ANÁLISIS DE TWITTER PARA COVID-19

HELENA GÓMEZ ADORNO, GEMMA BEL-ENGUIX, GERARDO SIERRA, GABRIEL CASTILLO.
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS, INSTITUTO DE INGENIERÍA, UNAM.

La comunicación a través de las redes sociales juega hoy en día un papel crucial en la vida de todos los sectores de la población. La información transmitida de estos medios proporciona descripciones y opiniones que pueden resultar valiosas para la toma de decisiones.

En la actualidad, las redes sociales son utilizadas para promocionar productos y servicios masivamente por diversas empresas. A su vez, las personas tienen la oportunidad de transmitir experiencias y opiniones por el mismo medio, dando lugar a una gran fuente de información textual. Por lo anterior, es posible utilizar estos conjuntos de datos textuales no estructurados para generar información sobre el comportamiento masivo, los pensamientos y las emociones en una amplia variedad de temas, como revisiones de productos, opiniones, tendencias políticas, y el sentimiento del mercado de valores.

En la situación de la pandemia actual de Coronavirus, se han decretado tiempos de cuarentena en diversos países del mundo. En particular, México inició la Fase 3 de la cuarentena el 21 de abril de 2020. Durante dicho periodo, buena parte del pueblo mexicano se ha quedado en sus casas para evitar el contagio masivo y entrar en condiciones críticas en la ocupación hospitalaria. Debido al confinamiento, se ha dado un crecimiento en el canal de comunicación mediante las redes sociales y los servicios digitales. Mediante las redes sociales, las personas, tanto en México como en el resto del mundo, pueden emitir comentarios de cómo están viviendo actualmente las consecuencias de la cuarentena, de cómo se está modificando el ritmo de vida, así como las actividades cotidianas de los habitantes.

El **objetivo** de este trabajo, es generar un sistema automático de vigilancia Covid-19, mediante el análisis de mensajes de la red social Twitter, para evaluar el comportamiento de las personas, estados de ánimos y la popularidad de las medidas dadas por el gobierno; además, monitorear usuarios con posibles síntomas de coronavirus. Esto responde a la iniciativa de la UNAM para desarrollar modelos y la visualización de información que puede apoyar en la toma de decisiones estratégicas[1].

1. https://gruposcovidunam.mx/

Metodología

La Figura 1 presenta la arquitectura general del sistema. Los *tweets* se descargan utilizando el API de Twitter por medio de un programa Python que recolecta los mensajes y los almacena en un servidor de bases de datos no relacionales (MongoDB). Utilizando un cron en Linux, realizamos el análisis (Sentimientos, Síntomas, Conteos/Estadísticas) automático diario de los *tweets* que se encuentran en la base de datos. Los resultados de los distintos análisis se almacenan en un formato JSON fácil de interpretar, luego se almacenan en otra base de datos no relacional MongoDB. La página web desarrollada consulta la base de datos de resultados de análisis y los despliega en los diversos formatos mencionados anteriormente.

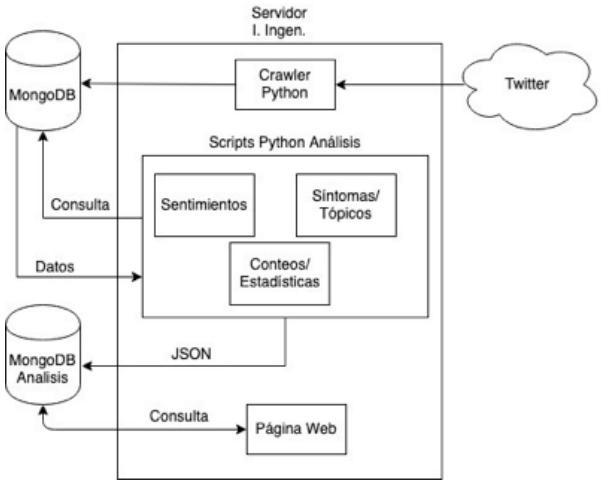


Figura 1. Arquitectura general del sistema

A partir del 1 de abril, el grupo empezó a recolectar mensajes de la red social Twitter. Hasta el 15 de septiembre, se han recogido más de 13,000,000 *tweets* que corresponden a todos los que retorna el *streaming* de Twitter, cifra que va en aumento cada día. Para el almacenamiento de estos datos, se puso en funcionamiento un servidor de bases de datos no relacionados, así como un programa que, con base en el API de Twitter, recolecta los mensajes. Todos los *tweets* han sido filtrados para que exclusivamente se consideren aquellos emitidos dentro de México.

Para el sistema automático de vigilancia, primero se han analizado de forma manual 12000 *tweets* con el fin de elaborar los siguientes recursos:

- Diccionario de variantes de términos relacionados con COVID-19, tanto científicas como coloquiales.
- Diccionario de *hashtags* relacionados con la enfermedad.
- Diccionario de sintomatología.
- Patrones que permitan reconocer estructuras que hacen referencia a síntomas o características de la enfermedad.

