ISSN 1870-4069

Minería de texto para identificar las principales preocupaciones de los usuarios de Twitter durante COVID-19 en la Ciudad de México

Anabel Pineda-Briseño[1], Josimar Edinson Chire Saire[2]

[1] Tecnológico Nacional de México campus Matamoros,
Departamento de Sistemas y Computación,
México
[2] Institute of Mathematics and Computer Science,
University of São Paulo,
Brazil

anabel.pineda@itmatamoros.edu.mx, jecs89@usp.br

Resumen. COVID-19 es un nuevo coronavirus que originó un brote epidemiológico a principios de año en China. Esta enfermedad se ha diseminado rápidamente por el mundo dando origen a la primera pandemia en la era digital. En este sentido, una de los principales retos de la comunnidad científica internacional es aprovechar de manera efectiva la tecnología para monitorear a la población con el fin de obtener retroalimentación (en tiempo real) de su comportamiento durante cualquier contingencia de salud pública. Una alternativa viable para materializar lo antes planteado son las redes sociales. En una red social las personas generalmente actúan como sensores que identifican e informan rápidamente acontecimientos originados en cualquier parte del mundo, además de compartir datos personales, derivados de su comportamiento e inlcusive de salud. En este trabajo se presenta un estudio a través de minería de texto para identificar las principales preocupaciones de los usuarios de Twitter en la Ciudad de México durante el desarrollo del COVID-19. Después de extraer, visualizar y analizar datos extraídos de Twitter se concluye que las principales prepcupaciones de la población están centradas en temas que tiene que ver con la cantidad de casos confirmados y con las medidas de seguridad sanitarias implementadas por el Gobierno de México.

Palabras clave: COVID-19, minería de texto, principales preocupaciones, procesamiento de lenguaje natural, redes sociales, twitter.

Text Mining for Identifying the Main Concerns of Twitter Users during COVID-19 in Mexico City

Abstract. COVID-19 is a new coronavirus that caused an epidemiological outbreak earlier in the year in China. This disease has spread

rapidly throughout the world, giving rise to the first pandemic in the digital age. In this sense, one of the main challenges of the international scientific community is to effectively take advantage of technology to monitor the population in order to obtain feedback (in real time) on their behavior during any public health contingency. A viable alternative to materialize the aforementioned is social networks. In a social network, people generally act as sensors that quickly identify and report events originating anywhere in the world, in addition to sharing personal data derived from their behavior and even their health. This work presents a study through text mining to identify the main concerns of Twitter users in Mexico City during the development of COVID-19. After extracting, viewing and analyzing data extracted from Twitter, it is concluded that the main concerns of the population are focused on issues that have to do with the number of confirmed cases and with the sanitary security measures implemented by the Government of Mexico.

Keywords: COVID-19, text mining, main concerns, natural language processing, social media, twitter.

1. Introducción

A finales del 2019 en la ciudad de Wuhan, China, se indentificó una nueva enfermedad infecciosa llamada 2019-nCoV, mejor conocida como COVID-19, la cual dio origen a un brote de neumonía viral a principios del 2020. Debido al alto nivel de contagio de esta enfermedad respiratoria su propagación se ha extendido por todos los continentes, motivo por el cual la Organización Mundial de la Salud (OMS) declaró oficialmente al COVID-19 una pandemia [19][17] el día 11 de marzo del 2020. Por lo antes mencionado, COVID-19 se ha convertido en emergencia de salud pública internacional ya que representa un alto riesgo para la humanidad y un impacto negativo en otras áreas como la economía, educación, etc. A nivel mundial actualmente esta enfermedad ha infectado a millones de personas y ha ocasionado la muerte de miles de ellas. Hasta el 15 de mayo del 2020 se han reportado ante la OMS 4,347.935 casos positivos y 297,236 defunciones por COVID-19 [18]. En México, los primeros dos casos de COVID-19 reportados a la OMS fueron del 29 de febrero del 2020, y para el 15 de mayo se reportaban 40,186 casos positivos y 4220 defunciones [10]. En relación a la Ciudad de México, el reporte hasta el 15 de mayo fue de 12,456 casos positivos y de 997 personas fallecidas a causa de esta pandemia [13]. Por otro lado, el crecimiento explosivo del uso de las redes sociales en los últimos años indica que cada vez más personas emplean esta tecnología para identificar e informar acontecimientos, además de compartir información personal, ideas, intereses, sentimientos, experiencias e inclusive información relacionada con la salud. En este contexto, las redes sociales representan una fuente de datos atractiva que puede ser explotada para propósitos de monitoreo y vigilancia de salud pública [6][9]. Entre las redes sociales más usadas destaca Twitter, un servicio gratuito para compartir mensajes de texto limitados a 280 caracteres.

