

Rastreado la pandemia: herramientas y experiencias internacionales

SEBASTIÁN HALPERÍN Y PEDRO ANTENUCCI

Breve reseña histórica del *Big Data*

Hasta hace poco tiempo, quienes trabajan en el área de investigación de opinión pública contaban con dos grandes vertientes, muy separadas por cierto, que eran las metodologías cualitativas y cuantitativas ya conocidas, en lo que tiene que ver con las encuestas, la realización de grupos focales, entrevistas y demás. Si bien hoy día se siguen utilizando estas metodologías, nos encontramos también con una tercera pata que no puede encuadrarse en ninguna de las dos categorías tradicionales porque tiene particularidades propias: se trata básicamente de lo que se conoce como el *Big Data*.

Luego de tener que lidiar con las dificultades vinculadas con las limitaciones para el procesamiento de datos en grandes volúmenes, el avance de la tecnología, que lógicamente impactó en la actividad de manera revolucionaria, hoy nos permite no solamente incrementar exponencialmente el tamaño de las muestras y variables con las que se puede trabajar, sino que, además, todo esto que se llama inteligencia artificial y *machine learning* nos ofrece la posibilidad de que, con una velocidad enorme de procesamiento, las computadoras desarrollen un aprendizaje automático.

Esto significa que, si bien no se llega a la computadora como una *tabula rasa*, cuando se analiza un problema, en este caso el Covid-19, se tiene una serie de hipótesis al momento de abordar el objeto de estudio, con base en resultados obtenidos a través de investigaciones ya realizadas.

Por ejemplo, a través de indicadores previos sabemos que hay una población que tiene más riesgo de contagio, la población de 65 años en adelante; se puede identificar que, en algunos países, la enfermedad se propagó con mayor velocidad; se puede incluso identificar que dentro de ciertas áreas, como los barrios vulnerables, la criticidad del problema es más aguda que en otros. Todo esto permite, en cierta forma, alimentar al sistema para que, a través de toda esa información, se puedan desarrollar modelos predictivos, tema central dentro del *Big Data*.

Es decir que con esa información se pueden identificar áreas de aspectos que permitan orientar el proceso de toma de decisiones. Por ejemplo, se pueden establecer, en función de estas variables, los puntos que se consideran críticos en cuanto a dónde podrá emerger con más fuerza la pandemia de cara al futuro y, en función de eso, establecer centros de salud en dichas áreas. También permite, por ejemplo, determinar patrones de comportamiento respecto a las características que tienen aquellos que han tenido una mejor *performance* en su tratamiento en cuanto a edad, sexo, nivel socioeconómico, lugares de residencia, etcétera. Además, puede contribuir a determinar el diseño y la elaboración de drogas y tratamientos, que son las medidas en las que se trabaja para el desarrollo de una posible vacuna.

A partir de estos ejemplos, se intenta demostrar que se trata de establecer correlaciones dentro del conjunto de información disponible que resulta inaprensible sin estas capacidades de procesamiento que nos propone la tecnología, y de poder establecer información útil para optimizar la toma de decisiones. De esto se trata el *Big Data*. Quiere decir que, desde la lógica conceptual, no es diferente a lo que se podía realizar con la encuesta tradicional, la diferencia está más bien en la capacidad de procesamiento y de establecimiento de modelos predictivos a partir del trabajo con un conjunto de variables que nos resultaba inmanejable hasta la irrupción de estas nuevas tecnologías.

La misma computadora, a través de la información con la que se alimenta, puede establecer modelos predictivos no solamente en puntos críticos de concentración de un problema, sino que podría también, en términos de una política pública, contribuir a determinar, por ejemplo, en qué zonas un programa social tiene mejores probabilidades de éxito por las características de su población, qué población potencial dentro de un programa social tiene mayor probabilidad de encuadrarse en dicho programa sobre la base de sus características, identificar distintos *clusters*, etcétera.