Con ayuda de los diccionarios recolectados, realizamos tres tipos de análisis automáticos sobre los *tweets* recolectados:

1. Análisis de contenido: Este tipo de análisis incluye el conteo de las etiquetas (#hash-tags), menciones (@usuario), y palabras más frecuentes y relevantes a la pandemia. El conteo se realiza por día, de manera que las palabras clave van cambiando; después, se presenta un total acumulado.
2. Análisis de emociones: Aquí se busca identificar el estado de ánimo de las personas mediante la clasificación de sus mensajes en las seis categorías básicas de Ekman (1992) (miedo, tristeza, asco, alegría, ira y sorpresa). Asimismo, se monitorean los sentimientos que genera la pandemia en la población, entre positivos y negativos. Se usó el Corpus de la Universidad de Jaén para el entrenamiento, el cual contiene un diccionario y varios *tweets* en español, anotados con emociones y sentimientos. Además, se utilizó el Lexicón de Emociones NRC, que contiene 14182 palabras clasificadas en positivas y negativas.
3. Monitor de síntomas: Con este análisis, se busca identificar personas que presenten o mencionen tener síntomas de Covid-19 y su correlación con la base de datos de infectados.

Entre los síntomas más comunes que presentan las personas que padecen Covid19 se encuentran: dolor de cabeza, cuerpo cortado, tos seca, fiebre, dolor de espalda, dolor de garganta, cansancio, diarrea, estornudo, hiperventilar, falta de aire, debilidad, irritación, dolor de cuerpo, cuerpo cortado, pérdida de los sentidos del olfato y del gusto, conjuntivitis, erupciones cutáneas, cambios de color, dedos, pies, manos, intenso dolor al pasar saliva, ardor de ojos, hipertensión, diabetes, cáncer, dolor en el pecho y dificultad para respirar.

Los síntomas que presentan las personas debido al aislamiento social principalmente son: depresión, cansancio, fatiga, insomnio, irritable, falta de apetito, frustración, trabajo para concentrarse, ansiedad, latidos muy rápidos del corazón, tensión, nerviosismo, miedo, taquicardia, tristeza, hipersensibilidad, aburrición, abuso de sustancias, estado de ánimo bajo, pocos deseos de hacer cosas, estar hambriento, tener pérdida de peso, engordar, desesperanza.

Resultados

En la página web[2] se reflejan en tiempo real todos los análisis realizados. La estructura de la página tiene un mapa interactivo de la República Mexicana, que permite la segmentación de información en los 32 estados que componen el país. Además, se incluyen componentes denominados tarjetas, los cuales facilitan la presentación de interés general.

En la Figura 2 se presentan los resultados del análisis de frecuencia de términos relevantes en una línea del tiempo que permite relacionarlos con las diferentes fases de la pandemia. La línea del tiempo presenta seis palabras clave (coronavirus, quedateencasa, @HLGatell, @lopezobrador, gripa y neumonía) dentro de los *tweets* obtenidos hasta el día actual, esto para conocer cómo se van mencionando a medida que pasa el

2. http://www.miopers.unam.mx/

tiempo y el virus avanza. Las primeras dos palabras son acerca del virus y el aislamiento social. Para el caso de **HLGatell** y **lopezobrador**, se buscaron menciones a sus cuentas oficiales. Finalmente, se agregan las palabras de **gripa** y **neumonía**, para conocer la frecuencia de estas enfermedades en los mensajes cortos.

El módulo de análisis de síntomas funciona como un sistema automático de vigilancia de síntomas de COVID19 para México, los síntomas presentados tienen dos categorías principales: los físicos, causados por COVID19, como fiebre, tos o gripe; y los trastornos, ocasionados por el aislamiento social, como ansiedad, depresión e insomnio, entre otros.

Los resultados de análisis de síntomas se muestran de dos maneras. La Figura 3a representa la cantidad de síntomas relacionados al COVID-19 y los estados de salud mental que se pueden presentar en la población debido al aislamiento social y demás factores. Por cada día se presentan dos valores, cada uno de ellos indica la cantidad de apariciones de síntomas de cada tipo (COVID-19 y estados de salud mental o psicológicos), esto con el fin de conocer cómo afecta el paso del tiempo a la frecuencia de estos dos valores. La Figura 3b representa la cantidad de *tweets* por Alcaldía de la Ciudad de México que contienen síntomas relacionados al COVID-19 y los estados de salud mental relacionados al distanciamiento social. Estos valores no son acumulativos, por lo que cada día cambian.

Es interesante notar que, incluso cuando la pandemia está creciendo, la cantidad de *tweets* relacionados está disminuyendo. Como era de esperar, la emoción más expresada dentro de los *tweets* es el miedo, mientras que la polaridad más común es negativa. En cuanto al análisis de los síntomas, los físicos son generalmente mayores que los trastornos. Vale la pena observar que la cantidad de *tweets* que contienen ambos tipos de síntomas se mantiene estable desde la relajación de las regulaciones de cuarentena en la Ciudad de México.

