Identification of Latent Topics in Patients Surviving COVID-19 in Mexico

Angélica Guzmán-Ponce, Ruben Fernandez-Beltran, *Senior Member, IEEE*, Rosa María Valdovinos-Rosas, Marcelo Romero-Huertas and J. Raymundo Marcial-Romero

***Abstract*—With the outbreak of the SARS-CoV-2 o COVID-19 pandemic, multiple studies of risk factors and their influence on patient deaths have been developed. However, little attention is often paid to analyzing patients in risk groups despite the fact that they have been infected and inpatients can survive. In this article, with the dataset available from the Ministery of the health of Mexico, this paper proposes the use of the latent topic extraction algorithm Latent Dirichlet Allocation (LDA) for the study of COVID-19 survival factors in Mexico. The results let us conclude that in the year before strategies for prevention and control of COVID-19, the latent topics support that patients without comorbidities have a low risk of death, compared with the period of 2021, wherein in spite of having some risk factors patients can survive.**

***Index Terms*—Latent topics, Latent Dirichlet Allocation (LDA), COVID-19, risk factors**

I. Introducción

El síndrome respiratorio agudo severo SARS-CoV-2, responsable de la pandemia de COVID-19 [1], ha generado más de 6 millones de muertes a nivel mundial desde su identificación inicial en Wuhan, China, en diciembre de 2019. En México, ha causado 322,072 muertes [2]. Siendo el COVID-19 el problema más crítico que ha afectado severamente la vida humana [3], dado que incide en esferas de salud, educación, transporte, política, cadena de suministro, etc.

Normalmente, las personas infectadas con COVID-19 experimentan problemas respiratorios y pueden recuperarse con un tratamiento intensivo y apropiado. Sin embargo, la progresión y muerte por las complicaciones de la enfermedad sigue siendo un desafío importante. Diversos estudios apuntan que las comorbilidades preexistentes en los pacientes infectados son los principales determinantes para un mal pronóstico a la recuperación por COVID-19. Estas comorbilidades incluyen principalmente hipertensión, diabetes, enfermedades vasculares, cáncer, obesidad, entre otras. [4], [5].

Para comprender mejor y dar respuestas al comportamiento de la pandemia por COVID-19, se han desarrollado investigaciones con soluciones eficientes basadas en Inteligencia Artificial (IA), aprendizaje automático (AA), reconocimiento de patrones (RP), así como en minería de datos (MD) [6].

Angélica Guzmán-Ponce is with the Department of Computer Languages and Systems, Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castelló de la Plana, Spain. e-mail: aguzman@uji.es

Rosa María Valdovinos-Rosas, Marcelo Romero-Huertas and J. Raymundo Marcial-Romero are with the Facultad de Ingeniería, Universidad Autónoma del Estado de México, Toluca, México. e-mail: rvaldovinosr@uaemex.mx, mromeroh@uaemex.mx, jrmarcial@uaemex.mx

Ruben Fernandez-Beltran is with the Departamento de Informática y Sistemas, Universidad de Murcia, Murcia, España. e-mail: rufernan@um.es

El uso de modelos de aprendizaje profundo para procesar imágenes médicas [7] han presentado estudios como el denominado *COVIDiagnosis-N* [8], en el cual se propone una metodología basada en redes neuronales pre-entrenadas y optimizadas con Bayes, utilizando imágenes de tórax de pacientes positivos de COVID-19 para construir un modelo de predicción robusto y sostenible. El empleo de redes neuronales se ha potenciado por los resultados obtenidos en la predicción de la enfermedad, en específico, el trabajo propuesto en [9] ha adecuado la red *AlexNet* para trabajar con imágenes en una sola intensidad, con el fin de clasificar si un individuo está enfermo o no mediante la extracción de biomarcadores, con una muestra de personas sanas y con efectos provocados por la COVID-19.

La minería de reglas de asociación (MRA) también ha sido aplicada para determinar la vinculación de enfermedades, por ejemplo, diabetes e hipertensión [10]. El uso del algoritmo Apriori fue usado para rastrear la propagación del virus mediante la exploración de los síntomas [11]. Incluso MRA es usada para comprender los efectos de diferentes intervenciones no farmacéuticas como el uso de cubre-bocas u otras órdenes gubernamentales para contener el virus [12].

Incluso existen otros trabajos que buscan puntualizar patrones de comportamiento asociados a la información derivada del COVID-19, especialmente en redes sociales. La mayoría de estas propuestas implementan técnicas basadas en el procesamiento de lenguaje natural y análisis de textos.Shurrab *et al.* [13] realizó un estudio basado en el modelo *Latent Dirichlet Allocation* (LDA) [14], con objeto de mostrar que, al inicio de la pandemia (Febrero del 2021), los tópicos de discusión predominantes en Twitter de Estados Unidos estaban orientados a economía, política y a la dispersión del virus; mientras que para Marzo y Abril del mismo año, los temas estaban fundamentalmente encauzados a discutir la prevención del virus. Por otro lado, Koukaras et al. [15] proponen combinar la asignación de LDA con MRA para identificar conjuntos de palabras frecuentes y generar reglas que infieren las actitudes de usuarios en redes sociales como Twitter. Pese la utilidad demostrada de LDA y otros modelos para la extracción de patrones ocultos más allá de corpus de texto [16], su uso para el análisis de factores de supervivencia al COVID-19 todavía no ha sido explorado.