Estrategias de implementación del *Big Data*

El *Big Data* nos permite trabajar con bases de datos muy distintas, y el gran desafío es hacer todo el trabajo de consistencia. Los analistas y programadores que trabajan en *Big Data* suelen decir que, muchas veces, es mucho más complejo el trabajo de consistencia de la información, lo que se llama el *cleaning*, que significa la limpieza de las bases de datos, la integración y la homogeneización de los datos para que puedan tener un tratamiento consistente. Eso, muchas veces, incluso les lleva el 80% más de trabajo que el procesamiento mismo de la información e, incluso, que el posterior análisis.

Las técnicas de *Big Data* nos permiten manipular información de diversa naturaleza. Con las encuestas tradicionales solo se podría trabajar con indicadores cuantitativos, pero a través del *Big Data* es posible fusionar bases de datos que tengan desde datos de encuestas hasta censos y otras fuentes, incluso datos de otra naturaleza, como los textuales, que, a través de distintos sistemas de procesamiento, se pueden cuantificar. Con lo cual las posibilidades de trabajo son absolutamente infinitas, se puede incorporar muchísima información de diversa naturaleza para su tratamiento y de ese modo enriquecer el análisis.

Cuando se trabaja con distintas organizaciones existe la posibilidad de que cuenten con una base de datos y que, a su vez, incorporen otras variables, como por ejemplo la exposición a medios, gustos y preferencias, que son todas aquellas cuestiones que se tienen por hipótesis y que, a través de estudios previos, cualitativos y cuantitativos, conviven y se articulan en esta tarea. Esto permite alimentar la base de datos y enriquecer también el *output* que se va a obtener en función del proceso de toma de decisiones.

Microtargeting

Una de las cuestiones que permite hacer los trabajos de *Big Data* es lo que se llama *microtargeting* o microsegmentación. La campaña de Obama en Estados Unidos ha sido pionera en esta metodología porque permitió determinar acciones, incluso dentro de lo que en Argentina conocemos como “radio censal”, de acuerdo con circunscripciones electorales. Ello permitió establecer un abordaje territorial en cuanto a acciones de impacto en la vía pública así como en cuanto a la posibilidad de segmentar los mensajes propuestos.

Hay un caso que se suele repetir en los análisis de *Big Data* a través de la posibilidad de microsegmentar la población y que se expone con el fin de mostrar, a modo de ejemplo, el alcance de esta herramienta: Trump, al asumir en su gestión, estableció 175.000 versiones distintas de un mismo mensaje con sistemas computarizados teniendo en cuenta los distintos perfiles poblacionales identificados dentro de los Estados Unidos.

Georreferenciación

La georreferenciación, como todas las herramientas, puede tener una manipulación de control social negativa, pero también permite incrementar la precisión de los

análisis para poder afinar la puntería en el desarrollo de tratamientos médicos, la optimización de la ubicación de un centro de salud, de la oferta de servicios de salud a través de las prestaciones que se realizan en un determinado hospital o clínica, por ejemplo. A través de sistemas de geolocalización, se puede monitorear el circuito de la gente que efectivamente se registra como contagiada e, inclusive, determinar, en función de su ámbito de circulación, con qué población pudo haber estado en contacto y, con base en ello, optimizar los mecanismos de testeos susceptibles de ser efectuados.

Análisis de redes sociales

A través de las redes sociales se puede implementar lo que se denomina “análisis de sentimientos” para, por ejemplo, observar la evolución en la *performance* registrada en torno a los discursos o la actuación de los distintos líderes o referentes de la política de un determinado país; también, se puede identificar cuáles son los aspectos que registran una mejor o peor evaluación en función de las expectativas y las demandas de la ciudadanía, etcétera. Todo este análisis se realiza en función de los permisos que establezcan las distintas redes sociales para acceder a la información. Cabe recordar el famoso caso de Cambridge Analítica donde se filtró información a través de Facebook. Hoy en día existen mecanismos de protección de datos personales y distintos niveles de autorización de la información que cada usuario puede presentar a través de las diferentes plataformas.

