Tendencias Sociales. Revista de Sociología, 7 (2021): 154-176
Facultad de Ciencias Políticas y Sociología

Temas latentes asociados a la aparición de las nuevas variantes del coronavirus

Latent themes associated with the appearance of new variants of the coronavirus

Farouk Ait Nasser
Universidad Nacional de Educación a Distancia (UNED)
farouk.aitnasser@hotmail.com

Resumen. En el presente trabajo realizamos un análisis que combina el enfoque cuantitativo y cualitativo. Esto nos ofrece resultados válidos que nos ayudan a evaluar la percepción de riesgo y el planteamiento de prioridades que los usuarios han manifestado en castellano vía Twitter, independientemente del tipo de perfil o de la ubicación, durante el periodo de estudio, que coincide con la aparición de las nuevas variantes del coronavirus. Por medio del modelado de temas, analizamos una base de datos de 108.134 tweets, con el objeto de identificar los temas latentes. Asimismo, al analizar la reacción pública empleando el método Asignación Latente de Dirichlet (LDA, por sus siglas en inglés), obtenemos unos resultados consistentes que facilitan la interpretación de la reacción de usuarios de Twitter de habla hispana. Entre los resultados cabe destacar la falta de interés por las variantes del coronavirus como cuestión científica. Asimismo, al registrarse un interés motivado por el peligro que representan, se observa que las reacciones se asocian más con la gestión política.

Palabras clave: COVID-19, LDA, modelado de temas, nueva cepa, Twitter.

Abstract. In the present work, we carry out an analysis that combines the quantitative and qualitative approach, capable of offering us valid results that help us to evaluate the perception of risk and the setting of priorities that every user has expressed in Spanish via Twitter, regardless of the type of profile or location, during the study period that coincides with the appearance of the new variants of the coronavirus. Through topic modeling, we analyzed a database of 108,134 tweets, in order to identify latent topics. Likewise, when analyzing public reaction using the latent Dirichlet allocation (LDA) method, we obtain consistent results that help us interpret the reaction of Spanish speakers Twitter users. Among the results, it is

worth highlighting the lack of interest in the variants of the coronavirus as a scientific matter. Likewise, when registering an interest motivated by the danger they pose, it is observed that the reactions are more associated with political management.

Keywords: COVID-19; LDA; new strain; topic modeling; Twitter.

INTRODUCCIÓN

El interés que promueve esta investigación es analizar la reacción de los usuarios de Twitter en relación con el surgimiento de las nuevas variantes que causan el COVID-19. Tomamos como punto de referencia en este estudio el anuncio oficial de la aparición de estas variantes, el primero el día 14 de diciembre de 2020, por parte de Reino Unido, y, poco después, el día 18 de diciembre, por parte de Sudáfrica. Ambos países comunicaron a la OMS (Organización Mundial de la Salud) el surgimiento de la variante conocida como SARS-CoV-2 VOC 202012/01 (en el caso de Reino Unido) y 501Y.V2 (en el caso de Sudáfrica). En ambos casos, se reconoce la rápida propagación de estas dos nuevas cepas (OMS, 2020).

Según el marco temporal establecido en este estudio, hemos propuesto analizar el modo en cómo perciben los usuarios de esta red social estas nuevas amenazas, tanto en el periodo anterior como en el posterior a las fechas de anuncio de las nuevas variantes. El motivo de la delimitación de este ámbito temporal es poder recolectar las reacciones, ya sea antes como después de que se confirmase la aparición de estas variantes, realizando un estudio que tiene como objetivo la conformación y análisis de una base de datos amplia que refleja la reacción frente a una amenaza que afecta a toda la humanidad, especialmente a los más vulnerables. De ahí, pues, hemos considerado conveniente la elección del siguiente periodo, del 1 de diciembre de 2020 al 9 de enero de 2021.

El anuncio de las nuevas mutaciones genéticas en los medios de comunicación en diciembre de 2020 generó mucha preocupación y se plantearon muchos interrogantes, como la posibilidad de aumento de riesgo de la enfermedad y la eficacia de las vacunas frente a las nuevas variantes que causan el COVID-19 (Johns Hopkins Medicine, 2021). El riesgo de padecer la enfermedad genera inquietud, y las medidas de contención del virus provocan consecuencias psicológicas de distinta índole. De hecho, existen varios estudios en castellano que tratan los efectos adversos en la salud mental que provocan estas medidas vinculadas al aislamiento social (Rojas Jara, 2020; Caballero Domínguez & Campo Arias, 2020; Palomino Oré & Huarcaya-Victoria, 2020). Nuestro interés es estudiar los factores que pueden contribuir en nuestra comprensión de cómo se reacciona y qué soluciones se aportan en un momento de gran tensión extendiéndose a los diferentes componentes sociales y suponiendo un reto de adaptación al nuevo paradigma impuesto por el COVID-19.

Entender los efectos de un problema de salud global sin repuesta médica, capaz de tambalear los cimientos de toda civilización humana, requiere de una proyección multidimensional que abarca el ámbito político, económico, sociológico, filosófico e, inclusive, el jurídico; además de otras perspectivas y enfoques de estudio (Belloso Martín, 2020). Sin embargo, en este trabajo, consideramos que una base de datos de gran volu-

men que almacena las reacciones en relación con un determinado fenómeno aporta una visión de conjunto acercándonos a la comprensión de la reacción inmediata que los usuarios en la plataforma Twitter han aportado. La base de datos que resulta de este trabajo está constituida por 108.134 tweets; se presenta como una muestra y un ejemplo que atestigua la reacción pública que utiliza la red social Twitter.

Conviene señalar que en todas las librerías mencionadas, así como los procesos automatizados seguidos para llevar a cabo esta investigación, se ha utilizado el lenguaje de programación Python versión 3.7.

Revisión de literatura

En nuestra revisión de la literatura hallamos, por consiguiente, pocos estudios en castellano que emplean la red social Twitter como objeto de estudio, empleando algún método de aprendizaje automático y sin limitarse a un perfil concreto de usuario. Nos referimos a la minería de grandes muestras de datos con el objeto de analizar un comportamiento colectivo que emplea las redes sociales como fuente de datos.

Por otra parte, hallamos varios estudios que comparten el siguiente enfoque metodológico, donde a menudo se seleccionan algunos perfiles muy concretos, ya sean periodistas (Daniela Barboza *et al*., 2016), partidos políticos (Jivkova Semova, Requeijo-Rey, & Padilla-Castillo, 2017), instituciones (Carrasco Polain, Villar-Cirujano, & Martín Cárdaba, 2019), etc.