Twitter cuenta a nivel mundial con alrededor de 340 millones de usuarios activos [7], mientras que en México el número de usuarios llega a 9.45 millones, siendo la Ciudad de México la que cuenta con la mayor cantidad de usuarios activos [16]. La principal contribución de este trabajo es identificar a través de minería de texto las principales preocupaciones de la población durante la pandemia COVID-19 en la Ciudad de México, en un esfuerzo por entender cómo se manifiesta y desarrolla una contingencia de salud pública en las redes sociales, específicamente empleando Twitter como fuente de datos.

El resto del artículo se organiza como sigue. En la sección 2 se presenta el trabajo relacionado al uso de datos de redes sociales para fines de vigilancia y monitoreo de enfermedades infecciosas. En la sección 3 se describe la metodología empleada para la extracción, analisis y visualización de los datos de este trabajo. En la sección 4 se presentan los resultados obtenidos. Por último, en la sección 5, se exponen las conclusiones y el trabajo futuro del presente estudio.

2. Trabajo relacionado

Existen diversas propuestas relacionadas al uso de datos de redes sociales y su valiosa contribución en el campo de la salud pública. En esta sección se presentan algunas de las propuestas más recientes. Los proyectos [14] y [2] emplean técnicas de Procesamiento de Lenguaje Natural para generar nueva información a partir de datos recolectados de Twitter, los cuales resultaron ser sumamente útiles para la investigación en salud pública. Las propuestas [3] y [9] presentan un análisis que muestra la efectividad del uso de la información de las redes sociales como estrategia de investigación en salud pública. Ambas propuestas recomiendan utilizar la combinación de redes sociales con otras técnicas para estudiar la enfermedad y su propagación. En [15] se presenta un sistema de tiempo real que ayuda a la predicción y detección de una epidemia a través de la identificación de tuits de enfermades por localización geográfica. El trabajo propuesto en [8], combina datos de Twitter con datos de Google Trends para realizar un seguimiento de la propagación de enfermedades infecciosas. Otro estudio que se presentó en [1], analizó datos de Twitter que fueron recolectados durante algunos brotes de enfermedades infecciosas. Los resultados experimentales ayudaron a entender cómo actúa la gente con transtorno de pánico en medio de una contingencia de salud. Existen también otras propuestas relacionadas a aplicaciones de monitoreo y vigilancia de enfermedades infecciosas que han empleado como fuente de datos Twitter, entre las que destacan: Monitor de la Pandemia N1H1 [4], Monitor del Dengue en Brasil [12], y Monitor del COVID-19 en Colombia [11] y Sudamérica [5]. Ahora bien, debido a que COVID-19 es una enfermedad de la cual se conoce muy poco, y que ha ocasionado un impacto negativo en diversos ámbitos en todos los países, en este artículo se presenta un estudio exploratorio encaminado a identificar las principales inquietudes de los usuarios de Twitter en México, específicamente de la Ciudad de México, una de las ciudades más pobladas del mundo y de América Latina.

3. Metodología

Este trabajo realiza experimentos sobre texto proveniente de Twitter, y aplica técnicas de Procesamiento de Lenguaje Natural. Básicamente la metodología se resume en:

- Selección palabras para la búsqueda en red social.
- Delimitación de parámetros para la búsqueda.
- Preparación del texto para ser procesado.
- Visualización.

3.1. Selección de palabras relevantes

Se consideran las palabras utilizadas en las noticias, experimentos previos de búsqueda en la red social. Se seleccionaron las siguientas palabras: coronavirus y covid19. Durante las búsquedas se encontraron variaciones como: corona-virus, corona_virus, covid-19, #covid_19, por tanto se crearon e incluyeron las combinaciones con caracteres: @, #, –, _.

