Modelo conceptual para el análisis de los factores asociados a las causas de la crisis de los contenedores aplicando técnicas de minería de texto

Jesús David Gómez Jaimes y Hugo Alexander Rincon Ballesteros

Trabajo de grado para optar al título de Ingeniero Industrial

Director

Henry Lamos Díaz

Ph.D en Matemática-Física

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2023

Agradecimientos

Agradezco a Dios por todas sus bendiciones, a la virgen maría quien siempre me acompaña y me guía. A mi mamá Elvia Lilia Ballesteros Parra por su apoyo incondicional, a mi papá Oliverio Rincon Monroy. Al profesor Henry Lamos por su paciencia y guía. A la profesora Yuly Andrea Ramírez Sierra por sus contribuciones importantes y al grupo de investigación OPALO por su apoyo a esta investigación.

Hugo Alexander Rincon Ballesteros

Agradecimientos

Agradezco al creador por este y todos los logros de mi vida, agradezco a mi familia y a mi pareja por su apoyo incondicional, al profesor Henry Lamos, a la profesora Yuli Ramírez y al grupo de investigación OPALO por su guía durante todo el desarrollo de este proyecto.

Jesús David Gómez Jaimes

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Los apéndices están adjuntos y pueden ser visualizados en la base de Datos de la Biblioteca UIS

Apéndice A. Base de datos de resúmenes y conclusiones de artículos científicos

Apéndice B. Código análisis exploratorio

Apéndice C. Código LDA

Apéndice D. Resultados de 8 y 12 topicos del modelo LDA

Apéndice E. Código K-Means

Apéndice F. Código Web Scraping

Apéndice G. Articulo publicable

Resumen

Título: Modelo conceptual para el análisis de los factores asociados a las causas de la crisis de los contenedores aplicando técnicas de minería de texto*

Autor(es): Jesús David Gómez Jaimes, Hugo Alexander Rincon Ballesteros**

Palabras Clave: Crisis de los contenedores, Minería de texto, Covid-19, Modelo Conceptual, K-Means, LDA, Causas.

Descripción: El presente trabajo de investigación propone identificar los factores asociados a las causas de la crisis de los contenedores, un fenómeno que ha impactado negativamente el comercio internacional y el transporte marítimo. Con este propósito, se emplearon técnicas de minería de texto con el fin de profundizar en la comprensión de la crisis y representarla mediante varios modelos conceptuales. En el proceso de recopilación de datos, se exploraron fuentes tanto de literatura gris como científica. Se aplicaron técnicas de minería de texto, como web scraping para la literatura gris y Latent Dirichlet Allocation (LDA) y k-means para la literatura científica. Estas dos técnicas fueron comparadas mediante sus resultados con el fin de seleccionar la herramienta adecuada para la investigación, finalmente se optó por el modelo LDA. Con esta técnica, se identificaron cuatro tópicos relevantes, donde por medio del análisis de bígamas se categorizaron de la siguiente manera: afectaciones en el transporte marítimo de mercancías, impacto en el flujo de contenedores, cambios en la logística de envíos de contenedores, y mitigación en la cadena de suministros, identificando a su vez las distintas fases de la crisis de contenedores siendo estas: 1) crisis de los contenedores 2) disrupción en la cadena de suministro y 3) resiliencia en la cadena de suministro. Los resultados obtenidos y su análisis permitieron la construcción de varios modelos conceptuales que ilustran los factores que causaron dicha crisis. En conclusión, la minería de texto no solo ratificó los factores previamente observados en la revisión de literatura científica y gris, también permitió identificar las fases que componen dicha crisis, proporcionando así una comprensión más completa y detallada del fenómeno.

*Trabajo de grado

** Facultad de Ingenierías Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: Henry Lamos Díaz. Ph.D en matemática-física.

Abstract

Title: Conceptual model for the analysis of the associated factors to the container crisis causes applying text mining techniques*

Author(s): Jesús David Gómez Jaimes, Hugo Alexander Rincon Ballesteros**

Key Words: Container crisis, Text mining, Covid-19, Conceptual model, K-Means, LDA, Causes.

Description: This research work proposes to identify the factors associated with the causes of the container crisis, a phenomenon that has negatively impacted international trade and maritime transport. For this purpose, text mining techniques were used in order to deepen the understanding of the crisis and represent it through a conceptual model. In the data collection process, both gray and scientific literature sources are explored. Text mining techniques were applied, such as web scraping for gray literature and Latent Dirichlet Allocation (LDA) and k-means for scientific literature. These two techniques were compared through their results in order to select the appropriate tool for the investigation, finally the LDA model was chosen. With this technique, four relevant topics were identified, where through the analysis of bigams they were categorized as follows: effects on the maritime transport of goods, impact on the flow of containers, changes in the logistics of container shipments, and mitigation . in the supply chain, identifying in turn the different phases of the container crisis, these being: 1) container crisis 2) disruption in the supply chain and 3) resilience in the supply chain. The results obtained and their analysis allowed the construction of several conceptual models that illustrate the factors that caused said crisis. In conclusion, text mining not only confirmed the factors previously observed in the review of scientific and gray literature, it also made it possible to identify the phases that make up said crisis, thus providing a more complete and detailed understanding of the phenomenon.

---

* Bachelor's degree

** Faculty of Physical Mechanical Engineering. School of Industrial and Business Studies. Director: Henry Lamos Díaz. Ph.D in mathematics-physics.

Introducción

La economía mundial a finales del siglo XX e inicios del siglo XXI se ha caracterizado por un proceso de globalización, transformándose en un factor determinante en el crecimiento económico internacional. En la actualidad, el comercio mundial de mercancías es fundamental en el estilo de vida de la población (Galeana, 2016). A diferencia del transporte aéreo, por carretera o ferroviario, el transporte marítimo es el principal método de distribución ya que permite enviar grandes cantidades de mercancía a un costo más económico, soportando un mayor movimiento y volumen tanto en contenedores, tanques y a granel. Este medio de transporte representa más del 80% del comercio internacional (Pérez, 2012). Siendo esta proporción especialmente alta para los países en desarrollo (Boiko & Getman, 2022). La calidad en conectividad de este método de distribución y la red globalizada entre los mercados internacionales es de vital importancia para el comercio contenerizado de un país (Tsantis et al., 2022). Por ejemplo, en Colombia el 98 % de las toneladas de carga exportadas e importadas del país, se realizan por vía marítima lo que destaca la importancia de esta actividad para la economía (Mindefensa de Colombia, 2023).

Si bien, la globalización ha estimulado aumentos sin precedentes en la escala y en la eficiencia de las organizaciones, también ha producido una mayor vulnerabilidad, ya que la interdependencia de los sistemas puede dañar el desempeño de la organización o afectar la economía de un país cuando se interrumpen las cadenas de suministro a nivel mundial (Shepherd & Williams, 2022).

Por otro lado, el advenimiento del siglo XXI ha traído diferentes desastres, algunos de los más reconocidos incluyen la ola de calor europea de 2003, la crisis económica del 2008-2009, el terremoto de 2010 en Haití, entre otros. Todos estos desastres resultaron en efectos sumamente

perjudiciales en la sociedad, junto con miles de millones de dólares en pérdidas económicas (Khan et al., 2022).Estas crisis demuestran las debilidades preexistentes de un sistema y ponen a prueba su resistencia (Notteboom et al., 2021).

La crisis de los contenedores, la cual afecto la cadena de suministros, se ve como resultado en parte por los paquetes de estímulo económico, los cambios en los patrones de gasto de los hogares, y el crecimiento ascendente resultante del comercio electrónico (Kent & Haralambides, 2022). El novedoso brote produjo un shock repentino en la economía global que interrumpió los ciclos de oferta y demanda de diferentes sectores económicos. Por ejemplo, las personas compraron compulsivamente más bienes esenciales de los necesarios, interrumpiendo el ciclo de oferta y demanda provocando escasez de alimentos y aumento de precios (Khan et al., 2022).

Debido a la importancia previamente mencionada del sector marítimo en la economía a nivel global, se presenta un análisis sobre los factores asociados al origen de esta crisis mundial, utilizando técnicas de minería de texto. Aunque ya se encuentran registros y literatura que documentan las causas más notorias, el objetivo de esta investigación es analizar la información disponible tanto en literatura gris como científica, con el fin de encontrar factores no evidentes o ratificar aquellos que ya han sido catalogados como los causantes de esta, dando claridad respecto a que origino la llamada “crisis de los contenedores”. Se presenta un modelo conceptual básico con el fin de consolidar los resultados obtenidos que faciliten la comprensión de las distintas partes interesadas, como lo pueden ser, empresas que dependan de este medio de transporte, la industria naviera, académicos, entre otros.

Tabla 1.

Cumplimiento de objetivos

Objetivos específicos	Numerales relacionados
1. Realizar una revisión de literatura sobre la crisis de los contenedores con el objetivo de identificar los factores que influyeron en ella.	Capítulo 03
2. Seleccionar la técnica apropiada de minería de texto para definir patrones hallados en los documentos científicos relevantes sobre la crisis de los contenedores.	Numeral 6.6
3. Aplicar el modelo de minería de datos no estructurados para la identificación y análisis de las causas principales de la crisis de los contenedores.	Numeral 6.5.1
4. Diseñar el modelo conceptual para las causas de la crisis de los contenedores con base en la revisión de la literatura y hallazgos obtenidos a partir de la minería de texto.	Capítulo 07
5. Elaborar un artículo académico de carácter publicable donde se presentan los resultados de la investigación.	Apéndice G

1. Planteamiento del Problema

El comercio internacional es fundamental en el crecimiento económico de cada país, se lleva a cabo por medio de negociaciones de mercancías que hay entre las naciones y el transporte de estas por vía marítima, aérea, terrestre o férrea. Siendo la más utilizada la vía marítima (Estrada Vidal & Reyes Hidalgo, 2017). Este deseo por llegar desde un mercado nacional a todos los rincones del mundo ha apoyado y se ha beneficiado por el desarrollo portuario, incrementando así los flujos comerciales, el desarrollo en otros sectores y la disminución de restricciones al realizar transacciones (Laxe, 2005). Aun así, los cargamentos siempre han sido vulnerables a factores externos como; el estado del comercio internacional, situaciones de carácter político, inclinaciones del mercado y limitaciones tecnológicas o legislativas en las partes involucradas (Jerebić & Pavlin, 2018).Otra problemática que afecta el comercio marítimo moderno es la planificación portuaria realizada en su mayoría con métodos empíricos, analíticos o de simulación, sin considerar escenarios o medidas ante una crisis (Rodríguez, 2013).

Durante periodos de pandemias, conflictos políticos y otros eventos especialmente a nivel mundial crecen las dificultades para la cadena global de suministros que abarcan la logística marítima, terrestre y aérea. Entre las diferentes problemáticas que han ocurrido durante la pandemia del covid-19, la crisis de la escasez de contenedores debido al cierre de fronteras y fábricas ha tenido como consecuencia la tendencia de la economía a paralizarse y con ella el transporte marítimo (Aguilar et al., 2020). Pese a esta grave problemática la pandemia del covid-19 también generó efectos inesperados, como el estímulo y rápido crecimiento de los e-commerce (Kent & Haralambides, 2022). Por otro parte, las restricciones tomadas para frenar de algún modo el avance del virus implicó una considerable disminución de la producción de miles de compañías alrededor del mundo o en algunos casos su paralización total, lo que implicó una drástica reducción en la compra de insumos impactando fuertemente la cadena de suministros.

El estancamiento de las actividades económicas y la demanda de productos en el 2020 presentó un reinicio exorbitante en sus primeros meses, rebasando la capacidad esperada de las cadenas de suministro y generando así saturación portuaria y escasez de contenedores (Díaz & Montealegre, 2022). La falta de estos generó serios problemas ya que, si un contenedor costaba 2.000 o 3.000 dólares, ahora se pagaban más de 15.000 dólares. Hay retrasos de ocho o nueve semanas, y en el peor de los casos no hay espacio, por lo tanto, los productos no pueden ser comercializados (Sánchez, 2021). El aumento en las tarifas de los contenedores tiene un impacto particular en el comercio a nivel mundial, ya que casi todos los productos manufacturados, los medicamentos y los productos alimentarios procesados, entre otros, se transportan en contenedores. Esta problemática afecta a todos los consumidores a nivel global.

Por lo anterior, el comercio marítimo se vio gravemente afectado al reducirse la demanda en el comercio internacional. La reacción de miles de compañías fue dejar de prestar ciertos

servicios o cancelar escalas de puertos, lo cual hizo aún más inestable el transporte marítimo de suministros, al mismo tiempo cosas como las restricciones o medidas de aislamiento relacionadas con el covid-19 también generaron congestiones y retrasos en las bahías de carga, debilitando la conectividad y las cadenas de suministro marítimas (Guerrero et al., 2022). A causa de la pandemia del covid-19 las compañías se detuvieron y dejaron de importar, esto generó que los contenedores se quedarán en los diferentes puertos evitando su flujo regular (Daza et al., 2022).

El gran impacto a nivel mundial de este evento ha generado mucha información referente a sus causas y efectos a nivel económico y social. Toda esta información puede ser útil, pero debido a la gran cantidad de textos es imposible para una persona o grupo de personas reunir toda la información y analizarla correctamente, los programas que filtran información poseen dificultades para interpretar fácilmente este tipo de textos ya sea por los modismos, las variaciones del idioma o jerga y el significado de términos basados en contexto, a diferencia de los humanos que pueden distinguir e interpretar estos patrones lingüísticos sin mayor inconveniente (Gupta & Lehal, 2009). Para realizar este proceso es necesario una herramienta especializada en la extracción de información de documentos, revistas, artículos y libros científicos, la cual facilita el proceso de análisis que se realizaría con los resultados arrojados.

Dentro de las herramientas disponibles, las más frecuentes que se encontraron son la minería de texto y la minería de datos. La minería de datos puede ser completamente caracterizada como la extracción de información implícita, desconocida y potencialmente útil a partir de los datos. Con la minería de texto, sin embargo, la información que se extrae es clara y explícitamente indicada en el texto (Arias et al., 2016). No es posible usar las mismas técnicas en el texto que se utilizaron en los datos estructurados almacenados en las bases de datos. Dado que los datos no estructurados son complejos, se requieren técnicas más potentes para extraer información de ellos.

Por lo tanto, fueron necesarias herramientas de minería de texto que puedan analizar estos datos. En otras palabras, las herramientas simples de minería de datos no son adecuadas para manejar texto (Kaur & Chopra, 2016).

El proyecto se centra en realizar un análisis de información semiestructurada y no estructurada, para ser más precisos, literatura científica y gris que abarca la situación económica y social actual que se generó por el covid-19 y otros factores relacionados, denominada la crisis de los contenedores. Este proceso se dará mediante la implementación de técnicas de minería de texto, abarcando así una mayor cantidad de documentos científicos y literatura gris, obteniendo información que pueda ser indicativa de las causas principales del efecto económico en cuestión.

Todo lo anteriormente mencionado se hará con la intención de esclarecer si las causas preconcebidas o que se especulan de manera superficial, son los factores definitivos que generaron este acontecimiento, o si realmente hay uno o más motivos que ocasionaron un impacto importante y no son fáciles de suponer debido a la diversidad y complejidad en la literatura respecto al tema. Por lo tanto, se desea esclarecer y mejorar el conocimiento mediante la aplicación de métodos multivariados en la minería de texto. Los resultados del presente proyecto se esperan ayuden a las diferentes partes interesadas, como lo pueden ser; empresas centradas en el comercio internacional (vía marítima), comunidad científica, entre otros, que buscan identificar y en la medida de lo posible mitigar dichos factores.

2. Objetivos

2.1 Objetivo General

Construir un modelo conceptual para el análisis de los factores asociados a las causas de la crisis de los contenedores aplicando técnicas de minería de texto.

2.2 Objetivos Específicos

Realizar una revisión de literatura sobre la crisis de los contenedores con el objeto de identificar los factores que influyeron en ella.

Seleccionar la técnica apropiada de minería de texto para definir patrones hallados en los documentos científicos relevantes sobre la crisis de los contenedores.

Aplicar el modelo de minería de datos no estructurados para la identificación y análisis de las causas principales de la crisis de los contenedores.

Diseñar el modelo conceptual para las causas de la crisis de los contenedores con base en la revisión de la literatura y hallazgos obtenidos a partir de la minería de texto.

Elaborar un artículo académico de carácter publicable donde se presentan los resultados de la investigación.

3. Revisión de Literatura

En el desarrollo del presente proyecto se realizó una investigación en diferentes bases de datos científicas como; Scopus, Elsevier, Springer y Scielo. Para ello se utilizaron las palabras clave “container”, “pandemic”, “crisis”, “causes”, “maritime transport”, “shipping”, “container crisis” y “text mining” filtrando la información respecto al periodo de tiempo comprendido entre los años 2020-2023, de igual manera las palabras claves mencionadas anteriormente fueron utilizadas en la búsqueda de información en la literatura gris. En el análisis de la literatura científica se evidencio que las causas de la crisis de los contenedores no eran el tema principal de múltiples documentos encontrados, ya que en su mayoría se hallaron aspectos como los efectos y consecuencias de la crisis, planes de mitigación, impacto ambiental y sostenibilidad. A pesar de ello, se detectaron de manera implícita causas que originaron la crisis de los contenedores, siendo estas de utilidad para el desarrollo de la presente investigación.

De esta manera se encontró que las cadenas de suministros están compuestas de una serie de etapas, desde la provisión de bienes intermedios hasta el consumo final de los bienes. Las interrupciones dentro de estas se manifiestan principalmente de tres maneras: interrupción en el suministro (falta de materiales, falta de mano de obra, etc), restricciones de distribución (restricciones comerciales, cierre de fábricas, ect) y una fuerte caída en la demanda (Notteboom et al., 2021). Una sola alteración de este tipo suele ser común, como cuando se produce un evento meteorológico o una huelga, dichos eventos están bien documentados ya que involucran un componente fácilmente identificable en la cadena de suministros. Sin embargo, los mecanismos de propagación y retro propagación que ocurren simultáneamente a gran escala en casos de desastres naturales significativos, revoluciones políticas, crisis financieras o las crisis sanitarias mundiales presentan mayor complejidad al momento de ser identificados.

La cadena de suministros de los contenedores se conforma de dos partes, el suministro hacia adelante, que son contenedores cargados de mercancía o materia prima que tienen como destino al cliente final siendo alquilados o vendidos por parte de las empresas que poseen dichos contenedores, y la cadena de suministro inversa o flujo de contenedores vacíos, que son posicionados estratégicamente por los cargadores para encontrar un mercado y realizar el proceso de devolución de los contenedores a las empresas adquiridas anteriormente, intentando retornarlos con mercancía (El Din et al., 2021). Durante una pandemia, la coevolución de los mecanismos de propagación y retro propagación crean varios shocks simultáneos (Notteboom et al., 2021).

Desde la propagación del virus covid-19 y tras su declaración como pandemia global, una de las actividades que se ha visto más afectada es la industria marítima y naviera (Narasimha et al., 2021). Muchas naciones en todo el mundo cerraron sus puertos marítimos y prohibieron la mayoría de sus actividades de exportación e importación (L. Liu et al., 2020). Causando interrupciones en la economía y el transporte de contenedores como medio predominante del flujo de mercancías (Kuźmicz, 2022).

Desde la perspectiva de la cadena de suministro, el covid-19 se desarrolló en fases secuenciales. La primera fase, a principios de 2020, consistió en una alteración de oferta en China, donde las medidas de bloqueo dieron como resultado una disminución drástica en la producción (Knower, 2020). La segunda fase comenzó a mediados de marzo de 2020 y consistió en un shock de demanda (global) con propagación hacia atrás a lo largo de las cadenas de suministro (Baschuk, 2020). Varias medidas de bloqueo implementadas en todo el mundo dieron como resultado una disminución en la demanda global. Posteriormente el cambio al comercio electrónico aceleró a los minoristas presionando con demandas adicionales, seguido por una recuperación que se arraigó en

el tercer trimestre de 2020, con condiciones altamente inciertas generadas por una nueva ola de casos de covid-19 y restricciones en países de todo el mundo (Notteboom et al., 2021).