México es uno de los países más afectados por la COVID-19, ocupando el tercer lugar en muertes en el 2021 [17]. Diversos estudios muestran que en 2020 los nuevos casos y muertes se centraron en la Ciudad de México [18]. Por ejemplo, Núñez*et al.* [19] presentan un estudio cuyo objetivo

es determinar las propiedades diagnósticas de sucesos locales en el área metropolitana de la Ciudad de México, mediante el cálculo de sensibilidad, especificidad, valor predictivo positivo y negativo, razón de verosimilitud positiva y negativa, en casos sospechosos y confirmados por vínculo epidemiológico de laboratorio. Sus resultados sugieren la evaluación formal de incidencias sospechosas, especialmente en lugares con alta carga de enfermedad y/o pruebas limitadas para identificar, aislar y rastrear contactos de manera efectiva. Adicionalmente, Guzmán-Torres *et al.* [20] publican un trabajo basado en los datos abiertos de México, a fin de establecer las primeras diez condiciones en pacientes infectados que causan su muerte. Con este propósito, los autores utilizan una técnica de *wrapping* para indicar las variables con mayor peso que provocan la enfermedad y un análisis de regresión logística con la finalidad de predecir la mortalidad del paciente con base a sus condiciones.

A pesar de todo el trabajo desarrollado por la comunidad científica, hasta el momento pocos esfuerzos han sido dirigidos a descubrir y estudiar los patrones de características que presentan las personas que han sobrevivido al COVID-19 en México, utilizando para ello técnicas de aprendizaje basadas en tópicos latentes. Por un lado, el análisis de los factores de sobrevivencia puede resultar de especial relevancia en un entorno de datos tan complejo como el correspondiente a la colección de acceso libre disponible en México. Por otro lado, el uso de métodos de extracción de tópicos, como LDA [14], puede permitir desacoplar con eficacia los patrones de factores de supervivencia al COVID-19 más relevantes en el país. A diferencia de otros tipos de técnicas como el pesado de variables o el análisis en componentes principales, los algoritmos de tópicos latentes permiten manejar datos complejos con un nivel mayor de abstracción para descubrir patrones de características ocultos (latentes) que no tienen por qué ser directamente observables en las muestras. En virtud de estos motivos, el presente trabajo se centra en analizar el grupo de características con LDA, en casos de personas positivas de COVID-19 con comorbilidades que fueron hospitalizadas y no llegaron a fallecer.

II. Tópicos Latentes

En general, los modelos de tópicos latentes pueden entenderse como esquemas gráficos probabilísticos que contienen una o más variables aleatorias ocultas útiles para descubrir la estructura subyacente de una colección de datos [21]. De esta forma, estos métodos son capaces de extraer los patrones generativos (tópicos) de una determinada colección, así como representar los propios datos según estos patrones.

Dentro de todos los algoritmos de tópicos existentes, uno de los más ampliamente utilizados por su versatilidad y eficiencia es Latent Dirichlet Allocation (LDA) [14]. Concretamente, LDA modela las instancias de datos como distribuciones multinomiales de ocurrencia de características que son generadas a partir de una distribución a priori de tipo Dirichlet. La Figura 1 muestra la representación gráfica del modelo, donde α representa la probabilidad a priori de la distribución Dirichlet, θ es la mezcla de K tópicos expresada como distribución multinomial, z es la variable latente que asocia las características

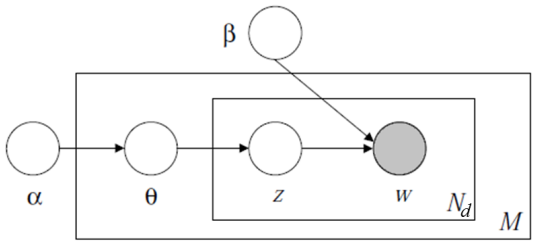

Fig. 1. Representación gráfica del modelo LDA.

visibles de los datos con los tópicos, β representa los K tópicos definidos como vectores de N términos del vocabulario de características y w son las características propias. Finalmente, M es el número de muestras de la colección y N_d el número de características de dichas muestras. Bajo este esquema, el proceso generativo según LDA puede describirse como sigue:

1. Seleccionar el número de características de la muestra, $N_d \sim Poisson(\xi)$.
2. Seleccionar el vector de parámetros de la mezcla de tópicos, $\theta \sim Dirichlet(\alpha)$. Notar que α un vector de K componentes (K es el número total de tópicos) donde $\alpha_k > 0$ y θ es un vector K-dimensional de tal forma que $\theta_k \geq 0$ y $\sum_{k=1}^{K} \theta_k = 1$. Lógicamente, $p(\theta|\alpha)$ representa la función de densidad de probabilidad de la distribución Dirichlet.
3. Para cada una de las N_d características de la muestra:
   a) Seleccionar un tópico $z_n \sim Multinomial(\theta)$.
   b) Seleccionar una palabra w_n de la distribución condicional $p(w_n|z_n, \beta)$, donde β es una matriz $K \times N$ (N es el máximo número posible de características únicas en los datos) tal que $\beta_{ij} = p(w_j|z_i)$ para todo $1 \leq j \leq N$ y $1 \leq i \leq K$.

