UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE QUITO

CARRERA DE INGENIERÍA DE SISTEMAS

MAPEO Y SÍNTESIS A TRAVÉS DE ESQUEMAS DE CLASIFICACIÓN DE UN DATASET REFERENTE A ARTÍCULOS CIENTÍFICOS SOBRE EL CORONAVIRUS A TRAVÉS DE LISTA Y NUBE DE PALABRAS DE LOS ARTÍCULOS (ETAPA 2)

Trabajo de titulación previo a la obtención del

Título de Ingeniera de Sistemas

AUTORA: Valeria Lizeth Pilacuan Erazo

TUTOR: Gustavo Ernesto Navas Ruilova

Quito – Ecuador

2022

CERTIFICADO DE RESPONSABILIDAD Y AUTORIA DEL TRABAJO DE TITULACIÓN

Yo, Valeria Lizeth Pilacuán Erazo con documento de identificación 1718566324 manifiesto que:

Soy la autora y responsable del presente trabajo; y, autorizo a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Quito, 8 de marzo del año 2022.

Atentamente,

Valeria Lizeth Pilacuán Erazo
1718566324

CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA

Yo, Valeria Lizeth Pilacuán Erazo con documento de identificación 1718566324 , expreso mi voluntad y por medio del presente documento cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que soy autora del Artículo Académico: "Mapeo Y Síntesis a Través de Esquemas de Clasificación de un Dataset Referente a Artículos Científicos Sobre el Coronavirus a Través de Lista y Nube de Palabras de los Artículos (Etapa 2)", el cual ha sido desarrollando para optar por el título de: Ingeniera de Sistemas, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Quito, 8 de marzo del año 2022.
Atentamente,

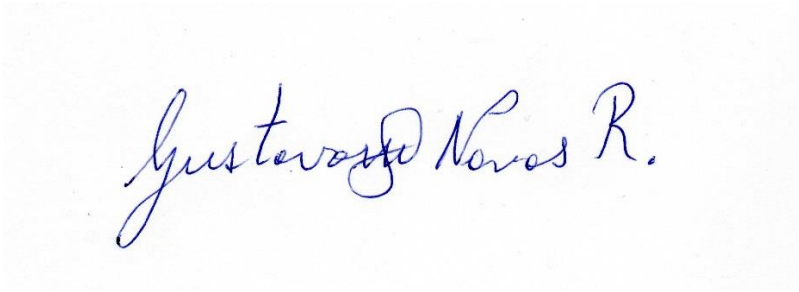
Valeria Lizeth Pilacuán Erazo
1718566324

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Gustavo Ernesto Navas Ruilova con documento de identificación N° 1705675625, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación : MAPEO Y SÍNTESIS A TRAVÉS DE ESQUEMAS DE CLASIFICACIÓN DE UN DATASET REFERENTE A ARTÍCULOS CIENTÍFICOS SOBRE EL CORONAVIRUS A TRAVÉS DE LISTA Y NUBE DE PALABRAS DE LOS ARTÍCULOS (ETAPA 2), realizado por Valeria Lizeth Pilacuán Erazo, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Quito, 8 de marzo del año 2022.

Atentamente,



Ing. Gustavo Ernesto Navas Ruilova, PhD
1705675625

DEDICATORIA

"Nada resulta más insoportable que tener que admitirse a uno mismo los propios errores".

Ludwig van Beethoven

Dedico con todo mi corazón mi tesis a Dios, mi Madre y mi Padre que son motivo de mi orgullo, quienes me han formado con reglas y algunas libertades, a pesar de todo me han ayudado hacer realidad mis sueños y anhelos, ellos que me han enseñado que el amor no puede ser medido en ningún tipo de magnitud.

A mis hermanos que son las personas que han estado a mí lado y han forjado mi camino a la superación todo este tiempo con su reflejo, siempre queriendo ser mi ejemplo.

A mis sobrinos Taiz, Nicolas e Iker que son uno de los motivos de mi sonrisa para mi inspiración de seguir adelante y fuente de formaleza cada día "no dejen de jugar por miedo a errar", es parte del camino del aprendizaje, son los sueños, dedicación y esfuerzo los que al final del día son recompensados.

AGRADECIMIENTO

Primero agradezco a Dios por darme la oportunidad de ser mejor cada día, por demostrarme su amor infinito dándome bendiciones y lecciones, por confiar en mí, que sería de mi si no me hubieras perdonado las veces que te ofendí y que pensé que no estabas ahí, ahora mi corazón puede sentir tu presencia.