En el presente estudio se parte del interés que genera el análisis de la percepción y la reacción a las variantes emergentes del coronavirus en las redes sociales. En cuanto a nuestra revisión de la literatura (realizada hasta febrero de 2021), no hemos hallado en castellano estudios que parten de un procedimiento similar al nuestro extrayendo un gran volumen de datos de las redes sociales en relación con la pandemia ocasionada por COVID-19, con el fin de optar a un análisis con la perspectiva de adentrarse en la exploración de las reacciones de los ciudadanos y los medios de comunicación.

Con el avance de la pandemia el orden y la convivencia a escala socioeconómica se han alterado y «la forma de convivencia de los grupos sociales y el funcionamiento de las organizaciones se ha visto desafiada» (Cuadra Martínez, Castro Carrasco, Sandoval Díaz, Pérez Zapata & Mora Dabancens, 2020: 1149). Si nos referimos a la reacción en base a las medidas de prevención impuestas como la cuarentena y el aislamiento, hay que tomar en consideración que se trata de un cambio brusco y repentino a nivel de comportamiento colectivo, habitualmente, ajeno a este tipo de restricciones. Conviene señalar que la reacción psicológica, que varía de un individuo a otro, surge a raíz del

riesgo de poder contraer la enfermedad y de propagarla infectando a los más vulnerables. Así pues, nos referimos, sobre todo, al miedo, la inseguridad, el temor y la preocupación que provoca un malestar (Robles Sánchez, 2020) y, en determinados casos, trastornos psicológicos (Cuadra Martínez *et al*., 2020).

El estudio llevado a cabo por Broche, Fernández y Reyes (2021) analiza este aspecto centrándose en el costo que supone en la salud mental del individuo. Al seguir un enfoque metodológico consistente en la revisión bibliográfica y documental, concluyen, subrayando, la importancia del acceso a la información y la mantención de canales de retroalimentación para minimizar el estrés provocado en los periodos de cuarentena. Chacón, Fernández y García (2020) señalan la frágil respuesta médica y precisamente la atención psicológica que no cubre la demanda psicológica incluso en un país desarrollado como España. Por otra parte, Urzúa, Vera, Caqueo y Polanco (2020) reivindican la importancia de la incorporación del conocimiento y de las técnicas psicológicas tanto en las campañas de prevención como en los medios de comunicación y en las nuevas tecnologías.

Desde un punto de vista social y político, cabe mencionar que las reacciones de los gobiernos, generalmente, ha sido similar, unánime y simplista en la toma de decisiones —independientemente de la eficacia de los mismos—. El confinamiento y la separación han sido las medidas más utilizadas, y la conciencia que deriva de la diversidad de casos y situaciones donde no es recomendable el confinamiento, a menudo, ha sido ignorada, a no ser que se trate de intereses económicos (Martuccelli, 2021).

En lo que se refiere a la cuestión de la percepción de la información procedente de los medios de comunicación en relación con la pandemia, el siguiente estudio realizado por varios autores peruanos (Mejia *et al*., 2020), plantea si existe una tendencia de exageración y magnificación del problema. El estudio revela la existencia de un vínculo asociativo que establece que la televisión y las redes sociales en Perú contribuyen a la percepción de exageración y a la generación del miedo entre algunos grupos de población.

Debido al carácter global de la pandemia, causada por COVID-19, podemos dar por hecho que el gran interés científico académico está reflejado por una amplia literatura que trata sus efectos desde varias áreas del conocimiento. Tomando en consideración que en nuestro estudio empleamos el modelado de temas, y más precisamente el modelo asignación latente de Dirichlet (ALD), o latent Dirichlet allocation (LDA), como modelo estadístico generativo con el propósito de hallar temas que expresan reacciones diversas respecto a la aparición de la nueva cepa del coronavirus, entendemos que nuestro estudio se presenta como una extensión de varios internacionales escritos en inglés que han profundizado en el análisis de las reacciones de los usuarios en Twitter.

En este epígrafe no podemos señalar todos los trabajos de investigación que comparten el mismo enfoque metodológico y temática con el presente estudio, debido a la

gran cantidad de producciones existente en inglés que toman las redes sociales como objeto de investigación con relación a la pandemia, analizando la reacción hacia la misma. De ahí, pues, consideramos apropiado mencionar los más representativos, que guardan cierta similitud con nuestro trabajo y que utilizan Twitter como fuente de información aplicando la minería de datos y procesándolos por medio del algoritmo LDA o incluyéndolo entre uno de los distintos métodos aplicados. De hecho, a diferencia de los estudios en castellano, en inglés hallamos varios que se acercan a nuestro planteamiento metodológico. El objetivo no difiere, consistiendo en saber los temas de discusión en Twitter, por medio del uso del modelado de temas, sin limitarse a la selección de un determinado perfil de usuario.

Así pues, hallamos el siguiente trabajo que coincide con el objetivo marcado en esta investigación, que es el análisis de la reacción pública respecto a COVID-19 (Xue *et al*., 2020). En él se plantea la identificación de temas de discusión por medio del uso de algunos términos de búsqueda como «virus», «distanciamiento social», «cuarentena», etc.; empleando los métodos de aprendizaje automático no supervisado, análisis de sentimientos y análisis cualitativo temático con el objeto de realizar una investigación sobre la evolución de las discusiones y los sentimientos públicos durante la pandemia de COVID-19.

En el siguiente se obtiene una base de datos formada por los tweets recolectados entre marzo y abril de 2020, se utiliza el modelado de temas para procesar los tweets sin limitarse a un perfil con la intención de analizar el discurso generado en torno a la pandemia, así como el análisis de la terminología empleada y su aproximación a un contexto bélico (Ordun, Purushotham, & Raff, 2020).

Esta investigación (Valdez, Thij, Bathina, Rutter, & Bollen, 2020) emplea una base de datos de 86.581.237 tweets para analizar los temas emergentes durante la pandemia por medio del uso del algoritmo LDA. Asimismo, se subrayan las consecuencias que las medidas de aislamiento provocan entre los internautas que buscan en las redes sociales una alternativa a su aislamiento. Además, documenta los negativos efectos entre la población estadounidense que derivan del uso intensivo de las redes sociales.

MATERIAL Y MÉTODOS

Twitter en la investigación académica

Twitter es una gran fuente de información y sus servicios no se usan únicamente por sus usuarios (Salvador, Pont-Sorribes, & Codina, 2017), sino, también, como herra-

mienta de marketing, educación e investigación. El motivo que explica el aumento de interés por el uso de las redes sociales y su análisis como metodología de investigación es la gran cantidad de datos que proporciona y, además, de la existencia de varias herramientas que facilitan la extracción, el procesamiento y la gestión de estos datos (Ricaurte Quijano & Ramos Vidal, 2015; Pérez Suasnavas, Karina, & Waldo, 2020).