Tomando en consideración este proceso generativo, LDA busca optimizar la probabilidad a posteriori del modelo dada una colección de datos específica, con objeto de estimar los parámetros θ y β de máxima verosimilitud. Este proceso de estimación suele llevarse a cabo mediante un algoritmo de inferencia variacional, donde se alivia la intratabilidad de marginalizar el espacio latente mediante un límite inferior basado en la desigualdad de Jensen [14]. De este modo, la entrada del modelo LDA es una colección de muestras definida en un determinado espacio de características y las salidas son una distribución de patrones de características o tópicos (identificada por β) y la distribución de muestras expresada en el espacio de tópicos (identificada por θ). Notar que, dentro del proceso de inferencia variacional, ambos parámetros son inicializados de forma aleatoria para ir actualizándose iterativamente hasta alcanzar la convergencia del modelo.

III. Estrategia metodológica

Tal y como ha sido comentando anteriormente, el presente trabajo tiene por objetivo estudiar el conjunto de datos del COVID-19 en México desde la perspectiva de las técnicas de tópicos latentes para poder descubrir nuevos patrones característicos de supervivencia a la enfermedad. En términos generales, el proceso metodológico implementado en este trabajo

está conformado por tres fases diferenciadas: la adquisición de las muestras, el pre-procesamiento y el modelado en tópicos latentes (Figura 2). Las siguientes sub-secciones detallan cada una de estas fases.

A. Adquisición de Datos

La base de datos de COVID-19 en México es de acceso abierto y se encuentra disponible para cualquier usuario interesado en el sitio Web de la Dirección General de Epidemiología[1]. La información contenida, es capturada por la red de centros de salud de México, la cual genera registros desde el 14 de abril del 2020, a inicios de la pandemia, hasta el día de hoy. Concretamente, contiene información general de pacientes como nacionalidad, sexo, edad, procedencia, entidad, sector de salud en el que el paciente fue atendido, conocimiento de enfermedades previas, como neumonía, EPOC (enfermedad pulmonar obstructiva crónica) y asma, además de, lógicamente, los respectivos resultados del paciente a la prueba del COVID-19.

Con los fines experimentales de este estudio, se trabajó con los datos de los pacientes positivos por COVID-19 del 14 de abril de 2020 al 31 de diciembre de 2020, y del 1 de enero de 2021 al 31 de diciembre de 2021. Además, el estudio se enfoca en dos estados céntricos de México: la Ciudad de México y el Estado de México, debido a que estudios previos sugieren que estos estados tuvieron un alto índice de muertes en comparación con el resto de la República Mexicana [22].

B. Pre-procesamiento de Datos

La Dirección General de Epidemiología proporciona un diccionario de datos, donde cada uno de los atributos que maneja el conjunto son de tipo categórico, por ejemplo, con valores numéricos 1 para mujeres, 2 para hombres, y 99 no especificado.

Dado que este estudio se centra en las personas que sobrevivieron al COVID-19, se tomaron únicamente casos positivos filtrados a partir de la clasificación final del propio conjunto de datos como casos confirmados SARS-COV2 por la Asociación Clínica Epidemiológica y por el Comité de dictaminación, conjuntamente con casos sospechosos. Asimismo, se consideraron casos con resultados positivos a pruebas de laboratorio y antígeno y que no tengan fecha de defunción. Además, se consideró el siguiente conjunto de características con base a su relevancia:[2]

- Sexo: *mujer/hombre*
- Intubado: *intubadoSI/NO/SIG*
- Neumonía:*neumoniaSI/NO/SIG*
- Embarazo: *embarazoSI/NO/SIG*
- Diabetes: *diabetesSI/NO/SIG*
- EPOC: *epocSI/NO/SIG*
- Asma: *asmaSI/NO/SIG*
- Enfermedades inmunosupresoras: *inmusuprSI/NO/SIG*
- Hipertensión: *hipertensionSI/NO/SIG*
- Enfermedad cardiovascular: *cardiovascularSI/NO/SIG*

[1]https://www.gob.mx/salud/documentos/datos-abiertos-152127

[2]Cuando se ignora la presencia de la característica, se agrega *SIG*.

TABLA I
DISTRIBUCIÓN DE DATOS DE LOS CONJUNTOS SELECCIONADOS.