Análisis de medios digitales

Otra de las cuestiones que han permitido el análisis de datos textuales es la cuantificación de la presencia en medios. Hoy se puede ver, entre otras cosas, qué

porcentaje de tiempo/espacio le dedican los distintos medios a un determinado descriptor, siendo estos palabras clave, contenidos asociados a la pandemia que se quieren ver o que se asume que pueden ser un indicador de una *performance* positiva o negativa en relación con la actuación de distintos líderes, interlocutores o referentes del mundo de la política, por ejemplo. Ello nos permite orientar la toma de decisiones respecto a qué tipo de discursos, en términos de comunicación política, generan más o menos pregnancia sobre la base de la posibilidad de promover una mayor o menor permeabilidad discursiva con respecto a ciertas cuestiones.

Uso de las herramientas en el caso de la pandemia y la actuación consecuente

A continuación, se expone de qué manera se han utilizado estas herramientas para gestionar la crisis del Covid-19 durante 2020. Para comenzar, y con el fin de dar un marco, cabe mencionar que cuando se piensa en el control de la pandemia hay un concepto fundamental de la epidemiología que es que la velocidad de reproducción de la pandemia depende de tres factores.

Por un lado, depende de la probabilidad de infección durante el contacto entre las personas, esto es, la probabilidad que tiene una persona de contagiarse si está con alguien que está infectado, lo cual varía en función de cómo se transmite el virus, por vía aérea en el caso del Covid-19. Para trabajar en este eje se han implementado medidas, como el uso de mascarillas, para reducir la probabilidad de que alguien quede infectado si está en contacto con una persona que lo está.

En segundo lugar, la velocidad de reproducción de la pandemia depende de la duración del período de infección, que significa durante cuánto tiempo una

persona que está infectada puede transmitir el virus. Para operar sobre esta dimensión, se recurre a medidas como la detección precoz de las personas infectadas, el testeo, el rastreo de quienes están infectados y el tratamiento.

Por otro lado, el tercer factor es la intensidad y frecuencia del contacto social, que es cuánto tiempo nos contactamos con otras personas, lo que va a influir sobre la probabilidad de contagio si alguien está infectado. Para operar sobre esta dimensión se han tomado políticas públicas, como la restricción de la circulación y la cuarentena, con el fin de reducir el contacto social y de esa manera tratar de controlar la velocidad de reproducción.

A continuación, se exponen aplicaciones del *Big Data* que tomaron distintos países fundados en las dimensiones mencionadas para disminuir la velocidad de reproducción de la pandemia. La selección se realizó de manera no exhaustiva priorizando los casos que presentaban más datos posibles de ser visualizados.

Detección precoz

Hay ciertas herramientas que permiten anticiparse a los casos. Por ejemplo, en China las búsquedas en Internet han permitido anticiparse a lo que fue efectivamente sucediendo en cuanto a los casos sospechosos y los confirmados. Al observar las búsquedas de palabras como “coronavirus” o “neumonía” en buscadores como Google o Baidu y Weibo (una red social), se destaca cómo las tendencias en las búsquedas se anticiparon a los casos sospechados y a los confirmados, lo que quiere decir que las búsquedas en Internet pueden funcionar como una herramienta para ayudar a la detección precoz.

Tal es así que en Italia las búsquedas en Internet del término “no puedo oler” reflejan un pico antes de que se

detectara que la limitación o incapacidad de oler era un síntoma propio del coronavirus (Stephens-Davidowitz, 2020).

Toda esta información, que puede parecer distante, es de fácil acceso y está disponible para distintos países y regiones. También existe para Argentina. Google Trends ofrece una evolución temporal de cómo fueron las búsquedas sobre el coronavirus y permite realizar comparaciones con otros temas de interés a nivel nacional y subnacional, ya que dispone de información desagregada por provincias y por ciudades que permite identificar dónde hay mayor interés por las distintas medidas paliativas o síntomas.

