BIG DATA EN SALUD: RETOS Y OPORTUNIDADES

ERNESTINA MENASALVAS
CONSUELO GONZALO
ALEJANDRO RODRÍGUEZ-GONZÁLEZ
Universidad Politécnica de Madrid

La informatización de los procesos ha provocado que empresas y organizaciones de todo tipo hayan acumulado una cantidad ingente de datos. Esto nos ha llevado a denominar los tiempos actuales como la era del «*Big Data*», donde se requieren nuevas tecnologías para gestionar y extraer el valor de los datos complejos que se generan en grandes volúmenes

a altas velocidades. Si bien este fenómeno está afectando a todos los sectores, el sector sanitario es una de las áreas en las que la incidencia de este fenómeno está siendo especialmente relevante, debido entre otros elementos a la implantación de la Historia Clínica Electrónica (EHR) (1), a la reciente explosión en lo referido a generación de datos de tipo genómico (gracias en parte a su abaratamiento), y al hecho de que la mayoría de datos que se generan en el sector son no estructurados.

Las aplicaciones de *Big Data* en el sector de la salud indican un alto potencial para mejorar la eficiencia y calidad de provisión de cuidados. En el estudio de McKinsey (McKinsey & Company, 2011) se afirma que si el sector de la salud en los Estados Unidos usara tecnologías de *Big Data* de manera efectiva para producir calidad el sector podría crear más de 300 billones de dólares cada año y dos tercios de esto sería en forma de reducción de gastos de salud en alrededor de un 8%. IBM proporciona cifras igualmente alarmantes en el informe de Korster y Seider de 2010, donde se analiza la ineficiencia del sistema de salud americano en 2,5 trillones de dólares malgastados anualmente y donde el grado de eficiencia se podría mejorar hasta un 35%, si se compara con otros grandes sectores industriales.

En concreto en España, el gasto en Tecnologías de Información y Comunicaciones (TIC) en la sanidad pública en 2013 ascendió a 624 millones de euros, lo que significa que menos del 1,20% del gasto sanitario público total se dedicó a las TIC. Estos datos proceden del último informe Índice SEIS (2), el cual recoge de forma objetiva datos reales del sector sobre el gasto TIC en sanidad en España. Del estudio se desprende que, en la fecha del estudio, las TIC se estaban utilizando para almacenar la información generada por el sistema sanitario de salud, pero no se estaba utilizando dicha información para mejorar la atención a los pacientes y la gestión y eficiencia de los servicios sanitarios.

La Historia Clínica Digital del Sistema Nacional de Salud (HCDSNS) surge para garantizar a los ciudadanos y a los profesionales sanitarios el acceso a la información clínica relevante para la atención sanitaria de un paciente desde cualquier lugar del SNS. La implantación de la Historia Clínica genera una enorme cantidad de datos no estructurados, casi 40 millones de Historias Clínicas

Electrónicas que se registraron en España en el 2013 y más de un 1 millón de GB que contienen las imágenes generadas en exploraciones médicas. A fecha de febrero de 2017, el informe de situación del proyecto (HCDSNS) refleja que los Servicios de Salud que tienen activado el perfil emisor (emite los distintos tipos de informes clínicos contenidos en el Conjunto Mínimo de Datos de los Informes Clínicos (CMDIC) para todos sus ciudadanos y se facilita el acceso al sistema a sus ciudadanos para la totalidad de los servicios disponibles) son 17 sobre un total de 18, con una cobertura de población con referencias HCDSNS en cifras absolutas de 35.726.423 ciudadanos que en relación con la población TSI (tarjeta sanitaria individual) supone un 77,48%.