Así pues, este aumento del uso de las redes sociales, y muy particularmente de Twitter, si bien ofrecen una fuente inagotable de información de acceso público y grandes oportunidades para los investigadores, por otra parte, su empleo también implica desafíos, retos técnicos y metodológicos cuya superación se presenta como esencial para los estudios vinculados al ámbito académico o empresarial. De hecho, existen varios debates y discusiones académicas sobre el manejo y la realización de estos métodos tanto a nivel conceptual como técnico. (Williams, Burnap, & Sloan, 2017).

El modelado de temas

Con el objeto de inferir los temas latentes en nuestra muestra, que es, a su vez, una colección de datos discretos formada por los tweets recolectados, recurrimos a una de las técnicas del procesamiento del lenguaje natural, el modelado de temas. Generalmente, se emplea en la minería de datos con el fin de hallar patrones ocultos que ayudan a revelar el contenido semántico no descubierto. Optamos por la elección del modelo LDA. Este método fue presentado por Blei, Ng y Jordan en 2003 (Blei, Ng, & Jordan, 2003). Hasta el momento, LDA se considera como una de las técnicas más utilizadas en el modelado de tópicos (Jelodar, *et al*., 2019).

El modelo LDA es un método estadístico generativo que ofrece la posibilidad de inferir los temas latentes difíciles de hallar al manejar una gran colección de documentos. Se utiliza para modelar documentos encontrando los distintos tópicos ocultos. El tópico, o tema «tendencia», se define como una distribución de términos significativos sobre un determinado vocabulario de términos fijos no significativos (Blei & Lafferty, 2009: 73).

Con el modelo LDA deducimos el tópico tratado de acuerdo a las probabilidades asociadas a los términos; a modo de ejemplo, si el tema gira en torno al arte, deporte o gastronomía, se podría saber al hallar un conjunto de términos con alta probabilidad de estar asociados con un tema u otro. Cabe señalar que el funcionamiento del método se basa en que dentro de una gran colección de documentos los que tratan temas similares utilizan un conjunto de palabras similares, si bien el mismo documento puede estar relacionado con varios temas y un mismo término puede señalar un tema o más de uno.

El modelo LDA es un modelo no supervisado; a diferencia de otros que se utilizan para el mismo propósito, admite la posibilidad de vincular un documento con varios tópicos (Alonso Berrocal, Figuerola, & Zazo Rodríguez, 2016). Se desarrolla para producir una representación en forma de palabras clave procedente de los documentos. El objetivo es descubrir las estructuras temáticas difíciles de hallar en una gran colección de documentos (Negara, Triadi, & Andryani, 2019).

Pasos seguidos en esta investigación

Se ha diseñado un método de investigación que parte de la obtención de tweets, su recopilación en una base de datos, su «preprocesamiento» y conversión a cada uno en bolsa de palabras para así poder emplear las técnicas propuestas con el objeto de hallar los tópicos ocultos. Se toma como periodo de estudio el comprendido entre el 1 de diciembre de 2020 y el 9 de enero de 2021.

Tras el inicio de la pandemia, hemos observado que «la nueva cepa» como expresión generaba cierto interés, puesto que la mutación del SARS-COV-2 era un hecho científico. Cabe señalar que durante la fase inicial de esta investigación (principios de diciembre de 2020), la aparición y la propagación de las nuevas variantes era una realidad, pero se cuestionaba el peligro que supone el comportamiento de las mismas[1], ya que las primeras variantes surgidas poseían únicamente nombres científicos; a modo de ejemplo, VOC 202012/01. Por tanto, optamos por «la nueva cepa» por ser una expresión frecuentemente empleada en este periodo por los usuarios de las redes sociales.

Ahora bien, como primera fase de este estudio, marcamos como objetivo obtener el máximo número de tweets que contienen los términos «nueva cepa» de acuerdo al periodo de estudio. De modo que, en el proceso de recolección de estos tweets, no se establece ningún criterio de búsqueda ni filtro alguno excepto la introducción de las palabras «nueva cepa» en el buscador y la delimitación por periodo de tiempo. En total se han obtenido 108.134.

Conviene aclarar que no se ha optado por una discriminación en la selección de usuarios: cualquier resultado obtenido vinculado a los términos de búsqueda era válido para su introducción en la base de datos. Tampoco hemos establecido preferencias en cuanto a la ubicación geográfica del usuario.

[1] En este artículo periodístico podemos observar una reacción donde se minimiza el peligro de las nuevas variantes (Villabona Arenas, 2020).

La recolección de datos se realiza de forma automática. Los datos recogidos son los siguientes: nombre de usuario, el usuario, Timestamp (registro de tiempo), texto (tweet), emoticono, número de comentarios, número de «likes» y número de «retweets», tal como se observa en la siguiente tabla:

TABLA 1
Las 9 primeras filas de la base de datos

Nombre de usuario	Usuario	Timestamp	Texto	Emoticono	Comentarios	Likes	Retweets
Sr. Mamami Condori	@MiguelZuu	12/1/2020 1:32	@atvpeOigan CARAJO EUROPA ESTA EN SEGUNDA OLA van a importar la nueva cepa de Covid				
Liverpoool	@liverpo87157298	12/1/2020 2:39	Siendo aquello así el resultado fue un lord ingles de pura cepa galega de la derecha más brutal, camuflado hipócritamente bajo su nueva y falsa identidad de 'mono cosmonauta soviético'. Hasta press metió a mi mamá.			1	
Camilo Solano	@Akustronique	12/1/2020 3:34	Y sobre lo de la nueva cepa, solo puede decirse que son muchas ganas de coproparlarArmando Benedetti @ AABenedetti · Nov 30, 2020 ¿Alerta por COVID-19 en Barranquilla? En cuatro días se registraron 1046 casos nuevos. Según médicos hay una nueva cepa que sería más fuerte que la anterior. ¿Hay desabastecimiento de insumos para atender pacientes hospitalizados? ¿Qué van a hacer antes que sea tarde?		1		3
Carlos	@Carlos63186641	12/1/2020 5:24	@negrodimarco and @RussianVolgaAcá también y hoy anunciaron una nueva cepa, más gasto público en políticas de género		1		
El insurrecto	@Elinsurrecto1	12/1/2020 9:52	@SilviaBueu and @ccarballo50Nadie ha dicho que el virus no exista. Sería de idiotas negarlo, lo que esta claro que las medidas no funcionan, los datos son falsos y mentiras, y hay una hipocresía, cobardía intrinseca en el covidiano medio. es una cepa de gripe nueva, y estamos aqui parando el mundo.		4		19
RositaV	@rousymariav	12/1/2020 12:00	@AABenedettiNueva cepa…, este virus muta cada media hora. Cómo todos los virus, cómo ha pasado toda la vida y no se ha extinguido la humanidad. Por qué nos tocaría está pandemia en manos de los políticos y ministros de salud más bestias de la historia?	🙄			3