Debido a esto, uno de los grandes problemas que ha dejado las medidas de restricción y como consecuencia una alteración en las cadenas de suministro global ha sido el atasco de contenedores en los puertos logísticos; una escasez de contenedores para transportar productos desde Asia a Occidente (Carga, 2022). De esta manera la escasez de contenedores vacíos se convirtió en una crisis mundial con efectos más devastadores que en períodos anteriores al combinarse con diversos problemas derivados del covid-19. La ausencia de contenedores vacíos en las regiones (donde se necesitaron) ralentizó las actividades industriales y bloqueó las redes de suministro globales, lo que exigió el uso de métodos alternativos que fueron ineficientes (Toygar et al., 2022).

Si bien el problema más crítico que ha enfrentado la industria del transporte marítimo ha sido como consecuencia de la pandemia esta no ha sido la única crisis a la que se ha enfrentado la industria naviera. La crisis económico-financiera, que comenzó en el otoño de 2008, muestra un claro ejemplo de las dificultades que ha enfrentado tradicionalmente la industria del transporte de contenedores, donde la pandemia tuvo lugar en medio de un proceso de lenta recuperación económica tras esta crisis del 2008-2009 (Notteboom et al., 2021).

El brote de covid-19 llegó cuando la economía aún no había logrado resolver problemas estructurales (R. Sánchez & F. Weikert, 2020). En la actual crisis las líneas navieras estuvieron mejor preparadas para enfrentar la pandemia debido a las lecciones aprendidas de la crisis financiera, gracias a una mejor gestión de capacidad más efectiva a través de alianzas (Notteboom et al., 2021). A pesar de la pandemia, las líneas navieras y sus alianzas mantuvieron la integridad de la red y recurrieron a más salidas en blanco para hacer frente a la disminución de la demanda

que se dio al inicio del covid-19, retirando hasta el 20% de su capacidad en las principales rutas comerciales (Kent & Haralambides, 2022). Sin embargo, una vez que pasó la primera ola de la pandemia, la demanda de bienes aumentó rápidamente, lo que se sumó al problema existente de acumulación de contenedores vacíos a nivel mundial (Sarmiento, 2022). El sistema no pudo adaptarse rápidamente al nuevo nivel de demanda y los contenedores quedaron abandonados o en lugares equivocados, debido a que muchos de estos se utilizaron en el primer semestre de 2020 para transportar equipos médicos a África y América Latina (Cullinane & Haralambides, 2021). Algunos exportadores llegaron a esperar semanas por contenedores disponibles para transportar su carga (Logimarex, 2021). El aumento de la demanda del contenedor vacío a nivel global afectó la oferta de éste, provocando el aumento de su precio y como consecuencia incrementando los fletes a nivel mundial (Orquera, 2022). Como resultado de este agravante, en 2021 el retraso promedio en el transporte de contenedores se duplicó en las rutas entre el Lejano Oriente y América del Norte. Entre el primer trimestre de 2020 y el último trimestre de 2021, los retrasos pasaron de 2 a 12 días (Naciones Unidas, 2022).

Debido al gran incremento en la demanda, se experimentaron retrasos cada vez mayores por las fuertes congestiones portuarias. La escasez de mano de obra, las interrupciones del trabajo en ciertos puertos y varios cuellos de botella agravaron la situación lo que ocasionó que los buques de carga tengan días y, en muchos casos, semanas de retraso (Chambers, 2021). Si antes de la pandemia el 62% de los barcos llegaba a tiempo, a finales de 2021 este valor había caído a sólo el 32% (Bnamericas, 2022).

Una vez se recuperó el flujo, los volúmenes de contenedores en el comercio transpacífico entre Asia y EE. UU llegaron a su límite provocando la congestión portuaria masiva en los puertos de Los Ángeles y Long Beach, lo que obligó a los transportistas marítimos a tomar medidas

extremas, una de esas medidas fueron los viajes "en blanco" (cancelando) no por falta de demanda, sino esta vez por falta de tonelaje, ya que los barcos estuvieron atascados esperando a su desembarque (Miller, 2021).

A medida que la pandemia obstaculiza las operaciones de las fábricas y sembró el caos en el transporte marítimo mundial, muchas economías de todo el mundo se vieron afectadas por la escasez de una amplia gama de productos, desde la electrónica hasta la madera y la ropa (Goodman & Chokshi, 2021). Esta crisis claramente reveló los límites de los principios de gestión de la cadena de suministro, como el justo a tiempo (JIT). Los propietarios de carga, los proveedores de servicios de logística, los transportistas y los operadores de terminales han experimentado el impacto de un sistema que lleva una década enfocado en la reducción de costos y procesos de cadena de suministro eficientes, lo que hace que cualquier interrupción sea un problema inmediato del transporte marítimo mundial. Este efecto ocasionó inventarios mínimos y una falta de capacidad de amortiguación para hacer frente a las interrupciones en la cadena de suministros global (Notteboom & Rodrigue, 2023).

Por otro lado, China es el mayor constructor naval del mundo. Sus fábricas constituyen más del 90% de los contenedores a nivel global y cuenta con siete de los 10 principales puertos de contenedores del mundo. Cuando China tropieza, también lo hace el transporte marítimo, tal como paso en 2022 cuando la demanda de importaciones de carga seca a granel se vio afectada por la crisis inmobiliaria y los cierres de covid-19 (Miller, 2022).

A pesar de un número relativamente bajo de casos de infectados por covid-19, varias ciudades chinas impulsaron estrictas medidas de restricción en respuesta al aumento de infecciones. Las autoridades han impuesto medidas en Shenzhen, Dongguan, Changchun y Shanghai, todos principales centros de fabricación. Además, Shanghái y Shenzhen son el primer y

tercer puerto más grande del mundo respectivamente. Si bien las medidas se han levantado en algunas ciudades, sus efectos se sintieron por la acumulación de carga y los continuos bloqueos localizados (Watt, 2022). Por ejemplo, en agosto de 2021, el puerto de Ningbo-Zhoushan se cerró después de que un empleado diera positivo por la variante del virus Delta. Un solo caso de covid-19 fue suficiente para poner en espera a todo el segmento de la industria marítima debido a una política de cero tolerancia por parte del país Asiático (Koshulko, 2023). Además de los impactos de la pandemia que perturbaron el flujo habitual del comercio y la disponibilidad de contenedores, otro factor importante fue el año nuevo chino, la festividad jugó un papel esencial en lo que respecta a la escasez de contenedores debido a que la mayoría de la población se encontraba de vacaciones, reduciendo así de manera significativa la mano de obra disponible en los puertos y centros de fabricación (Youd, 2021).

Si bien China es fundamental para el transporte marítimo internacional, otros eventos a nivel global también agravaron la crisis; por ejemplo, el Canal de Suez, una ruta comercial marítima que separa a África de Asia, por el cual se desplaza aproximadamente el 12% del comercio mundial (LaRocco, 2021). Sufrió un atasco el 23 de marzo de 2021 debido al buque portacontenedores “Ever Given” que bloqueo la importante vía fluvial durante seis días, lo cual aumentó los retrasos e interrupciones en la cadena de suministro global (Visco R&D, 2023). En su momento más de 450 barcos esperaban cerca del canal, este retraso fue una tensión adicional para el sistema de transporte de mercancía marítimo (La República, 2021).

Según Olaf Meker, la interrupción de la cadena de suministro marítimo no se trata de un megabuque atascado en el Canal de Suez ni de la falta de infraestructura adecuada en los puertos; en última instancia, se trata de la falta de una política de competencia efectiva para el transporte marítimo de línea global (Merk, 2021).

Coincidiendo con los brotes de covid-19 en China, la guerra en Ucrania agravo las presiones de oferta/demanda en curso para el transporte marítimo, lo que ha resultado nuevamente en congestión portuaria, tarifas de flete más altas y tiempos de tránsito más prolongados (Allianz, 2022). La industria naviera se ha visto afectada frente a la crisis mundial que se da como consecuencia de esta guerra, este conflicto tiene impactos en los diferentes sectores del transporte marítimo mundial ya que los puertos ucranianos están cerrados. En cuanto al transporte marítimo de portacontenedores, el conflicto ha impactado al sector en el aumento de los precios del petróleo y en las tarifas de fletes (Comunicaciones Puerto de Santa Marta, 2022). El mayor impacto de la guerra hasta ahora ha sido en los barcos que operan en el Mar Negro y/o comercializan con Rusia. Los principales puertos de Ucrania, incluido el de Odessa, se cerraron debido al conflicto y al bloqueo naval ruso a Ucrania (Allianz, 2022). Por ejemplo, Más de 200 barcos esperaron para cruzar el Estrecho de Kerch, que conecta el Mar Negro y el Mar de Azov (Paris & Faucon, 2022).

Toda la información anteriormente mencionada fue el resultado de una investigación centrada en las diferentes causas de la crisis de los contenedores y sus agravantes presentes en la literatura gris y científica, ahora, debido a que el objetivo es usar una metodología para analizar esta información, se amplió la búsqueda revisando los modelos de decisión que se han construido para la comprensión del fenómeno, encontrando así varios estudios.

El primero examina el impacto en la interrupción de los principales puertos de las rutas comerciales de contenedores por medio de pruebas de estrés utilizando la regresión por cuantiles de la cópula de Vine. De esta manera los autores identificaron la interconexión entre los puertos chinos y las principales rutas comerciales de contenedores en situaciones extremas. Dicho estudio proporciona una comprensión más profunda de cómo reacciona el mercado de transporte de contenedores ante desastres que afectan a varios puertos. Además, identificaron el efecto dominó

y de red en situaciones caóticas portuarias, donde se examina cómo cambiarían los volúmenes en diferentes rutas en caso de acontecimientos agravantes como lo es la pandemia del covid-19. Mediante los resultados, los autores afirman que “se pueden plantear estrategias de gestión de la capacidad de las líneas de contenedores con respecto a la planificación de rutas y el despliegue de la flota en caso de interrupciones en los principales puertos” (Xiao & Bai, 2022).

Se destacó un documento centrado en el contexto general de la pandemia del covid-19 y los estragos de la crisis de escasez de contenedores, en el cual se centran en explicar el impacto al igual que plantear un modelo que mezcla el método SWARA (Step-wise Weight Assessment Ratio Analysis) para encontrar los efectos más prominentes y ARAS (Additive Ratio Assessment) para las soluciones adecuadas entorno a la crisis. De esta manera las causas más representativas, según los resultados del estudio son; el aumento de costos en el transporte de contenedores, incertidumbre en la cadena de suministro, pérdida de volumen en el transporte de contenedores, aumento en la navegación en blanco. Además, las soluciones más adecuadas son; reserva garantizada de envío, tecnologías de comunicación de la información, contenedor propiedad del remitente, llamadas de incentivo y servicios de entrega E2E (Toygar et al., 2022).

Otro estudio relacionado propone una formulación de programación lineal entera para el problema de la ubicación confiable de instalaciones con contenedores plegables mediante una función objetivo. Esto debido a que los contenedores plegables se consideran una opción de costo-beneficio para solventar la crisis en la cadena de suministro global causada por la pandemia del covid-19. El estudio aborda el problema de la ubicación confiable asumiendo escenarios de crisis en los puertos. De acuerdo con los resultados presentados por los autores, cuando la probabilidad de interrupción en los puertos es demasiado alta, los contenedores plegables no otorgan beneficios operativos y monetarios, en cambio, deben tratarse con contenedores estándar, en este sentido, una

combinación adecuada de contenedores estándar y plegables juega un papel clave en la operación confiable de estos últimos para facilitar la circulación del flujo en escenarios de crisis (Jeong & Kim, 2023).

La revisión de la literatura pretendía conocer los diferentes modelos que hay para explicar la crisis y cómo ésta ha sido estudiada, se llegó a la conclusión de que no se evidencia información alguna referente al uso de técnicas de minería de texto para el análisis de los factores asociados a las causas de la crisis de los contenedores. Dicho lo anterior, se realizó una recopilación de documentos por medio de palabras clave y ecuaciones de búsqueda tanto en literatura gris como científica para implementar técnicas de minería de texto que permitan comprender de una mejor manera las causas de dicho fenómeno mediante el desarrollo de un modelo conceptual.

4. Marco de Antecedentes

Se realizó una búsqueda para la elaboración del marco de antecedentes como primera instancia en la biblioteca de la Universidad Industrial de Santander (UIS), con el fin de identificar si anteriormente se habían realizado trabajos de grado enfocados en la crisis de los contenedores la cual no arrojó resultados relacionados con la presente investigación. Posteriormente se realizó una consulta en diferentes universidades a nivel nacional. En la búsqueda efectuada se utilizaron palabras claves como: factores, crisis de los contenedores, causas, efectos.

Dentro de los documentos que se encontraron, el más alineado con la investigación fue “Crisis de los contenedores, una mirada desde el contexto global y sus implicaciones en Colombia” de Daniela Garzón Rivero y Erika Tatiana Espitia Forero de la Institución Universitaria de Colegios de Colombia (UNICOC), donde se menciona que al inicio de la pandemia y debido a las restricciones causadas por el covid-19, en marzo del 2020 el mundo entró en un confinamiento

que afecto las cadenas de suministro. Debido a esto, muchos buques portacontenedores quedaron estancados o navegando cerca a puertos de carga, lo que generó que los contenedores, ahora vacíos, no regresaran a su lugar de origen. Dada la problemática, el sector se vio gravemente afectado tras el cierre de algunos puertos y la falta de infraestructura con la que no contaban otros, sin embargo, los problemas se fueron agravando, ya que como consecuencia del confinamiento miles de personas empezaron a realizar compras en línea, lo que causó que el sistema comercial experimentará una sobre-demanda, ya que las fábricas estaban produciendo mucho menos de lo normal debido a la falta de personal. Esto ocasionó que los intercambios comerciales empezarán a detenerse, generando a su vez un estancamiento masivo de contenedores. Además, en la investigación realizada por las autoras se menciona, como dato importante, que la cadena de suministros ya estaba tensionada antes de la crisis de los contenedores debido a la falta de buques y espacio en los puertos (Garzón R. & Espitia F., 2022).

Ampliando la búsqueda se encontraron dos (02) trabajos de grado relevantes de instituciones de educación superior, el primero siendo de Guayaquil, Ecuador, publicado en el 2022 por los autores Jonathan Delgado y Mery Alvarado, describiendo y explicando tanto las causas como efectos de la pandemia, planteando también posibles soluciones y mitigaciones de estos por parte de las empresas y el gobierno ecuatoriano. De este proyecto se hizo especial enfasis en las causas que mencionan, describiendo sus efectos y reafirmando los factores encontrados en el anterior documento, como lo son, la disminución de la mano de obra, las restricciones del país asiático y la volatilidad de la demanda, que como consecuencia de un sistema de precios variable y dependiente de varios factores globales, aumentaron los costes de transporte o fletes y la producción de materias primas y productos, siendo los más afectados, los países en vía de desarrollo, pues su infraestructura y logística no pudo adaptarse o resistir estos duros golpes como

otras naciones del primer mundo. Otras causas también son mencionadas, como la temporada de tifones en China, que sigue siendo el país más importante en el comercio naviero, el embotellamiento en los puertos de Long Beach y Los Ángeles, ocasionando esperas de hasta 60 buques, y el conflicto político de Rusia y Ucrania que afecta la cadena de suministros de múltiples países implicados y no implicados en este conflicto, causando a su vez tiempos de espera prolongados, pues muchos buques presentes en el mar negro quedaron atrapados y otros deben tomar rutas alternativas (Alvarado & Delgado, 2022).

El segundo documento, que corresponde al país de Honduras publicado en el 2022 por la autora Rosa Adriana Tadeo Guzmán, menciona algunas causas agravantes de la crisis de los contenedores, como el estancamiento del portacontenedores Ever Given que encalló en el Canal de Suez, lo que empeoró la congestión de cientos de otros buques que transitan por esta importante vía fluvial, aumentando la escasez de contenedores e incrementando el precio de los fletes a nivel global. En el documento también se menciona otro factor influyente, debido a que los mayores importadores se resistieron al incremento en el precio del flete de las distintas compañías navieras alquilando sus propios barcos para transportar su mercancía, lo que condujo a un mayor tránsito de barcos portacontenedores agravando así la congestión (Tadeo, 2022).

La investigación reveló la escasez de documentación sobre la crisis de los contenedores en las instituciones de educación superior. La mayoría de la documentación sobre este tema se encuentra en artículos científicos, revistas y periódicos.

5. Marco Teórico

5.1 Transporte Contenerizado

Los transatlánticos de contenedores se han convertido en el modo de transporte principal en el transporte marítimo de mercancías desde la década de 1950. El transporte marítimo de línea significa que los buques portacontenedores viajan a lo largo de rutas regulares con tarifas fijas de acuerdo con horarios regulares. El transporte marítimo de contenedores y la red globalizada de transporte marítimo de contenedores permiten a los importadores y exportadores de bienes intermedios y manufacturados comercializar con socios remotos en países extranjeros (Tsantis et al., 2022).

5.2 Cadena de Suministros

Las cadenas de suministros o abastecimiento se describen como los recursos interconectados y las actividades necesarias para crear y entregar productos y servicios a los clientes, por lo cual se extienden desde el punto donde se extraen los recursos naturales hasta el consumidor (Vianchá, 2014). Una de las bases en la cadena de suministro es su orientación hacia una filosofía que gestiona la coordinación del flujo total de un canal de distribución desde el proveedor hasta el cliente final (Sablón-Cossío et al., 2021). La gestión integral de la cadena de suministro abarca tres fases fundamentales: aprovisionamiento, producción y distribución/comercialización. La efectiva gestión de todos los procesos vinculados con la cadena de suministro repercute directa o indirectamente en la calidad de los productos y en la optimización de los recursos disponibles (Nugent et al., 2019).

5.3 Contenedores

Según la agencia de aduanas de Colombia, Junior aduanas S.A. un contenedor es: “Un elemento de equipo de transporte reutilizable, que consiste en un cajón portátil, tanque movible u otro elemento análogo, total o parcialmente cerrado, destinado a contener mercancías para facilitar su transporte por uno o varios modos de transporte, sin manipulación intermedia de la carga, de fácil llenado y vaciado y de un volumen interior de un metro cúbico, por lo menos.” (Junior Aduanas S.A., s. f.).

5.4 Crisis de los Contenedores

El crecimiento del tráfico de carga marítima se desaceleró debido a la pandemia en China a principios de 2020. Se cerraron una gran cantidad de fábricas en los EE. UU y Europa, y se suspendieron muchas instalaciones de producción. Todo esto llevó a que los contenedores quedaran atascados en diferentes puertos. La crisis se manifestó en forma de escasez de contenedores en los países desde los que se exportaban las mercancías y su exceso en aquellos países a los que se entregaban las mercancías (Tumanovich, 2022).

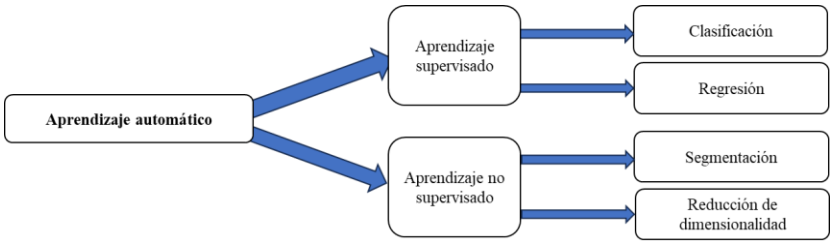
5.5 Aprendizaje Automático (Machine Learning, ML)

El aprendizaje automático (ML, por sus siglas en inglés) es el estudio científico de algoritmos y modelos estadísticos que utilizan los sistemas informáticos para realizar una tarea específica sin estar programados explícitamente. El aprendizaje automático se utiliza para enseñar a las máquinas cómo manejar los datos de manera más eficiente. En algunos casos, después de ver los datos, no es posible interpretar la información extraída de estos. En estas situaciones es conveniente el uso del aprendizaje automático (Mahesh, 2018). Los algoritmos de Machine Learning más utilizados son “clasificador lineal”, “regresión logística”, “Naïve Bayes (NB)”, “red

bayesiana", "máquinas de vectores de soporte (SVM)", "árbol de decisión", "bosque aleatorio", "AdaBoost", "agregación bootstrapped (embolsado)", "k-vecino más cercano. (k-NN)" y "red neuronal artificial (ANN)" (Shinde & Shah, 2018). Aunque comúnmente se asocia como sinónimo de inteligencia artificial, es crucial destacar que la inteligencia artificial constituye una categoría más amplia. En ella se engloban no solo técnicas para el análisis de datos estructurados, como el aprendizaje automático, sino también aquellas destinadas al procesamiento de datos no estructurados, como el procesamiento de lenguaje natural (Pedrero et al., 2021).