Como consecuencia del almacenamiento de estos datos, los hospitales cuentan con una ingente cantidad de información no estructurada en forma textual o visual, que puede ser explotada con tecnologías de *Big Data*. Y es que *Big Data* representa el paradigma perfecto para el desarrollo de lo que se conoce como medicina basada en la evidencia. La Medicina Basada en la Evidencia es una expresión que se ha generalizado en castellano como equivalente a la expresión inglesa *Evidence-Based Medicine* (EBM). La definición clásica de EBM es «el uso consciente, explícito y juicioso de las mejores y más actuales evidencias o pruebas en la toma de decisiones sobre el cuidado de los pacientes». Es un mecanismo efectivo no sólo para mejorar la calidad de los cuidados de salud, sino también para reducir los errores clínicos y la variabilidad en la práctica clínica, e influye directamente en la capacitación para la aplicación de lo que se denomina medicina personalizada. Para llevar a cabo la EBM, es necesario que se integre la experiencia clínica de los profesionales y las mejores evidencias disponibles procedentes de la investigación científica, siempre teniendo en cuenta las características y valores de cada paciente.

No obstante, la aplicación de técnicas de *Big Data* en el sector salud todavía tiene que afrontar algunos retos tecnológicos causados por los avances de los mecanismos de almacenamiento y gestión de datos, computación en la nube y por los resultados derivados del IoT (3) (Internet de las cosas) que permite adquirir, almacenar y procesar datos de todo tipo (sensores, clínica, comportamiento, genómica, proteómica, imagen, texto, …). Solo por mencionar algunos de estos retos, destacamos aquí:

1. las necesidades de integración de información;
2. las necesidades de documentar de manera digital sin que se requiera un esfuerzo extra a los profesionales de la salud;
3. análisis de contenidos de datos no estructurados de salud (imagen, texto, ...) que se tendrán que mejorar con anotaciones semánticas;
4. silos de datos cuya integración habrá que afrontar;
5. medios técnicos y legales que aseguren la compartición y el intercambio de datos;
6. medios para asegurar la calidad.

En este artículo cubrimos algunos de los aspectos que suponen hoy en día un reto en salud. En particular abordamos, en primer lugar, un análisis del problema de descubrimiento de conocimiento en bases de datos; con posterioridad analizamos el reto de las redes sociales en salud y después abordaremos algunos de los retos que la explotación de la información contenida en la imagen médica conlleva.

BIG DATA EN SALUD

Desde la introducción del concepto *Big Data* se han asociado una serie de características clave que lo definen, llamadas las «V» de *Big Data*. Inicialmente, estas características eran tres: volumen, velocidad y variedad. Posteriormente se han añadido otras características, como la veracidad de los datos, añadiendo así cuatro dimensiones y marcando las características que tradicionalmente se están asociando a *Big Data*, y otra quinta dimensión, el valor de los datos, que se ha añadido recientemente, generando así las denominadas cinco V del *Big Data*.

Esta definición es válida en todos los sectores en los que se dé la confluencia de esta serie de características en mayor o menor medida, y así ocurre en el dominio de la salud, por lo que estos términos también son relevantes en el sector.

En el campo de la medicina el enorme volumen de datos de asistencia sanitaria existente incluye registros médicos personales, imágenes médicas, datos de ensayos clínicos, altas, datos a nivel genético, las secuencias genómicas de datos de población, etc. Más recientemente, están alimentando este exponencial crecimiento las imágenes en 3D, así como las lecturas de los sensores biométricos o los dispositivos *wearables*. Afortunadamente, los avances en la gestión de datos, en particular la virtualización y el *Cloud Computing*, están facilitando el desarrollo de plataformas para la captura más eficaz, almacenamiento y manipulación de estos grandes volúmenes de datos (Bonnie Feldman, Ellen M. Martin, & Tobi Skotnes, 2012).

Por otra parte, la mayor parte de datos de la salud han sido tradicionalmente estáticos: registros médicos, rayos X, pruebas de laboratorio, etc. Pero esta tendencia está cambiando y ahora la velocidad de generación va en aumento con datos como por ejemplo los relacionados con la supervisión periódica, tales como mediciones de glucosa en diabéticos (control continuo por las bombas de insulina), la presión arterial, electrocardiogramas, y toda la monitorización realizada en las unidades de cuidados intensivos, por mencionar sólo algunos ejemplos. El análisis de estos datos en tiempo real podría servir para identificar y aplicar los tratamientos adecuados que podrían ayudar a reducir la morbilidad y la mortalidad de los pacientes e incluso prevenir los brotes hospitalarios.

