página

Tabla B.1 – *Continúa en la página previa*

Título	**Cluster Esperado**	**Cluster Predicho**	**Error**
Energy efficient statistical cooperative spectrum sensing in cognitive radio networks	63	63	0
Computational modelling for dish-to-dish coupling investigations on MeerKAT telescope	222	222	0
Non-decimated Wavelet Transform for a Shift-invariant Analysis	145	145	0
Fuzzy Linear Automata and Some Equivalences	21	28	1
Different Numerical Inversion Algorithms of the Laplace Transform for the Solution of the Advection-Diffusion Equation with Non-local Closure in Air Pollution Modeling	229	229	0
Adapted Fuzzy Integral: An Application in the Finite Element Method	315	230	1
Arboreal Identification Supported by Fuzzy Modeling for Trunk Texture Recognition	97	340	1
Trusses Nonlinear Problems Solution with Numerical Methods of Cubic Convergence Order	243	243	0
Homogenization of a Continuously Microperiodic Multidimensional Medium	232	232	0
Shifts in Online Consumer Behavior: A Preliminary Investigation of the Net Generation	257	257	0
Behavioral Customer Loyalty in Online Shopping: The Role of E-Service Quality and E-Recovery	77	75	1
Electronic Commerce: Factors Involved in its Adoption from a Bibliometric Analysis	306	306	0
A Novel Approach to Find Pseudo–peripheral Vertices for Snay's Heuristic	304	304	0
Parallel Implementation of a Two-level Algebraic ILU(k)-based Domain Decomposition Preconditioner	66	2	1
Editorial: Usage of Social Neuroscience in E-Commerce Research - Current Research and Future Opportunities	178	178	0
Mobile Shopping Consumers' Behavior: An Exploratory Study and Review	135	135	0

Continúa en la siguiente página

Tabla B.1 – *Continúa en la página previa*

Título	**Cluster Esperado**	**Cluster Predicho**	**Error**
The Cultivation of Information Infrastructures for International Trade: Stakeholder Challenges and Engagement Reasons	178	178	0
Views on Open Data Business from Software Development Companies	19	19	0
Integration of social media with healthcare big data for improved service delivery	86	86	0
Turning mirrors into windows: Knowledge transfer among indigenous healers in Limpopo province of South Africa	87	87	0
Analysing information literacy practices at selected academic libraries in Zimbabwe	238	238	0
Research data collection in challenging environments: Barriers to studying the performance of Zimbabwe's Parliamentary Constituency Information Centres (PCICs)	253	253	0
Challenges, benefits, and adoption dynamics of mobile banking at the base of the pyramid (BOP) in Africa: A systematic review	343	343	0
Some password users are more equal than others: Towards customisation of online security initiatives	306	306	0
An investigation on e-resource utilisation among university students in a developing country: A case of Great Zimbabwe University	322	322	0
The role of information and communication technologies in the delivery of health services in rural communities: Experiences from Malawi	278	278	0
The role of demographic and motivational factors on mobile commerce usage activities in South Africa	309	309	0
Application of the Technology Acceptance Model and the Technology-Organisation-Environment Model to examine social media marketing use in the South African tourism industry	194	194	0
Knowledge management as a strategic tool for human resource management at higher education institutions	334	334	0
The theory of planned behaviour and user engagement applied to Facebook advertising	177	177	0

Continúa en la siguiente página

Tabla B.1 – *Continúa en la página previa*

Título	**Cluster Esperado**	**Cluster Predicho**	**Error**
An information security behavioural model for the bring-your-own-device trend	315	300	1
Determinants of social media usage among a sample of rural South African youth	298	5	1
Knowledge retention in a platinum mine in the North West Province of South Africa	56	56	0
Perceptions of scratch programming among secondary school students in KwaZulu-Natal, South Africa	221	221	0
Enforcement of the Protection of Personal Information (POPI) Act: Perspective of data management professionals	131	131	0
Capturing tacit knowledge: A case of traditional doctors in Mozambique	202	202	0
The uncertain path to enterprise architecture (EA) maturity in the South African financial services sector	253	253	0
Information and knowledge sharing within virtual communities of practice	161	10	1
Treatment of Kenya's internet intermediaries under the Computer Misuse and Cybercrimes Act, 2018	211	9	1
Factors influencing e-health implementation by medical doctors in public hospitals in Zimbabwe	331	331	0
A South African university's readiness assessment for bringing your own device for teaching and learning	341	341	0
A multilevel approach to big data analysis using analytic tools and actor network theory	102	102	0
A Run-Time Algorithm for Detecting Shill Bidding in Online Auctions	39	45	1
Influencing Factors Analysis for a Social Network Web Based Payment Service in China	141	50	1
The Collaborative Economy Based Analysis of Demand: Study of Airbnb Case in Spain and Portugal	270	53	1
Editorial: Journal of Theoretical and Applied Electronic Commerce Research - CiteScore Metric from Scopus	256	256	0

Continúa en la siguiente página

Tabla B.1 – *Continúa en la página previa*

Título	**Cluster Esperado**	**Cluster Predicho**	**Error**
Social Agents to Enable Pervasive Social Networking Services	136	136	0
Managers/Owners' Innovativeness and Electronic Commerce Acceptance in Chilean SMEs: A Multi-Group Analysis Based on a Structural Equation Model	157	157	0