Figura 1.

Tipos de aprendizaje automático



Nota. Adaptado de *Introducing Machine Learning*. Math Works (2006, p.4)

5.5.1 Aprendizaje Supervisado

En el aprendizaje supervisado, los predictores y las variables de respuesta son conocidos, lo que permite la construcción de modelos matemáticos con el objetivo de predecir o clasificar observaciones futuras. Estos métodos se denominan supervisados porque el modelo se desarrolla utilizando los valores conocidos de las observaciones, es decir, la máquina aprende de los datos ya conocidos con la finalidad de predecir resultados en el futuro (Mora, 2020). A medida que el conjunto de datos aumenta, el algoritmo tiene la capacidad de aprender más sobre el tema relacionado con esos datos (Ortegón et al., 2023).

5.5.2 Aprendizaje no Supervisado

En esta categoría, no se tienen previamente las "etiquetas" de los datos, lo que significa que no se cuenta con información preexistente sobre los grupos a los que pertenecen las muestras. En este contexto, el algoritmo debe aprender a describir la estructura de los datos. Los métodos principales en esta categoría incluyen la agrupación (clustering) y los métodos de análisis de componentes principales (PCA, por sus siglas en inglés) (Ortegón et al., 2023). A diferencia del aprendizaje supervisado, no hay respuestas correctas ni guiadas. Los algoritmos funcionan sin intervención de los investigadores para descubrir y presentar la estructura interesante de los datos. (Mahesh, 2018).

5.6 Minería de Texto

La minería de texto permite explorar y extraer contenido de documentos textuales, con el objetivo de obtener información relevante. “Text Mining” es el descubrimiento por computadora de información nueva, previamente desconocida, mediante la extracción automática de información de diferentes recursos escritos. Un elemento clave es la vinculación de la información extraída para formar nuevos hechos o nuevas hipótesis que serán exploradas más a fondo por medios de experimentación más convencionales (Hearst, 2003).

Según Ángel Freddy Godoy Viera en su artículo “Técnicas de aprendizaje de máquina utilizadas para la minería de texto”. Esta se enfoca en reconocer y extraer características representativas para documentos en lenguaje natural. Estas características pueden ser palabras clave, identificación de nombres de personas, organizaciones, etcétera. El objetivo del preprocesamiento es transformar datos no estructurados que se encuentran en documentos a un formato más explícito (Godoy, 2017).

5.7 Metodología KDT

La extracción de conocimiento de bases de datos y textos, que se conoce como "Descubrimiento de conocimiento de datos" (KDD) y "descubrimiento de conocimiento de textos" (KDT) respectivamente, son una de las formas de revisión bibliográfica y documentos más utilizados cuando se analiza una gran cantidad de datos e información de texto. El conocimiento de su proceso de trabajo, así como las herramientas utilizadas para tal fin dan libertad al investigador para producir conocimientos previamente no visibles (Da Silva, 2020).

De forma general, se pueden resumir las fases principales del KDT en estas tres: preprocesamiento, minería de textos y visualización (Torre, 2017).

5.8 Proceso de Descubrimiento de Conocimiento en Textos (KDT)

Para realizar adecuadamente un proceso de minería de texto se requiere una metodología que permita un correcto desarrollo, la más comúnmente utilizada para este tipo análisis es la metodología KDT (por sus siglas en ingles). A continuación, se describen cada una de sus etapas.

5.8.1 Preprocesamiento de Texto

El preprocesamiento de texto se aplica a la compilación de documentos que contienen datos no estructurados o semiestructurados. El preprocesamiento de texto convierte un archivo de esta índole sin procesar en una secuencia bien definida de unidades lingüísticamente significativas (Gohil, 2015).

5.8.2 Tokenización

Es el proceso de dividir un flujo de texto en palabras, frases, símbolos u otros elementos significativos llamados tokens. El objetivo de la tokenización es la exploración de las palabras presentes en una oración. En general, estos tokens se agrupan en una lista que se convierte en la

entrada para procesamiento posterior, como análisis o extracción de texto (Kannan & Gurusamy, 2014).

5.8.3 Limpieza de Texto

La fase de limpieza en el análisis de texto consiste en la eliminación de tokens que no aportan valor significativo como signos de puntuación o palabras poco representativas, conocidas como "stopwords". Para llevar a cabo esta etapa de limpieza de manera efectiva, es necesario contar con un diccionario o listado de términos clave que sirva como referencia durante el proceso (Núñez et al., 2021). Este proceso realiza a su vez labores como eliminación de anuncios de páginas web, eliminación de tablas, figuras, entre otros (Gohil, 2015).

5.8.4 Lematización

El proceso de lematización es un método de análisis morfológico que se utiliza para obtener o identificar el tronco o raíz de una palabra. Por ejemplo, la palabra “autos” corresponde al plural de la palabra “auto”, y a su vez, “autito” es un diminutivo de la palabra “auto”, ambas variantes morfológicas de la palabra “auto” tienen interpretaciones semánticas similares y se pueden considerar como equivalentes en el uso de herramientas de Minería de Texto. Los algoritmos de lematización evitan que palabras con variantes morfológicas se consideren como palabras diferentes (Pérez Cárcamo, 2010). De esta manera, La lematización busca normalizar automáticamente los términos que pertenecen a una misma familia y comparten significado, reduciéndolos a una forma canónica o lema. En este proceso, los sustantivos se transforman al masculino singular, y los verbos se llevan al infinitivo (Torres-Moreno, 2010).

5.8.5 Etiquetado de Parte del Discurso (POS-Tagging)

En esta etapa se determina la categoría lingüística de la palabra. Donde se asigna cada clase de palabra a cada token. En inglés, hay ocho clases de lexemas: sustantivo, pronombre, adjetivo,

verbo, adverbio, preposición, conjunción e interjección. Lo anterior debido a que las librerías de programación están constituidas en este idioma. Además, las técnicas para el etiquetado de puntos de venta son enfoques basados en modelos ocultos de Markov y enfoques basados en reglas (Gohil, 2015).

5.8.6 Transformación de Texto

Este proceso realiza la generación de características seguida de la tarea de selección. La generación de características representa documentos por las palabras que contienen y sus ocurrencias donde el orden de las palabras no es significativo. Utiliza una bolsa de palabras o modelo de espacio vectorial. Reduciendo así la dimensionalidad al eliminar características redundantes e irrelevantes (Gohil, 2015). De esta manera el proceso consiste en transformar datos no estructurados en un formato estructurado para identificar patrones significativos y nueva información. Mediante la aplicación de técnicas analíticas avanzadas, como Naïve Bayes, máquinas de vectores de soporte (SVM, por sus siglas en inglés) y otros algoritmos (IBM, 2018).

5.9 N-Gramas

Las técnicas basadas en N-gramas (N-Grams en inglés) son predominantes en el procesamiento de lenguaje natural y sus aplicaciones. Comúnmente, se emplean como características esenciales en la representación del modelo de espacio vectorial. En términos generales, los N-gramas consisten en secuencias de elementos tal como se disponen en los textos, los cuales son generalmente palabras, caracteres u otros componentes dispuestos uno tras otro. La convención común es que "n" en n-gramas corresponde al número de elementos en una secuencia (Sidorov et al., 2014). Se distinguen principalmente tres tipos de N-grams, presentados a continuación (Tiffani, 2020).

- Unigrama: Token formado por una sola palabra
- Bigrama: Token formado por dos palabras
- Trigrama: Token formado por tres palabras

5.10 Matiz TF-IDF

La técnica TF-IDF se implementa ya que esta elimina los términos más comunes y extrae solo los términos más relevantes del corpus (Bafna et al., 2016). Para determinar el valor, el método utiliza dos elementos: TF - frecuencia del término i en el documento j e IDF - frecuencia inversa del documento del término i (Liberatore et al., 2018) Esta técnica se representa en la siguiente ecuación:

$$a_{ij} = tf_{ij} idf_{ij} = tf_{ij} * \log_2(\frac{N}{df_i})$$

donde a_{ij} es el peso del término i en el documento j, N es el número de documentos en la colección, tf_{ij} es la frecuencia del término i en el documento j y df_i es la frecuencia del documento del término i en la colección (Nielsen & Nock, 2015). TF-IDF considera el peso de cada palabra mediante el uso de dos enfoques, la frecuencia de un término y en cuántos archivos se puede encontrar dicho término. De esta manera se logra convertir datos no estructurados en datos estructurados que puedan ser reconocidos por los algoritmos (Hakim et al., 2014).

5.11 Clustering

La agrupación o clustering es uno de los principales modelos en minería de datos ya que se utiliza de diversas formas para agrupar datos en información significativa (Saputra et al., 2020). Esta técnica juega un papel importante en aplicaciones como exploración de datos científicos, recuperación de información y minería de texto, así como en aplicaciones relacionadas con bases

de datos espaciales, marketing, diagnóstico médico, análisis de ADN en biología computacional, entre otras. (Garre et al., 2007).

La tarea de agrupación, conocida como clustering, es una actividad descriptiva que busca organizar a los individuos disponibles en grupos o clústers con comportamientos similares. A diferencia de la tarea de clasificación y las distintas técnicas utilizadas en ella, en el análisis de clúster y en otras técnicas de agrupación, los grupos son desconocidos a priori y es precisamente lo que se busca determinar. De esta manera, el objetivo es obtener agrupaciones (clusterings, o clusters), mediante un análisis de carácter exploratorio (Martínez, 2012). Esta técnica se caracteriza por la mayor similitud dentro del mismo conglomerado y la mayor disimilitud entre diferentes conglomerados (Sinaga & Yang, 2020). Este divide un conjunto de objetos en grupos, donde los objetos presentes en un grupo son muy similares entre sí, y al mismo tiempo, son diferentes respecto a otros grupos (Pérez Ortega et al., 2018). En la investigación se identificaron varios métodos de agrupación como; métodos jerárquicos, basados en particiones, basados en densidad y maximización de expectativas, los cuales se describen a continuación:

- Clúster basado en métodos jerárquicos: En el método jerárquico se realizan particiones sucesivas a diferentes niveles de agregación o agrupamiento. Estos métodos suelen dividirse en dos categorías: los aglomerativos (ascendentes), que fusionan grupos progresivamente en cada paso, y los divisivos (descendentes), que descomponen el conjunto total de datos en grupos cada vez más pequeños. La representación de la jerarquía de clusters obtenida, se realiza comúnmente a través de un diagrama en forma de árbol invertido denominado dendrograma (Martínez, 2012).
- Clúster basado en particiones: El clustering basado en particiones es un modelo basado en centroides y el valor del clúster. El primer paso en la agrupación basada en particiones es

determinar el valor K. Después de eso, el algoritmo de agrupamiento basado en particiones formará diversos grupos. Determinar el valor de K es uno de los mayores problemas en la agrupación basada en particiones, no existe una forma universal de determinar este valor. Uno de los algoritmos más utilizados de este tipo de agrupación es el algoritmo K-means. La principal diferencia entre este método y el método jerárquico es que en este no se genera un cluster como subgrupo (Saputra et al., 2020).

- Clúster basado en densidad y maximización de expectativas: Otros métodos de clustering son el agrupamiento espacial basado en densidad o Expectation-Maximization (EM). El primero de ellos intenta buscar regiones densas en las que un conjunto de puntos es alcanzable secuencialmente (según su proximidad) y a partir de estas regiones organiza la agrupación. El segundo, utiliza un algoritmo en dos fases (fase esperanza, fase maximización) pretendiendo estimar la esperanza de la verosimilitud de que un elemento pertenezca a cada posible grupo y, a partir de ahí agrupar de forma que se maximice la esperanza de pertenencia a cada grupo candidato (Martínez, 2012).

5.11.1 K-Means

El k-means es uno de los algoritmos de aprendizaje no supervisado más populares que resuelven el conocido problema de agrupamiento (Sinaga & Yang, 2020). Fue creado por MacQueen en 1967 siendo reconocido como uno de los algoritmos más simples de clustering de minería de texto (Sánchez Álvarez, 2021). Consiste en iniciar k semillas distintas, que se establecen como los centros iniciales de los grupos. Luego, de manera iterativa, se asignan los puntos a sus centros más cercanos y se actualizan los centros de los grupos tomando los centroides de estos (Nielsen & Nock, 2015). Para su aplicación, se necesitan como parámetros de entrada el número de grupos (k) y una métrica de distancia. En la fase inicial, cada punto de datos se asigna

a uno de los k grupos según su proximidad a los centroides (los centros de los grupos) (Rodriguez et al., 2019).

Este algoritmo permite agrupar datos D-dimensional que constan de N muestras x_n en grupos de acuerdo con sus distancias entre muestras, donde n = 1, ..., N. El objetivo es construir centroides $\{\mu_k\}$, donde k = 1, ..., K, de modo que el centroide μ_k pertenezca al k-ésimo grupo. Luego, el punto de datos x_n se asigna al grupo cuyo centroide está a la distancia más pequeña de x_n, comúnmente utilizando la métrica euclidiana:

$$\|X_n - \mu_k\|^2$$

Durante el procedimiento de búsqueda de centroides óptimos μ_k, un objetivo de optimización viene dado por

$$J = \sum_{n,k} r_{nk} \|x_n - \mu_k\|^2$$

donde $r_{nk} = 1$, si la muestra de datos x_n se asigna al k-ésimo grupo; de lo contrario, $r_{nk} = 0$. Por lo tanto, el término J representa la suma de los cuadrados de las distancias entre la muestra y su grupo asignado con centroide μ_k. Para minimizar J, es necesario encontrar conjuntos de $\{r_{nk}\}$ y $\{\mu_k\}$. Esto se hace mediante un procedimiento de optimización iterativo donde cada iteración consta de dos pasos. En el primer paso, la minimización se realiza con respecto a $\{r_{nk}\}$, lo que significa asignar muestras de datos x_n a sus centroides de grupo más cercanos μ_k. En el segundo paso, la minimización se realiza con respecto a $\{\mu_k\}$, el valor de μ_k se calcula como la media de las muestras de datos asignadas al k-ésimo grupo. Esta optimización se repite hasta la convergencia. Por lo general, la inicialización del conjunto $\{\mu_k\}$ se realiza mediante la asignación de muestras de datos seleccionadas aleatoriamente a los centroides del grupo μ_k (Křupka, 2013).

La fortaleza de K-means radica en su forma sencilla de agrupar datos en función de su centroide y la distancia a cada dato (Saputra et al., 2020).

Las métricas de distancia permiten calcular el grado de similitud entre dos puntos de datos, cuanto mayor sea la distancia, más diferentes son los objetos y menor la probabilidad de que los métodos de clasificación los asigne en el mismo grupo o cluster (Malki et al., 2016). La métrica de distancia más utilizada es la euclidiana, mientras que otras dos métricas populares son la distancia de Manhattan y la de Minkowski. En general, la distancia euclidiana y la de Manhattan se pueden considerar como variantes de la distancia de Minkowski, donde el cuadrado y la raíz cuadrada o p en la ecuación (1), son 1 para la distancia de Manhattan y 2 para la distancia euclidiana (Saputra et al., 2020).

5.11.1.1 Distancia Euclidiana. Matemáticamente, la distancia euclidiana es una forma de medir una distancia entre dos puntos en una dimensión que arroja resultados como la ecuación de Pitágoras. La distancia euclidiana se obtiene generalmente a partir de la raíz cuadrada de la suma de las diferencias al cuadrado entre dos puntos (Saputra et al., 2020). Lo dicho anteriormente se muestra en la ecuación (1).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

5.11.1.2 Distancia de Manhattan. La distancia de Manhattan se conoce como distancia de la cuadra de la ciudad y se utiliza para medir la distancia de un dato determinado a otro dato. La distancia de Manhattan refleja la distancia entre puntos de una vía urbana dentro de 01 cuadra. Una ecuación matemática de la distancia de Manhattan se observa en la ecuación (2) y se calcula sumando los resultados absolutos de la reducción entre puntos (Saputra et al., 2020).

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (2)$$

5.11.1.3 Distancia Minkowski. La distancia de Minkowski es una métrica en el espacio vectorial normalizado que podría considerarse una generalización de una distancia euclidiana y una distancia de Manhattan. La distancia de Minkowski del orden de p entre 2 puntos podría definirse mediante la ecuación (3).

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \quad (3)$$

La distancia de Minkowski se usa generalmente con p igual a 1 o 2, que corresponde a la distancia de Manhattan y Euclidiana. En el caso de limitar p desde un valor infinito, se obtiene la distancia de chebyshev (Saputra et al., 2020).

5.11.1.4 Distancia de chebyshev. También conocida como distancia del tablero de ajedrez, se define como la mayor de las distancias entre cualquiera de los valores a lo largo de sus ejes de coordenadas. La comparación con el ajedrez se utiliza debido a que, en un plano bidimensional, la distancia entre dos casillas sería equivalente al número mínimo de movimientos que realizaría el rey para llegar de una a la otra (Heras Calvo, 2023).

Esta métrica está definida en un espacio vectorial, donde la distancia entre dos vectores es la mayor de sus diferencias a lo largo de cualquier dimensión de coordenadas (Arriaga, 2015). Expresada en la ecuación (4).

$$D(i,j) = max\left|x_{ki} - x_{jk}\right| \quad (4)$$

5.12 Modelamiento de Tópicos

El modelamiento de tópicos es una forma de analizar texto, cuyo fin es encontrar estructuras en una colección de documentos con el objetivo de generar clusters o agrupación de documentos en base a uno o varios tópicos (Sepúlveda Ramírez, 2019). El modelado de tópicos descubre automáticamente los temas en grandes corpus de datos no estructurados que forman un tema en común; los temas descubiertos son conjuntos de palabras y cada documento puede contener varios temas (Smith et al., 2018). Esta técnica puede considerarse como una metodología para presentar la gran cantidad de datos generados por el avance tecnológico, con el fin de exponer conceptos ocultos (Kherwa & Bansal, 2018).

El modelado de tópicos se desarrolló originalmente en la década de 1980 y se separó del área temática de "modelado probabilístico generativo". Este tipo de modelado supone que las variables observadas interactúan con parámetros no observados o latentes en una relación probabilística específica, que luego genera y agrupa los datos dentro de un conjunto. El desarrollo de estos procesos surgió de la necesidad de describir elementos dentro de grandes colecciones de datos, sin comprometer las relaciones estadísticas necesarias para completar análisis más sencillos, como clasificación y resumen. Este modelo es una herramienta analítica popular para evaluar datos basados en texto, así como una variedad de otras fuentes de datos. Se han desarrollado numerosos métodos de modelado de tópicos que consideran todo tipo de relaciones y restricciones dentro de conjuntos de datos (Vayansky & Kumar, 2020). Este modelo se da con el fin de encontrar la semántica oculta en la colección de documentos y agrupar los temas como tópicos. El modelo distribucional como el espacio vectorial, el análisis semántico latente, el modelo semántico latente probabilístico y la asignación de Dirichlet latente se pueden utilizar para derivar la representación del significado de las palabras basándose en el análisis estadístico (Kherwa & Bansal, 2018).

5.12.1 Latent Dirichlet Allocation (LDA). La asignación de Dirichlet latente es un algoritmo ampliamente utilizado en la minería de texto, basado en modelos matemáticos estadísticos bayesianos (Alghamdi & Alfalqi, 2015). Basado a su vez en un modelo de variables ocultas que surge de la interacción de los datos observados con estas. Los datos observados son los textos individuales, mientras que las variables ocultas representan los tópicos de cada documento (Gómez & Fuentes, 2018).

LDA es un modelo probabilístico generativo de un corpus, y las revisiones se representan como mezclas aleatorias sobre K dimensiones latentes, donde cada dimensión se caracteriza por una distribución de palabras (Guo et al., 2017). El modelo descrito es considerado un tipo de modelo de aprendizaje no supervisado. Esto quiere decir que la asignación inicial de los tópicos no depende de una persona que agrupe los datos, si no que dependerá únicamente de la estructura de estos (Sepúlveda Ramírez, 2019).