En la misma medida en la que evoluciona la naturaleza de los datos de salud lo tienen que hacer las técnicas de análisis. Ya no podemos hablar solo de los datos estructurados como los recogidos en los historiales médicos electrónicos. Cada vez más, los datos están en formatos multimedia y no están estructurados. La enorme variedad de datos estructurada, no estructurada y semi-estructurada es lo que hace que los datos de salud sean al mismo tiempo interesantes y desafiantes.

Los datos estructurados son aquellos que se pueden almacenar, consultar, analizar y manipular por ordenador fácilmente. Históricamente, en la asistencia sanitaria, los datos estructurados y semi-estructurados incluyen las lecturas de los dispositivos, y los datos de la historia clínica digital. Se generaban también datos no estructurados (que se siguen generando allí donde no se ha adoptado la historia clínica digital): los registros médicos de la oficina, notas manuscritas, ingresos en el hospital, así como los registros de altas, las recetas de papel, las radiografías, Imagen por Resonancia Magnética (MRI) (4), Tomografía Computarizada (CT) (5) y otras imágenes.

Hoy en día, además de todos estos datos, el sector médico se inunda cada día con los datos generados por dispositivos móviles, la genética y la genómica, los medios sociales, literatura y otras fuentes. Sin embargo, relativamente pocos de estos datos en la actualidad se pueden adquirir, almacenar y organizar de tal manera que se pueda analizar por los ordenadores para obtener información útil. Se precisan técnicas y herramientas más eficientes que permitan combinar y convertir todos estos datos en datos estructurados para su posterior análisis.

Los datos estructurados de los registros médicos digitales incluyen datos como el nombre del paciente, datos de nacimiento, dirección, nombre, nombre del hospital del médico y dirección, tratamientos u otras informaciones relativamente sencillas de codificar y automatizar en bases de datos. Pero la historia clínica digital contiene información en lenguaje natural cuyo análisis podría desvelar información de interés. El potencial de *Big Data* en la asistencia sanitaria radica en la combinación de datos tradicionales con las nuevas formas de datos, tanto de forma individual como poblacional.

Por otra parte, hemos de hablar también de la calidad de los datos. Los problemas de calidad de datos son especialmente importantes en el sector de la salud por dos razones: las decisiones de vida o muerte dependen de tener la información precisa, y la calidad de los datos de salud, especialmente de datos no estructurados, es altamente variable y con demasiada frecuencia incorrecta o incompleta. La veracidad asume escalabilidad en granularidad y en el rendimiento de las arquitecturas y plataformas, algoritmos, metodologías y herramientas para que responda a las exigencias de *Big Data*. Se requieren nuevas arquitecturas para procesar los datos y nuevos algoritmos.

Finalmente, con la evolución surgida especialmente en los últimos años, se considera muy relevante que todos los proyectos de *Big Data* tengan como enfoque el aporte de valor durante o al final del proceso elaborado como parte del proyecto. De esta forma, los proyectos *Big Data* no sólo se enfocan en el análisis de grandes cantidades de datos, que se generan rápidamente, con unas características muy heterogéneas y que sean correctos, sino que además el análisis de estos datos debe arrojar un beneficio para las entidades implicadas. En el caso del dominio sanitario, los beneficios pueden ser económicos por reducción de costes, eficiencia en la gestión de farmacia, disminución del número de ingresos o estancias hospitalarias, disminución en el número de consultas sucesivas o una mayor capacidad y mejor calidad de atención de pacientes.