A mis padres Sabina y Pedro de los cuales soy la niña de sus ojos, recuerden que mi corazón late por ustedes, siempre el silencio de mi voz dice que los amo papitos. Me convertí en un ser hermoso gracias a ustedes, ojalá pudiera no crecer más, me asusta mucho el mundo exterior, pero me siento preparada para volar, el de arriba me ha dado demasiado, son el regalo que no merecía, espero ser siempre su pequeña princesa.

Mis hermanos Mónica y Rubén que son mi regalo más grande, son de los que se abren y encuentra una cantidad imaginable de magia, a veces sus gestos duros esconden toda su fragilidad, pero son mis ángeles que más de una vez me han enseñado innumerables cosas y no solo de la vida sino también profesionalmente, amigos incondicionales en los buenos y malos momentos, mis hermanos "coraje".

Madrina Melita que Dios la bendiga siempre, usted que a sido mi segunda mamá, gracias por todos sus consejos y sobre todo por siempre estar junto a mí.

Agradezco a mi Universidad, mi tutor de tesis Ing. Gustavo Navas que confió en mí y sobre todo siempre estuvo dispuesto a ayudarme, mis de más profesores y compañeros por darme infinitas alegrías y enseñanzas, ser una buena cristiana y honrada ciudadana, gracias por todos los conocimientos obtenidos durante este camino.

Un Agradecimiento especial a mis mejores amigos de la Universidad Stalin y Christopher que son las personas en las que me he apoyado este tiempo, son mis super héroes chiquitos, los buenos recuerdos se quedan en mi cabeza pero tienen mi corazón, gracias por mirarme a los ojos a pesar de verme llorar porque nunca me falto un abrazo suyo para decirme que no me dé por vencida, los quiero mucho, han sido parte de este hermoso sueño que cada día lo vamos construyendo, jamás me olvidare de las risas en los pasillos.

Mayra y amigas queridas, Dani, Joss, Diana, Evelyn y demás amigos, les agradezco la paciencia, a unas más que a otras, gracias por los consejos y por cada día preguntarme por la tesis, a veces necesitamos esa presión psicológica.

Amigos de la EEQ gracias por todo esa enorme enseñanza y amistad, siempre me han dado un empujoncito para adelante.

No podía faltar un agradecimiento a mí, por no rendirme porque siempre trate de superarme, me siento muy satisfecha con mi desempeño, amigos disculpen mi narcisismo, pero no podía faltar este pequeño párrafo.

MAPEO Y SÍNTESIS A TRAVES DE ESQUEMAS DE CLASIFICACIÓN DE UN DATASET REFERENTE A ARTÍCULOS CIENTÍFICOS SOBRE EL CORONAVIRUS A TRAVÉS DE LISTA Y NUBE DE PALABRAS DE LOS ARTÍCULOS (ETAPA 2)

MAPPING AND SYNTHESIS THROUGH CLASSIFICATION SCHEMES OF A DATASET REFERRING TO SCIENTIFIC ARTICLES ON THE CORANAVIRUS THROUGH LIST AND WORD CLOUD OF ARTICLES (STAGE 2)

Valeria L. Pilacuán 1[1], Gustavo E. Navas[2]

Resumen

El intento de optimización de un Systematic Mapping Study (SMS) a través de Machine Learning tiene dos etapas, en la primera etapa de este estudio que tomó el nombre de "Clasificación y Mapeo de un dataset de Artículos Científicos sobre SARS-CoV2 a través de Lista y Nube de Palabras", el autor del trabajo Moromenacho muestra las etapas de: Definición de preguntas de investigación, Ejecución de la búsqueda y Selección de artículos relevantes, en la cual involucra las herramientas Python y Excel, para la automatización se utilizó el Dataset de CORD-19 que fue descargado desde Kaggle, en la implementación se obtuvo un archivo con extensión .xlsx donde se encuentran títulos, abstracts y palabras más usadas, total de palabras del texto de los artículos.