Nombre de usuario	Usuario	Timestamp	Texto	Emoticono	Comentarios	Likes	Retweets
Carlos Escobar	@carl2010	12/1/2020 13:06	@CRAMSVNo es nada nuevo. Quizás se vivió una pausa con el fmln o al fue una nueva «cepa» del virus.				
Nicolas Mohrez Muvdi	@NicolasMM17	12/1/2020 13:53	Sería amenaza de una nueva cepa de Covid 19, más de 1000 nuevos casos en Barranquilla los últimos días, más de 500 casos en cuatro días en el cesar y todos frescos de concierto en concierto con las casas llenas bailes por montón y de la ocupación de UCI poco o nada se habla		7	24	50
Gerardo E. Orta	@gerardoorta	12/1/2020 15:30	Ha un año en #Wuhan #China inició lo que se convertiría en una #pandemia de una nueva cepa del #Coronavirus que, hasta hoy, parece no tener fin. El primer caso confirmado en esa región asiática, dejaría estragos en todo el mundo en este atípico año.	🔪			

Fuente: elaboración propia.

Como segunda fase, procesamos el corpus. Para ello empleamos la librería Gensim (Rehurek & Sojka, 2010).

Preprocesamiento del corpus

Para mayor aclaración de la terminología empleada conviene indicar que:

Stopwords: es un proceso de eliminación de las palabras que no aportan un significado relevante («palabras vacías»). Dicho proceso se realiza por medio de la librería NLTK (https://www.nltk.org), que, a su vez, se basa en la base de datos WordNet, desarrollada por la Universidad de Princeton.

Bolsa de términos: palabras sin estructura sintáctica. Cada palabra se separa por comas por medio de un proceso automático denominado «tokenización», donde empleamos el módulo Word_Tokenize de la librería NLTK.

Documento: un documento está formado por una bolsa de términos que a su vez forma parte de un corpus de documentos. Cada documento procede de un tweet que tiene un límite máximo de extensión de 280 caracteres.

Tópico: se trata de un conjunto de términos que habitualmente acaecen juntos dentro del corpus (Negara, Triadi, & Andryani, 2019), del cual podemos deducir el tema tratado.

Por lo general, cuando nos referimos a los tweets, sabemos que encontraríamos palabras mal formadas o no incluidas en el diccionario, expresiones irregulares, frases incompletas o de una estructura sintáctica incorrecta, etc. (Jianqiang & Xiaolin, 2017).

Para convertir los tweets en documentos válidos para su procesamiento automático necesitamos normalizar la bolsa de palabras existente en cada tweet descartando algunas expresiones. Detallamos el proceso como sigue:

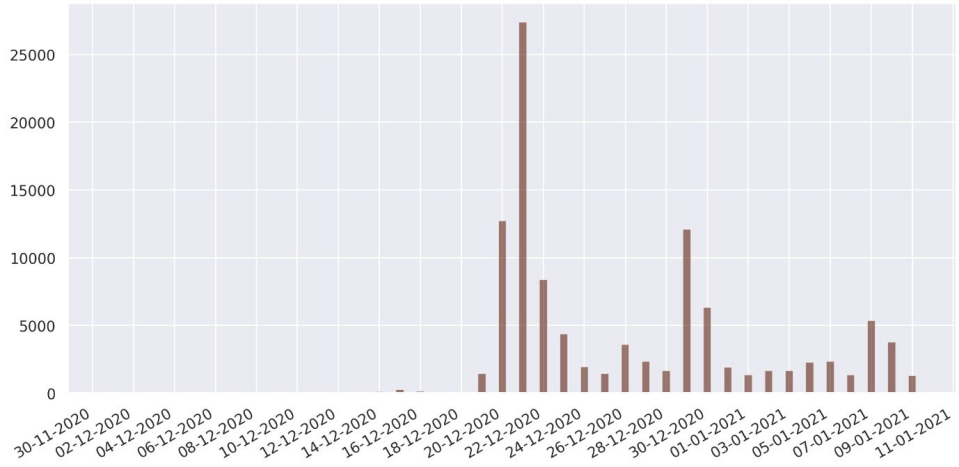
- Convertir los caracteres a minúscula.
- Eliminar los símbolos, las puntuaciones y los URL.
- Eliminar los términos de menor valor significativo por medio de la implementación de Stopword —tanto en inglés como en castellano—.
- Eliminar las siguientes expresiones: «nueva», «cepa», «corona», «coronavirus», «covid19», «virus», «covid», «nuevacepa».
- Tras revisar los doscientos términos más frecuentes, optamos por eliminar los siguientes: «q», «cada», «da», «mismo», «cómo», «hora», «jan», «después», «van», «través», «aún», «tan», «según», «mas», «dos», «así», «gran», «tras», «ahora», «dec», «si».

El motivo de la eliminación de las palabras «coronavirus», «COVID-19», «nueva» y «cepa» es limitar los resultados a aquello que pueda mostrar la reacción pública. De este modo, aumentamos las posibilidades de obtener un resultado de mayor relevancia.

RESULTADOS

El Gráfico 1 constituye una representación cuantitativa de la frecuencia de los tweets por días. Observamos que no hubo una reacción clara en Twitter en cuanto a la aparición de las nuevas variantes —y más precisamente la variante británica— hasta el 21 de diciembre, cuando se observa que el número de tweets superó los 25.000. La aparición de la variante británica se anunció el 14 de diciembre por parte del ministro británico de Sanidad, Matt Hancock. Podemos indagar sobre los motivos detrás de la demora en la reacción, ya que los tweets no superaron la barrera de 5.000 hasta el 20 de diciembre. Una reacción que se traduce en un aumento muy considerable de frecuencia de tweets el 14 y el 18 de diciembre por el anuncio de las dos variantes explicaría la existencia de una participación ciudadana más atenta e interesada por cuestiones científicas y no políticas; sin embargo, observamos lo contrario. En términos generales, podemos atribuir esta lentitud por el interés por la nueva cepa al optimismo por el inicio de las campañas de vacunación (en Reno Unido empezó el 8 de diciembre) y la respuesta de la OMS.