3.2. Delimitación de parámetros para la búsqueda

La recolección de datos se realiza a través de la Interfaz de Programación de Aplicaciones(API) de Twitter, los parámetros usados son:

- Fechas: 13-03-2020 a 20-03-2020 y 01-05-2020 a 09-05-2020.
- Palabras claves: Mencionados en la subsección anterior.
- Geolocalización: Latitud y Longitud: (19.4333,-99.1333), ver Fig.1 1.
- Idioma: Español.
- Radio: 50 kilómetros.

Fig. 1. Geolocalización de la Ciudad de México.

3.3. Preparación del texto para ser procesado

Este paso es importante para limpiar el texto de símbolos, palabras que no añaden valor al análisis del presente trabajo, etc. El proceso consiste en términos generales en:

- Convertir de mayúsculas a minúsculas.
- Eliminar signos de puntuacón como: comas, puntos, puntos y comas, signos de interrogación, signos de exclamación, entre otros.
- Excluir palabras con un tamaño menor a 3.
- Agregar excepciones, por ejemplo: https, url, etc.

3.4. Visualización

Los datos pre-procesados son empleados para responder preguntas durante la exploración de la información. Una adecuada visualización de los valores resultantes a través de gráficas ayuda de manera significativa en la realización de un buen análisis de la información. Este trabajo se enfoca en graficar información relacionada a:

- Mostrar el progreso de la pandemia graficando el número casos positivos por día y acumulados.
- Frecuencia de publicación de los usuarios de Twitter tanto del inicio del brote de la pandemia como durante el pico máximo de brotes de ésta.
- Histograma de los términos más publicados al inicio del brote de la pandemia y durante el pico máximo de brotes de ésta.
- Nubes de palabas (unigramas) más publicadas por día, tanto del inicio del brote de la pademia como del pico máximo de brotes de la misma.
- Nubes de pares de palabras (bigramas) más publicados por día, tanto del inicio del brote de la pademia como del pico máximo de brotes de la misma.

4. Resultados

En esta sección se presentan los resultados del estudio propuesto en este artículo. Para una mejor organización esta sección se ha dividido en cuatro subsecciones. La primera despliega a manera de introducción el progreso de la pandemia en la Ciudad de México. En la segunda subsección se muestran las gráficas de los datos extraídos correspondientes al inicio de brote de la pandemia. La tercera subsección despliega las gráficas de los datos extraídos correspondientes al pico máximo de brotes de la misma. Finalmente, la última subsección presenta un análisis general de los hallazgos del estudio.

4.1. Progreso de la pandemia COVID-19

En la Fig. 2 se presenta el progreso de los casos confirmados, por día y acumulados, del COVID-19 en la Ciudad de México. Los datos fueron extraídos del sitio oficial Coronavirus de la Secretaría de Salud del Gobierno de México [13]. De acuerdo a la información disponible en el portal, los primeros casos del COVID-19 en la ciudad se reportaron a finales del mes de febrero del 2020, mientras que para el periodo entre el 6 y el 10 de mayo, la Secretaría de Salud proyectaba el pico máximo de brotes de la pandemia. Hasta el 15 de mayo, fecha que se consultó el portal del Coronavirus de la Secretaría de Salud del Gobierno de México [13], se tenían registrados un total de 12,456 casos del COVID-19 confirmados, lo cual representa un promedio 95.8 contagios diarios.

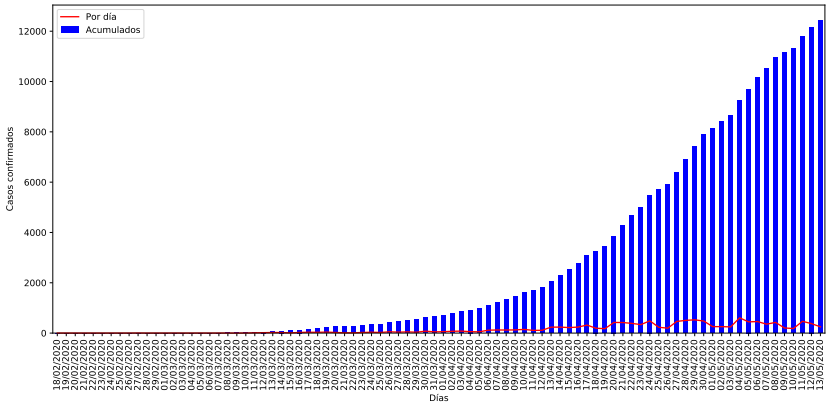

Fig. 2. Casos confirmados del COVID-19 en la Ciudad de México.