Para esta segunda se realiza una validación de datos, trataremos las etapas de: Selección de artículos relevantes, búsqueda de palabras clave, proceso de mapeo y extracción de datos, las herramientas que se

Abstract

The optimization attempt of a Systematic Mapping Study (SMS) through Machine Learning has two stages, in the first stage of this study that took the name of "Classification and Mapping of a dataset of Scientific Articles on SARS-CoV2 through List and Cloud of Words", the author of the work Moromenacho shows the stages of: Definition of the investigation, Search of articles, keywords and Selection of articles, in which the Python and Excel tools are involved, for the automation the Dataset was used. of CORD-19 that was downloaded from Kaggle, in the implementation a file with extension .xlsx was obtained where titles, abstracts and most used words, total words of the text of the articles are found.

For this second one, a data validation is carried out, we will deal with the stages of: Selection of relevant articles, keyword search, mapping process and data extraction, the tools used with Python were Apache Spark and the Non-Negative Factorization method of

[1] Estudiante de Ingeniería de Sistemas – Universidad Politécnica Salesiana, Egresada – UPS – Sede Quito-
Autora para correspondencia: vpilacuan@est.ups.edu.ec

[2] Máster Universitario en Ciencias y Tecnología de la Computación, Master Universitario en Software Libre, Ingeniero Mecánico – UPS – sede Quito.
Autor por correspondencia: gnavas@ups.edu.ec

utilizó con Python fueron Apache Spark y el método de Factorización No Negativa de Matrices (NMF) de Machine Learning, para lograr esta clasificación se tuvo que partir desde los criterios de inclusión y exclusión, generar palabras clave a través del modelo y finalmente como objetivo principal poder ver hacia que aspecto social se esté generando una tendencia.

Palabras Clave: Coronavirus, SARS-Cov-2, Covid-19, Apache Spark, clasificación de documentos, Factorización de matrices no negativa, SMS, ML.

Matrices (NMF) of Machine Learning, to achieve this classification it was necessary to start from the inclusion and exclusion criteria, generate keywords through the model and finally as the main objective to be able to see towards which social aspect a trend is being generated.

Keywords: Coronavirus, SARS-Cov-2, Covid-19, Apache Spark, document classification, non-negative matrix factorization, SMS, ML.

1. Introducción

El brote de COVID 19 empezó hace aproximadamente dos años, en Wuhan se tomaron medidas de prevención y control a nivel nacional, a pesar de varias medidas tomadas en China, hay distintos factores que influyen a la propagación de COVID-19 [1], tenemos varia información sobre la enfermedad, pero aún no se ha logrado obtener una cura total para este síndrome respiratorio agudo de Coranonavirus-2 (SARS-CoV-2), se sabe que la enfermedad es altamente infecciosa, por el momento encontramos varios géneros de coronavirus; alfa, beta, gamma, delta y ómicron, los cuales tienen estructuras genéticas diferentes [2].

A más de los síntomas por síndrome respiratorio a este se suma la cantidad de personas que aumentaron su ansiedad debido al COVID-19, se percibió un incremento en mujeres y en personas con asma no controlada [3].

Hace un tiempo no había vacunas y tratamientos, ahora la intención es disminuir la cantidad de personas infectadas aun cuando en la actualidad existen variantes de este virus y crea incertidumbre a nivel mundial sobre lo que significa tener una cura [4].

Según la Universidad de Medicina Johns Hopkins, los datos acerca del COVID-19 son: Tabla 1.

Tabla 1: Datos Mundiales de Covid-19
Fuente: [27]

Datos de Covid-19	
376.313,301	Total de Casos.
5.668.707	Total de Muertes.
9.974.345.828	Total de dosis de vacunas administradas.

Varios investigadores han utilizado Machine Learning (ML) y el COVID-19 para identificar datos relacionados con política, cultura, educación y más temas que son varios factores controversiales en el mundo [5], también se habla sobre modelos logísticos para identificar, notificar y analizar.

China es uno de los países que absorbe datos muy completos de la población, Apple y Google quieren realizar sistemas donde se identifique a las personas que estuvieron en contacto con aquellas que tengan COVID-19, sin embargo, se dieron cuenta que hay datos incompletos y son cambiantes para saber el comportamiento del mundo frente a esta enfermedad [6].

Según [7] es importante el papel que realiza un científico de datos, lo que hace la minería de datos es integrar los datos y la tecnología, en este caso como se lo realizara con el virus lo que ayuda a planear acciones y a la toma de decisiones.

En este artículo quiero presentar un intento de optimización de Systematic Mapping Study (SMS) con Machine Learning debido a que en la actualidad se las realiza de manera manual aproximadamente desde el 2008, varios de los estudios se elaboran con esta metodología, se utiliza para estructurar un área de investigación por lo que se propone mejorar las revisiones sistemáticas de la literatura (SLR) [8].