Estado	Año	Vivos	Muertos
Ciudad de México	2020	360,902	25,626
	2021	980,215	51,836
Estado de México	2020	90,571	16,951
	2021	213,254	30,583

- Obesidad: *obesidadSI/NO/SIG*
- Enfermedad renal crónica: *renal_cronicaSI/NO/SIG*
- Tabaquismo: *tabaquismoSI/NO/SIG*
- UCI: *uciSI/NO/SIG*
- Edad: De acuerdo al rango establecido

En el caso del atributo edad, se consideraron los siguientes rangos para completar el vocabulario de características utilizado:

- Mayores de 60 (edad $>= 60$): *adulto60*
- Adulto de 50 (edad $<= 59$ y edad $>= 50$): *adulto50*
- Adulto de 40 (edad $<= 49$ y edad $>= 40$): *adulto40*
- Adulto de 30 (edad $<= 39$ y edad $>= 30$): *adulto30*
- Adulto de 20 (edad $<= 29$ y edad $>= 18$): *adulto20*
- Menor (edad $<= 17$ y edad $>= 12$): *menor*
- Niño (edad $<= 11$): *nino*

Considerando todos estas características, el resto de atributos no considerados (por ejemplo, la fecha de actualización, la entidad de nacimiento y nacionalidad, entre otros) no son objeto de estudio para este trabajo. Los datos se preprocesaron mediante la tokenización de las muestras, teniendo en cuenta que en ningún caso se consideran los valores que aparecen como no especificados en los atributos. Seguidamente, cada instancia del conjunto se convierte a tipo texto, incluyendo lógicamente cada uno de sus atributos, para poder generar los histogramas de características (también llamadas palabras) que servirán como entrada del algoritmo LDA. Tabla I muestra los detalles de la distribución de datos de los cuatro subconjuntos considerados.

C. Modelado en Tópicos Latentes de Casos Positivos

Una vez que las muestras han sido pre-procesadas y traducidas a formato texto, cada instancia se representa como un histograma de ocurrencia de características según el modelo de codificación denominado *bolsa de palabras*. En este esquema, cada muestra es codificada como un vector del tamaño del vocabulario de características (según la notación $N_d = N = 15$), donde cada posición representa una característica concreta y el valor contenido representa la aparición de dicha característica en la muestra. Posteriormente, la matriz de cada conjunto de datos (tamaño $M \times N$, siendo M el número de pacientes supervivientes de cada conjunto) es utilizada como entrada del algoritmo LDA para obtener como resultado K tópicos latentes (según Tabla II) que expresan la probabilidad de aparición de las características consideradas.

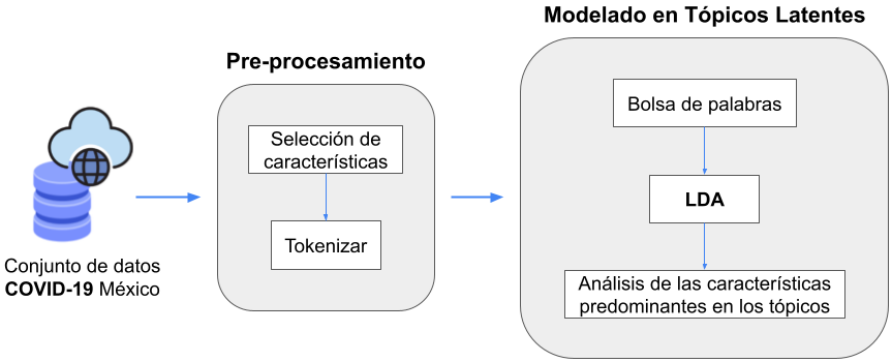

Fig. 2. Metodología usada en el presente trabajo para la identificación de tópicos latentes en pacientes sobrevivientes al COVID-19 en México.

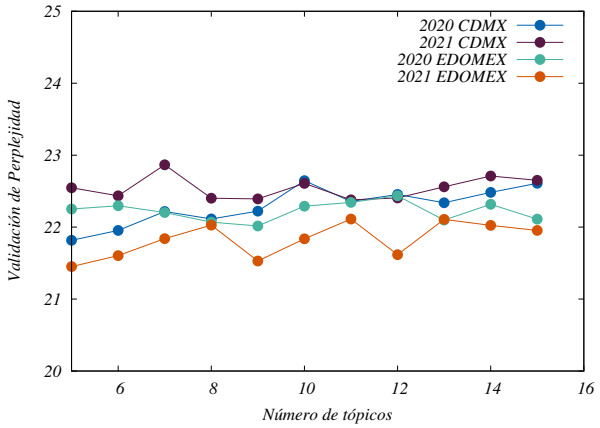

Fig. 3. Perplejidad de conjuntos de datos según el número de tópicos.

IV. Escenario Experimental

Para determinar el número de tópicos más adecuado para los conjuntos de datos, se aplicó un proceso llamado análisis de perplejidad que es típicamente usado en el contexto de las técnicas de tópicos [21]. Concretamente, la perplejidad es una medida predictiva para modelos probabilísticos, de forma que un valor bajo indica qué tan bueno es el modelo a la hora de representar la muestra completa de datos [23].