Para generar cada documento, primero se toma como muestra un K-vector θ que representa la proporción de mezcla de k temas de una distribución previa de Dirichlet $p(\theta|\alpha)$. La variable k definirá la dimensión de esta distribución y, por tanto, también la dimensión de la variable del tema z, pero también representa el número total de temas que se devolverán en el modelo. Es importante señalar que, en este enfoque, se supone que este valor es estático y conocido. Además, una matriz β con dimensiones $k \times V$ parametriza probabilidades de palabras tales que $\beta_{ij}=p(w_j=1|z_i=1)$ donde $i=0, 1, \ldots, K$ y $j=0, 1, \ldots, V$. Cuando $\theta_i \geq 0$ y $\Sigma\, \theta_i = 1$ kí $=1$, una variable de Dirichlet θ de k-dimensionalidad puede ocupar valores en el $(k-1)$-símplex y tiene una densidad de probabilidad en este símplex determinada por la siguiente ecuación (Vayansky & Kumar, 2020).

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{z_1-1} \ldots \theta_k^{z_k-1}$$

LDA se ha utilizado en una variedad de contextos para investigar tendencias en el contenido textual a lo largo del tiempo y está diseñado para analizar una gran cantidad de documentos, cada uno de los cuales contiene potencialmente múltiples temas (Dyer et al., 2017). La razón de la aparición del modelo LDA es mejorar la forma de mezclar modelos que capturan la intercambiabilidad de las palabras y los documentos (Alghamdi & Alfalqi, 2015).

5.13 Text Classification

La clasificación de texto es el proceso de clasificar documentos de texto en un número fijo de clases predefinidas. La aplicación de clasificación de texto incluye filtrado de spam, enrutamiento de correo electrónico, identificación de idioma, etc (Vijayan et al., 2017). En algunas situaciones, definir un conjunto de reglas lógicas utilizando técnicas de ingeniería del conocimiento y basadas en opiniones de expertos para clasificar documentos ayuda a automatizar la tarea de clasificación. Esta clasificación podría dividirse en tres categorías: clasificación de texto supervisada, no supervisada y semi-supervisada basada en el principio de aprendizaje seguido por el modelo de datos (Kowsari et al., 2019).

En general, una técnica de clasificación podría dividirse en enfoques estadísticos y de aprendizaje automático (machine learning). Las técnicas estadísticas satisfacen puramente las hipótesis proclamadas manualmente, por lo tanto, la necesidad de algoritmos es pequeña, pero las técnicas de machine learning se inventaron especialmente para la automatización (Thangaraj & Sivakami, 2018).

La tarea de clasificación de texto consiste en determinar la categoría de cada documento en la colección de acuerdo con categorías predefinidas. Con el aumento de artículos científicos, información de Internet y otros datos en formato de texto, la categorización automática desempeña

un papel importante en la recuperación de información, la extracción de datos y el aprendizaje automático. Los métodos de clasificación comúnmente utilizados son la red neuronal de propagación hacia atrás, los árboles de decisión, regresión logística, random forest, naive bayes y support vector machine (SVM), y especialmente SVM logra un buen rendimiento en la efectividad y estabilidad de la clasificación (Kowsari et al., 2019). Como se describen a continuación.

5.14 Análisis de Sentimientos

El campo del análisis de sentimientos se ocupa del análisis de las opiniones que se encuentran en los documentos. Una de las tareas básicas es la clasificación de sentimientos, donde las unidades de texto se ordenan en clases que corresponden a la positividad, negatividad o neutralidad de la opinión expresada (Radovanović & Ivanović, 2008). Técnicas avanzadas permiten analizar gramaticalmente y descomponer la oración. La minería de opiniones tiene un mercado ávido de conocer, indexar y resumir opiniones en grandes volúmenes de texto con fines de mercadeo y manejo de imagen (Hernández & Gómez, 2013). Muchas de las técnicas de minería de texto presentes en la clasificación del texto también son aplicables para el análisis de sentimientos.

5.15 Modelo Conceptual

Un modelo conceptual es una representación gráfica que expresa el significado de términos y conceptos a través de herramientas descriptivas que permite la comprensión de las variables más influyentes en el comportamiento de un sistema. Los modelos conceptuales se derivan de un proceso de documentación de un problema con el propósito de entenderlo y comunicarlo a las partes interesadas, haciendo uso de textos y gráficos que resuman el modelo en sí (Arango Serna et al., 2016).

6. Metodología

La metodología KDT (Knowledge Discovery in Text) fue utilizada para el desarrollo de este proyecto con el fin de cumplir los objetivos propuestos, ya que estructura el proceso de extracción de información de manera ordenada y simple (Feldman & Dagan, 1995). La metodología presentada incluye una fase adicional centrada en la selección y construcción del modelo conceptual. A continuación, se detallan las fases que la componen.

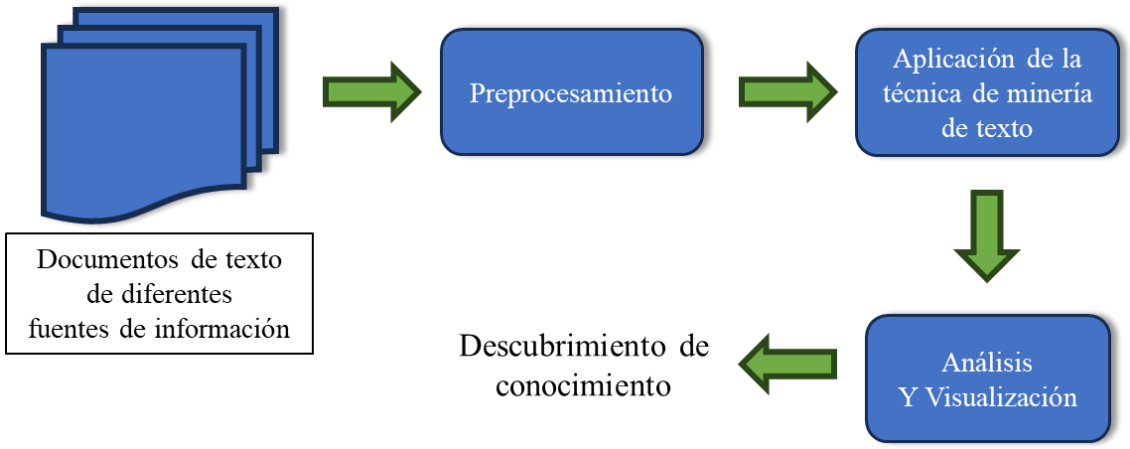
- Recopilación de información: Esta etapa consiste en la creación de una ecuación de búsqueda para la literatura científica. Por otro lado, las palabras clave utilizadas para la investigación en la literatura gris fueron crisis, crisis de los contenedores, causas, pandemia, escasez y transporte marítimo.
- Preprocesamiento de texto: Luego de obtener los documentos, se aplican los procesos de preprocesamiento de texto: tokenización y remoción de stopwords para reducir el ruido en los diferentes documentos.
- Transformación de texto: Tras el preprocesamiento de los documentos, se eliminan las características irrelevantes o redundantes que se han hallado en fases previas, en esta fase es fundamental transformar los datos no estructurados en estructurados para que puedan ser codificados por el algoritmo de minería de texto a utilizar.
- Aplicación de técnicas de minería de texto: La técnica de minería de texto se aplica en esta etapa para obtener información sobre los factores que causaron la crisis de los contenedores. Esta información se utiliza para crear diversos modelos conceptuales.
- Diseño del modelo conceptual: Luego de obtener los resultados de la minería de texto, se diseñan varios modelos conceptuales para representar de forma clara el fenómeno de estudio. Estos modelos buscan ilustrar y comprender con mayor profundidad la crisis de

los contenedores. Su diseño se basa en tres pasos: identificación de los conceptos relevantes, definición de las relaciones entre ellos y representación gráfica del modelo.

En general la metodología KDT se puede resumir en tres fases; preprocesamiento, minería de texto y visualización (Torre, 2017). Como se puede observar en la siguiente figura.

Figura 2.

Representación gráfica, metodología KDT



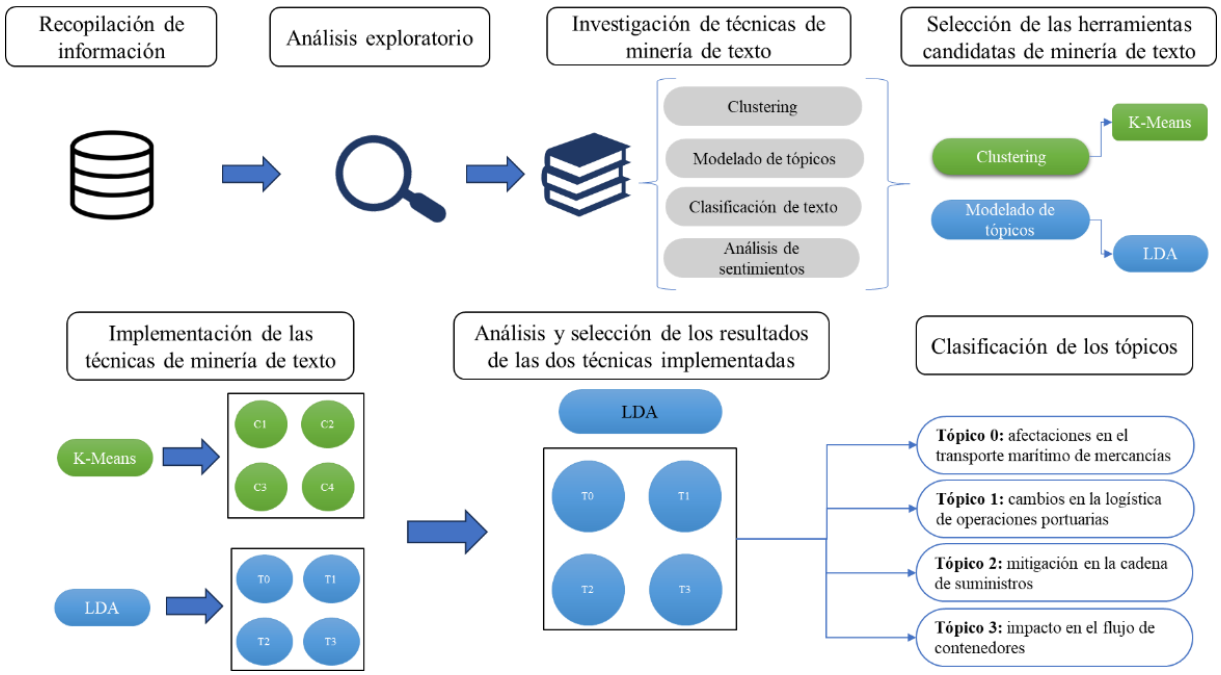
Nota. Adaptado de (ShrihariR & Desai, 2015).

7. Aplicación y Comparación de Técnicas de Minería de Texto

En esta fase, se recopila la información, se implementan y comparan las herramientas de minería de texto para encontrar la técnica más adecuada para la investigación. Posteriormente, se realiza un análisis y categorización de los resultados obtenidos de la técnica seleccionada. En la figura 4 se ilustran las diferentes etapas de este proceso.

Figura 3.

Framework del desarrollo de las técnicas de minería de texto



7.1 Recopilación de la Información

La investigación recopiló documentos científicos de diversas bases de datos de la biblioteca virtual de la UIS, publicados entre 2020 y 2023. Se planteó la ecuación de búsqueda preliminar "container crisis" pero la baja cantidad de resultados (3 en ScienceDirect) llevó a la elección de una ecuación más amplia. Esta última consistió en agregar términos relacionados, tales como "shortage", "pandemic", "shipping", entre otros. La formulación de la ecuación final utilizó ecuaciones específicas para cada recurso, como se muestra en la siguiente tabla:

Tabla 2.

Ecuaciones de búsqueda aplicadas en bases de datos científicas

Base de datos	Ecuación de búsqueda	Número de resultados recuperados	Fecha
Link spring	[(“Container AND (crisis OR shortage) AND shipping AND (covid OR pandemic) AND (Factors OR causes OR reasons)”]]	53	12/10/2022
Knovel	"CONTAINER" AND ("CRISIS" OR "SHORTAGE")	21	12/10/2022
ESCO host: Business source ultime	((container crisis) OR (container shortage)) AND (covid OR pandemic)	63	13/10/2022
Scopus	((container AND crisis OR container AND shortage) AND (covid 19 OR pandemic))	16	13/10/2022
Sciencia direct	container AND (crisis OR shortage) AND shipping AND (covid OR pandemic) AND (factors OR causes OR reasons)	76	19/10/2022

Tras la aplicación de las ecuaciones de búsqueda, se logró recopilar un total de 229 documentos para la investigación. Es importante resaltar que la mayoría de los artículos científicos recopilados se encuentran en inglés, ya que la mayor cantidad de información relevante para el tema de análisis está presente en este idioma. De la colección completa de documentos se llevó a cabo un proceso de selección, excluyendo campos relacionados con la salud, medio ambiente y desarrollo sostenible, dado que estos documentos se centran en temas no pertinentes para la investigación, como los efectos del covid-19 en las personas, niveles de contaminación presentes en la industria naviera, métodos ferroviarios de transporte de contenedores, entre otros. Tras este proceso, se obtuvo un conjunto de 100 documentos científicos, de los cuales se extrajeron sus resúmenes y conclusiones con el propósito de recopilar la información más relevante. Posteriormente, se creó un archivo .csv con una columna denominada 'text' y 100 filas, donde cada fila contiene el resumen y la conclusión de los documentos, esta información se encuentra disponible en el apéndice A.

7.2 Análisis Exploratorio

Después de recopilar la información, se llevó a cabo un análisis exploratorio para evaluar la situación actual. Este análisis se realizó utilizando el lenguaje de programación Python 3 a través de la herramienta en línea Google Colab, haciendo uso principalmente de las bibliotecas de procesamiento de lenguaje natural “NLTK”, así como de las bibliotecas “NumPy” y “Pandas”. El análisis exploratorio, que se encuentra en el apéndice B, se efectuó de la siguiente manera:

7.2.1 Extracción de la Información

Una vez que se obtiene la base de datos en un archivo .csv se extrae la información teniendo en cuenta la codificación Latin-1, que es la predeterminada en este tipo de archivos y la que se debe especificar para Python. Toda la información se extrae como una cadena de caracteres y se almacena en una variable denominada 'data_frame’.

7.2.2 Preprocesamiento

Después de extraer la información de la base de datos y almacenarla en una variable llamada 'data_frame,' se convierte cada palabra en tokens, obteniendo así un total de 68,744. Posteriormente, se efectúa un proceso de limpieza que incluye la eliminación de espacios en blanco, números, símbolos y palabras irrelevantes conocidas como 'stopwords', las cuales están predefinidas en la biblioteca NLTK. Este proceso resulta en un conjunto final de 37,436 tokens.

7.2.3 Análisis y Resultados

Con la información procesada en el paso anterior, se crea un diagrama de frecuencias que muestra las 45 palabras más comunes a través de un gráfico de barras y un gráfico de líneas. Esto permite analizar la información recopilada, como se observa en las figuras 5 y 6:

Figura 4.

Gráfico de líneas, frecuencia de términos del análisis exploratorio

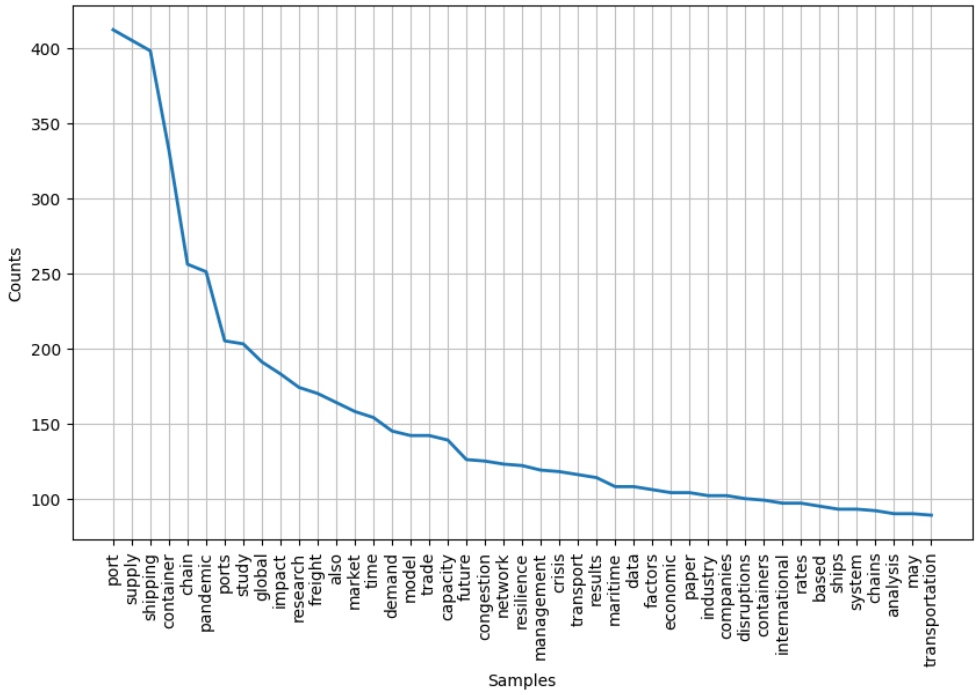
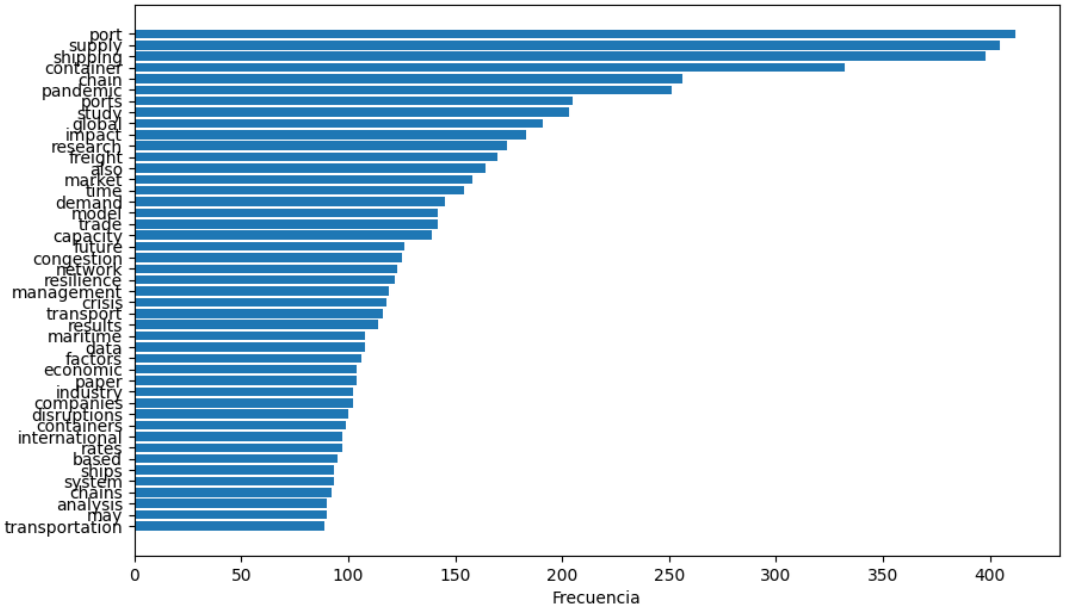


Figura 5.

Gráfico de barras, frecuencia de términos del análisis exploratorio



De los resultados obtenidos, se derivan las siguientes conclusiones; Se observa que las palabras más frecuentes son "port," "supply", "shipping”, "container," y "chain" siendo estas demasiado amplias, la palabra "supply", por ejemplo, podría relacionarse con varios eventos o circunstancias, por este motivo no se logra asociar directamente a un factor que haya causado la crisis de los contenedores. Dicho esto, se observa que, por defecto, un token se asocia a una palabra, lo que se conoce en la literatura como unigramas. Aunque este enfoque podría ser apropiado en otros contextos de investigación, el objetivo es identificar factores relacionados con las causas de la crisis de los contenedores. Por lo tanto, se llegó a la conclusión de que es necesario implementar el uso de bigramas, es decir, dos palabras por token y de esta manera se logra un mayor contexto de la información obtenida.

Por otro lado, la implementación de la minería de texto se lleva a cabo, entre otras aplicaciones, con el propósito de descubrir información no evidente en datos no estructurados. Uno de los propósitos fundamentales de la investigación consiste en explorar la posibilidad de identificar factores relacionados con las causas de la crisis de los contenedores que no hayan sido evidentes en la revisión de literatura. Por lo tanto, a raíz de los resultados obtenidos en el análisis exploratorio, se planteó una revisión de diversas técnicas de minería de texto con el objetivo de seleccionar la técnica adecuada que permita identificar factores no observados, como se describe a continuación.