Este estudio se realiza en 2 etapas, la primera etapa de este estudio tomó el nombre de "Clasificación y Mapeo de un Dataset de Artículos Científicos sobre SARS-CoV2 a través de Lista y Nube de Palabras", Moromenacho en este artículo nos muestra un intento de optimización en las etapas de: Definición de la investigación, Búsqueda de artículos, palabras clave, en la cual involucra herramientas de Python y Excel para automatizar el Dataset de CORD-19 que fue descargado desde Kaggle de esta automatización se obtuvo un .xlsx donde se encuentran títulos, abstracts, palabras más utilizadas, conteo de palabras, de los diferentes artículos [9].

Para esta etapa se partió desde la etapa de Selección de artículos con las herramientas de Python con Apache Sparck, este utiliza memoria computacional, tiene un conjunto de datos distribuidos resistentes (RDD) que son la

estructura de datos subyacente y sirve para la abstracción de datos [10], tiene mejor escalabilidad, tiempo de ejecución más rápida [11], con esta herramienta se volvió a procesar los datos para descartar artículos que se hubieran clasificado mal, el resultado fue un CSV con 629 artículos con títulos y abstracts, para palabras clave y proceso de mapeo y extracción de Datos se utilizó el modelo no supervisado de Factorización Matricial no Negativa (NMF) que categoriza y crea resultados óptimos y aceptables, impone restricciones como variables regulatorias [12], agrupa textos y presenta en orden secuencial las palabras que aparecen, este descompone la matriz palabra-contexto [13].

2. Método

Esta investigación presenta la segunda etapa de intento de optimización del Systematic Mapping Study (SMS), para este proceso se realizó una categorización de artículos sobre CORD-19 para el aprendizaje del modelo MNF.

SMS cuenta con 6 etapas que son las siguientes: Figura 1.

Figura 1. Etapas de SMS [8]

A continuación, se hace una presentación sumaria de la elaboración de la etapa 1 que lleva el nombre de "Clasificación y Mapeo de un Dataset de Artículos Científicos sobre SARS-CoV-2 a través de Lista y Nube de palabras de los artículos (etapa 1)" [9].

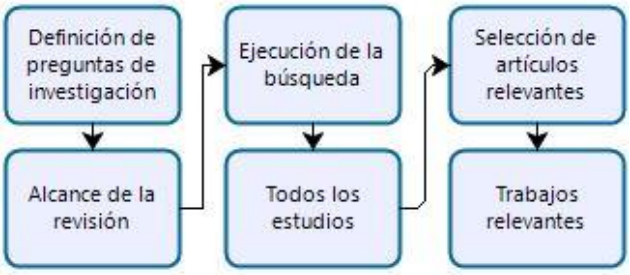


Figura 2. Proceso de Systematic Mapping Study Etapa 1 [8]

3. Primera etapa

2.1.1. Definición de Preguntas de investigación

Las preguntas de investigación se realizan de acuerdo con el tema de estudio [14], se toma una exploración de idioma, categorización y técnicas para barrido de temas [15].

Las preguntas que se presentan en la etapa 1 son:

PL1. ¿Cuáles son los principales tópicos que cubren los estudios del CORD-19?
PL2. ¿Cuáles son los factores principales que ocasionaron el virus COVID-19?
PL3. ¿Cuál es el estado actual del conocimiento sobre coronavirus?

2.1.2. Ejecución de la búsqueda

Se establece una cadena de búsqueda que contenga sinónimos, palabras clave, limitación y objetivo de la búsqueda [16], Moromenacho nos menciona que el dataset de CORD-19 tenía cerca de 500.000 artículos académicos y que se relacionaban con el tema COVID-19, SARS-Cov-2 y Coronavirus, el dataset elaborado por el autor de la etapa 1 contiene 19 columnas y nos dice que cada una de ellas representa un dato único para los artículos [9].

2.1.3. Selección de Artículos relevantes

Aplicada la cadena de búsqueda en las fuentes seleccionadas se aplican criterios de inclusión y exclusión [14], se busca los estudios que cumplan con los criterios y se adaptó el proceso de

búsqueda [15], el resultado de esta etapa fueron artículos académicos potencialmente ligados a nuestro estudio de COVID-19.