Anticipación a los casos

El *Big Data* también ha permitido que se desarrollen aplicaciones que posibilitan anticiparse a los casos. Por ejemplo, a partir de la información recabada a través de lo que se denomina el *Internet de las cosas* (IoT), que son elementos vinculados a Internet que permiten tener y registrar información que de otra manera no sería accesible, la empresa Kinsa desarrolló un mapa del estado de salud de la población (*HealthWeather*¹). Dicho mapa utiliza la información brindada por los registros de temperaturas de los termómetros que la gente tiene en sus casas, y las aplicaciones en los teléfonos móviles a los cuales estos están vinculados, para visibilizar áreas en las que se concentran lecturas de temperatura elevadas y así predecir focos de infección e identificar zonas de riesgo en los Estados Unidos.

Testeo/tratamiento: capacidad estatal

Otras dos dimensiones que pueden ser optimizadas mediante el uso del *Big Data* son el testeo y el tratamiento. Frente a la pandemia, muchos países han implementado medidas para aumentar la capacidad estatal de testeo y de

¹ <https://healthweather.us/>.

tratamiento de la enfermedad. Sin embargo, un problema fundamental que han debido afrontar radica en la identificación de las áreas en las que es más eficiente instalar centros de testeo o tratamiento adicionales, sobre todo en aquellos países donde las estimaciones poblacionales que se utilizan para tomar decisiones basadas en evidencia están desactualizadas. En Argentina la última medición exhaustiva disponible de la distribución de la población y sus características sociodemográficas es el censo de 2010. Indudablemente, los cambios poblacionales ocurridos desde entonces afectan las conclusiones a las que se pueda arribar a partir de estos datos. Sin embargo, existen herramientas que se pueden utilizar para tratar de mitigar esta dificultad.

El Departamento de Inteligencia Artificial de Facebook, mediante el procesamiento de imágenes satelitales, ha actualizado las estimaciones censales. Para ello se utilizaron herramientas de aprendizaje automático y semiautomático que permitían visualizar las imágenes satelitales disponibles en Internet para detectar dónde había nuevos asentamientos y, con base en ello, actualizar la información disponible sobre la distribución de la población². Una vez que se dispone de esa información, sumado a la ubicación de los centros de testeo actuales y de la distribución de la población, se puede estimar, por ejemplo, cuál es la distancia de cada hogar al centro de testeo más cercano. A partir de ello se pueden observar áreas donde se necesitan centros de testeo adicionales por estar densamente pobladas y tener poca oferta de centros de testeo.

Rastreo

Otra de las dimensiones sobre las que se puede operar para tratar de reducir la velocidad de transmisión de la pandemia es el rastreo. Corea del Sur, a partir de

² <https://bit.ly/3FLOnWB>.

sistemas de información que compilan datos de transacciones bancarias, geolocalización en teléfonos celulares, distintos datos de ubicación e incluso encuestas realizadas a pacientes o a casos positivos, utilizó y puso a disposición³ esa información para que, desde Internet, se pueda identificar en qué lugares estuvieron las personas infectadas. De esta manera, se puede saber por ejemplo dónde estuvo un individuo y hace cuántos días pasó por allí. Con esta información, se pudieron mapear las áreas donde se ubicaron las personas infectadas⁴. Esta herramienta sirve, por ejemplo, para identificar áreas en donde podría haber apariciones de nuevos casos.

Hong Kong, por estar muy cerca de China, que es donde ha comenzado la circulación del virus, era una de las zonas que se estimaba podía estar más afectada. Para evitarlo, se implementaron pulseras o bandas que se les daban a todos los visitantes que entraban al país y a quienes resultaban casos positivos de coronavirus para poder rastrearlos. Estas pulseras estaban vinculadas con una aplicación en el teléfono, por lo que se podía saber dónde estaba esa persona durante el período de las dos semanas necesarias para conocer su evolución, los distintos lugares que visitó y con quién se relacionó. También está disponible la información sobre cuántas personas están haciendo cuarentena, dónde se encuentran y la fecha de finalización de estas⁵.

Otro ejemplo de utilización del *Big Data* para el rastreo de los casos es el de China, donde, a través de la aplicación *Alipay*, que es una billetera digital, se utilizó un sistema de aprendizaje automático para identificar a aquellos individuos que podrían tener distintos grados de riesgo de estar contagiados⁶. En función del grado

³ <https://bit.ly/30g1Fdx>.