En nuestro caso, se empleó el análisis de perplejidad para estimar el rendimiento de los tópicos extraídos por LDA a medida que vamos variando su número (K), programado en MATLAB. En más detalle, usamos LDA con valores de K que fueron gradualmente incrementados de 5 a 15 para encontrar el número más apropiado. Lógicamente, se generan índices de perplejidad para cada ejecución de LDA. La Figura 3 muestra la perplejidad que resulta de las ejecuciones de LDA sobre los datos con un número variable de tópicos. Cada valor de perplejidad representa la media ponderada de la inversa del logaritmo de la verosimilitud para un conjunto externo con el 10 % de las muestras. Con base a estos resultados, en este trabajo se estableció para cada conjunto datos el número de tópicos indicado en Tabla II.

TABLA II
NÚMERO DE TÓPICOS CONSIDERADOS PARA CADA CONJUNTO DE DATOS.

Estado	Año	Número de tópicos
Ciudad de México	2020	5
	2021	11
Estado de México	2020	9
	2021	5

V. Resultados

En esta sección, se reportan los resultados obtenidos de los principales tópicos y el análisis de las características de las personas que presentaron COVID-19 y sobrevivieron. Al analizar las peculiaridades de los pacientes de COVID-19 en México, se resume el significado de los tópicos de la siguiente manera (Figura 4).

En 2020, un tópico latente en Ciudad de México indica un 14 % de probabilidad de ser mujer y sobrevivir cuando se presenta un 29 % de no ser intubada en una misma probabilidad de no presentar alguna enfermedad inmunosupresora. Además, de no presentar una enfermedad cardiovascular en un 27 %. Otro tópico latente, se encuentra en un 20 % de probabilidad de ser hombre y con un 3 % de ser fumador, sobrevivirá cuando se presente un 40 % de no tener asma y en un 36 % no presentar obesidad.

Para el Estado de México, un tópico presenta en un 30 % de probabilidad de ser hombre y sobrevivir cuando en un 58 % no manifiesta asma, pero en un 13 % presenta neumonía. Otro caso peculiar en el Estado de México se encuentra en los adultos de 40 años (17 % de probabilidad) cuando se tiene un 65 % de probabilidad de no tener obesidad, pero en un 8 %, 5 % y 1 % de presentar, diabetes, ser fumador y tener una enfermedad inmunosupresora, respectivamente.

El año 2021 fue el periodo en el cual la vacunación se llevó a cabo en toda la República Mexicana en los diversos sectores de población. Derivado de esto, el análisis de tópicos latentes se resume de la siguiente manera (Figura 5).

Para la Ciudad de México, una persona sobrevivirá a pesar de presentar en un 7 % y 3 % de probabilidad, hipertensión y neumonía. Además, sobrevivirá si, en un 89 % de probabilidad no tiene asma. Un tópico distintivo, se generó a partir de tener

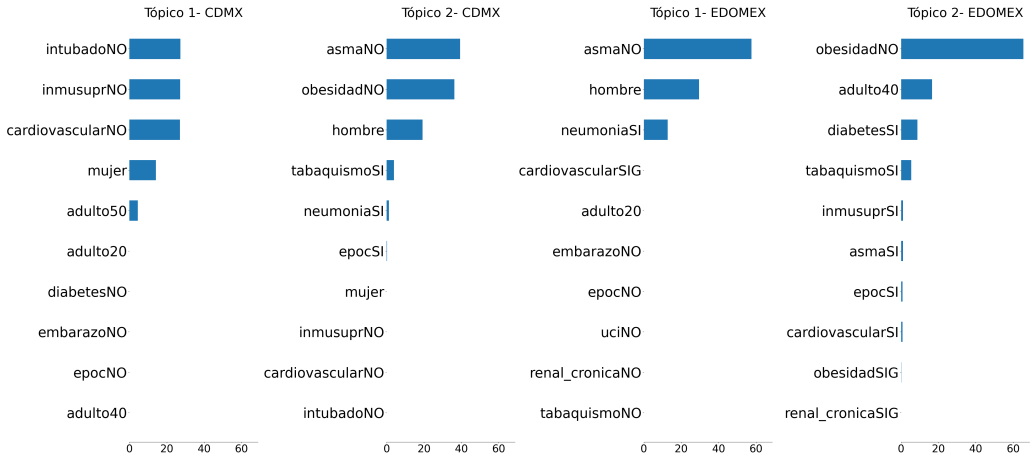

Fig. 4. Análisis de tópicos latentes para 2020, donde cada columna representa un tópico y las filas muestran sus características más probables. Notar que el eje horizontal expresa el valor de probabilidad en %.