En este punto se realiza la unión entre la etapa 1 y etapa 2, para lo cual se volvió a realizar optimización del SMS con el dataset de CORD-19.

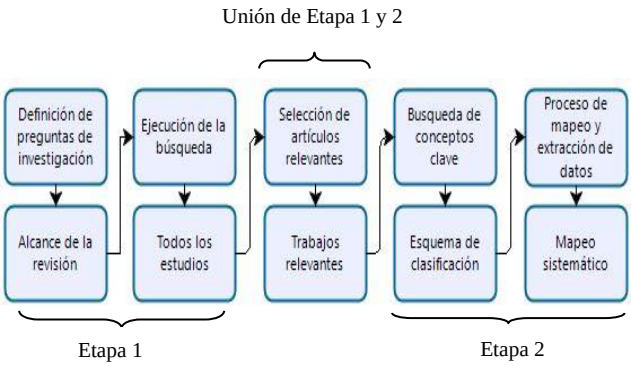


Figura 3. Unión de etapa 1 y 2
Fuente: Autores

En la primera etapa uno de los procesos de filtrado se lo observa en la Figura 4, este se lo realizó en Excel y Python.

```
.......Unicos......

hello how are you
Bien hecho  es mejor que bien dicho
hola como estas
Success in management requires learning as fast as the world is changing world is
Where there is love there is life  is love there is life
Once you choose hope, anything's possible anything's possible
Try to be a rainbow in someone's cloud Try to be a rainbo
Honesty is the first chapter in the book of wisdom Honesty is the first
yellow Blue

.......Repetidos......

hello how are you
Where there is love there is life  is love there is life
Try to be a rainbow in someone's cloud Try to be a rainbo
```

Figura 4. Muestra de resultados Obtenidos en Etapa 1 [9]

Para iniciar la segunda etapa correspondiente al presente documento se tomaron los 630 artículos académicos seleccionados por Moromenacho, se validó utilizando Apache Spark (Figura 5) y con los mismos criterios de inclusión y exclusión se realizó una nueva optimización de los artículos, dando lugar a 627 artículos, que es el 99.52% de similitud hallado por Moromenacho [9].

```
An Interprofessional Approach in Caring for a Patie
RATIONALE: Hemodialysis patients are at significant
RVENTIONS: An interprofessional team was establishe
SEIR model for COVID-19 dynamics incorporating the
OBJECTIVE: Coronavirus disease 2019 (COVID-19) is a
spread on novel coronavirus;

"COVID-19: Results of a national survey of United K
OBJECTIVE: COVID-19 has caused a global healthcare
of the respondents (95.23%) had direct patient cont
```

Figura 5. Ejemplo de muestra de artículos obtenidos en Apache Spark
Fuente: Autores

2.2. Segunda Etapa

La segunda etapa de esta investigación se enfoca en los siguientes procesos.

Figura 6. Procesos para Etapa 2 [8].

2.2.1. Búsqueda de palabras clave

Las palabras clave ayudan a reducir el tiempo necesario para el desarrollo con el contexto de la investigación, se logra un alto nivel de comprensión acerca de la naturaleza y la contribución de la investigación [17]. En [18] se escogió temas que con frecuencia han sido utilizados en dos criterios: la opinión pública y mediante minería de datos, [19] utiliza un esquema clásico para el procesamiento de datos en documentos: tokens, palabras vacías y lematización.

Estas perspectivas se consideraron para las palabras clave que se utiliza en el aprendizaje no supervisado con NMF, con el procesamiento de

Apache Spark se obtuvo una columna con palabras que se repetían con frecuencia en el dataset (ver tabla 2).

2.2.2. Proceso de mapeo y extracción de datos.

El esquema de clasificación se lo lleva a cabo con la extracción de datos reales, para la documentación de este proceso se lo realiza en Excel, las tablas tienen una categoría y al ingresar un artículo al esquema lo debería colocar en una determinada categoría [8].

A demás de haber encontrado las palabras clave también determinamos las categorías para el modelo, según [18] se debe identificar los temas con mayor frecuencia, palabras más utilizadas en el conjunto de datos y realizar una nube de palabras.