4.2. Con cuánta frecuencia y qué publicaron los usuarios de Twitter acerca del COVID-19 durante su brote en la Ciudad de México

Con el propósito de saber con qué frecuencia y qué publicaron los usuarios de Twitter acerca del COVID-19 durante el brote de casos en la Ciudad de México, se extrajeron datos de Twitter del periodo comprendido entre el 13-03-2020 y el 20-03-2020. Este periodo fue seleccionado ya que en este se identificó un incremento constante de casos positivos, tal como se puede constatar en la Fig.2. Para identificar con qué frecuencia se publicó acerca del COVID-19, en la Fig. 3 se despliega el número de tuits por día. Se puede observar en la gráfica que conforme se fueron confirmando casos, la cantidad de publicaciones incrementó, concentrándose la mayor parte en la segunda mitad del periodo. En total se contabilizaron un total de 22,755 mensajes de texto, que representan en promedio 2844 mensajes por día.

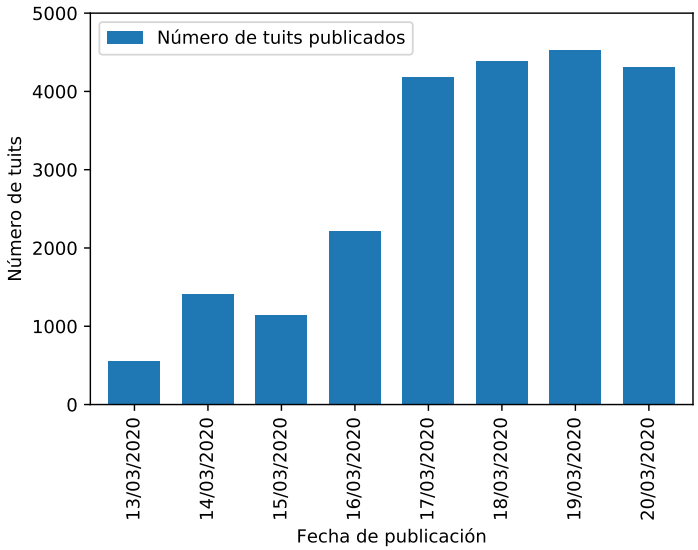

Fig. 3. Publicaciones diarias acerca del COVID-19 en la Ciudad de México al inicio del brote de la pandemia.

Por otro lado, con el fin de caracterizar el comportamiento o las preocupaciones de los usuarios de Twitter durante este periodo, fue interesante identificar y analizar las palabras más empleadas y su frecuencia de publicación. En la Fig.4 se despliegan las 30 palabras más publicadas durante el inicio del brote del COVID-19 en la Ciudad de México. Como se puede observar en la lista clasificada, las cinco palabras más populares fueron "casos", "presidente", "salud", "medidas" y "gobierno". Ahora bien, con el fin de poder tener un panorama general de lo que más se publicó en este periodo por día, se utilizaron nubes de palabras para representar gráficamente las palabras (unigramas) y pares de palabras (bigramas) más publicadas. Las Fig. 5 y 6 muestran este conjunto de nubes de palabras y pares de palabras construidas en base a los datos recolectados. En ambas representaciones, palabras y pares de palabras, se confirma la popularidad por día de la palabra "casos" y del par de palabras "casos confirmados", respectivamente.

4.3. Con cuánta frecuencia y qué publicaron los usuarios de Twitter acerca del COVID-19 durante el pico máximo de brotes en la Ciudad de México

Con el propósito de saber con qué frecuencia y qué publicaron los usuarios de Twitter acerca del COVID-19 durante el pico máximo de brotes en la Ciudad de México, se extrajeron datos de Twitter del periodo comprendido entre el 01-05-2020 y el 09-05-2020.