GRÁFICO 1
Distribución de los tweets vinculados a la «nueva cepa»



Fuente: elaboración propia.

Desde la OMS aseveraron —el mismo día de la confirmación de la mutación por parte de Mike Ryan, experto en emergencias— que son conscientes de la existencia de la variante genética reportada en mil individuos en Inglaterra, pero manifestaron la ausencia de la evidencia científica que demuestra que la variante descubierta se comportara de manera distinta (Reuters, 2020).

Según varios medios de comunicación consultados, se interpretó que la OMS había disminuido la importancia del hallazgo y se supuso que no hubo suficientes pruebas para que la nueva mutación generase más riesgo. A nuestro juicio, este factor fue determinante a la hora de explicar la falta de interés de los usuarios de Twitter, ya que, a partir del 19 de diciembre, Reino Unido confirmó que la variante descubierta se propagaba más rápidamente, y se anunciaron fuertes medidas de confinamiento por parte del primer ministro británico, Boris Johnson. Tras el anuncio, observamos un claro aumento de interés por este fenómeno, tal como muestra el Gráfico 1.

A partir del 21 de diciembre, donde distinguimos el mayor nivel de reacción, observamos fluctuaciones respecto a la frecuencia de tweets; existe una tendencia a la baja excepto en algunas fechas. No obstante, para identificar los temas latentes se procesa todo el corpus por medio de LDA; en la Figura 1 se muestra el resultado.

Conviene señalar que hemos seleccionado tres tópicos. La decisión responde al hecho de que estamos analizando una temática bastante delimitada; además, aumentar su número no nos garantiza poder discernir entre ellos fácilmente.

FIGURA 1
Tópicos identificados por medio de LDA

Fuente: elaboración propia.

La Figura 1 muestra los tres tópicos. Cada tópico se ve identificado por un conjunto de pablaras que podemos observar por medio de la «nube de términos»; para ello se ha utilizado la librería Wordcloud. Se han seleccionado 150 palabras en cada tópico.

Tal como se puede observar, existen varios términos que podrían haberse eliminado de cara a la obtención de mejores resultados. Sin embargo, contando con un amplio conjunto de palabras, esto hace posible prescindir de esta práctica. Porque, al fin y al cabo, nuestro objetivo es identificar los temas ocultos.

Tópico 1: hallamos palabras como «medidas», «confinamiento», «cuarentena», «alerta», «viajeros», «frontera», etc.; esto indica que tratamos el tema de «el anuncio de las medidas».

Tópico 2: los términos como «irresponsable», «irresponsabilidad», «vacaciones», «culpa», y otros que podemos ver de forma clara, como «claudialopez[2]», «gobierno», «país», apuntan hacia la «gestión política» como tema dominante.

Tópico 3: de las palabras como «caso», «detectó», «detectada», «mutación», «evidencia», «reino», «unido», «casos», «oms», podemos inferir que estamos tratando el tema de «el surgimiento de la nueva variante».

Al repasar varios tweets asociados al tópico 2, apreciamos un alto interés por la gestión política. El contenido de estos tweets refleja la necesidad de una evaluación crítica de las políticas seguidas, insistiendo en su fragilidad, arbitrariedad y falta de sentido común en la toma de decisiones. Cuantas más críticas, indignación y descontento social de unos ciudadanos en relación con un dirigente y su gestión política, mayor es el nivel de frecuencia y participación (en un entorno democrático). Esto es lo que explica, según el resultado obtenido, la aparición del nombre de la alcaldesa de Bogotá, Claudia López, en una investigación que tiene como objetivo el análisis de todo el contenido recolectado en Twitter en castellano de acuerdo al marco metodológico planteado. Sorprende el grado de participación de los ciudadanos colombianos y su interés por influir en la política local para hallar soluciones a la pandemia, así como las complicaciones que conllevan las medidas de confinamiento planteadas.

Por otra parte, cabe afirmar que en cada documento existe un porcentaje que lo vincula a cada uno de los tres tópicos que hemos seleccionado. A modo de ejemplo, en un mismo documento podemos hallar mayor o menor probabilidad de coincidir con algún tópico. Esto no nos garantiza que un documento pertenezca a un solo tópico de forma

[2] Claudia Nayibe López Hernández asumió el puesto de alcalde de la ciudad de Bogotá desde el año 2000.

exclusiva. Los resultados están supeditados a los hallazgos del modelo LDA, tal como podemos observar en la Tabla 2, que muestra los primeros nueve documentos y la probabilidad de pertenecer a un tópico u otro.

TABLA 2
Relación porcentual entre documentos y tópicos

	Documentos	Tópico 1	Tópico 2	Tópico 3
Doc. 1	[‘atvpeoigan’, ‘carajo’, ‘europa’, ‘segunda’, ‘ola’, ‘importar’]	*0,26*	*0,68*	*0,05*
Doc. 2	[‘siendo’, ‘aquello’, ‘resultado’, ‘lord’, ‘ingles’, ‘pura’, ‘galega’, ‘derecha’, ‘brutal’, ‘camuflado’, ‘hipócritamente’, ‘bajo’, ‘falsa’, ‘identidad’, ‘mono’, ‘cosmonauta’, ‘soviético’, ‘press’, ‘metió’, ‘mamá’]	*0,08*	0,89	0,02
Doc. 3	[‘solo’, ‘puede’, ‘decirse’, ‘muchas’, ‘ganas’, ‘coproparlararmando’, ‘benedetti’, ‘aabenedetti’, ‘nov’, ‘barranquilla’, ‘cuatro’, ‘días’, ‘registraron’, ‘casos’, ‘nuevos’, ‘médicos’, ‘fuerte’, ‘anterior’, ‘desabastecimiento’, ‘insumos’, ‘atender’, ‘pacientes’, ‘hospitalizados’, ‘hacer’, ‘tarde’]	*0,08*	0,65	0,27
Doc. 4	[‘negrodimarco’, ‘russianvolgaacá’, ‘hoy’, ‘anunciaron’, ‘gasto’, ‘público’, ‘políticas’, ‘género’]	*0,25*	0,7	0,05
Doc. 5	[‘silviabueu’, ‘dicho’, ‘exista’, ‘idiotas’, ‘negarlo’, ‘claro’, ‘medidas’, ‘funcionan’, ‘datos’, ‘falsos’, ‘mentiras’, ‘hipocresía’, ‘cobardía’, ‘intrinseca’, ‘covidiano’, ‘medio’, ‘gripe’, ‘aqui’, ‘parando’, ‘mundo’]	*0,02*	0,96	0,02
Doc. 6	[‘aabenedettinueva’, ‘muta’, ‘media’, ‘pasado’, ‘toda’, ‘vida’, ‘extinguido’, ‘humanidad’, ‘tocaría’, ‘pandemia’, ‘manos’, ‘políticos’, ‘ministros’, ‘salud’, ‘bestias’, ‘historia’]	*0,02*	0,95	0,02
Doc. 7	[‘cramsvno’, ‘nuevo’, ‘quizás’, ‘vivió’, ‘pausa’, ‘fmln’]	*0,51*	0,43	0,06