2.2.2.1. Optimización basada en modelo MNF.

Non-Negative Matrix Factorization o en español Factorización Matricial no Negativa es un algoritmo no supervisado que proyecta los datos inferiores, reduce la cantidad de características y reconstruye los datos originales [20], Se muestra la ecuación para el proceso:

$$V \approx W \times H \quad (1)$$

Variables Visibles $\approx$ $Pesos$ $\times$ $Variables\ ocultas$

Esta fórmula se descompone en matrices más pequeñas que son:

$$\text{V} = Documento \times Matriz\ de\ terminos$$

$$V = n \times m \quad (2)$$

$$W = Documento \times Matriz\ de\ temas$$

$$W = n \times p \quad (3)$$

$$H = Temas \times Matriz\ de\ terminos$$

$$H = p \times m \quad (4)$$

En la Matriz V, las filas representan un documento o lo que es la aparición de cada palabra y las columnas son variables visibles.

En la Matriz W o de pesos, las filas representan el documento en las cuales tenemos los temas no normalizados y las columnas son las repeticiones de características semánticas.

La Matriz H o de variables ocultas, las filas son un tema o características semánticas y las columnas una variable visible.

La matriz W y H pueden reconstruir V con la multiplicación de sus matrices. El resultado de NMF es cambiante de acuerdo con la ejecución y se debe elegir un número de temas, al actualizar las matrices W y H minimizamos la función de error [21].

En [19] afirma que se puede construir grupos que cumplan las delimitaciones que se requieren con conjuntos de datos pequeños.

Para el algoritmo utilizamos las palabras claves como categorías del modelo (ver Tabla 2), a su vez encontraremos la cantidad de artículos académicos que fueron clasificados según su categoría.

Tabla 2: Categorías utilizadas en modelo NMF
Fuente: Autores

ID	Categoría	# artículos académicos
0	Patients	45
1	COVID-19.	76
2	Diabates	41
3	MERS	25
4	SARS-CoV-2	53
5	Coronavirus	35
6	Respiratory	128
7	Hospital	35
8	Symptoms	13
9	Vaccines	70
10	Diseases	15
11	Politic	10
12	Pandemic	40
13	Clinical	32
14	Education	8

3. Resultados y Discusión

En esta sección se describe una muestra de los experimentos realizados con el modelo NMF, se utiliza el dataset obtenido en el apartado de

"Selección de artículos relevantes" obtenido en esta etapa 2 que contiene los artículos académicos que serán utilizados para el proceso de clasificados.

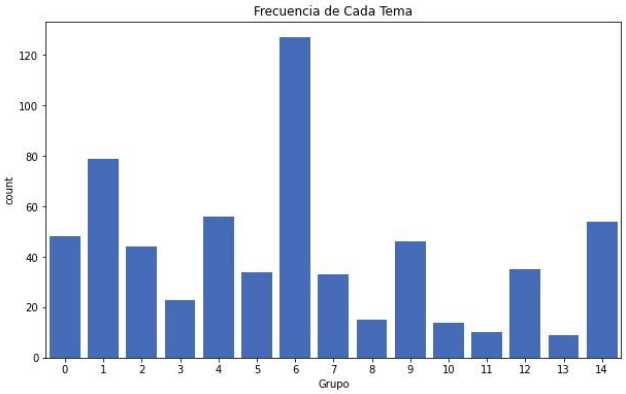


Figura 7. Muestra de Ejecución realizada en NMF
Fuente: Autores

Se evaluó los términos según el dataset que se obtuvo en una de las etapas con la selección de palabras clave que eran similares y repetidas, además se obtuvo 15 categorías y 627 artículos académicos para clasificación.

NMF es uno de los modelos eficiente en procesamiento de datos de texto, esta captura la información semántica y las conecta, cada iteración va reforzando su aprendizaje, la Figura 7 muestra la clasificación general de los 627 artículos con su categoría correspondiente, el resumen de los resultados de dicha clasificación arrojó que los artículos académicos se relacionan en gran cantidad con la respiratory, seguido de COVID-19 y SARS-CoV-2 (ver Tabla 2).

En la Figura 8 se observa la muestra de un artículo académico que fue escogido al azar como ejemplo, en este caso el artículo 200 señala las categorías (Tabla 2) probables para su clasificación, medido mediante un coeficiente de 0 a 1.