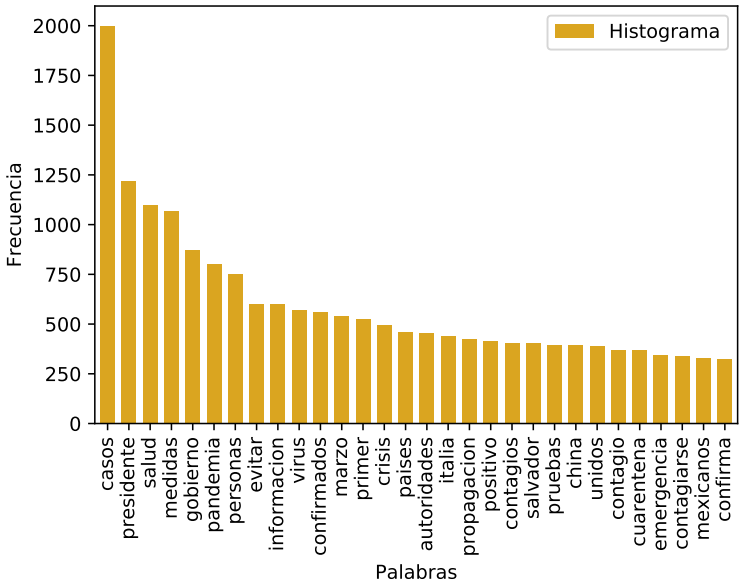

Fig. 4. Frecuencia de las palabras más publicadas acerca del COVID-19 al inicio del brote en la Ciudad de México.

Fig. 5. Nubes de palabras (unigramas) más publicados por día durante el inicio del brote del COVID-19 en la Ciudad de México.

Fig. 6. Nubes de pares de palabras (bigramas) más publicadas durante el inicio del brote del COVID-19 en la Ciudad de México.

De acuerdo a la Secretaría de Salud del Gobierno de México, dentro de este periodo de fechas se proyectó de manera inicial el pico máximo de brotes del COVID-19 en México. De manera similar a la subsección anterior, para identificar con qué frecuencia se publicó acerca del COVID-19 se graficó la cantidad de tuits diarios. Se puede observar en la Fig.7 que en términos generales la distribución de la frecuencia de publicación fue homogénea. Para este periodo fueron recuperados un total 20,703 tuits, que representan en promedio 2300 mensajes publicados por día.

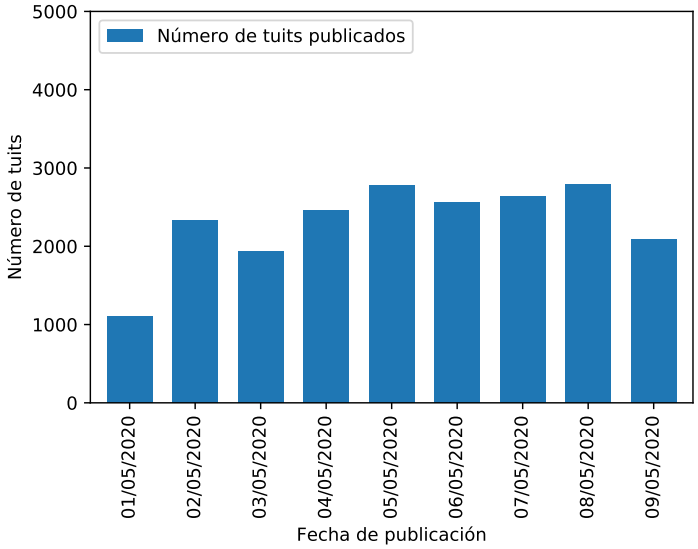

Fig. 7. Publicaciones diarias acerca de COVID-19 en la Ciudad de México durante el pico máximo de bronte de la pandemia.

En relación a las palabras más publicadas por la comunidad de Twitter en este periodo de exploración, en la Fig.8 se presenta un histograma de estas. Las palabras más populares de acuerdo a la lista clasificada son: “casos”, “gobierno”, “pandemia”, “pacientes” y “salud”. De igual manera son presentadas las gráficas de nubes de palabras más publicadas por día y nubes de pares de palabras más publicados por día. En la Fig. 9 se puede observar que la palabra más popular fue “casos”, mientras que en la Fig. 10 se identificó que en general el par de palabras más empleado por día fue “casos confirmados”.

4.4. Análisis comparativo de resultados

En esta sección se presenta un análisis comparativo de los resultados obtenidos en ambos periodos de estudio.