Con el conocimiento de estas características, se observa que la mejora en el punto de recogida de datos para tratar de evitar errores en la propia recogida y reducir costes son cruciales, pero el aumento de variedad y de alta velocidad obstaculizan la capacidad de limpiar los datos antes de analizarlos y consiguientemente impactan a la toma de decisiones puesto que se pone en duda la «confianza» de los mismos.

No obstante, si las cinco V analizadas son un punto de partida apropiado para una discusión acerca del análisis de grandes volúmenes de datos en la asistencia sanitaria, no debemos olvidar otras cuestiones como arquitecturas y plataformas, herramientas, metodologías y la necesidad de interfaces construidas por y para los profesionales de la salud.

Todos estos temas se deben abordar para aprovechar y maximizar el potencial del análisis de datos en la asistencia sanitaria.

DATA MINING

El campo de la Informática de la Salud está en la cúspide de su período más emocionante hasta la fecha, entrando en un nueva era donde la tecnología está empezando a manejar grandes volúmenes de datos, dando lugar a un potencial ilimitado para el crecimiento de la información. La minería de datos y análisis de datos grandes están ayudando a tomar decisiones relativas a diagnóstico, tratamiento, … Y todo finalmente enfocado a una mejor atención al paciente.

El uso de minería de datos en Estados Unidos en salud puede ahorrar a la industria de la salud hasta 450 mil millones de dólares cada año (Basel Kayyali, David Knott, & Steve Van Kuiken, 2013). Esto se debe a los volúmenes crecientes de datos generados y de las tecnologías para analizarlos.

El explosivo crecimiento de datos generó, ya en la década de los 80, la aparición de un nuevo campo de investigación que se denominó KDD (*Knowledge Discovery in Databases*). Bajo estas siglas se esconde el proceso de descubrimiento de conocimiento en grandes volúmenes de datos (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). El proceso de KDD ha servido para unir a investigadores de áreas como la inteligencia artificial, la

estadística, las técnicas de visualización, el aprendizaje automático o las bases de datos en la búsqueda de técnicas eficientes y eficaces que ayuden a encontrar el potencial conocimiento que se encuentra inmerso en los grandes volúmenes de datos almacenados por las organizaciones diariamente.

Si bien el nombre con el que apareció esta área de investigación fue el de KDD, en la actualidad este nombre ha sido sustituido por el de *Data Mining*. En un principio, *Data Mining* fue tan sólo usado para referirse a la etapa del proceso en la que se aplican las técnicas y algoritmos de descubrimiento, no obstante, ahora se usa para referirse al proceso global de descubrimiento.

Aunque no hay una única definición de *Data Mining*, la siguiente es, posiblemente, la más aceptada: «Proceso de extracción de información desconocida con anterioridad, válida y potencialmente útil de grandes bases de datos para usarla con posterioridad para tomar decisiones importantes de negocio» (Ian H. Witten, Eibe Frank, & Mark A. Hall, 2011).

El término proceso implica que la extracción de conocimiento es la conjunción de muchos pasos repetidos en múltiples iteraciones. Se dice, por otra parte, que es no trivial, porque se supone que hay que realizar algún tipo de proceso complejo. Los patrones deben ser válidos, con algún grado de certidumbre, y novedosos, por lo menos para el sistema y, preferiblemente, para el usuario, al que deberán aportar alguna clase de beneficio (útil). Por último, está claro que los patrones deben ser comprensibles, si no de manera inmediata, sí después de ser pre-procesados.

Como consecuencia de la complejidad del desarrollo de proyectos de minería de datos, a comienzos de los 90 surge el estándar de modelo de proceso denominado CRISP-DM que divide el proceso en las siguientes fases:

- Comprensión del negocio: Se pretende aquí comprender los objetivos del proyecto y sus requerimientos desde la perspectiva del negocio, convirtiendo este conocimiento en un problema de *Data Mining* y un plan preliminar para cumplir dichos objetivos.
- Comprensión de los datos: Se cuenta en un principio con una colección de datos, se deben identificar los problemas de calidad de los datos, detectar subconjuntos de interés, etc.
- Preparación de los datos: Mediante esta fase se construye el conjunto de datos final obtenido de la colección inicial de datos que será proporcionada a las herramientas de modelado.
- Modelado: Se seleccionan y aplican varias técnicas de modelado, ajustándolas para obtener valores óptimos.
- Evaluación: Una vez construido un modelo se debe evaluar y revisar los pasos ejecutados para construir un modelo que consiga los objetivos de negocio.
- Desplegado: Generalmente se deben aplicar modelos en procesos de toma de decisiones de una organización.