⁴ <https://coronamap.site/>.

⁵ <https://bit.ly/3mRZu7G>.

⁶ <https://nyti.ms/3vjxUv>.

de riesgo asignado a cada individuo, determinado por la zona en la que había estado, las personas con las que estuvo en contacto y otros factores, a cada individuo se le asignaba un código QR que podía ser verde, amarillo o rojo y que determinaba la posibilidad o no de circular. Si el código QR que tenía una persona en su celular era verde, podía circular libremente; si el código era amarillo, tenía que tener 7 días de cuarentena; y si era rojo debía cumplir 14 días de cuarentena.

En Singapur, por su parte, se utilizó una aplicación basada en tecnología Bluetooth. Dicha aplicación registraba, a través de la información que se intercambiaba con otros teléfonos móviles por Bluetooth, cuándo un individuo se cruzaba con otra persona con la misma aplicación, quedando así registradas todas las interacciones que hubo entre las personas⁷. En función de eso, se arman mapas de interacciones entre individuos y, cuando un caso se confirma positivo, es posible rastrear con quiénes estuvo en contacto, hacer los tests y aislar a los contactos estrechos.

Contacto social

Para disminuir el contacto social se tomaron medidas como la cuarentena. Las herramientas del *Big Data* permiten observar en qué medida son efectivas estas estrategias para reducir la movilidad. Por ejemplo, en Estados Unidos, a través de la información brindada por los celulares cuando una persona otorga el permiso a una aplicación para que esta pueda acceder a su ubicación, la empresa *Cuebiq* observó en qué lugares la gente tenía más desplazamiento⁸. En aquellos estados que tenían algún tipo de cuarentena la movilidad fue menor que en los que no implementaron estas medidas.

⁷ <https://bit.ly/3vjzcf5>.

⁸ <https://nyti.ms/3p2NmU0>.

En Argentina y otros países de América Latina, una iniciativa conjunta del PNUD y Grandata, que utiliza datos recabados a partir de las antenas que brindan cobertura a los celulares, estimó cuánta gente tuvo desplazamientos a lo largo del tiempo. Como observan Terradez *et al.* (2020), a partir de estos datos, las variaciones en la movilidad en el conurbano bonaerense y en el cumplimiento de la cuarentena se explican fundamentalmente por el nivel socioeconómico de la población, observándose mayor cumplimiento a medida que las necesidades sociales y medioambientales de los ciudadanos están mejor satisfechas.

En esta misma línea, otra herramienta disponible, nuevamente con datos de Facebook, son los datos de movilidad recabados por un proyecto llamado *Covid-19 Mobility Data Network*⁹, en donde se expone cuánto se ha movido la gente durante la pandemia y en qué lugares se evidencia mayor movimiento. Esta herramienta está disponible para distintos países del mundo y permite estudiar y comparar estructuras de movilidad.

Reapertura

Al ingresar en una instancia en donde se evalúa la posibilidad de reapertura, existen otras herramientas complementarias que facilitan la toma de decisiones basadas en evidencia.

Los reportes de movilidad de Google¹⁰ en Argentina registran cómo fue cambiando y evolucionando la movilidad en distintos rubros, tanto en espacios de recreación y lugares como *shoppings* y restaurantes como en farmacias y locales de venta, parques, estaciones de tránsito o lugares de trabajo, y cómo ello se fue reduciendo durante la pandemia. Como contrapartida, se observa un incremento en

⁹ <https://bit.ly/3BTLIHO>.

¹⁰ <https://bit.ly/3j13z8G>.

la movilidad residencial ya que la gente pasó más tiempo en sus hogares.

Estos datos se han utilizado para “modelar la curva”, que no es otra cosa que tratar de estimar cuál va a ser la velocidad de reproducción de la pandemia. Por ejemplo, utilizar el número de fallecimientos registrados, la cantidad de nuevos casos o la distribución de las hospitalizaciones y, en función de ello, determinar cuándo es conveniente, o no, comenzar la reapertura.