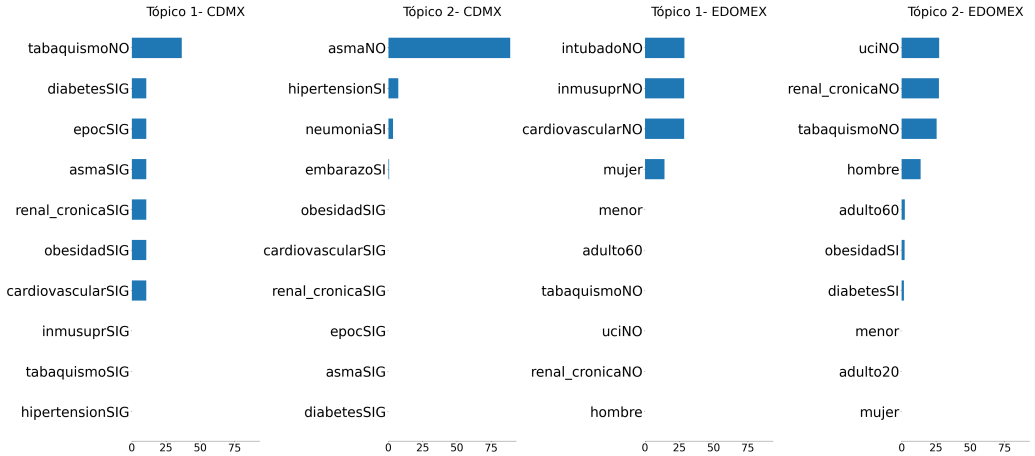

Fig. 5. Análisis de tópicos latentes para 2021, donde cada columna representa un tópico y las filas muestran sus características más probables. Notar que el eje horizontal expresa el valor de probabilidad en %.

un 55 % de probabilidad de no ser fumador, pero se ignora en su mayoría en un 10 % el tener diabetes, EPOC, asma, algún problema renal crónico, obesidad o una enfermedad cardiovascular, este tópico sugiere que la información está comprometida por diversas razones, por ejemplo, por el propio conocimiento de los pacientes, respecto a padecer alguna enfermedad o que el personal de salud no logró un seguimiento oportuno.

Por otro lado, en el Estado de México, un tópico latente de sobrevivientes se encuentra en un 13 % de ser hombre con un 3 % de ser adulto de 60 años, sobrevivirá cuando hay una probabilidad del 28 % de no estar en UCI, así como en un 27 % de probabilidad de no tener una enfermedad crónica renal. Asimismo, en un 25 % no fumar, pero si en un 2 % de tener obesidad y diabetes. Otro tópico latente de sobrevivir se encuentra cuando en un 14 % de probabilidad de ser mujer y de igual forma en un 29 % no ser intubada, además de manifestar en un 28 % no tener una enfermedad inmunosupresora, de igual manera no tener una enfermedad cardiovascular.

Con la llegada de las vacunas a México en el 2021 [24], el panorama de supervivencia posterior al contagio cambio. Los datos analizados permitieron identificar los tópicos latentes que sugieren que una persona positiva a COVID-19 a pesar de tener asociado algún factor de riesgo, ser un adulto mayor o inclusive presentar complicaciones intrahospitalarias graves como la neumonía, ha podido sobrevivir. Esta situación es posible observarse cuando se comparan los resultados con las muestras del 2020.

VI. Conclusiones

En este trabajo se analizaron los registros de pacientes hospitalizados posterior al contagio por el COVID-19 en el Estado de México y la Ciudad de México durante el año 2020 y 2021. Con la ayuda del algoritmo de tópicos latentes LDA, se extrajeron y analizaron patrones de características distintivas de los pacientes positivos a COVID-19 que no fallecieron con la intención de brindar un panorama amplio del comportamiento de la enfermedad en estos escenarios.

Para cada escenario, el número de tópicos fue seleccionado de acuerdo al análisis de perplejidad cuyo objetivo es determinar las mejores configuraciones. Una vez analizados 1,341,117 registros correspondientes a la Ciudad de México del periodo 2020-2021 y 303,825 registros del Estado de México en el

mismo periodo, ha sido posible identificar cómo los tópicos latentes variaron entre cada uno de los años y pudieron haber influido en la supervivencia de los pacientes.

En relación con identificar las condiciones de las personas que sobrevivieron al COVID-19, los resultados mostraron que el no presentar obesidad, así como, ninguna enfermedad cardiovascular fueron las condiciones más favorables en los casos positivos registrados en el 2020. Contrario a lo identificado en el 2021, en los que una gran cantidad personas que sobrevivieron al padecimiento lo hicieron aun cuando presentaron diabetes, hipertensión u obesidad y siendo adultos mayores. Esto último es verdaderamente importante, ya que puede estar dando evidencia de la efectividad de las vacunas y podrían reforzar la hipótesis vertida en múltiples fuentes de información, en la que se afirma que la vacunación favorece la supervivencia, pese a contar con comorbilidades y ser del grupo de riesgo.

De los tópicos resultantes del 2020 se resalta un par de características tanto para la Ciudad de México como para el Estado de México, y son la presencia de neumonía y que los pacientes sean hombres. Esto último sugiere que la población que sobrevivió en ambos estados a pesar de tener complicaciones por COVID-19 como es la neumonía, no fallecen siendo hombres. De igual modo, en el Estado de México, se observa que, en los adultos de 40 años, la presencia de enfermedades inmunosupresoras, así como de enfermedades pulmonares como EPOC o asma, no causó muerte por COVID-19, este comportamiento no se refleja en la Ciudad de México.