Los problemas que se pueden abordar desde la perspectiva de *Data Mining* a menudo se agrupan en las siguientes categorías:

1. Los problemas predictivos cuyo objetivo es predecir el valor de un atributo en particular basado en los valores de otros atributos. El atributo que se predice se denomina comúnmente como atributo objetivo (o variable dependiente), mientras que los atributos que se utilizan para la predicción son conocidos como atributos explicativos (o variables independientes). Destacan aquí los problemas de clasificación o de estimación de valor y como técnicas podemos destacar los enfoques basados en estadística, regresión, árboles de decisión y redes neuronales.
2. Los problemas descriptivos cuyo objetivo es derivar patrones (correlaciones, tendencias, agrupaciones o clústeres, trayectorias y anomalías) que resuman las características inherentes a los datos. Dentro de este grupo, cabe destacar el análisis de reglas de asociación para el que el algoritmo «*A priori*» es el más conocido, así como los problemas de segmentación o *clustering*.

BIG DATA Y *SOCIAL MEDIA* EN SALUD

La evolución de la Web ha sufrido un cambio dramático en los últimos años que ha dado lugar a la explosión de los datos. El inicio de esta transformación viene de la evolución de la web 1.0 a la 2.0. En la web 1.0 hablábamos de una web prácticamente estática, donde los contenidos eran generados principalmente por corporaciones, y donde los usuarios finales apenas generaban contenido, solo lo consumían.

En la web 2.0 el modelo empieza a cambiar, dando lugar a una generación de contenido por parte de los usuarios y con lo que sería la explosión de los datos en Internet. Este fenómeno comienza a pequeña escala con la introducción de elementos como los *blogs* o los foros, primeros antecesores de las redes sociales modernas, y se convierte finalmente en un elemento de escala global con la generación de redes sociales como Facebook, Twitter o Instagram por destacar algunas de las principales.

La combinación de uso de este tipo de redes sociales ha hecho que el volumen de datos que se genera cada día haya crecido de forma exponencial, dando lugar a que las cantidades de datos que se generan sean inmensas. Algunos de estos números pueden visualizarse en la infografía publicada por DOMO llamada *Data Never Sleeps* (6). Ejemplos de eventos que suceden cada minuto son: Google realiza casi 70 millones de traducciones de palabras; Instagram recibe casi 2,5 millones de '*likes*' a los *posts* de sus usuarios; Siri responde casi 100.000 peticiones; los usuarios suben más de

800.000 nuevos ficheros a Dropbox; se publican casi 10.000 *tweets* con *emojis*, etc. Estos números son solo una pequeña muestra de las cantidades de datos que son generadas por las redes sociales. En realidad, estos volúmenes de datos son mayores ya que ni contemplan todas las redes sociales ni todos los tipos de datos concretos que son susceptibles de análisis o gestión.

Redes sociales y salud

Dentro del entorno de salud, las redes sociales han emergido con un fuerte y claro propósito: compartir información de salud y conectar con otros enfermos, doctores o profesionales de la salud. Las personas tienen cada vez un *expertise* mayor en el uso de las nuevas tecnologías lo que da lugar a la proliferación de este tipo de redes sociales orientadas al campo de la salud. Aunque existen diferentes tipos de redes sociales en salud, según su orientación o tipo de usuarios registrados se pueden dividir fundamentalmente entre tres tipos: redes sociales orientadas a profesionales (conectan profesionales de salud entre sí mismos), redes sociales orientadas a pacientes (conectan pacientes entre sí mismos) y redes mixtas (conexión paciente-médico).