	Documentos	Tópico 1	Tópico 2	Tópico 3
Doc. 8	[‘amenaza’, ‘nuevos’, ‘casos’, ‘barranquilla’, ‘últimos’, ‘días’, ‘casos’, ‘cuatro’, ‘días’, ‘cesar’, ‘frescos’, ‘concierto’, ‘concierto’, ‘casas’, ‘llenas’, ‘bailes’, ‘montón’, ‘ocupación’, ‘uci’, ‘habla’]	*0,3*	0,56	0,14
Doc. 9	[‘año’, ‘wuhan’, ‘china’, ‘inició’, ‘convertiría’, ‘pandemia’, ‘hoy’, ‘parece’, ‘tener’, ‘fin’, ‘primer’, ‘caso’, ‘confirmado’, ‘región’, ‘asiática’, ‘dejaría’, ‘estragos’, ‘mundo’, ‘atípico’, ‘año’]	*0,02*	0,61	0,37

Fuente: elaboración propia.

Tomando en consideración el funcionamiento del modelo LDA, cada documento es una mezcla de tópicos, cada término puede atribuirse a uno o varios tópicos y cada tópico representa un conjunto de términos que a menudo acaecen juntos. Sin embargo, nos interesan únicamente aquellos documentos que representan un determinado tópico en un porcentaje igual o superior al 90 %. Queremos identificar los documentos vinculados en gran medida con algún tópico específico para evaluar así su nivel de frecuencia en todo el periodo de estudio y obtener su media de «likes» y «retweets».

Así pues, hemos filtrado los documentos. A modo de ejemplo, de los nueve primeros mostrados en la Tabla 2 se eligen únicamente 2, ya que solamente el quinto y el sexto documento cumplen con esta condición. El primero pertenece al segundo tópico en un porcentaje del 96 %. Y el siguiente, al mismo tópico al 95 %. El resultado se puede apreciar en el Gráfico 2. Por otra parte, conviene señalar que, de un total de 108.134 documentos, obtenemos este resultado que satisface nuestra exigencia:

TABLA 3
El total de documentos y su media de «retweets y «likes»

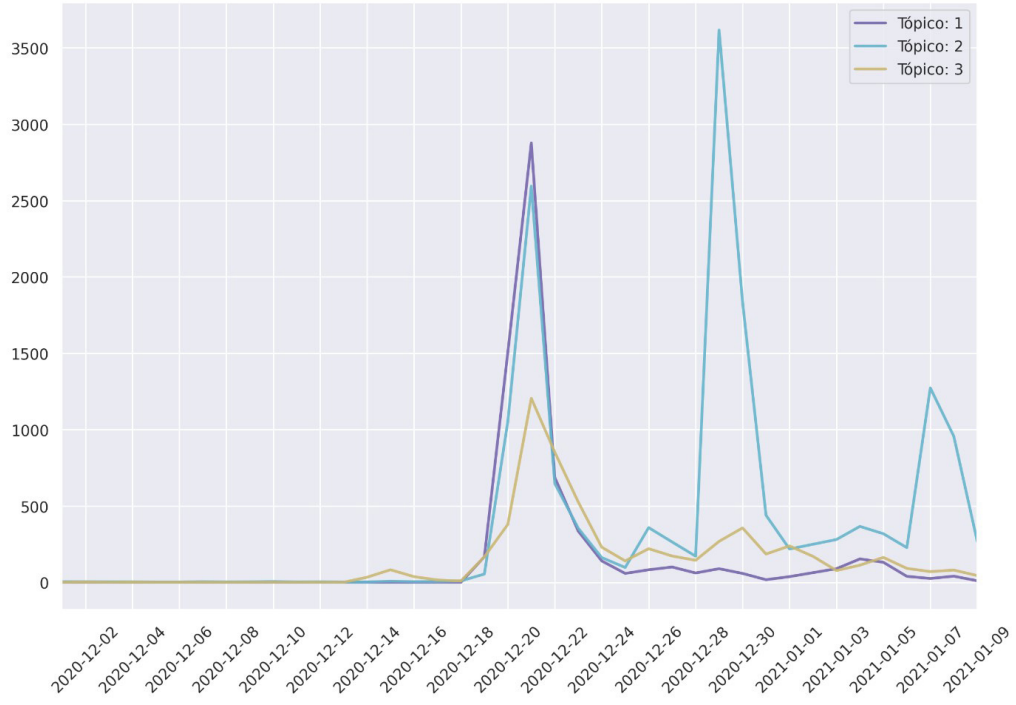
	Total	Media de «retweets»	Media de «likes»
Tópico 1	6.766	11,60	7,01
Tópico 2	15.864	14,30	4,68
Tópico 3	6.068	5,85	4,09

Fuente: elaboración propia.

Según se desprende de la Tabla 3, el número de documentos asociados al tópico 2 supera el total de documentos asignados a cada uno de los tópicos restantes. Observamos, asimismo, que el tópico vinculado a la gestión política es más «retweeteado», pero obtiene una media de «likes» no tan elevada. Generalmente, en comparación con otros tópicos, los tweets vinculados al surgimiento de la nueva variante (tópico 3) son los que menos «likes» y «retweets» tienen. En cuanto al anuncio de las medidas, aunque este tópico no ofrece una media alta en relación a «retweets», en cambio observamos más «likes». De esto deducimos que la gestión política obtiene una media alta de «retweets» y que el anuncio de las medidas alcanza una media de «likes» mayor que en el resto de los tópicos.

A diferencia del Gráfico 1, en esta última gráfica advertimos un cambio de tendencia en los niveles de frecuencia de los tópicos.

GRÁFICO 2
Distribución de los tweets vinculados a cada tópico



Fuente: elaboración propia.