7.3 Selección de las Herramientas Candidatas de Minería de Texto

En el proceso de investigación sobre las técnicas de minería de texto se realizó una búsqueda mediante una revisión de literatura gris y científica, donde se encontraron diferentes métodos, siendo estos: análisis de sentimiento, clustering, clasificación de texto y modelamiento de topicos.

De los mencionados anteriormente, se descartan las herramientas; clasificación de texto y análisis de sentimiento. Debido a que son herramientas supervisadas en las cuales se establecen palabras predefinidas, posiblemente sesgando y limitando los resultados que se obtendrían en la investigación a parámetros ya seleccionados previamente, evitando encontrar nueva información o profundizar en el tema.

De esta manera se seleccionó el modelo Latent Dirichlet Allocation (LDA), el cual pertenece al modelamiento de topicos, ya que puede identificar temas presentes en los documentos y asignar palabras a esos tópicos de manera probabilística. Lo cual permite que los tópicos emerjan de los datos sin limitarlos por categorías predefinidas. La elección de k-means, la cual pertenece a la técnica de clustering, se fundamenta en su capacidad para agrupar documentos según similitudes, utilizando métricas de distancia para asignar documentos a clústers. Esta elección facilita la formación de agrupaciones de documentos que comparten características en común, permitiendo la posibilidad de obtener factores asociados a las causas de la crisis de contenedores al vincular cada clúster con un tema en particular.

Una vez seleccionadas las herramientas candidatas, se procede a su ejecución y comparación con el objetivo de seleccionar la más adecuada para llevar a cabo la investigación.

7.4 Implementación de las Técnicas de Minería de Texto

En el desarrollo de ambas técnicas de minería de texto, al igual que el análisis exploratorio, se usó la herramienta en línea de Google Colab (Python 3), para la programación de los algoritmos. A su vez, para el desarrollo de las técnicas se implementó la metodología descubrimiento de conocimiento en texto (KDT, por sus siglas en inglés).

7.4.1 Aplicación del Modelo Latent Dirichlet Allocation

En esta fase se describen los diferentes pasos que se implementaron para el desarrollo del algoritmo LDA. Este código se encuentra disponible en el anexo C.

7.4.1.1 Preprocesamiento. Siguiendo la metodología KDT, se inicia por la preparación de los datos para el posterior desarrollo del modelo, los cuales se describen a continuación.

7.4.1.1.1 Extracción de la información. El proceso de extracción de datos no estructurados se realizó de la misma manera que en el análisis exploratorio, tomando la información del archivo .csv, almacenando los datos en una variable “df”.

7.4.1.1.2 Remover puntos, comillas, símbolos y conversión de las palabras en tokens. Para llevar a cabo el proceso de LDA, es necesario realizar una preparación de datos que incluye la eliminación de símbolos y caracteres como puntos, espacios, comas, guiones, paréntesis, comillas simples y dobles, entre otros. Una vez completado este proceso de limpieza, cada palabra se convierte en tokens.

7.4.1.1.3 Remoción de stopwords. Tras la estructuración de las palabras en tokens, se realizó una identificación y eliminación de palabras innecesarias como “tea”, “re”, edu”, entre otros. La biblioteca NLTK cuenta a su vez con una lista de stopwords predefinidas, la cual se fue expandiendo posteriormente con la revisión de las nubes de palabras.

7.4.1.1.4 Proceso de lematización. Se llevó a cabo el proceso de lematización, ya que este permite evitar la aparición de palabras con el mismo significado en su forma singular y plural como "congestion" y "congestions". Esto permite prevenir la duplicidad y mejorar la calidad de los resultados.

7.4.1.1.5 Construcción de bigramas. Dado que hasta el momento la información se encuentra en unigramas y basados en las conclusiones del análisis exploratorio, se realiza la construcción de bigramas como se observa en la figura 7.

Figura 6.

Código de construcción de bigramas en LDA

```
# for Bigram formation
res = [i for j in data_ready
       for i in zip(j[:-1], j[1:])]

# Convertir bigramas a una sola palabra separada por "_"
bigramas_procesados = ['_'.join(bigrama) for bigrama in res]

# Convertir bigramas procesados a un array of unicode tokens
bigramas_procesados_unicode = [token.strip() for token in bigramas_procesados]

# Convertir bigramas procesados a una lista de listas
bigramas_procesados_unicode_list = [[token] for token in bigramas_procesados_unicode]
```

Posteriormente es importante transformar la lista de bigramas procesados a un texto Unicode, ya que LDA solo acepta este estándar de codificación. Además, debido a que los bigramas son una lista de dos tokens es necesario convertirlos a una lista de lista para que el modelo LDA los comprenda. Este proceso se detalla en la figura 8.

Figura 7.

Código de conversión de bigramas a unicode y a lista de listas

```
# Convertir bigramas procesados a un array of unicode tokens
bigramas_procesados_unicode = [token.strip() for token in bigramas_procesados]

# Convertir bigramas procesados a una lista de listas
bigramas_procesados_unicode_list = [[token] for token in bigramas_procesados_unicode]
```

7.4.1.1.6 Creación del diccionario de términos y corpus. Dado que LDA es un modelo matemático, en esta fase se creó un diccionario que represente numéricamente cada uno de los

tokens. Posteriormente, este diccionario se transformó en un corpus mediante una lista de listas de duplas, siendo este el formato requerido para su implementación. Como se observa en la siguiente figura.

Figura 8.

Creación del diccionario de términos y corpus

```
# Crear un diccionario de términos a partir de tus bigramas procesados
id2word = corpora.Dictionary(bigramas_procesados_unicode_list)

# Crear el corpus en el formato adecuado para LDA (una lista de listas de tuplas)
corpus = [id2word.doc2bow(text) for text in bigramas_procesados_unicode_list]
```

7.4.1.1.7 Construcción de matriz tf-idf. Con la intención de mejorar la calidad de los resultados, se implementó la matriz tf-idf para reducir las palabras poco relevantes al implementar su construcción, ya que tf-idf considera el peso de cada palabra mediante el uso de dos enfoques, la frecuencia de un término y en cuántos archivos se puede encontrar dicho término (Hakim et al., 2014).

7.4.1.3.1 Construcción de la métrica coherencia de tópicos. La medición de coherencia se realiza a través de diferentes caminos; a través de un experto que determina la coherencia del resultado, técnicas de análisis cuantitativo y juicio humano dado por un grupo de personas conocedoras del tema (Rincón Ruiz, 2021).

En el modelamiento de tópicos, no existe un resultado único de solución ideal al que aproximarse, por lo que la evaluación requiere del empleo de herramientas matemáticas de aproximación, y por ello se exploran varias alternativas (perplejidad, coherencia). Uno de los estudios más realizados consiste en pruebas sobre la coherencia de los términos que caracterizan un tópico, esta se aplica a las “n” palabras principales dentro de un tópico. Típicamente se define

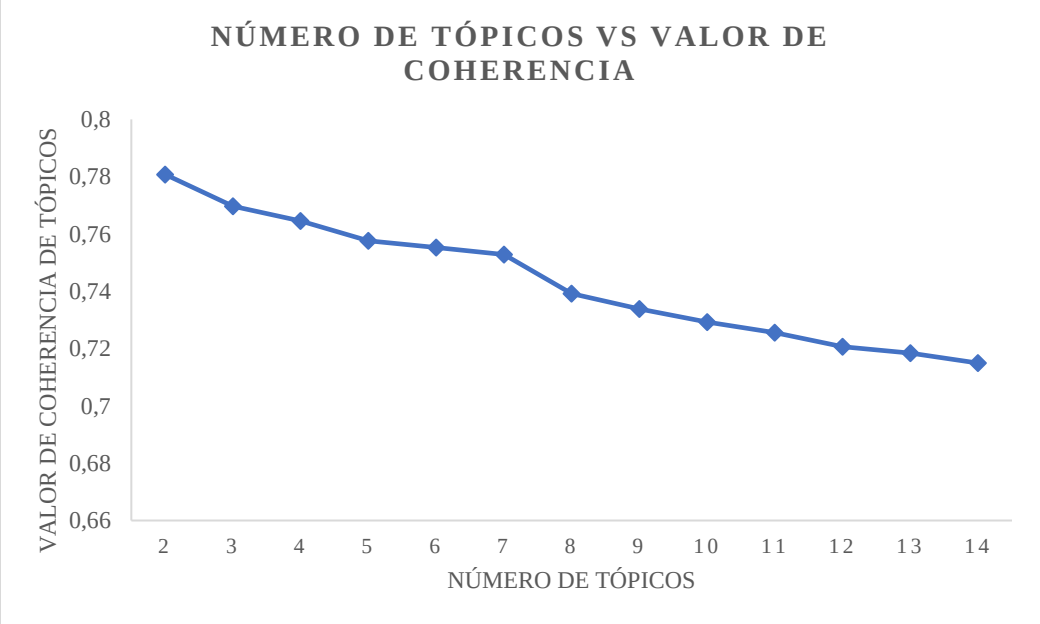
como el promedio de las medidas de similitud de las palabras de un documento, obtenidas por pares (cada palabra con todas las demás). Aquellos modelos que tengan unos topicos más relacionados, tendrán unas medidas de coherencia más altas (Opacua Lomoschitz, 2019). La puntuación oscila entre cero y uno; uno significa que el modelo de tema es perfectamente coherente, y cero significa que no es coherente en absoluto (Rijcken et al., 2021).

Dentro de las métricas de coherencia de tópicos se encontró lo sigueinte: Medición intrínseca C_{umass}, Medida extrínseca C_{uci}, Información mutua normalizada por puntos C_{npmi} y Vector de coherencia C_{v}. Sin embargo, esto no es factible ya que la puntuación de coherencia es una medida no supervisada de temas basada en una fuente de conocimiento a gran escala, no existe una verdad fundamental “mejores temas” (Gui et al., 2019).

Teniendo en cuenta esta información se construyó una gráfica con los valores de coherencia de tópicos, haciendo uso de la medida C_{v}, como se muestra en la siguiente figura.

Figura 9.

Grafica de coherencia de tópicos



7.4.1.1.8 Decisión número de tópicos. Para determinar el número óptimo de tópicos, se llevó a cabo una comparación entre los resultados de 2, 4 y 6 tópicos, buscando un número par que permitiera presentar de manera clara y resumida la información relacionada con los factores asociados a la crisis de los contenedores. Se descartaron valores impares por estar en la vecindad del número par seleccionado. Asimismo, se excluyeron valores superiores a seis, ya que, según la gráfica de coherencia de tópicos, en estos casos la coherencia entre ellos disminuye de manera constante, indicando una menor consistencia temática. A continuación, se presenta el análisis detallado de estas tres opciones:

En el resultado de dos (2) tópicos, si bien en la gráfica de coherencia de tópicos se observa un valor de 0,7808 siendo este el mayor, se concluye que la información es insuficiente, puesto que, aunque es posible categorizarlos al observar que el primer tópico hace referencia a la congestión de los puertos y el segundo a la disrupción en la cadena de suministro, no se lograría obtener suficiente información que permita identificar temas adicionales.

Figura 10.

Resultados de LDA con 2 tópicos

En el resultado de cuatro (4) tópicos, se destaca una mayor claridad en relación con los temas presentes en cada clúster. Estos tópicos abarcan una cantidad suficiente de temas que permiten identificar de manera más precisa y resumida los factores asociados a la crisis de contenedores. Además, la métrica de coherencia de tópicos indica un desempeño similar al de 2 tópicos, con un valor de 0,7697.

Figura 11.

Resultados de LDA con 4 tópicos

En los resultados de seis (6) tópicos, se observa que la información no es tan fácil de categorizar, ya que no se puede discriminar con la misma facilidad en comparación a los casos anteriores. Cabe mencionar que, si bien se analizaron una mayor cantidad de tópicos (8, 10 y 12), estos tópicos se encuentran en el apéndice D, se observó el mismo inconveniente de categorización

de temas que en los 6 tópicos, además el nivel de significancia de los bigramas presentes en varias nubes de palabras se hacían cada vez menor, siendo estos más pequeños.

Figura 12.

Resultados de LDA con 6 tópicos

Debido a esto, se optó por tomar 4 tópicos para ser analizados mediante el modelo LDA, ya que la información se presenta de manera más clara, facilitando a su vez la comprensión de los temas.

7.4.1.2 Aplicación de la técnica de minería de texto. Con la elección de cuatro (4) tópicos se presenta a continuación como se construyó el modelo.

7.4.1.2.1 Construcción del modelo. Con los datos preprocesados y en el formato adecuado se construye el modelo LDA teniendo en cuenta los siguientes parámetros:

- Num_topics: Este parámetro representa el número de tópicos del modelo, se establece en un total de 4 tópicos.
- Passes: Es la cantidad de veces que el modelo se entrena en el corpus con el fin de estabilizar los resultados, de esta manera se elige una iteración de 100 veces.
- Random_state: Es la semilla aleatoria para el entrenamiento del modelo, en este caso se eligió la semilla cero.

Figura 13.

Código del modelo LDA

```
# modelo LDA
lda_model = gensim.models.LdaMulticore(corpus=corpus,
                                       id2word=id2word,
                                       num_topics=4,
                                       passes=100,
                                       random_state=0)
```

7.4.1.3 Resultado del modelo LDA. Los resultados del modelo LDA son los pesos asignados a cada bigrama en cada uno de los 4 tópicos, como se puede observar en la siguiente figura.

Figura 14.

Pesos asignados a los bigramas

```
(0,
 '0.008*"covid_pandemic" + 0.002*"long_term" + 0.002*"port_resilience" + 0.002*"import_export" + 0.002*"international_trade" + 0.002*"chain_disruption" +
0.001*"chain_resilience" + 0.001*"shipping_market" + 0.001*"automate_terminal" + 0.001*"shipping_link"')
(1,
 '0.004*"global_supply" + 0.003*"shipping_network" + 0.003*"freight_rate" + 0.002*"shipping_industry" + 0.002*"port_operation" + 0.002*"decision_make" +
0.002*"container_port" + 0.002*"hub_port" + 0.001*"container_logistic" + 0.001*"cause_covid"')
(2,
 '0.019*"supply_chain" + 0.007*"container_shipping" + 0.003*"shipping_company" + 0.002*"global_economy" + 0.002*"shipping_line" + 0.001*"shipping_capacity" +
0.001*"foreign_trade" + 0.001*"liner_company" + 0.001*"demand_shock" + 0.001*"risk_mitigation"')
(3,
 '0.005*"port_congestion" + 0.003*"impact_covid" + 0.002*"container_shortage" + 0.002*"empty_container" + 0.001*"disruption_impact" + 0.001*"container_ship" +
0.001*"negative_impact" + 0.001*"liner_shipping" + 0.001*"risk_management" + 0.001*"ship_port"')
```

Según la figura se puede concluir que, por ejemplo, la palabra más dominante en el tópico "0" es "covid_pandemic" con un peso de 0.008, seguido de "long_term" con un peso de 0.002 siendo el mismo valor para "port_resilience", "import_export", "international_trade" y "chain_disruption". Para el tópico "1" la palabra con mayor peso es "global_supply" con un valor de 0.004, seguido de "shipping_network" y "freight_rate" con un peso de 0.003. En el caso del tópico "2" la palabra más dominante con un peso de 0.019 es "supply_chain", el siguiente valor es de 0.007 asignado a "container_shipping" y luego "shipping_company" con un peso de 0,003. Finalmente, en el tópico "3" la palabra más relevante es "port_congestion" con un peso de 0.005 seguido de "impact_covid" con un valor de 0.003 y de "container_shortage" y "empty_container" con un peso de 0.002. Los valores que se asignan a cada palabra indican su importancia en cada uno de los topicos y por ello son representados con el mayor tamaño en las nubes de palabras, como se observa en la siguiente figura.

Figura 15.

Nubes de palabras, resultado del modelo LDA

Los resultados obtenidos a través de la aplicación del modelo LDA revelaron asignaciones de tópicos coherentes en los documentos, permitiendo una identificación precisa de las temáticas predominantes. Además, el análisis de las distribuciones de palabras en los tópicos proporcionó una comprensión detallada de las palabras clave asociadas a cada tema, contribuyendo así a una mejor interpretación de la estructura latente del corpus.

7.4.2 Aplicación Técnica K-Means

En esta fase se describen los diferentes pasos que se implementaron para el desarrollo del algoritmo k-means. Este código se encuentra disponible en el apéndice E.

7.4.2.1 Preprocesamiento. Al igual que en el algoritmo LDA, es necesario realizar este proceso, teniendo en cuenta las necesidades de este modelo, las cuales se describen a continuación.

7.4.2.1.1 Extracción de la información. El proceso de extracción de la información de k-means se realizó de la misma manera que LDA y el análisis exploratorio, almacenando la información en una variable llamada “df”.

7.4.2.1.2 Creación de Tokens. Con la información extraída y almacenada en la variable “df”, se procede a la creación de tokens con las palabras obtenidas de la base de datos.

7.4.2.1.3 Remover puntos, comillas y símbolos. Cabe mencionar que, aunque esta fase en el proceso LDA se lleva a cabo previamente a la creación de tokens, en el caso del proceso de k-means, realizarla de esta manera alteraba el orden establecido de los documentos almacenados en la variable “df”. Por lo tanto, la eliminación de caracteres como puntos y símbolos ocurre después del proceso de transformar las palabras en tokens.

7.4.2.1.4 Remoción de las stopwords. En este proceso se conservó la misma lista de stopwords definida en la técnica de LDA. Ya que la base de datos utilizada sigue siendo la misma.

7.4.2.1.5 Proceso de lematización. Este proceso se lleva a cabo con la misma intención que fue desarrollada en la técnica LDA, evitar la duplicidad y mejorar la calidad de los resultados.

7.4.2.1.6 Creación de bigramas. Para trabajar con los datos es esencial transformar los tokens, originalmente unigramas, en bigramas y mantener estos últimos en una lista por cada documento, este proceso se detalla en la siguiente figura.

Figura 16.

Creación de bigramas en K-Means

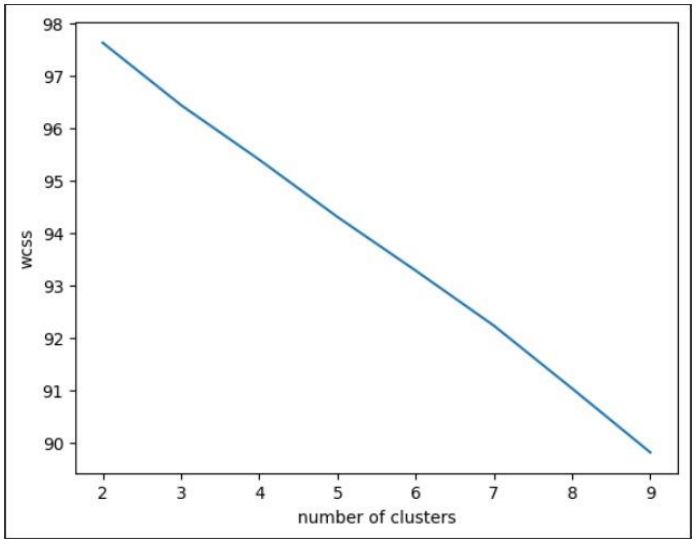
```
# Formación de Bigramas
res = ['_'.join(bigrama) for bigrama in zip(dfa_l2[:-1], dfa_l2[1:])]

# Convertir bigramas a una sola palabra separada por "_"
bigramas_procesados = [''.join(bigrama) for bigrama in res]
```

7.4.2.1.7 Construcción de matriz tf-idf. Durante el desarrollo de la técnica de k-means la generación de la matriz tf-idf es un paso esencial, ya que por medio de esta se transforman las palabras en números, permitiendo al algoritmo su interpretación. Además, tf-idf permite mejorar la calidad de los resultados, filtrando los tokens irrelevantes y acentuando la importancia de aquellos términos que son más distintivos en cada documento.

7.4.2.1.8 Decisión inicial número de clústers. Durante el proceso para determinar la cantidad adecuada de clústers, se exploraron diversas herramientas, siendo el método del codo el más comúnmente utilizado en la literatura. Sin embargo, en la construcción del código, este método no pudo aplicarse de manera efectiva, ya que en la gráfica resultante no se evidencio ningún punto de inflexión, como se puede observar en la figura 18.