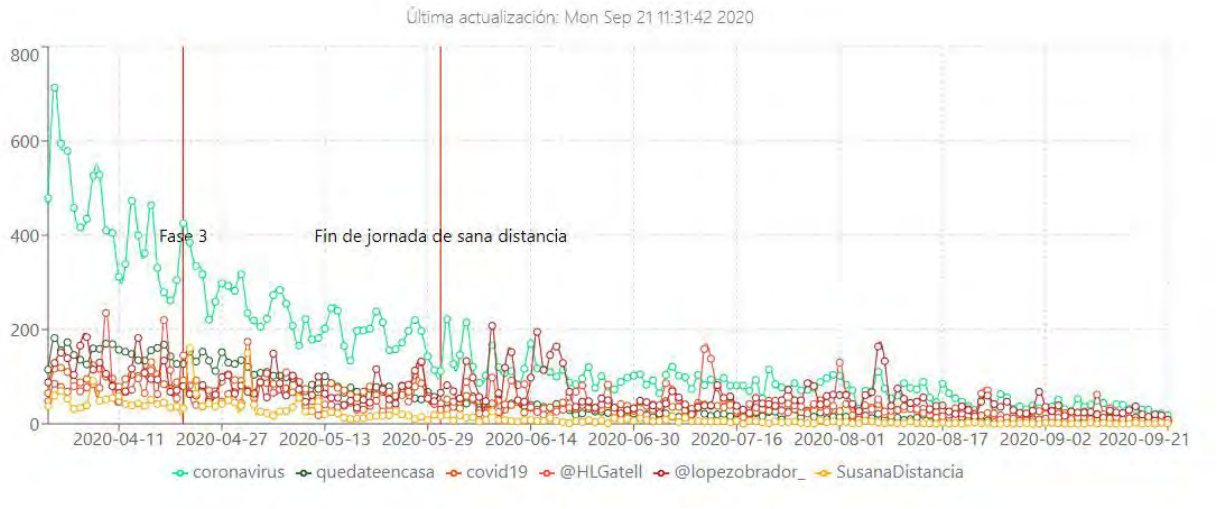
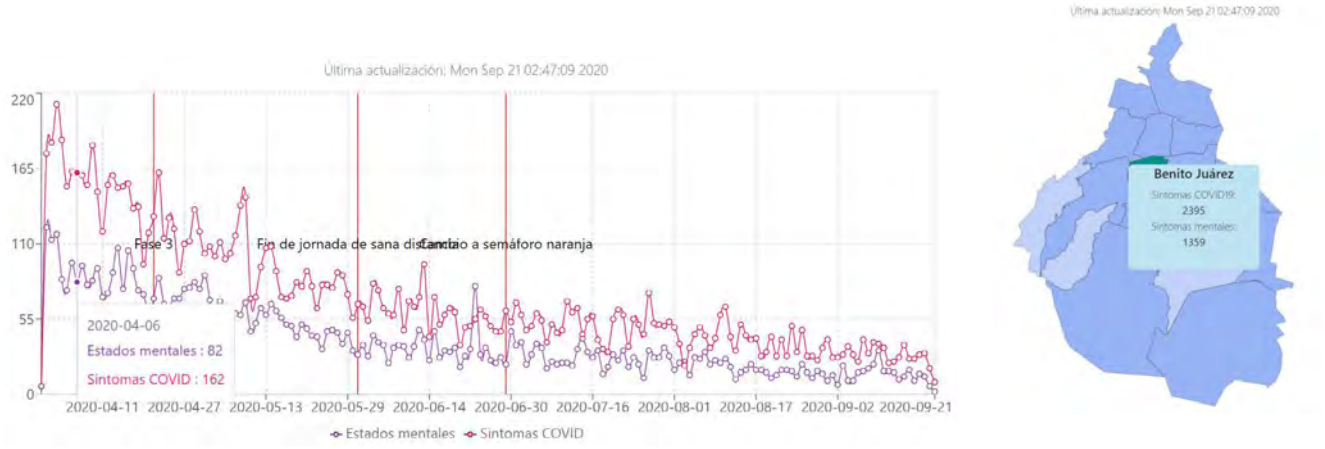


Figura 2. Cantidad de tweets con 6 palabras clave por día



a) Línea de tiempo

b) Mapa

Figura 3. a) Cantidad de tweets con síntomas por día. b) Mapa de la Ciudad de México coloreado de acuerdo a la cantidad de *tweets* con síntomas por alcaldía

Grupo de Trabajo

Coordinadores:

- Helena Gómez Adorno (IIMAS)
- Gemma Bel Enguix (II)
- Gerardo Sierra (II)

Desarrolladores y analistas:

- Gabriel Castillo Hernández (II)
- Ricardo Cruz
- Jesús Germán Ortiz
- Jessica Méndez
- José Armando López
- Pablo Camacho

PONTE PUMA

PONTE EL CUBREBOCAS

Referencias

Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3-4):169–200.