Hallamos un nivel alto de frecuencia observado en los tres tópicos el 19 de diciembre de 2020. Sin embargo, en fechas posteriores observamos un cambio de interés por

un tema u otro. No obstante, se muestra evidente más actividad registrada por la frecuencia existente en el tópico 2. Son tres picos que reflejan una reacción vinculada a la gestión política. El segundo pico refleja el nivel más alto, que traspasa los 3.500. Tomando en consideración que estamos analizando los tweets de usuarios de habla hispana, no nos sorprende que el nivel más alto se sitúe en la fecha 27 de diciembre debido a la posibilidad de extensión del virus a algunos países hispanos.

DISCUSIÓN

Como limitación, señalamos que al no delimitar los tweets por zonas geográficas no se obtiene una idea clara sobre cómo se asocia la evolución de los acontecimientos relativos a la pandemia con el interés por un tópico u otro. Esto se debe a que en cada país se presenta una situación distinta en cuanto a la aparición de las nuevas variantes. Sin embargo, esto indica que, aunque surja y se anuncia la existencia de una emergencia sanitaria, no se percibe como tal, registrándose una reacción importante hasta que se anuncian los primeros casos en el mismo país, y, por ende, la reacción se asocia, sobre todo, con el tópico 2.

Por otra parte, se podría interpretar como irrelevante la extracción de los tweets durante el periodo que antecede el anuncio de la primera variante por Reino Unido (del 1 al 13 de diciembre), sin embargo, recordemos que anteriormente ya se han observado muchas voces advirtiendo sobre el posible peligro que representa el comportamiento de una o varias mutaciones genéticas del coronavirus si aparecen; y otras descartando tal hipótesis. En cuanto a la frecuencia de los tweets registrados en este periodo, aunque no se pueden observar en la Figura 1, generalmente oscilan entre 8 y 26 tweets al día. La inclusión de este periodo se presenta como necesaria para ilustrar la falta de atención hacia un eventual empeoramiento de la situación asociada a la evolución de la pandemia.

En línea con los hallazgos de esta investigación, hallamos en el estudio realizado por Xue *et al*. (2020), como ya se ha señalado, similitudes tanto en el planteamiento metodológico como en el objetivo de investigación. En cuanto a las conclusiones, aunque el estudio parte del análisis de los tweets sin enfocarse en las variantes, sino en la situación anterior, sus autores concluyen que, con la rápida evolución de la pandemia, los temas que más se repiten se asocian con los nuevos casos, la mortalidad, las medidas preventivas, las políticas gubernamentales, las autoridades sanitarias, el estigma social relacionado con COVID-19 y los efectos psicológicos. Observamos que los tópicos 1 y 2 figuran entre los temas más reiterados, aun tratándose de dos muestras distintas. En cuanto a la ausencia del tópico 3, se debe al periodo de investigación elegido.

CONCLUSIONES

No podemos considerar toda decisión implementada al inicio de la pandemia como altamente eficiente, deducimos que las sociedades encuentran grandes dificultades para reaccionar. Ya que, tal como hemos constatado, no hubo suficiente interés público por la «nueva cepa» como cuestión científica, teniendo en cuenta que es un tema muy vinculado a la realidad actual impuesta por la pandemia. Esto indica que si las pandemias son fenómenos cíclicos, sus consecuencias, a menudo, serán desastrosas debido a la lentitud que se observa en la reacción. Con esto queremos decir que cuando el peligro no se percibe como tal, pocas oportunidades se tiene de prepararse correctamente para minimizar sus consecuencias.

Un interés alto por la gestión política y las medidas impuestas en la pandemia puede tener varias explicaciones, pero a menudo destaca la cuestión del desacuerdo en la aplicación e intensificación de las medidas o la falta de aplicación de las mismas. En todo caso, difícilmente podemos calificar un proceso democrático de toma de decisiones vinculadas a las medidas preventivas frente al COVID-19 como altamente efectivo.

Las redes sociales desencadenan un rol de agitación política, y en relación con el contexto de la pandemia, sostenemos que la reacción de los ciudadanos en masa, vía Twitter, puede contribuir a la toma de decisiones políticas, pero no debe ejercerse bajo una influencia sin fundamento científico. Por otra parte, comprendemos que las medidas no son fáciles de adoptar debido al alto coste económico que entrañan. Sin embargo, la toma de decisiones que pueden resultar eficaces para limitar las negativas consecuencias no son fácilmente deducibles en un momento idóneo y a veces tampoco es factible su aplicación por falta de consenso; esto ralentiza la capacidad de reacción y, al no actuar con diligencia, el tiempo transcurre y las malas predicciones se convierten en consecuencias.

A nuestro parecer, la falta de atención respecto a las nuevas variantes del virus que ha causado la pandemia vinculada al COVID-19 podría deberse no solamente a la falta de interés o cansancio provocado por la situación de estrés causada por las medidas preventivas como el aislamiento o las noticias sobre la creación de las vacunas, sino también por intereses políticos con el ánimo de aliviar las dificultades vinculadas con la situación económica del momento. Debido a ello, no fue sorprendente hallar en los medios de comunicación un contenido que cuestiona la aparición de las nuevas variantes o el peligro que suponen.

En resumidas cuentes, y en relación con lo tratado en este estudio, cabe indicar que la pandemia ha supuesto un enorme desafío social, y en torno a ella seguirá habiendo un gran debate científico. Se trata de un tema que generará muchas producciones científi-

cas, y sabemos que en un futuro cercano la interpretación en relación con el presente artículo podría no tomar en cuenta que todo el proceso de redacción del mismo finalizó en marzo de 2021. Durante este periodo no se ha observado la existencia de estudios en castellano que analizan la pandemia y su impacto en las redes sociales empleando la minería de grandes colecciones de datos y el procesamiento natural del lenguaje o el aprendizaje automático como método de investigación para analizar la reacción de los usuarios. Asimismo, con la confirmación de las nuevas mutaciones genéticas por parte de la comunidad científica no se ha generado un interés significativo en la red social Twitter por parte de los usuarios hispanohablantes. Tras anunciarse las medidas preventivas vinculadas a este fenómeno por parte de Reino Unido, se empieza a percibir una reacción asociándose, sobre todo, con la gestión política.

Por otra parte, con la consolidación de las técnicas del aprendizaje automático, el paradigma cuantitativo y cualitativo como métodos de investigación están adquiriendo más relevancia debido a las distintas aportaciones en varias disciplinas. Trabajar con estas técnicas en el ámbito de las ciencias sociales nos proporciona varias ventajas que podrían beneficiar cualquier investigación.

REFERENCIAS BIBLIOGRÁFICAS

Alonso Berrocal, J., Figuerola, C., & Zazo Rodríguez, Á. (2016). Análisis de temas emergentes a través de Twitter. *Scire: Representación y organización del conocimiento*, 67-73.