En primer lugar, se puede observar en la Fig.3 que durante el inicio del brote, como era de esperarse, las publicaciones diarias acerca de COVID-19 fueron

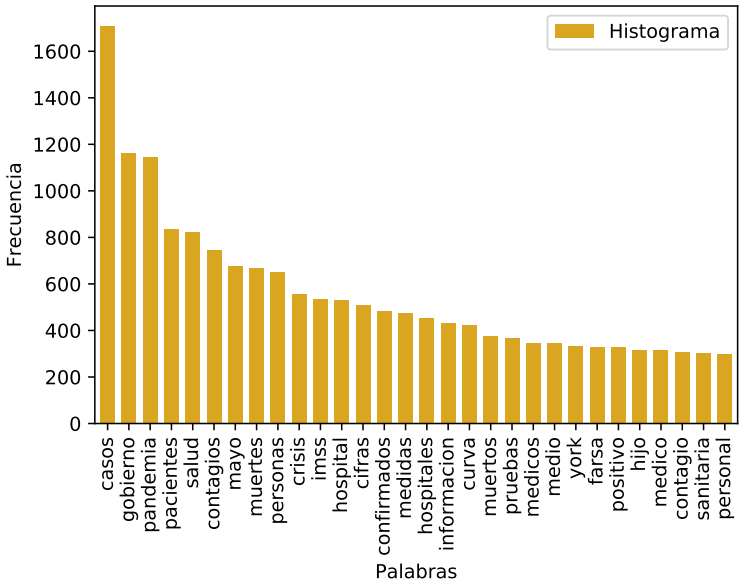

Fig. 8. Frecuencia de palabras más publicadas durante pico máximo de brotes de COVID-19 en la Ciudad de México.

Fig. 9. Nubes de palabras (unigramas) más publicados por día durante el pico máximo de brotes de COVID-19 en la Ciudad de México.

incrementando conforme se fueron reportando casos positivos. Sin embargo, de acuerdo con la Fig.7, durante el pico máximo de contagios esto no fue así, en este segundo periodo de estudio la frecuencia de publicación acerca del COVID-19 mostró en términos generales una distribución homogénea y un porcentaje menor (19.13 %) de publicaciones con respecto al inicio del brote de la pandemia.

Algo importante a destacar es que durante el pico máximo de contagios, la

Fig. 10. Nubes de pares de palabras (bigramas) más publicadas por día durante el pico máximo de brotes de COVID-19 en la Ciudad de México.

población de la Ciudad de México se encontraba en cuarentena, por lo que los datos sugieren que el exceso de información relacionada al COVID-19 pudo haber ocasionado la disminución de publicaciones debido al estrés, ansiedad y temor de la población por la emergencia sanitaria. Como información de referencia, el gobierno de México declaró cuarentena a partir del 20 de marzo del 2020, esto como parte de las estrategias para contener los contagios una vez que el país entró en una etapa de emergencia sanitaria.

Para distinguir de mejor manera cuáles fueron los temas que más han preocupado a la población de la Ciudad de México durante el inico de la pandemia y durante el pico máximo de brotes, se identificaron las 30 palabras más publicadas por periodo, destacando que coinciden un 43.33 % de las palabras más publicadas. Las tres palabras más publicadas que coinciden en el top 5 de ambas listas clasificadas fueron: “casos”, “gobierno” y“salud”, que sugieren temas relacionados con el número de casos confirmados y con las medidas de seguridad sanitarias implementadas por el Gobierno de México. En relación al resto de palabras que discreparon (56.66 %), y con el apoyo de las nubes de palabras (unigramas) y pares de palabras (bigramas), se logró identificcar otros temas de interés de la población. Por ejemplo, al inicio del brote de la pandemia la población además mostraba su preocupación publicando sobre las formas y la posibilidad de contagio; la carencia de equipos de protección médica; y la difusión de noticias falsas.

Por su parte, durante el pico máximo de brotes la población también manifestó preocupación comentando sobre la crisis sanitaria y económica; el apoyo de la Secretaría de la Defensa Nacional (Sedena) y la Secretaría de la Marina-Armada de México (Marina) con el “Plan de Auxilio a la población civil en casos de desastre” (Plan DN-III); las acusaciones lanzadas por cuatro diarios interna-

cionales a la Secretaría de Salud sobre el presunto ocultamiento de cifras de decesos y casos positivos COVID-19 en México; y por la compra de ventiladores para el COVID-19 a precios excesivos por parte del Gobierno Mexicano al hijo del director de la Comisión Federal de Electricidad (CFE) de México.