Figura 17.

Método del codo

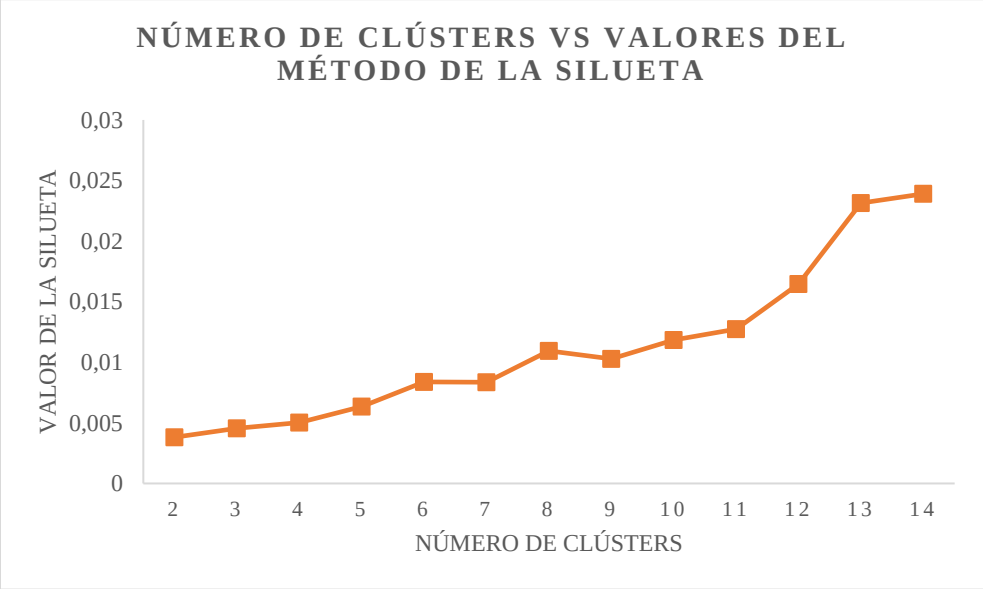


De esta manera, se optó por una alternativa conocida como el método de la silueta. Este método se utiliza para ver la calidad del grupo y su fuerza, qué tan bueno es un objeto para colocarlo en un grupo. Se trata de una combinación de métodos de cohesión y separación. El valor de la silueta varía entre 0 y 1; cuanto menor sea el valor obtenido, más cerca estará el objeto del grupo inapropiado. El conjunto de datos del grupo apropiado debe tener un valor cercano a 1 (Humaira & Rasyidah, 2020).

Aunque el método de la silueta sugirió un número óptimo de 14 clústers, como se observa en la figura 21, se seleccionó una cantidad de cuatro clústers con el objetivo de mantener una relación coherente con la cantidad de tópicos obtenidos mediante la técnica LDA.

Figura 18.

Método de la silueta



7.4.2.2 Aplicación de la técnica de minería de texto. Con la elección de cuatro (4) clústers se presenta a continuación como se construyó el modelo.

7.4.2.2.1 Construcción del modelo. Durante el desarrollo del modelo de k-means, es necesario configurar ciertos parámetros de ejecución, como el número de clústers, el número de iteraciones y la semilla aleatoria. La configuración seleccionada consta de cuatro (04) clústers, conforme a lo planteado anteriormente, con un número de iteraciones establecido en 50. Esta elección se fundamenta en la necesidad de un número significativo de iteraciones para estabilizar los resultados del algoritmo, y tomando como semilla aleatoria el número 0, como se observa en la figura 20.

Figura 19.

Construcción del modelo K-Means

```
Numero= 4

#k means clustering
kmeans = KMeans(n_clusters = Numero, n_init=50, random_state=0 )
kmeans.fit(conteo)
km_result=kmeans.predict(conteo)
df['cluster_group']=pd.Series(km_result)
```

Es importante mencionar que la distancia utilizada para la construcción del modelo fue la distancia euclidiana, la cual es la predeterminada en la importación para la clase k-means del módulo sklearn.cluster.

7.4.2.3 Resultado del modelo k-means. Después de ejecutar el modelo de k-means, los resultados se observan inicialmente a través de la siguiente figura.

Figura 20.

Resultados de asignación de los clústers a cada documento

	text	token	token_list	cluster_group
0	The coronavirus disease 2019 or Covid-19 pande...	coronavirus_disease disease_covid covid_pandem...	[coronavirus_disease, disease_covid, covid_pan...	1
1	Facing the shortage of storage space of contai...	facing_shortage shortage_storage storage_space...	[facing_shortage, shortage_storage, storage_sp...	2
2	Currently the shipping industry is at a cross...	currently_shipping shipping_industry industry_...	[currently_shipping, shipping_industry, indust...	1
3	Due to the COVID-19/Omicron pandemic and the t...	due_covid covid_omicron omicron_pandemic pande...	[due_covid, covid_omicron, omicron_pandemic, p...	3
4	US AGRICULTURE EXPORTERS are find- ing it hard...	us_agriculture agriculture_exporters exporters...	[us_agriculture, agriculture_exporters, export...	3
...	...	...	...	...
95	Shipping lines face various marketing risks w...	shipping_lines lines_face face_various various...	[shipping_lines, lines_face, face_various, var...	3
96	This study deals with the dynamic interactions...	deals_dynamic dynamic_interactions interaction...	[deals_dynamic, dynamic_interactions, interact...	3
97	A container shipping network connects coastal ...	container_shipping shipping_network network_co...	[container_shipping, shipping_network, network...	2
98	The COVID-19 outbreak has had a serious effect...	covid_outbreak outbreak_serious serious_effect...	[covid_outbreak, outbreak_serious, serious_eff...	3
99	A case study in the logistics department of an...	logistics_department department_original origi...	[logistics_department, department_original, or...	3

100 rows × 4 columns

Los resultados de la anterior figura evidencian la asignación de cada documento a uno de los cuatro (04) clústers, que van desde el numero “0” al “3” como se observa en la columna

"cluster_group". Por ejemplo, el documento "0" es asignado al clúster "1", el documento "1" es asignado al grupo "2" y así sucesivamente hasta el último documento denominado "99" asignado al clúster "3".

Figura 21.

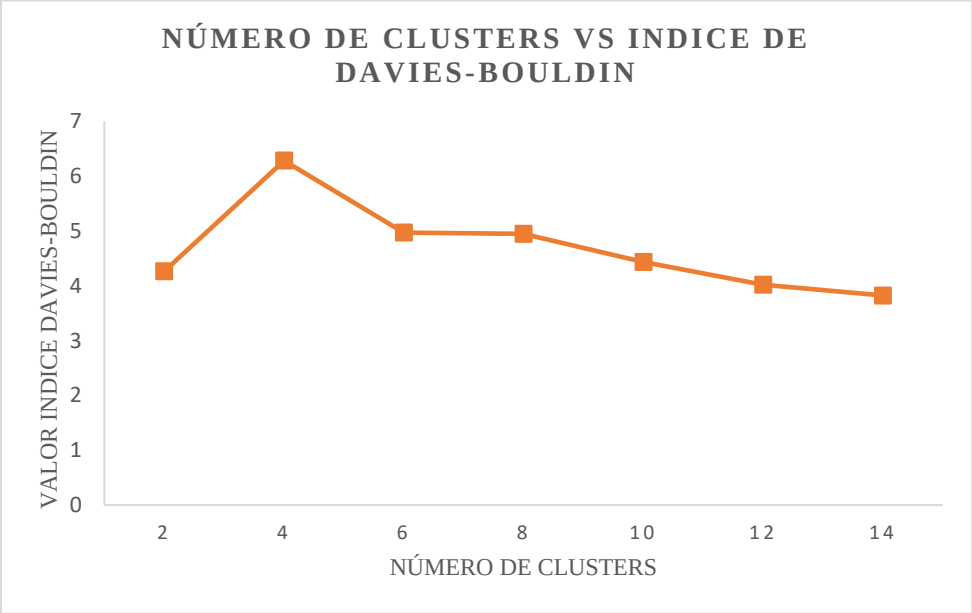
Nubes de palabras, resultado del modelo K-Means

Por otro lado, al examinar los resultados presentes en la figura 22, se observa principalmente duplicidad de palabras en distintos clústers. Esta redundancia se puede explicar debido al funcionamiento de k-means. En este contexto, cada documento posee una serie de características representadas por bigramas, y debido a que la agrupación en clústers se realiza por documento y no por bigramas, si varios documentos en diferentes clústers comparten los mismos bigramas, y estos tienen un peso significativo, se reflejarán en múltiples clústers. Esta conclusión se basa en los resultados observados en las nubes de palabras de la figura anterior.

7.4.2.3.1 Construcción de la métrica index Davies Bouldin. El índice de Davies Bouldin se emplea como medida de la validez de los clústers en un método de agrupamiento como k-means (Mughnyanti et al., 2020). El índice DB es una función de la relación entre la suma de la dispersión dentro del grupo y la separación entre conglomerados (Y. Liu et al., 2011). Si la distancia entre grupos es máxima, significa que las características de cada grupo son pequeñas, por lo que las diferencias entre grupos son más evidentes. Si la distancia es mínima dentro del grupo, significa que cada objeto en el grupo tiene un alto nivel de similitud. Un valor pequeño de esta evaluación indica un buen resultado de agrupación (Y. Liu et al., 2011). De esta manera se construye la gráfica de valores de este índice, que se ilustra a continuación.

Figura 22.

Gráfica Índice Davies Boulding



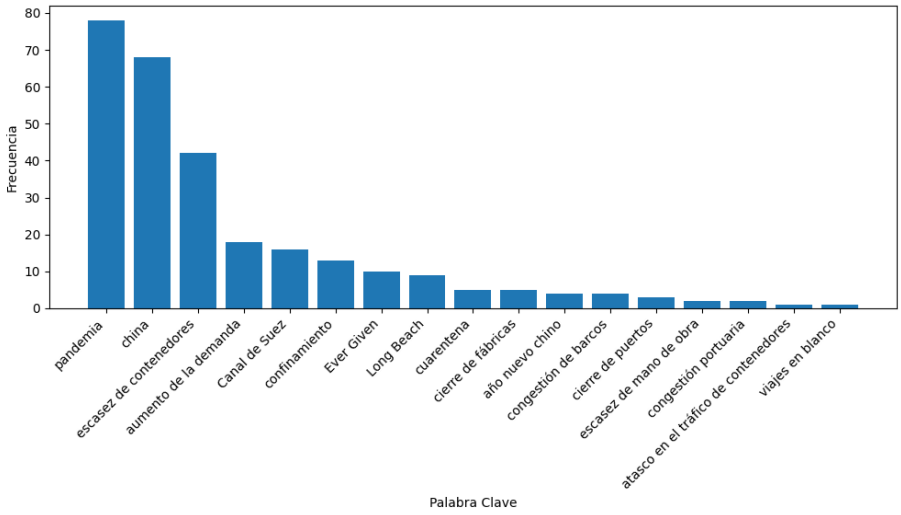
De acuerdo con la gráfica, se observa que para un valor de 4 clusters, el índice de Davies-Boulding es igual a 6,2815 lo cual es un valor bastante alejado de cero (siendo este último valor el óptimo) lo cual indica que no es un buen resultado de agrupamiento.

7.4.3 Web Scraping

Para analizar la información que se encuentra en la literatura gris, se realizaron búsquedas en diferentes sitios web en español acerca de la crisis de los contenedores, una vez identificadas estas páginas se almacenaron sus links en un archivo .txt con un total de 100 enlaces, presente en el apéndice F. Tras la recolección de la información y a diferencia de las técnicas de LDA y k-means, se establecen palabras predefinidas con el fin de elaborar un diagrama de frecuencia con los factores ya encontrados en la revisión de literatura como “aumento de la demanda”, “cierre de fábricas”, “congestión portuaria”, “pandemia”, “escasez de contenedores”, entre otros. Como se ilustra en la siguiente figura.

Figura 23.

Gráfico de frecuencias de Web Scraping



Cabe destacar que para la implementación de esta técnica a diferencia de las anteriores no es necesario la fase de preprocesamiento, ya que consiste en un algoritmo más simple en el que se itera a través de los links, encontrando palabras definidas y calculando su frecuencia. De esta

manera, se busca observar cuáles fueron los factores que más se mencionan en la literatura gris. Esto se puede observar en la siguiente nube de palabras.

Figura 24.

Nube de palabras del Web Scraping

A partir de los enlaces recopilados de la literatura gris y, mediante el análisis de la nube de palabras y el gráfico de frecuencia, se puede concluir que los factores predominantes son claros. La pandemia (del covid-19) se ratifica como el principal desencadenante de la crisis de los contenedores. En los resultados también se hace evidente la participación de China, siendo el país donde se originó la pandemia y se implementaron la mayor cantidad de restricciones, a su vez los puertos chinos representan la mayor cantidad de exportaciones de mercancías por contenedores a nivel mundial (Orús, 2023). Estos factores no solo afectaron a dicho país, también a la economía y el comercio de manera global. Estos hechos, junto con la notable escasez de contenedores, como se corrobora en los hallazgos, generaron la crisis de los contenedores.

7.5 Selección de la Herramienta de Minería de Texto

Con la intención de seleccionar la herramienta adecuada para llevar a cabo la investigación entre k-means y LDA, se propuso una comparación de la calidad de los resultados de ambas técnicas. Es importante destacar que conceptualmente ambas técnicas son diferentes, siendo k-means un modelo determinístico que emplea un análisis dimensional basado en distancias. Mientras que LDA es un modelo probabilístico que utiliza la variable Dirichlet para asignar pesos a diferentes tokens y agruparlos en distintos tópicos. La diferencia en el funcionamiento de ambos modelos dificulta encontrar una métrica en común para su comparación. En consecuencia, se optó por identificar métricas de calidad específicas para cada modelo y evaluar su desempeño en relación con dichas métricas individuales. Esta comparación se complementa por medio de un análisis visual de las nubes de palabras de ambas técnicas, donde se observó la duplicidad de bigramas presente en los clústers de k-means aun cuando se realizó el proceso de lematización. El modelo LDA por su parte, no presentó esta duplicidad.

En el análisis de las métricas de ambos métodos se observó que, en el caso de k-means el índice de Davies Bouldin arrojó un valor de 6.2815 para cuatro clústers, el cual es considerablemente distante de cero, este índice como se mencionó anteriormente permite evaluar la similitud de las palabras dentro de un grupo, refleja que valores más pequeños indican una mayor similitud entre las palabras que componen un clúster. En contraste, un valor más alto sugiere que los clústeres no están bien definidos y separados. La obtención de un valor tan elevado en esta métrica revela una notable falta de cohesión y separación entre los clústers en el método k-means.

Por otro lado, el resultado 0.7646 obtenido de la métrica de coherencia de tópicos en el modelo LDA, evidencia un valor significativamente alto. Esta métrica, que oscila entre 0 y 1, dictamina que, a medida que el valor aumenta, las palabras dentro de un tópico presentan fuertes

conexiones. Este resultado indica la calidad en la identificación y definición de los tópicos generados por el modelo, ya que la coherencia semántica entre las palabras facilita una interpretación más clara y significativa de cada tópico.

Teniendo en cuenta el desempeño de la métrica de LDA y que esta técnica no presenta duplicidad en sus resultados, se optó por seleccionar el modelo LDA como la herramienta de minería de texto para la presente investigación.

7.6 Clasificación de los tópicos

Tras la selección de la técnica LDA se procede a categorizar cada uno de los tópicos, esto debido a que, al ser un algoritmo no supervisado el programa asocia los diferentes datos por parámetros ocultos al encontrar similitudes. El algoritmo LDA asocia dichos bigramas, pero es deber de los investigadores categorizar cada uno de los resultados. De esta manera se clasifican a continuación.

7.6.1 Tópico I

Los principales bigramas presentes en este tópico como “covid_pandemic”, “chain_disruption e “impact_pandemic”, demuestran que hay una tendencia hacia efectos negativos globales, otros bigramas más específicos como “long_term”, “freight_transport”, “hub_port”, “port_logisitc”, “port_shipping”, “international_maritime”, “transportation_cost”, “interntaional_cost”, “demand_supply”, permiten identificar una tendencia hacia el transporte marítimo de mercancías, y palabras como “impact_pandemic”, “significant_impact”, “cause_disruption”, se pueden asociar a diversas afectaciones. De esta manera, el tópico 01 se etiquetó como “afectaciones en el transporte marítimo de mercancías”.

Figura 25.

Tópico I, Afectaciones en el transporte marítimo de mercancías

7.6.2 Tópico II

Los bigramas presentes en este tópico, como “global_supply”, “shipinnig_indsutry”, “freigth_rate”, “shopping_line”, “port_operation”, “import_export”, si bien, todos muestran factores importantes a nivel global sobre las operaciones portuarias de envío de mercancías, no tienen una connotación negativa. Al observar bigramas más específicos aparecen términos como “disruption_supply”, “adverse_event”, “adverse_efffect”, “covid_outbreak”, “affect_port” representando un cambio negativo, a su vez términos como “desicion_make”, “policy_maker”, “make_estrategic”, “resilience_performance”, “capacity_management”, “shipping_link”, se pueden asociar hacia la logística. Por tanto, el tópico 02 se etiquetó como “cambios en la logística de operaciones portuarias”.

Figura 26.

Tópico II, Cambios en la logística de operaciones portuarias

7.6.3 Tópico III

Palabras como “supply_chain”, “conteiner_shipping” y “shipping_network” son los bigramas principales de esta nube de palabras y se centran en la cadena de suministros, estos se pueden alinear hacia la mitigación de los efectos causados por la crisis gracias a bigramas más específicos como, “chain_resilience”, “capacity_port”, “risk_mitigation”, “network_conectivity”, “optimal_solution”, “resilencie_covid”. De esta manera, el tópico 03 se categoriza como “mitigación en la cadena de suministros”.

Figura 27.

Tópico III, Mitigación en la cadena de suministros

7.6.4 Tópico IV

En el análisis de este tópico y basados en la información obtenida de la revisión de literatura, palabras como “port_congestion”, “impact_covid”, son factores principales que afectaron la crisis. Observando otros bigramas como “empty_container”, “container_shortage”, “blank_sailing” que reflejan un impacto directo en el flujo de contenedores. Se puede categorizar el tópico 04 como “impacto en el flujo de contenedores”.

Figura 28.

Tópico IV, Impacto en el flujo de contenedores

8. Modelo Conceptual

Con los resultados y la categorización de cada uno de los topicos del modelo LDA se procede al desarrollo del modelo conceptual, el cual se planteó de diversas maneras con la intención de reflejar los resultados hallados de la minería de texto y la revisión de la literatura, permitiendo así una mejor comprensión del fenómeno. Se realizo un total de cuatro modelos conceptuales los cuales se describen a continuación.

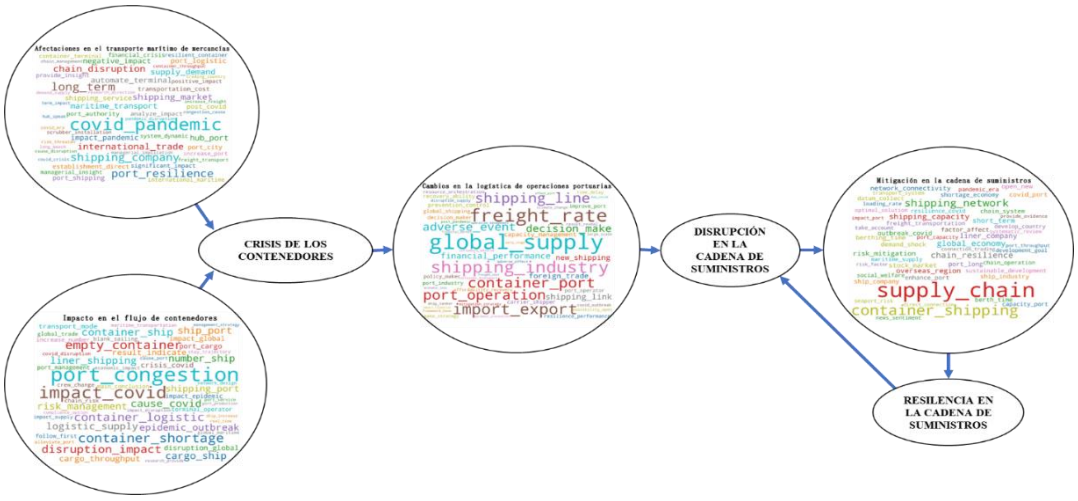
8.1 Construcción del primer modelo conceptual

Tras etiquetar cada una de las nubes de palabras del modelo LDA se encontró que los tópicos “impacto en el flujo de contenedores” y “afectaciones en el transporte marítimo de mercancías” generaron o afectaron la crisis, por tanto, se representa esta influencia por medio de una flecha que tiene conexión directa desde cada uno de los tópicos hacia la primera fase, denominada “crisis de los contenedores”.