Belloso Martín, N. (2020). La multidimensionalidad de una pandemia. Sociedad y Derecho en la era del post-coronavirus. *Cuadernos Electrónicos de Filosofía del Derecho* (43).

Blei, D., & Lafferty, J. (2009). Topic models. En *Text Mining: Classification, Clustering, and Applications.* Mineáapolis: University of Minnesota.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of machine Learning research,* 3 (Jan), 993-1022.

Broche-Pérez, Y., Fernández-Castillo, E., & Reyes Luzardo, D. (2021). Consecuencias psicológicas de la cuarentena y el aislamiento social durante la pandemia de COVID-19. *Revista Cubana de Salud Pública*.

Caballero Domínguez, C., & Campo Arias, A. (2020). Problemas de salud mental en la sociedad: Un acercamiento desde el impacto del COVID 19 y de la cuarentena. *Duazary: Revista internacional de Ciencias de la Salud, 17*(3), 1-3. doi:https://doi.org/10.21676/2389783X.3467

Carrasco Polain, R., Villar-Cirujano, E., & Martín Cárdaba, M. (2019). Redes, tweets y engagement: análisis de las bibliotecas universitarias españolas en Twitter. *El profesional de la información,* 28(4). doi:https://doi.org/10.3145/epi.2019.jul.15

Chacón Fuertes, F., Fernández Hermida, J., & García Vera, M. (2020). La psicología ante la pandemia de la COVID-19 en España. La respuesta de la organización colegial. *Clínica y Salud*, 119-123.

Cuadra Martínez, D., Castro Carrasco, P., Sandoval Díaz, J., Pérez Zapata, D., & Mora Dabancens, D. (2020). COVID-19 y comportamiento psicológico: revisión sistemática de los efectos psicológicos de las pandemias del siglo XXI. *Revista Médica de Chile*, 1139-1154.

Daniela Barboza, M., Blanco, L., Meleán, R., Páez Moreno, Á., Silva, P., & Villasmil, M. (2016). Una lectura cibergráfica sobre la interactividad en las cuentas de Twitter que manejan los periodistas venezolanos. *Temas de comunicación*, 162-188.

Hidayatullah, A., & Ma'arif, M. (2017). Road Traffic Topic Modeling on Twitter. *International Conference on Sustainable Information Engineering and Technology (SIET)* (pp. 47-52). IEEE.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications,* 78(11), 15169-15211.

Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access, 5*, 2870-2879.

Jivkova Semova, D., Requeijo-Rey, P., & Padilla-Castillo, G. (2017). Usos y tendencias de Twitter en la campaña a elecciones generales españolas del 20D de 2015: hashtags que fueron trending topic. *El profesional de la información*, 824-837.

Johns Hopkins Medicine. (22 de 2 de 2021). Obtenido de Johns Hopkins Medicine: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/a-new-strain-of-coronavirus-what-you-should-know

Martuccelli, D. (2021). La gestión anti-sociológica y tecno-experta de la pandemia del Covid-19. *Papeles del CEIC*, 1-16. doi:http://dx.doi.org/10.1387/pceic.21916

Mejia, C., Rodriguez Alarcon, J., Garay Rios, L., Enriquez Anco, M., Moreno, A., Huaytan Rojas, K., … Curioso, W. (2020). Percepción de miedo o exageración que transmiten los medios de comunicación en la población peruana durante la pandemia de la COVID-19. *Revista cubana de investigaciones biomédicas*.

Negara, E., Triadi, D., & Andryani, R. (2019). Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. En 2019 International Conference on Electrical Engineering and Computer Science (ICECOS) (pp. 386-390). IEEE.

OMS. (31 de 12 de 2020). *Organización Mundial de la Salud*. Obtenido de https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en/

Ordun, C., Purushotham, S., & Raff, E. (2020). Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PLOS ONE, 15*(9). doi:https://doi.org/10.1371/journal.pone.0240010

Palomino Oré, C., & Huarcaya-Victoria, J. (2020). Trastornos por estrés debido a la cuarentena durante la pandemia por la COVID-19. *Horizonte médico*, 61-66.

Pérez Suasnavas, A., Karina, C., & Waldo, H. (2020). Beneficios del uso de técnicas de minería de datos para extraer y analizar datos de twitter aplicados en la educación superior: una revisión sistemática de la literatura. 32(2), 181-218. doi:http://dx.doi.org/10.14201/teri.22171

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. En *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (págs. 45-50). Valleta: ELRA.

Reuters. (14 de 12 de 2020). Obtenido de Reuters.com: https://www.reuters.com/article/health-coronavirus-who-idINKBN28O2DK

Ricaurte Quijano, P., & Ramos Vidal, I. (2015). Investigación en redes sociales digitales: consideraciones metodológicas desde el paradigma estructural. *Virtualis: revista de cultura digital,* 6(11), 165-194.

Robles Sánchez, J. (2020). La psicología de emergencias ante la COVID-19: enfoque desde la prevención, detección y gestión operativa del riesgo. *Clínica y Salud*, 115-118.

Rojas Jara, C. (2020). Cuarentena, aislamiento forzado y uso de drogas. *Cuadernos de Neuropsicología/Panamerican Journal of Neuropsychology,* 14(1), 24-28. doi:10.7714/CNPS/14.1.203

Salvador, P.-M., Pont-Sorribes, C., & Codina, L. (2017). A sample design proposal for the analysis of Twitter in political communication. *El profesional de la información (EPI),* 26(4), 579-588.

Urzúa, A., Vera Villarroel, P., Caqueo Urízar, A., & Polanco Carrasco, R. (2020). La Psicología en la prevención y manejo del COVID-19. Aportes desde la evidencia inicial. *Terapia psicológica*, 103-118.

Valdez, D., Thij, M., Bathina, K., Rutter, L., & Bollen, J. (2020). Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data. *Journal of Medical Internet Research,* 22(12). doi:10.2196/21418

Villabona Arenas, J. (6 de noviembre de 2020). Motivos para no alarmarse por la mutación del coronavirus. *elespectador.com*. Obtenido de https://www.elespectador.com/salud/motivos-para-no-alarmarse-por-la-mutacion-del-coronavirus-article/

Williams, M., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology,* 51(6), 1149-1168.

Wilson, A., & Chew, P. (2010). Term weighting schemes for latent dirichlet allocation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* (pp. 465-473). Los Ángeles: Association for Computational Linguistics.

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter Discussions and Emotions About the COVID-19 Pandemic:Machine Learning Approach. *Journal of medical Internet research,* 22. doi:http://dx.doi.org/10.2196/20550