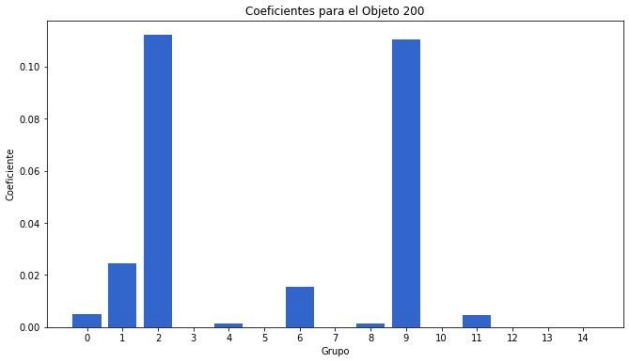


Figura 8. Muestra individual del tema 200
Fuente: Autores

Al tener una matriz de términos podemos descartar aquellas palabras que son irrelevantes para el aprendizaje, por lo que se utiliza un vector llamado "stop_words" al que se tuvo que ir reajustando en el proceso.

Las palabras clave son dicotómicas y otras se escogieron por la frecuencia en la cual son utilizadas, en NMF se tomó en cuenta el desafío que está afrontando el mundo con la pandemia de COVID-19, y cuáles son los temas que se relacionan. A medida que el algoritmo siga aprendiendo hará que los investigadores tengan un panorama más claro del contenido de los documentos.

El aprendizaje del modelo es 56.28% sin embargo, este varia constantemente debido a que con cada ejecución el modelo aprende.

3.1. Respuestas a preguntas de investigación

PL1. ¿Cuáles son los principales tópicos que cubren los estudios del CORD-19?
Se cubre 15 temas del dataset de CORD-19, estas palabras se repiten a lo largo del dataset, las cuales son: patients, COVID-19, diabates, MERS, politic, coronavirus, respiratory, hospital, symptoms, vaccines, diseases, SARS-CoV-2', pandemic, clinical, education.

PL2. ¿Cuáles son los factores principales que ocasionaron el virus COVID-19?
Varios de los estudios sugieren que de los murciélagos se originó esta enfermedad y el pangolín al ser un huésped intermedio facilitó la mutación del coronavirus.

PL3. ¿Cuál es el estado actual del conocimiento sobre coronavirus?
En la actualidad se conoce muchos datos acerca del coronavirus, pero sigue siento incierto el estado actual, cada día aparece una variante diferente de COVID-19, lo que es cierto es que se debe seguir manteniendo las medidas de bioseguridad que se ha dispuesto por la OMS, cada día se van desarrollando diferentes medicamentos para detener esta enfermedad.

4. Conclusiones

El objetivo de este artículo académico fue tratar de optimizar el Estudio de Mapeo Sistemático el cual se ha intentado resolver con una capa de Machine Learning aplicada al dataset de Cord-19 el cual ha sido analizado y procesado durante las etapas de palabras clave y proceso de mapeo y extracción de datos.

En este artículo analizó con Apache Spark la comprobación de similitud dando como resultado el 99,52%, para lo que se utilizó los mismos criterios de inclusión y exclusión para el dataset de CORD-19 de la primera etapa, los 627 artículos arrojaron la probabilidad más alta de palabras clave en cada tema, y estas reflejan la eficiencia de este primer método.

De acuerdo con los objetivos planteados NMF es un modelo que debe irse mejorando, ya no como un modelo no supervisado ya que este método puede no ser tan confiable como otros, a pesar de haber agrupado y clasificado los artículos académicos se comprende los errores que se pueda obtener al momento del aprendizaje, sin embargo, estos pueden ser corregidos al tener una base de datos con mayor número de artículos académicos.

Finalmente se obtiene que los 627 artículos son clasificados según la categoría más probables en el algoritmo NMF, lo cual ayudará a los investigadores a reducir la cantidad artículos académicos según el tema de estudio que se plantea, el aprendizaje del modelo fue 56.28% que puede mejorar en el proceso de entrenamiento.

Referencias

[1] J. H. R. L. Q. S. D. H. X. L. J. Q. Yifei Han, «Impact analysis of environmental and social factors on early-stage,» *Science Direct*, p. 9, 2021.

[2] B. C. ,.-C. C. Jie-Ming Qu, Respiratory virus and COVID-19, Shanghai: ELSEVIER, 2021, pp. 1-6.

[3] I. M. A. G. M. W. C. A. E. M. A. B. I. K. Sandra Ekström, «General Stress Among Young Adults with Asthma During the COVID-19 Pandemic,» *Science Direct*, p. 8, 2021.

[4] N. P. B. L. Vardavas Raffaele, «Modeling COVID-19 Nonpharmaceutical Interventions: Exploring periodic NPI strategies,» *medRxiv*, p. 45, 2021.