Posteriormente, la crisis afecta directamente la cadena de suministros, generando una serie de efectos negativos representados en el tópico “cambios en la logística de operaciones portuarias”, el cual se asocia directamente por medio de una flecha a la segunda fase, denominada “disrupción de la cadena de suministros”. Finalmente, el tópico “mitigación en la cadena de suministros” nace como una respuesta a los efectos negativos presentados en la disrupción, el cual se relaciona directamente y por medio de una flecha con la última fase, denominada “resiliencia en la cadena de suministros”. Dicha resiliencia contrarresta los efectos de la disrupción, representado por medio de una flecha desde esta fase hacia la anterior. Como se observa en la siguiente figura.

Figura 29.

Primer modelo conceptual



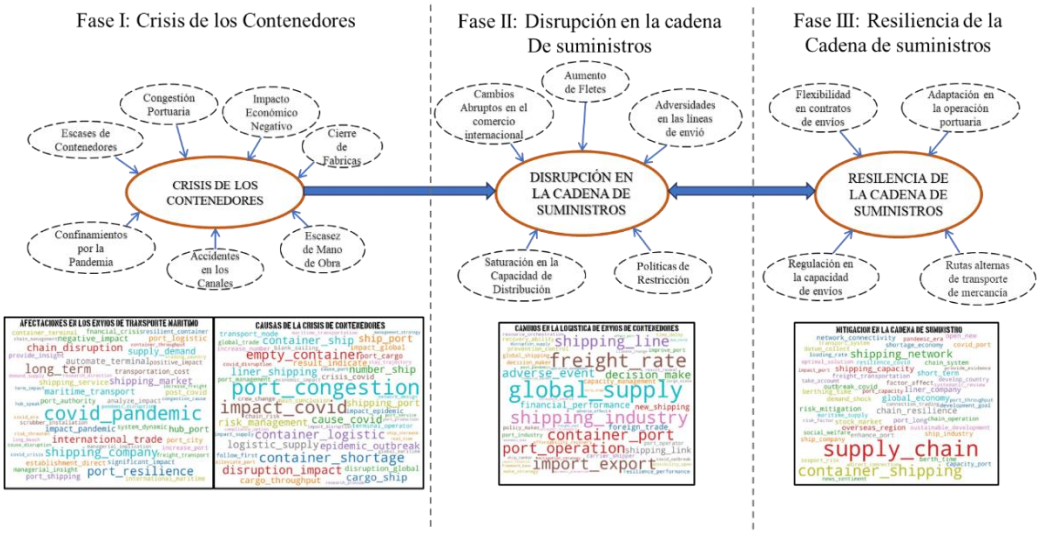
8.2 Construcción del segundo modelo conceptual

Teniendo en cuenta la relación que existe entre la resiliencia y la disrupción de la cadena de suministros descrita en el modelo anterior, se observó un bucle entre las dos fases y el tópico "mitigación en la cadena de suministros", esto se intercambió por una flecha de sentido doble, debido a la relación que ambas fases tienen, ya que la respuesta a la disrupción género directamente la resiliencia, y esta a su vez, reduce los efectos nocivos de la disrupción en la cadena de suministros.

Por otro lado, con la finalidad de enfocar los distintos factores relevantes en cada fase, se han propuesto una serie de nodos, representados por elipses mediante líneas punteadas, respaldados por la revisión de la literatura. Cada nodo está directamente vinculado con sus respectivas fases a través de flechas. Este modelo se planteó como una necesidad de ilustrar de una manera más clara los factores y la separación de las fases, conservando los resultados de la minería de texto en la parte inferior y manteniendo así la relación existente en el modelo anterior. Como se observa en la siguiente figura.

Figura 30.

Segundo modelo conceptual

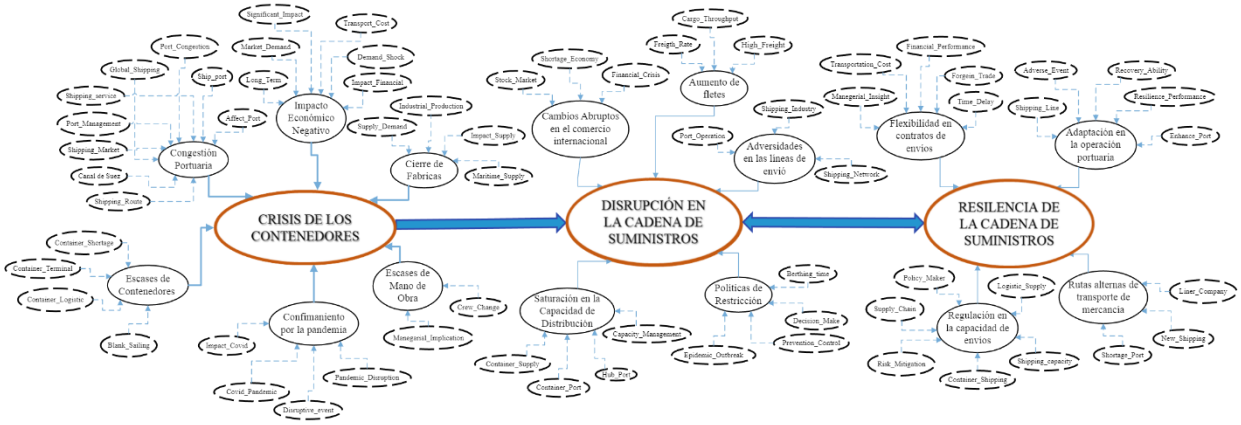


8.3 Construcción del tercer modelo conceptual

Con el fin de generar una relación directa con los resultados de la minería de texto (modelo LDA y webscrapping) se extrajeron los bigramas presentes en cada uno de los tópicos, vinculándolos directamente a los nodos y dichos nodos guardan su relación directa con las distintas fases, como en el anterior modelo. Esto se realizó ya que la minería de texto transmite información más detallada del fenómeno. Los bigramas se ilustraron como elipses con línea punteada, transformando los nodos (del modelo anterior) a elipses de línea continua. Como se observa a continuación.

Figura 31.

Tercer modelo conceptual



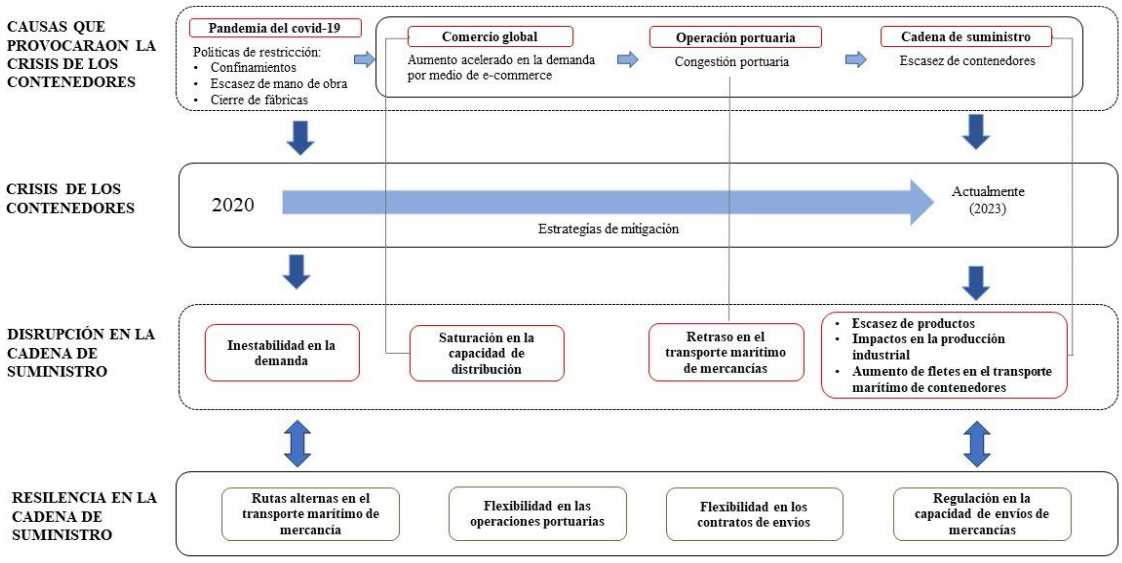
8.4 Construcción del cuarto modelo conceptual

Este modelo conceptual se planteó con el fin de sintetizar los modelos anteriores. Se utilizaron fechas para indicar el flujo de los eventos presentes en la crisis y a quienes se afectaron directamente, manteniendo los mismos conceptos previamente planteados y describiendo cómo la pandemia del covid-19 afectó directamente el comercio global, la operación portuaria y por ende la cadena suministros. De esta manera generando la crisis de los contenedores provocando a su vez la disrupción en la cadena de suministros. Efectos como el comercio global son directamente

asociados con la saturación en la capacidad de distribución, la pandemia del covid-19 con la inestabilidad de la demanda, la operación portuaria con el retraso en el transporte marítimo de mercancías y la cadena de suministros con efectos como; la escasez de productos, los impactos en la producción industrial y el aumento de fletes en el transporte de contenedores. Manteniendo la relación recíproca entre la disrupción en la cadena de suministros y su resiliencia. Como se ilustra a continuación.

Figura 32.

Cuarto modelo conceptual



Nota: Adaptado de Notteboom et al. (2021)

8.5 Análisis de las Distintas Fases que Comprenden el Fenómeno.

Tras la identificación de las diferentes fases presentes en los modelos conceptuales, y con el objetivo de profundizar en la comprensión y contextualización de los elementos, se plantea un análisis de cada una de ellas, las cuales se describen a continuación.

8.5.1. Crisis de los Contenedores

La pandemia ha frenado la actividad económica a nivel mundial, y uno de los sectores más afectados es el comercio internacional. La interrupción en las actividades de los trabajadores portuarios y camioneros debido al covid-19 ha dificultado la descarga y distribución de productos, lo que ha provocado un retraso en la entrega de bienes y ha impactado negativamente en la economía (Goodman & Chokshi, 2021). Lo anteriormente mencionado generó una mezcla de problemas que alteraron el flujo económico global, la demanda y toda la cadena de suministros, provocando amplios déficits para las industrias navieras y dificultades para los compradores de materias primas y consumidores a nivel internacional, generando así la disrupción en la cadena de suministros. Todo esto se denominó la crisis de los contenedores.

8.5.2 Disrupción de la Cadena de Suministros

Siendo esta un efecto directo y un agravante de la crisis de los contenedores, esta fase también tiene sus propios efectos asociados a ella. La alta demanda de contenedores vacíos ha provocado una escasez que ha elevado su precio. Esto ha encarecido el transporte de mercancías, lo que ha impactado negativamente en el consumidor final (Orquera, 2022). El alto costo en los fletes marítimos de envió de contenedores terminaron perjudicando la cadena de suministros global, perpetuando y agravando esta crisis. Las políticas de restricción junto con las adversidades en las líneas de envío aumentaron el riesgo de que los cargamentos pudieran quedar en espera por un tiempo indefinido (Watt, 2022), generando una incertidumbre en toda la cadena de suministros.

8.5.3 Resiliencia en la Cadena de Suministros

Tras el impacto generado por la crisis de los contendores, las empresas involucradas en toda la cadena de suministros empezaron a generar diferentes planes de mitigación, teniendo presente a su vez la crisis financiera del 2008 y 2009. De esta manera se ejecutaron planes de

mitigación como lo fueron la flexibilidad de contratos de envíos y la regulación en la capacidad de envíos, reduciendo la cantidad de cargamentos que se generaban con el fin de mitigar el riesgo y evitar pérdidas. Otras medidas de resiliencia fueron la flexibilidad de contratos de envió, las cuales se plantearon con el fin de dar un margen más amplio de entrega a los clientes afectados por la crisis, y la implementación de rutas alternas de transporte de mercancía (Notteboom et al., 2021).

8.6 Conclusiones Referentes a los Modelos Conceptuales

Finalmente, en el desarrollo de los modelos conceptuales se identificó todo el proceso de la crisis de los contenedores siendo estos; su concepción, su efecto directo y su respuesta. Mediante el proceso de minería de texto se obtuvieron resultados como, “blank sailings”, “port congestión”, “container shortage”, entre otros, cada uno de ellos siendo asociado a la crisis de los contenedores o algunos de los efectos que se encontraron. Mediante la revisión de literatura realizada en las etapas tempranas del proyecto, se tenía un conocimiento de los factores que se asociaban a las causas de la crisis de los contenedores, de ellos fue evidente su conexión con los resultados obtenidos de la técnica LDA. Basado en esto, se plantearon los tres primeros modelos ilustrando la conexión entre los resultados obtenidos de la minería de texto y la información mediante la revisión de literatura. El último modelo se planteó con la intención de sintetizar la información obtenida.

9. Conclusiones

Tras el análisis de este fenómeno se puede concluir que la presencia de diversas fases y factores contribuyen al origen de esta crisis en el trasporte marítimo de contenedores. La comprensión de estos patrones es fundamental para que las investigaciones futuras consideren este

modelo y puedan generar técnicas o procedimientos para preverlo o en la medida de lo posible mitigarlo.

En la revisión de literatura científica y gris se identificaron factores como: aumentó del e-commer, pandemia del covid-19, congestión portuaria, aumento de fletes, cierre de fábricas, entre otros. Los cuales permitieron contar con un contexto amplio sobre la crisis de los contendores, esto fue útil para la construcción de las ecuaciones de búsqueda obteniendo así la base de datos utilizada para el desarrollo de esta investigación.

Realizar un análisis exploratorio permitió alinear la investigación hacia el uso de bigramas y una búsqueda de herramientas de minería de texto, seleccionando la técnica Latent Dirilech Allocation (LDA), ya que esta mostró mejores resultados en cuanto a métricas de validación al compararlo con el modelo k-means, permitiendo así la posibilidad de identificar patrones ocultos en los documentos científicos sobre la crisis de los contenedores.

Tras la aplicación del modelo LDA, si bien no se encontraron factores no previstos se lograron identificar las diferentes fases de este fenómeno, las cuales son: crisis de los contenedores, disrupción de la cadena de suministros y resiliencia de la cadena de suministros, permitiendo la construcción de diversos modelos conceptuales.

Estos modelos ayudan a comprender como los bigramas del resultado de Latent Dirilech Allocation (LDA) y la revisión de literatura se asocian a los diferentes factores que causaron este fenómeno y las fases que lo componen, donde estos elementos interactúan y se afectan mutuamente, proporcionando una comprensión más profunda de las causas y consecuencias de este evento.

10. Recomendaciones

A lo largo de la investigación se realizó una búsqueda sobre las diferentes técnicas de minería de texto con el objetivo de seleccionar la más viable, al ejecutar las técnicas k-means y LDA en la investigación se encontró que el modelamiento de tópicos tuvo un mejor desempeño, de esta manera se propone ejecutar y comprar diferentes técnicas de esta categoría con el fin de determinar cuál de ellas generaría mejores resultados.

Por otro lado, ya que actualmente si bien la crisis se ha logrado mitigar, al momento de realizar esta investigación algunos efectos siguen presentes, como altos fletes de transporte de mercancía marítima, debido a esto se propone generar una investigación de carácter similar en el futuro con el fin de recopilar más información científica ya que actualmente en el año 2023 se continúa publicando documentos al respecto. Por último, se propone para futuras investigaciones plantear un modelo conceptual que correlacione estadísticamente los factores encontrados.

Referencias Bibliográficas

Aguilar, D. A., Romero, J. N., & Leon, L. A. (2020). *Análisis de la escasez de contenedores en el transporte marítimo a nivel mundial año 2020*. *8*. https://www.dominiodelasciencias.com/ojs/index.php/es/article/view/2540

Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, *6*(1). https://doi.org/10.14569/IJACSA.2015.060121

Allianz. (2022). *Impact of Ukraine war on global shipping | AGCS*. AGCS Global. https://www.agcs.allianz.com/news-and-insights/expert-risk-articles/shipping-safety-22-ukraine-war.html

Alvarado, M. B., & Delgado, J. J. (2022). *Complejidades del comercio internacional. Análisis de la crisis de los contenedores y sus efectos en el crecimiento económico*. http://repositorio.ug.edu.ec/handle/redug/59290

Arango Serna, M. D., Ruiz Moreno, S., Ortiz Vásquez, L. F., & Zapata Cortes, J. A. (2016). Modelo conceptual para la administración de los recursos operacionales en las empresas transportadoras de carga terrestre en Colombia. *Universidad, Ciencia y Tecnología*, *20*(79), 75-86.

Arias, A., Mattos, Y., Heredia, J., & Heredia, D. (2016). Minería de texto como una herramienta para la búsqueda de artículos científicos para la investigación. *..pdf*. https://revistas.unisimon.edu.co/index.php/identic/article/view/2502/2403

Arriaga, M. (2015). *Comparación de métricas de distancia en el algoritmo K-vecinos más cercanos para el problema de reconocimiento automático de dígitos manuscritos*. Pontificia Universidad Católica del Paraíso.

Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 61-66. https://doi.org/10.1109/ICEEOT.2016.7754750

Baschuk, B. (2020, marzo 26). *World Trade Hit by Virus Sees Worst Collapse in a Generation—Bloomberg*. Bloomberg. https://www.bloomberg.com/news/articles/2020-03-26/world-trade-rocked-by-virus-sees-worst-collapse-in-a-generation#xj4y7vzkg

Bnamericas. (2022). *Crisis en transporte marítimo se debe a infraestructura deficiente y concentración de mercado*. BNamericas.com. https://www.bnamericas.com/es/noticias/crisis-en--transporte-maritimo-se-debe-a-infraestructura-deficiente-y-concentracion-de-mercado

Boiko, I. V., & Getman, A. G. (2022). Modern Challenges for International Sea Freight. *Administrative Consulting*, *8*, 36-45. https://doi.org/10.22394/1726-1139-2022-8-36-45

Carga. (2022, febrero 17). Crisis de los contenedores, ¿cómo afecta al mundo? *Carga*. https://carga.com.co/crisis-de-los-contenedores-como-afecta-al-mundo/

Chambers, S. (2021, febrero 4). *Californian port congestion spreads north to Oakland*. Splash247. https://splash247.com/californian-port-congestion-spreads-north-to-oakland/

Comunicaciones Puerto de Santa Marta. (2022, abril 28). Impacto de la guerra Rusia-Ucrania en el transporte marítimo mundial. *Noticias Puerto de Santa Marta*. https://noticiaspuertosantamarta.com/impacto-de-la-guerra-rusia-ucrania-en-el-transporte-maritimo-mundial/

Cullinane, K., & Haralambides, H. (2021). Global trends in maritime and port economics: The COVID-19 pandemic and beyond. *Maritime Economics & Logistics*, *23*(3), 369-380. https://doi.org/10.1057/s41278-021-00196-5

Da Silva, L. R. A. B. (2020). Inteligência artificial em processos de extração de conhecimento KDD e KDT. *Revista de Estudos Universitários - REU*, *46*(1), 161-180. https://doi.org/10.22484/2177-5788.2020v46n1p161-180

Daza, E. A., Mateus, L. J., & Patiño, J. M. (2022). *Plan de Marketing para el Posicionamiento de la Marca Kodak Alaris en Escáneres Documentales para El Mayorista.pdf*. http://repository.unipiloto.edu.co/bitstream/handle/20.500.12277/11444/Plan%20de%20Marketing%20para%20el%20Posicionamiento%20de%20la%20Marca%20Kodak%20Alaris%20en%20Escáneres%20Documentales%20para%20El%20Mayorista.pdf?sequence=1&isAllowed=y

Díaz, J. G., & Montealegre, R. J. (2022). Crisis Internacional de Contenedores en las Exportaciones de Banano desde Ecuador. *Economía y Negocios*, *13*(2), 124-132. https://doi.org/10.29019/eyn.v13i2.1008

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, *64*(2), 221-245. https://doi.org/10.1016/j.jacceco.2017.07.002

El Din, M. A., Reason, M., & Ncube, M. (2021). *The Impact of Post-Covid-19 Container Shortage Crisis on Global Supply Chains.pdf*. http://www.ieomsociety.org/china2021/papers/211.pdf

Estrada Vidal, A. C., & Reyes Hidalgo, N. Y. (2017). *Factores que generaron la crisis en el sector naviero de transporte de contenedores y los cambios en la configuración de las líneas navieras entre los años 2014 y 2016*.