5. Conclusiones y trabajo futuro

En este artículo se presentó un estudio basado en minería de texto con el objetivo de identificar las principales inquietudes de los usuarios de Twitter durante la propagación del COVID-19 en la Ciudad de México. Los escenarios de exprimentación se enfocaron en extraer, visualizar y analizar datos durante el inicio del brote de la pandemia y durante el pico máximo de brotes de la misma. De acuerdo a los resultados se concluye que los temas centrales están fuertemente relacionados con el número de casos confirmados y con las medidas de seguridad sanitarias implementadas por el Gobierno de México, estas últimas a su vez generando el miedo por una posible crisis económica en la población de la Ciudad de México, una de las ciudades más pobladas del mundo y de América Latina. Como trabajo futuro derivado de esta investigación se pretende ampliar el campo de estudio a otras ciudades de México. Asimismo, medir el impacto mediático y el impacto que la pandemina está generando en otras áreas como la educación y la salud mental. Además de un estudio dedicado exclusivamente a monitorear el desarrollo del COVID-19 en la frontera México-USA.

Referencias

1. Ahmed, W., Bath, P.A., Sbaffi, L., Demartini, G.: Moral panic through the lens of Twitter: An analysis of infectious disease outbreaks. In: Proceedings of the 9th International Conference on Social Media and Society. pp. 217–221 (2018)
2. Breland, J.Y., Quintiliani, L.M., Schneider, K.L., May, C.N., Pagoto, S.: Social media as a tool to increase the impact of public health research. American journal of public health 107(12), 1890 (2017)
3. Charles-Smith, L.E., Reynolds, T.L., Cameron, M.A., Conway, M., Lau, E.H., Olsen, J.M., Pavlin, J.A., Shigematsu, M., Streichert, L.C., Suda, K.J., et al.: Using social media for actionable disease surveillance and outbreak management: a systematic literature review. PloS one 10(10) (2015)
4. Chew, C., Eysenbach, G.: Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. PloS one 5(11) (2010)
5. Chire Saire, J.E.: Infoveillance based on social sensors to analyze the impact of Covid19 in South American population (2020)
6. Christaki, E.: New technologies in predicting, preventing and controlling emerging infectious diseases. Virulence 6(6), 558–565 (2015)
7. Clement, J.: Global social networks ranked by number of users 2020. Accedido en 16-04-2020 a url https://www.statista.com/ (2020)
8. Hong, Y., Sinnott, R.O.: A social media platform for infectious disease analytics. In: International Conference on Computational Science and Its Applications. pp. 526–540. Springer (2018)

9. Paul, M.J., Sarker, A., Brownstein, J.S., Nikfarjam, A., Scotch, M., Smith, K.L., Gonzalez, G.: Social media mining for public health monitoring and surveillance. In: Biocomputing 2016: Proceedings of the Pacific symposium. pp. 468–479. World Scientific (2016)
10. Rios Montanez, A.M.: Mexico: COVID-19 cases and deaths 2020. URL: https://www.statista.com/ [accessed 2020-05-15] (2020)
11. Saire, J.E.C., Navarro, R.C.: What is the people posting about symptoms related to coronavirus in Bogota, Colombia? arXiv preprint arXiv:2003.11159 (2020)
12. Saire, J.E.C.: Building intelligent indicators to detect dengue epidemics in Brazil using social networks. In: 2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI). pp. 1–5. IEEE (2019)
13. Secretaría de Salud, México: Sitio oficial del gobierno de México sobre el coronavirus. URL: https://coronavirus.gob.mx/ [accessed 2020-05-15] (2020)
14. Sinnenberg, L., Buttenheim, A.M., Padrez, K., Mancheno, C., Ungar, L., Merchant, R.M.: Twitter as a tool for health research: a systematic review. American journal of public health 107(1), e1–e8 (2017)
15. Sivasankari, S., Kavitha, M., Saranya, G.: Medical analysis and visualisation of diseases using tweet data. Research Journal of Pharmacy and Technology 10(12), 4306–4312 (2017)
16. Statista Research Department: Latin America: Twitter users 2020, by country. Accessed on 16-04-2020, url https://www.statista.com/ (2020)
17. Wang, C., Horby, P.W., Hayden, F.G., Gao, G.F.: A novel coronavirus outbreak of global health concern. The Lancet 395(10223), 470–473 (2020)
18. WHO: Coronavirus (COVID-19). URL: https://covid19.who.int/ [accessed 2020-05-15] (2020)
19. WHO: WHO statement regarding cluster of pneumonia cases in Wuhan, China. Beijing: WHO 9 (2020)