[5] Z. Z. W. C. Y. L. B. D. C. C. Q. L. N. U. S. J. C. C. Y. Z. X. W. Mengyuan Li, «Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach,» *Science Direct*, p. 14, 2021.

[6] M. T. R. J. B. O'Neill Patrich H., «MIT technology review,» 7 03 2020. [En línea]. Available: https://www.technologyreview.com/2020/05/07/1000961/launching-mittr-covid-tracing-tracker/.

[7] P. G. P. P. S. K. Gupta Rajan, «Machine Learning Models for Government to Predict COVID-19 Outbreak,» *ACM Digiral Library*, p. 6, 2021.

[8] R. F. S. M. M. M. Kai Petersen, «Systematic Mapping Studies in

Software Engineering,» *scienceopen*, pp. 1-10, 2008.

[9] E. Moromenacho, «"Clasificación y mapeo de un Dataset de Artículos Científicos sobre SARS-CoV2 a través de Lista y Nube de Palabras",» p. 18, 2021.

[10] A. K. A. K. L. Amritpal Singh, «Performance Comparison of Apache Hadoop and Apache Spark,» *ACM Digital Library*, p. 5, 2019.

[11] W. E. Wijayanto Ardhi, «Implementation of Multi-criteria Collaborative Filterring on Cluster Using Apache Spark,» *IEEE*, p. 5, 2016.

[12] Z. M. D. Aghdam Mehdi Hosseinzadeh, «A novel regularized asymmetric non-negative matrix factorization,» *Science Direct*, p. 16, 2021.

[13] S. S. N. M. Ailem Melissa, «Non-negative Matrix Factorization Meets Word Embedding,» *ACM Digital Library*, p. 4, 2017.

[14] U. M. Arshad Ali, «Security at Software Architecture Level: A Systematic Mapping Study,» *IEEE*, pp. 164-168, 2011.

[15] M. M. W. N. C. O. A. AlOmar Eman Abdullah, «On preserving the behavior in software refactoring: A systematic mapping study,» *Science Direct*, p. 20, 2021.

[16] L. P. S. M. Waseem Muhammad, «A Systematic Mapping Study on Microservices Architecture in DevOps,» *Science Direct*, p. 30, 2020.

[17] S. Nature, «Titles, Abstracts & Keywords,» 2021. [En línea]. Available: https://www.springernature.com/gp/authors/campaigns/writing-a-manuscript/titles-abstracts-keywords.

[18] O. A. C. F. Arango Pastrana Carlos Alberto, «Aislamiento social obligatorio: analisis de sentimientos mediante machine learning,» *Suma de Negocios*, p. 13, 2020.

[19] M. P. F. C. Vallejo Huanga Diego, «Semi-Supervised Clustering Algorithms for agrouping Scientific Articles,» *Science Direct*, p. 10, 2017.

[20] R. L. N. Bernadete, «Machine Learning for adaptive Many - Core Machines - A practical Approach,» *Springer Link*, pp. 127-154, 2015.

[21] G. Anupama, «NMF- A visual explainer and Python Implementation,» 17 03 2021. [En línea]. Available: https://towardsdatascience.com/nmf-a-visual-explainer-and-python-implementation-7ecdd73491f8.

[22] I. V. Di Nella Dino, «Causas y consecuencias de la pandemia COVID-19. De la inmovilidad de la humanidad a la circulación desconcentrada de personas,» *REDEA*, p. 71, 21 03 2020.

[23] Kaggle, «COVID-19 Open Research Dataset Challenge (CORD-19),» Marzo 2020. [En línea]. Available: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

[24] M. U. Muhammad Manan Qadir, «Software Engineering Curriculum: ASystematic Mapping Study,» *ACM* , pp. 1-6, 2011.

[25] C. R. Center, «COVID-19 by the Center for Systems Science and Engineering,» JOHNS HOPKINS UNIVERSITY MEDICINE, 31 01 2022. [En línea]. Available: https://origin-coronavirus.jhu.edu/map.html. [Último acceso: 31 1 2022].

[26] H. M. Koury Juan M., «Acta Odontológica Venezolana,» *Science Direct*, pp. https://www.actaodontologica.com/ediciones/2020/especial/art-2/, 21 03 2020.

[27] M. D. J. Peters, «DEAKIN UNIVERSITY,» 18 11 2021. [En línea]. Available: https://deakin.libguides.com/systematicreview/step3.