Feldman, R., & Dagan, I. (1995). *Knowledge Discovery in Textual Databases (KDT)*. *95*, 112-117.

Galeana, A. K. (2016). *La importancia del transporte marítimo y el desarrollo de su infraestructura dentro de su infraestructura dentro de la logística comercial internacional* [Universidad nacional autónoma de méxico]. http://132.248.9.195/ptd2016/octubre/0751043/0751043.pdf

Garre, M., Cuadrado, J. J., Sicilia, M. A., Rodríguez, D., & Rejas, R. (2007). *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software*. *1*.

Garzón R., D., & Espitia F., E. T. (2022). *Crisis de los contenedores una mirada desde el contexto global y sus implicaciones en colombia.* http://repositorio.unicoc.edu.co:8080/xmlui/handle/1/945

Godoy, A. F. (2017). *Técnicas de aprendizaje de máquina utilizadas para la minería de texto*. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-358X2017000100103#fn22

Gohil, L. (2015). Text Mining: Process and Techniques. *International Journal of Innovative Research in Computer Science & Technology*, *3*(3). https://ijircst.org/DOC/16_irp380906f6fff-4b9f-4686-a4a5-cfa526dd4c0b.pdf

Gómez, M. S., & Fuentes, J. J. M. (2018). *Implementación de Asignación Jerárquica Latente de Dirichlet para Modelado de Temas*. Universidad de Sevilla.

Goodman, P. S., & Chokshi, N. (2021, junio 3). Cómo el mundo se quedó sin nada. *The New York Times*. https://www.nytimes.com/es/2021/06/03/espanol/justo-a-tiempo-cadena-suministro.html

Guerrero, D., Letrouit, L., & Pais-Montes, C. (2022). The container transport system during Covid-19: An analysis through the prism of complex networks. *Transport Policy*, *115*, 113-125. https://doi.org/10.1016/j.tranpol.2021.10.021

Gui, L., Zhou, Y., Xu, R., Leng, J., & Pergola, G. (2019). *Modelo de tema neuronal con aprendizaje por refuerzo*.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467-483. https://doi.org/10.1016/j.tourman.2016.09.009

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76. https://doi.org/10.4304/jetwi.1.1.60-76

Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2014). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 1-4. https://doi.org/10.1109/ICITEED.2014.7007894

Hearst, M. (2003). *What Is Text Mining?* https://people.ischool.berkeley.edu/~hearst/text-mining.html

Heras Calvo, D. (2023). *Biblioteca para la evaluación sistemática de algoritmos de clustering*. Universidad Polítecnica de Madrid.

Hernández, M., & Gómez, J. (2013). Aplicaciones de Procesamiento de LenguajeNatural. *Revista Politécnica*, *31*(1).

Humaira, H., & Rasyidah, R. (2020). Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. *Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*. Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia, Padang, Indonesia. https://doi.org/10.4108/eai.24-1-2018.2292388

IBM. (2018). *¿Qué es la minería de texto? | IBM*. https://www.ibm.com/es-es/topics/text-mining

Jeong, Y., & Kim, G. (2023). Reliable design of container shipping network with foldable container facility disruption. *Transportation Research Part E: Logistics and Transportation Review*, *169*, 102964. https://doi.org/10.1016/j.tre.2022.102964

Jerebić, V., & Pavlin, S. (2018). Global Economy Crisis and its Impact on Operational Container Carrier's Strategy. *PROMET - Traffic&Transportation*, *30*(2), 187-194. https://doi.org/10.7307/ptt.v30i2.2440

Junior Aduanas S.A. (s. f.). *Glosario—- Contenedor*. https://junioraduanas.com/herramientas-de-consulta/glosario/71-contenedor

Kannan, D. S., & Gurusamy, V. (2014). *Preprocessing Techniques for Text Mining*.

Kaur, A., & Chopra, D. (2016). Comparison of text mining tools. *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 186-192. https://doi.org/10.1109/ICRITO.2016.7784950

Kent, P., & Haralambides, H. (2022). A perfect storm or an imperfect supply chain? The U.S. supply chain crisis. *Maritime Economics & Logistics*, *24*(1), 1-8. https://doi.org/10.1057/s41278-022-00221-1

Khan, J., Ishizaka, A., & Mangla, S. K. (2022). Assessing risk of supply chain disruption due to COVID-19 with fuzzy VIKORSort. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04940-9

Kherwa, P., & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. *ICST Transactions on Scalable Information Systems*, *0*(0), 159623. https://doi.org/10.4108/eai.13-7-2018.159623

Knower, G. (2020, marzo 4). *Extent of Chinese factory slump supports fears over inventory levels | Journal of Commerce*. Journal of commerce. https://www.joc.com/article/extent-chinese-factory-slump-supports-fears-over-inventory-levels_20200304.html

Koshulko, A. (2023). *Shipping Container Crisis 2021*. GMDH. https://gmdhsoftware.com/shipping-container-crisis-2021/

Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, & Brown. (2019). Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

Křupka, A. (2013). The Use of Co-occurrence Features and Clustering Techniques for Definition of Typical Textures of Sedimentary Grains. *International Journal of Advances in Telecommunications, Electrotechnics, Signals and Systems*, *2*(2). https://doi.org/10.11601/ijates.v2i2.47

Kuźmicz, K. A. (2022). Impact of the COVID-19 Pandemic Disruptions on Container Transport. *Engineering Management in Production and Services*, *14*(2), 106-115. https://doi.org/10.2478/emj-2022-0020

La República, S. A. S. (2021). *El buque Ever Given atrapado en el Canal de Suez ha sido reflotado parcialmente*. Diario La República. https://www.larepublica.co/globoeconomia/el-buque-ever-given-atrapado-en-el-canal-de-suez-ha-sido-parcialmente-reflotado-3145968

LaRocco, L. A. (2021, marzo 25). *Suez Canal blockage is delaying an estimated $400 million an hour in goods*. CNBC. https://www.cnbc.com/2021/03/25/suez-canal-blockage-is-delaying-an-estimated-400-million-an-hour-in-goods.html

Laxe, F. G. (2005). Puertos y Transporte Maritimo: Ejes de una nueva articulación global. *Revista de economia mundial*, *12*, 123-148.

Liberatore, G., Voutto, A., & Fernández, G. vanessa. (2018). *Desarrollo de una herramienta para el análisis y representación semántica de colecciones documentales a través del factor TF-IDF*. http://humadoc.mdp.edu.ar:8080/bitstream/handle/123456789/631/ponencia_vuottoliberatorefernandez.pdf?sequence=1

Liu, L., Miller, H. J., & Scheff, J. (2020). The impacts of COVID-19 pandemic on public transit demand in the United States. *PLOS ONE*, *15*(11), e0242476. https://doi.org/10.1371/journal.pone.0242476

Liu, Y., Wu, X., & Shen, Y. (2011). Automatic clustering using genetic algorithms. *Applied Mathematics and Computation*, *218*(4), 1267-1279. https://doi.org/10.1016/j.amc.2011.06.007

Logimarex. (2021, diciembre 16). The container crisis and its effects. *Logimarex*. https://logimarex.com/en/the-container-crisis-and-its-effects/

Mahesh, B. (2018). *Machine Learning Algorithms—A Review*. *9*(1).

Malki, A. A., Rizk, M. M., El-Shorbagy, M. A., & Mousa, A. A. (2016). Hybrid Genetic Algorithm with K-Means for Clustering Problems. *Open Journal of Optimization*, *5*(2), Article 2. https://doi.org/10.4236/ojop.2016.52009

Martínez, I. (2012). *Mineria de datos y negocios*. https://www.uv.es/mlejarza/datamine/

Merk, O. (2021, abril 1). Atrapado en el Canal de Suez: Las lecciones que deja el atasco. *Trade News*. https://tradenews.com.ar/atrapado-en-el-canal-de-suez-las-lecciones-que-deja-el-atasco/

Miller, G. (2021, enero 27). *Trans-Pacific trade crashes into max-capacity ceiling*. FreightWaves. https://www.freightwaves.com/news/trans-pacific-trade-crashes-into-max-capacity-ceiling

Miller, G. (2022, diciembre 19). *Principales noticias marítimas de 2022: El auge de los contenedores llega a su fin, la guerra aviva los petroleros*. FreightWaves. https://www.freightwaves.com/news/principales-noticias-maritimas-de-2022-el-auge-de-los-contenedores-llega-a-su-fin-la-guerra-aviva-los-petroleros

Mindefensa de Colombia. (2023). *Estadísticas de Transporte Marítimo Internacional | Portal Marítimo Colombiano—Dimar*. Estadísticas de Transporte Marítimo Internacional. https://www.dimar.mil.co/operaciones-estadisticas/estadisticas-de-transporte-maritimo-internacional

Mora, A. H. (2020). *Algoritmos de aprendizaje supervisado utilizando datos de monitoreo de condiciones: Un estudio para el prono´stico de fallas en m´aquinas*.

Mughnyanti, M., Efendi, S., & Zarlis, M. (2020). Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation. *IOP Conference Series: Materials Science and Engineering*, *725*(1), 012128. https://doi.org/10.1088/1757-899X/725/1/012128

Naciones Unidas. (2022). *Informe sobre el Transporte Marítimo 2022*. https://unctad.org/system/files/official-document/rmt2022overview_es.pdf

Narasimha, P. T., Jena, P. R., & Majhi, R. (2021). Impact of COVID-19 on the Indian seaport transportation and maritime supply chain. *Transport Policy*, *110*, 191-203. https://doi.org/10.1016/j.tranpol.2021.05.011

Nielsen, F., & Nock, R. (2015). Total Jensen divergences: Definition, Properties and k-Means++ Clustering. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016-2020. https://doi.org/10.1109/ICASSP.2015.7178324

Notteboom, T., Pallis, T., & Rodrigue, J.-P. (2021). Disruptions and resilience in global container shipping and ports: The COVID-19 pandemic versus the 2008–2009 financial crisis. *Maritime Economics & Logistics*, *23*(2), 179-210. https://doi.org/10.1057/s41278-020-00180-5

Notteboom, T., & Rodrigue, J.-P. (2023). Maritime container terminal infrastructure, network corporatization, and global terminal operators: Implications for international business policy. *Journal of International Business Policy*, *6*(1), 67-83. https://doi.org/10.1057/s42214-022-00142-z

Nugent, M. A. L. M., Quispe, J. T., Llave, A. M. T., & Morales, J. A. F. (2019). Gestión de cadena de suministro: Una mirada desde la perspectiva teórica. *Revista Venezolana de Gerencia*, *24*(88), 1136-1146.

Núñez, N. A., Crisóstomo, R. A., Sánchez, S. A., Núñez, N. A., Crisóstomo, R. A., & Sánchez, S. A. (2021). Uso de minería de textos para comparar los contenidos relacionados a calidad y acreditación generados en redes sociales por universidades de Perú y Chile. *Formación universitaria*, *14*(1), 111-120. https://doi.org/10.4067/S0718-50062021000100111

Opacua Lomoschitz, M. I. (2019). *Named Entity Recognition y Topic Modeling: Metodología y aplicaciones al procesamiento de texto*. Universidad Carlos III de Madrir.

Orquera, M. O. (2022, marzo 28). *Escasez global de contenedores marítimos*. La comunidad logística. https://logistica.enfasis.com/logistica-y-distribucion/escasez-global-de-contenedores-maritimos/

Ortegón, S., Salamanca, B., & Julian, A. (2023). *Aprendizaje supervisado para clasificación de experiencias de viajes turísticos en Colombia*.

Orús, A. (2023). *Ranking de los puertos contenedores más grandes del mundo en 2022, según el rendimiento*. Statista. https://es.statista.com/estadisticas/635312/principales-puertos-de-contenedores-a-nivel-mundial-por-volumen-de-carga-manipulada/

Paris, C., & Faucon, B. (2022, marzo 1). War in Ukraine Disrupts Ships Around the Globe. *Wall Street Journal*. https://www.wsj.com/articles/war-in-ukraine-disrupts-ships-around-the-globe-11646138737

Pedrero, V., Reynaldos-Grandón, K., Ureta-Achurra, J., Cortez-Pinto, E., Pedrero, V., Reynaldos-Grandón, K., Ureta-Achurra, J., & Cortez-Pinto, E. (2021). Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. *Revista médica de Chile*, *149*(2), 248-254. https://doi.org/10.4067/s0034-98872021000200248

Pérez Cárcamo, C. A. (2010). *Evaluación de Reglas de Asociación en Text Mining Utilizando Métricas Semánticas y Estructurales*. Universidad de Concepción.

Pérez, J. (2012, diciembre 14). *El transporte marítimo*. El Orden Mundial - EOM. https://elordenmundial.com/el-transporte-maritimo/

Pérez Ortega, J., Hidalgo Reyes, M., Castro Sánchez, N. A., Pazos Rangel, R., Díaz Parra, O., Olivares Peregrino, V., & Almanza-Ortega, N. (2018). *Una heurística eficiente aplicada al algoritmo K-means para el agrupamiento de grandes instancias altamente agrupadas*.

https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462018000200607#B15

R. Sánchez & F. Weikert. (2020). *Logística internacional pospandemia: Análisis de las industrias aérea y de transporte marítimo de contenedores*.

Radovanović, M., & Ivanović, M. (2008). TEXT MINING: APPROACHES AND APPLICATIONS. *Text Mining*.

Rijcken, E., Scheepers, F., Mosteiro, P., Zervanou, K., Spruit, M., & Kaymak, U. (2021). A Comparative Study of Fuzzy Topic Models and LDA in terms of Interpretability. *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1-8. https://doi.org/10.1109/SSCI50451.2021.9660139

Rincón Ruiz, J. H. (2021). *Estudio comparativo de técnicas tradicionales del modelado de tópicos frente a redes neuronales artificiales tomando como contexto el discurso digital del autor en la red social Twitter y otras publicaciones*.

Rodriguez, M. Z., Comin, C. H., Casanova, × Dalcimar, Bruno, O. M., Amancio, D. R., da F Costa, L., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLoS One*, *14*(1). https://doi.org/10.1371/journal.pone.0210236

Rodríguez, W. (2013). *Indicadores de gestión portuaria. Aplicación al sistema portuario español*. https://upcommons.upc.edu/handle/2099.1/22951

Sablón-Cossío, N., Crespo, E. O., Pulido-Rojano, A., Acevedo-Urquiaga, A. J., & Ruiz Cedeño, S. D. M. (2021). Análisis de integración de la cadena de suministros en la industria textil en Ecuador. Un caso de estudio. *Ingeniare. Revista chilena de ingeniería*, *29*(1), 94-108. https://doi.org/10.4067/S0718-33052021000100094

Sánchez, Á. (2021, octubre 24). *Atasco global: Obras paradas, videoconsolas agotadas y nueve meses para recibir un coche*. El País. https://elpais.com/economia/2021-10-24/la-crisis-de-suministros-atasca-la-globalizacion-obras-paradas-videoconsolas-agotadas-y-nueve-meses-para-recibir-un-coche.html

Sánchez Álvarez, R. (2021). Clasificación no supervisada de imágenes médicas y minería de datos. Algoritmo S3 vs K-medias. *Revista Cubana de Investigaciones Biomédicas*, *40*. http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S0864-03002021000200006&lng=es&nrm=iso&tlng=es

Saputra, D. M., Saputra, D., & Oswari, L. D. (2020). Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method. *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*. Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019), Palembang, Indonesia. https://doi.org/10.2991/aisr.k.200424.051

Sarmiento, A. E. (2022). *COVID-19 Y SU IMPACTO EN LA ESCASEZ GLOBAL DE CONTENEDORES MARÍTIMOS*. Logistec. https://www.revistalogistec.com/logistica/freight-management-2/4172-covid-19-y-su-impacto-en-la-escasez-global-de-contenedores-maritimos

Sepúlveda Ramírez, M. F. (2019). *Herramientas de analítica visual para modelos de tópicos sobre colecciones de documentos* [Magíster en Ciencias de la Ingeniería]. https://doi.org/10.7764/tesisUC/ING/22979

Shepherd, D. A., & Williams, T. A. (2022). Different response paths to organizational resilience. *Small Business Economics*. https://doi.org/10.1007/s11187-022-00689-4

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1-6. https://doi.org/10.1109/ICCUBEA.2018.8697857

ShrihariR, C., & Desai, A. (2015). A Review on Knowledge Discovery using Text Classification Techniques in Text Mining. *International Journal of Computer Applications*, *111*(6), 12-15. https://doi.org/10.5120/19542-0784

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, *41*(3), 853-860. https://doi.org/10.1016/j.eswa.2013.08.015

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716-80727. https://doi.org/10.1109/ACCESS.2020.2988796

Smith, A., Kumar, V., Boyd-Graber, J., Seppi, K., & Findlater, L. (2018). Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. *23rd International Conference on Intelligent User Interfaces*, 293-304. https://doi.org/10.1145/3172944.3172965

Tadeo, R. A. (2022). *Estudio de caso: Impacto de la crisis de contenedores en la empresa AGRONEGOCIOS de El Salvador*. https://bdigital.zamorano.edu/items/0daac9a3-ebea-4ad5-8c01-96398c9f1f83/full

Thangaraj, M., & Sivakami, M. (2018). Text Classification Techniques: A Literature Review. *Interdisciplinary Journal of Information, Knowledge, and Management*, *13*, 117-135. https://doi.org/10.28945/4066

Tiffani, I. E. (2020). Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review. *Journal of Soft Computing Exploration*, *1*(1). https://doi.org/10.52465/joscex.v1i1.4

Torre, M. C. J. de la. (2017). *NUEVAS TÉCNICAS DE MINERÍA DE TEXTOS: APLICACIONES* [Tesis doctoral, Universidad de Granada]. http://hdl.handle.net/10481/46975

Torres-Moreno, J.-M. (2010). Reagrupamiento en familias y lexematización automática independientes del idioma. *INTELIGENCIA ARTIFICIAL*, *14*(47), 643. https://doi.org/10.4114/ia.v14i47.1570

Toygar, A., Yildirim, U., & İnegöl, G. M. (2022). Investigation of empty container shortage based on SWARA-ARAS methods in the COVID-19 era. *European Transport Research Review*, *14*(1), 8. https://doi.org/10.1186/s12544-022-00531-8

Tsantis, A., Mangan, J., Calatayud, A., & Palacin, R. (2022). Container shipping: A systematic literature review of themes and factors that influence the establishment of direct connections between countries. *Maritime Economics & Logistics*. https://doi.org/10.1057/s41278-022-00249-3

Tumanovich, A. V. (2022). *Global Container Crisis*. https://rep.bntu.by/bitstream/handle/data/120835/90.pdf?sequence=1

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Vianchá, Z. H. (2014). *Modelos y configuraciones de cadenas de suministro en productos perecederos*. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0122-34612014000100009

Vijayan, V. K., Bindu, K. R., & Parameswaran, L. (2017). A comprehensive study of text classification algorithms. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1109-1113. https://doi.org/10.1109/ICACCI.2017.8125990

Visco R&D, T. (2023, enero 5). *The Suez Canal Crisis—How can we be better prepared next time?* Visco Software. https://viscosoftware.com/the-2021-suez-canal-crisis-how-can-we-be-better-prepared-next-time/

Watt, A. (2022). *La política de «COVID cero» de China exacerba los problemas de la cadena de suministro | crisis24*. https://crisis24.garda.com/insights-intelligence/insights/articles/chinas-zero-covid-policy-exacerbates-supply-chain-issues?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es-419&_x_tr_pto=wapp

Xiao, Z., & Bai, X. (2022). Impact of local port disruption on global container trade: An example of stressing testing Chinese ports using a D-vine copula-based quantile regression. *Ocean & Coastal Management*, *228*, 106295. https://doi.org/10.1016/j.ocecoaman.2022.106295

Youd, F. (2021, abril 29). Global shipping container shortage: The story so far. *Ship Technology*. https://www.ship-technology.com/features/global-shipping-container-shortage-the-story-so-far/