Es importante mencionar que una de las limitaciones de este trabajo es que el conjunto es alimentado por el personal de salud y en algunos casos se ignora el seguimiento oportuno a los pacientes. Es decir, se indica que fueron hospitalizados o se les realizó la prueba para determinar COVID-19, pero no se reportan los resultados. Esta situación se presenta en al menos 13 características del conjunto usado. Desafortunadamente, la falta de monitoreo tanto de las altas médicas, así como del esquema de vacunación, se convierten en limitantes del conjunto. Pues, para el último caso, a pesar de que la estrategia de vacunación para la población en general inició a principios del 2021, hasta el momento en las bases de datos oficiales, no existe un atributo que permita dar seguimiento.

Las líneas abiertas de estudio se orientan a la aplicación de diversas técnicas de limpieza en los conjuntos para eliminar datos superfluos, así como la aplicación de técnicas de imputación para identificar los valores faltantes y disminuir el ruido que las muestras no especificadas podrían estar generando. Además, resulta interesante plantear la realización de un análisis de manera georreferenciada de la información, en la que se incluyan todos los estados de la República Mexicana, así como los fenómenos migratorios presentados durante el 2020 con las caravanas migrantes. Por último, una línea de estudio sugerente, es considerar pacientes fallecidos, de este modo ajustar un modelo que permita visualizar el panorama general de supervivientes y fallecidos.

AGRADECIMIENTOS

Este trabajo fue validado en su interpretación con el apoyo del Dr. Julián Martínez Navarro, Especialista en urgencias médicas (Cédula 09270193) adscrito al Hospital General Sahuayo, Michoacán, México, en el área de atención crítica a pacientes infectados por SARS-CoV-2. También fue parcialmente apoyado por el proyecto 6364/2021SF de la UAEM. Angélica Guzmán-Ponce contó con el apoyo del contrato postdoctoral Margarita Salas MGS/2021/23(UP2021-021) financiado por la Unión Europea-NextGenerationEU.

REFERENCIAS

[1] W. Liu, C. Yang, Y. Liao, F. Wan, L. Lin, X. Huang, B. Zhang, Y. Yuan, P. Zhang, X. Zhang, Z. She, L. Wang, and H. Li, "Risk factors for COVID-19 progression and mortality in hospitalized patients without pre-existing comorbidities," *J. Infect. Public Health*, vol. 15, no. 1, pp. 13–20, 2022.

[2] "https://datos.covid-19.conacyt.mx/,"

[3] J. Tardif, M. Cossette, M. Guertin, N. Bouabdallaoui, M. Dubé, and G. Boivin, "Predictive risk factors for hospitalization and response to colchicine in patients with COVID-19," *Int. J. Infect. Dis.*, vol. 116, pp. 387–390, 2022.

[4] P. Liu, A. Blet, D. Smyth, and H. Li, "The Science Underlying COVID-19 Implications for the Cardiovascular System," *Circulation*, vol. 142, no. 1, pp. 68–78, 2020.

[5] J. Tian, X. Yuan, J. Xiao, Q. Zhong, C. Yang, B. Liu, Y. Cai, Z. Lu, J. Wang, Y. Wang, S. Liu, B. Cheng, J. Wang, M. Zhang, L. Wang, S. Niu, Z. Yao, X. Deng, F. Zhou, W. Wei, Q. Li, X. Chen, W. Chen, Q. Yang, S. Wu, J. Fan, B. Shu, Z. Hu, S. Wang, X. Yang, W. Liu, X. Miao, and Z. Wang, "Clinical characteristics and risk factors associated with COVID-19 disease severity in patients with cancer in Wuhan, China: a multicentre, retrospective, cohort study," *Lancet Oncol.*, vol. 21, no. 7, pp. 893–903, 2020.

[6] Q. Pham, D. Nguyen, T. Huynh-The, W. Hwang, and P. Pathirana, "Artificial intelligence (ai) and big data for coronavirus (covid-19) pandemic: A survey on the state-of-the-arts," *IEEE Access*, vol. 8, pp. 130820–130839, 2020.

[7] Z. KARHAN and F. AKAL, "Covid-19 Classification Using Deep Learning in Chest X-Ray Images," in *TIPTEKNO*, (Virtual), pp. 1–4, 2020.

[8] F. Ucar and D. Korkmaz, "COVIDiagnosis-Net: Deep Bayes-Squeezenet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images," *Med. Hypotheses*, vol. 140, p. 109761, 2020.

[9] E. Cortés and S. Sánchez, "Deep learning transfer with alexnet for chest x-ray covid-19 recognition," *IEEE Lat. Am. Trans.*, vol. 19, pp. 944–951, Jun. 2021.

[10] M. Tandan, Y. Acharya, S. Pokharel, and M. Timilsina, "Discovering symptom patterns of COVID-19 patients using association rule mining," *Comput. Biol. Med.*, vol. 131, p. 104249, 2021.

[11] N. Uma, A. Arulanandham, G. Keerthy, and N. P. B, "A novel approach for tracking the spread of covid-19 disease and discovering the symptom patterns of covid-19 patients using association rule mining," in *ICONAT*, pp. 1–6, 2022.

[12] S. Katragadda, R. Gottumukkala, R. Bhupatiraju, A. M. Kamal, V. Raghavan, H. Chu, R. Kolluru, and Z. Ashkar, "Association mining based approach to analyze covid-19 response and case growth in the united states," *Sci. Rep.*, vol. 11, no. 1, p. 18635, 2021.