La capacidad de predicción de los modelos realizados a partir de datos de movilidad de Google ha sido considerable si se observa, por ejemplo, la cantidad de muertes que hubo en distintos países de Europa (Bryant *et al.* 2020).

Visualización de la información para la toma de decisiones

Las herramientas de visualización son tan importantes para la toma de decisiones como la disponibilidad de información. En este sentido, resulta pertinente mencionar la utilidad de diversos tableros de gestión que sintetizan información sobre la evolución de la pandemia, el número de casos, su distribución geográfica, la cantidad de casos hospitalizados o los fallecimientos. El monitoreo de estos indicadores puede asistir, por ejemplo, en la estrategia de reapertura o en el monitoreo de la evolución de la pandemia para realizar las adecuaciones necesarias. A modo de ejemplo, el tablero de monitoreo del Estado de Nueva York¹¹ ha permitido realizar reaperturas escalonadas por regiones acorde con la situación epidemiológica en cada una de ellas.

¹¹ <https://on.ny.gov/3AEXt3K>.

Nuevos horizontes en el uso del *Big Data*

Tras los procesos de reapertura y la gradual vuelta a la normalidad, se plantea el desafío de mejorar la velocidad de detección de nuevos casos y rebrotes. Para ello, entre otros desarrollos incipientes, se han comenzado a utilizar nuevas cámaras que integran sensores de temperatura con identificación facial y, mediante la aplicación de inteligencia artificial, incluso cuando los individuos tienen la cara cubierta por un tapabocas. El gobierno de China y Panasonic han realizado avances considerables en este sentido. Cabe recordar que hasta hace poco en lugares como Hong Kong estaba prohibido usar máscaras porque no se podía identificar a la gente que asistía a las protestas.

Reflexión final

A pesar de disponer de todas estas herramientas, existe un *trade-off*¹² entre libertades individuales y salud pública, y así es como aparecen titulares como “Corea del Sur está mirando a sus ciudadanos en cuarentena con una aplicación de Smartphone”¹³, “En Polonia hay una aplicación que ayuda a la policía a custodiar la cuarentena en las casas”¹⁴, “En Moscú la policía puede utilizar el reconocimiento facial para atrapar a más de 200 personas que violaron la cuarentena”¹⁵, entre otros similares. Las sociedades, en sus marcos institucionales, deben dar el debate y resolver hasta qué punto se ceden libertades individuales o hasta dónde se quiere, como sociedad, avanzar sobre dichas libertades

¹² En economía, *trade-off* es la relación costo-beneficio.

¹³ <https://bit.ly/3p7HBol>.

¹⁴ <https://privacyinternational.org/examples/3473/poland-app-helps-police-monitor-home-quarantine>.

¹⁵ <https://news.trust.org/item/20200319132732-2c3av>.

en el marco de objetivos vinculados con la vigilancia de la salud pública.

Otro interrogante sobre el que cada sociedad debe reflexionar se refiere a si es posible hacer vigilancia sanitaria que respete los estándares de privacidad de los individuos. En este sentido, la Unión Europea ha hecho avances en la definición de los estándares que serían necesarios para operar en estos contextos invadiendo lo mínimo posible las libertades individuales. En este sentido la Comisión Europea ha identificado una serie de parámetros deseables¹⁶:

- Garantizar que la información que se utilice sea agregada y que esté anonimizada.
- Asegurar que haya consentimiento libre e informado de las personas al ceder la información.
- Atender a que se limite la recolección de la información a aquella que es necesaria en lugar de extenderse a relevar todo lo que potencialmente podría ayudar.
- Preguntarse si las herramientas de vigilancia van a ser efectivas y si van a tener un impacto significativo para resolver la crisis, ya que no tiene sentido plantear herramientas de vigilancia que no tengan un impacto.
- Observar que los datos que se recolectan tengan una fecha de expiración, es decir que en algún momento caduquen y sean borrados.
- Procurar que existan restricciones para el uso de la información para otros propósitos.

¹⁶ <https://bit.ly/30rgVEF>.