[13] S. Shurrab, Y. Shannak, A. Almshnanah, H. Khazaleh, and H. Najadat, "Attitudes evaluation toward covid-19 pandemic: An application of twitter sentiment analysis and latent dirichlet allocation," in *12th ICICS*, (Valencia, Spain), pp. 265–272, 2021.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journ. mach. Learn. research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[15] P. Koukaras, C. Tjortjis, and D. Rousidis, "Mining association rules from covid-19 related twitter data to discover word patterns, topics and inferences," *Inf. Syst.*, vol. 109, p. 102054, 2022.

[16] R. Fernandez-Beltran and F. Pla, "Latent topics-based relevance feedback for video retrieval," *Pattern Recognit.*, vol. 51, pp. 72–84, 2016.

[17] H. Paiva, R. Magalhães-Afonso, D. Sanches, and F. Ribeiro-Pelogia, "COVID-19 Trend Analysis in Mexican States and Cities," in *43rd EMBC*, (Virtual), pp. 1820–1823, 2021.

[18] V. Suárez, M. Suarez-Quezada, S. Oros-Ruiz, and E. Ronquillo-De-Jesús, "Epidemiology of COVID-19 in Mexico: from the 27th of February to the 30th of April 2020," *Rev. Clin. Esp.*, vol. 220, pp. 463–471, 2020.
[19] I. Núñez, Y. Caro-Vega, and P. Belaunzarán-Zamudio, "Diagnostic precision of local and world Health Organization definitions of symptomatic COVID-19 cases: an analysis of Mexico's capital," *Public Health*, vol. 205, pp. 187–191, 2022.
[20] J. A. Guzmán-Torres, E. M. Alonso-Guzmán, F. J. Domínguez-Mota, and G. Tinoco-Guerrero, "Estimation of the main conditions in (SARS-CoV-2) Covid-19 patients that increase the risk of death using Machine learning, the case of Mexico," *Res. Physics*, vol. 27, p. 104483, 2021.
[21] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.
[22] S. Dahal, R. Luo, M. H. Swahn, and G. Chowell, "Geospatial Variability in excess death rates during the COVID-19 pandemic in Mexico: Examining Socio Demographic, Climate and Population Health Characteristics," *Int. Jour. Infec. Diseases*, vol. 113, pp. 347–354, 2021.
[23] V. Principe, R. de Souza-Vale, J. de Castro, L. Carvano, R. Henriques, de V.J. Almeida e Sousa Lobo, and de R. Alkmim Moreira Nunes R., "A Computational Literature Review of Football Performance Analysis through Probabilistic Topic Modeling," *Artif. Intell. Rev*, p. 1–21, 2021.
[24] M. Sánchez-Talanquer, E. González-Pier, J. epúlveda, L. Abascal-Miguel, J. Fieldhouse, C. del Río, and S. Gallalee, "La respuesta de méxico al covid-19: Estudio de caso," *Institute for Global Health Sciences*, 2021.

Angélica Guzmán-Ponce She was conferred a Ph.D. degree in Computer Science, from the Autonomous University of the State of Mexico in 2021. She is a member of the National System of Researchers (CONACYT) Level I. She is currently a postdoctoral research in the University Jaume I (Castellon de la Plana, Spain) and the Universitat Politècnica de València. Her research interests lie in Machine Learning and Graph Theory.

Ruben Fernandez-Beltran (SM'20) earned a B.Sc. degree in Computer Science, a M.Sc. in Intelligent Systems and a Ph.D. degree in Computer Science, from the University Jaume I (Castellon de la Plana, Spain) in 2007, 2011 and 2016, respectively. He is currently an Assisstant Professor within the Department of Computer Science and Systems at the University of Murcia, Spain. His research interests lie in multimedia retrieval and spatio-spectral image analysis.

Rosa María Valdovinos-Rosas She was conferred a Ph.D. degree in Computer Science. She is a member of the National System of Researchers (CONACYT) Nivel II and the AMEXCOMP. She has been a Full-Time Lecturer-Researcher at the Autonomous University of the State of Mexico (UAEMex). She contributes to the strengthening and consolidation of the scientific community through the training of quality human resources and disseminating knowledge and science in academic-scientific events at a national and international level.

Marcelo Romero-Huertas He was conferred the degree of Philosophy Doctor in Computer Science in 2011 from the University of York (England). He has been a Full-Time Lecturer-Researcher at the Department of Engineering of the Autonomous University of the State of Mexico (UAEMex) since 2011. He is a member of the IEEE, the National System of Researchers (CONACYT) and the Mexican Academy of Computing. His research interest includes image processing and pattern recognition, data network communication protocols.

J. Raymundo Marcial-Romero received his PhD in Computational Science from Birmingham University in 2005. He has been a Full-Time Lecturer-Researcher at the Department of Engineering of the Autonomous University of the State of Mexico. He is a member of the National System of Researchers (CONACYT) and Nivel I. His research interest includes approximation theory, computational complesity and graph theory.