Un estudio realizado por AMN *Healthcare* (7) estimó que aproximadamente un tercio de los profesionales de la salud usan redes sociales específicas. Sin embargo, los datos reflejados el informe de AMN *Healthcare* solo hablan del uso de las redes sociales por parte de los profesionales médicos, y tal y como analiza la infografía se centra en un uso muy concreto de estas redes sociales por parte de éstos. No obstante, el uso de este tipo de redes por parte de los propios profesionales es más que interesante y relevante de cara a los datos que se comparten en ella. Algunos ejemplos de este tipo de redes sociales y del potencial que se puede derivar de los datos que contienen son iniciativas como Doximity (8), una red social que permite a los profesionales de la medicina comunicarse entre ellos con el objetivo de compartir información sobre posibles diagnósticos. Doximity contiene más de 250.000 miembros en Estados Unidos, que representan aproximadamente el 40% de los médicos de este país.

Otra red social de interés similar dentro de los Estados Unidos es Sermo (9), que se autodefine en su propia página web como la red social de médicos más grande del país, y actualmente, del mundo. Esta red también permite, al igual que Doximity, las discusiones sobre cualquier tema relacionado con la salud de una forma abierta y colaborativa.

Otras redes sociales de temática similar a nivel internacional son Figure 1 (10) (enfocada en diagnóstico colaborativo), SharePractice (11) (similar a Figure 1 pero orientada a tratamientos), WeMedUp (12) o Doc2Doc (13) entre muchas otras.

A nivel nacional o que puedan tener una orientación al mercado hispano hablante destacan iniciativas como Ippok (14), red social orientada a profesionales de la salud que pueden pedir u ofrecer consejos o ayuda en términos de tratamientos, casos clínicos o similar, además de acceder a documentos y ofertas de trabajo; DoctorDice (15) considerada como la red social exclusiva para médicos más grande de México y que permite el intercambio de todo tipo de información clínica dentro de la misma; MedCenter (16) que es un portal dirigido a la comunidad médica con un objetivo prioritario centrado en la educación o Medicalia (17). Aparte de estas iniciativas existen muchas otras con orientaciones similares a las ya descritas.

En lo que se refiere a redes sociales orientadas a pacientes hay diferentes alternativas y con diferentes ámbitos. Las redes sociales de este tipo pueden buscar diferentes objetivos: desde simplemente compartir experiencias con la idea de buscar apoyo o recomendaciones hacia determinadas enfermedades o tratamientos, hasta buscar segundas opiniones o alternativas incluso de tratamiento en base a las opiniones o recomendaciones de otros usuarios. De cualquier forma, lo que las une es la interacción paciente-paciente, sin que el rol del profesional de la medicina tenga por qué estar necesariamente presente como elemento principal de la red social.

Existen diferentes redes sociales que comparten el objetivo de conectar pacientes, siendo una de las principales y más importantes PatientsLikeMe (18). El objetivo con el que se define esta red social es el de permitir a sus usuarios compartir tratamientos o síntomas de sus respectivas enfermedades con el fin de poder hacer un seguimiento y aprender de los resultados de otros. Esta red social tiene diferentes comunidades para diferentes enfermedades entre las que destacan algunas patologías como la Esclerosis Lateral Amiotrófica (ELA), la Esclerosis Múltiple (EM), parkinson, fibromialgia, VIH, síndrome de fatiga crónica o trastorno de ánimo entre algunas de las más habituales, aunque también dispone de comunidades para algunas enfermedades raras.

Otras redes sociales con objetivos similares, pero en muchas ocasiones muy focalizadas a una patología concreta son por ejemplo tudiabetes.org (19), stupidcancer (20), curetogether (21) o Aorana (22) entre otras.

Finalmente, en el grupo mixto que conecta a pacientes y médicos, aunque este tipo de redes sociales están más limitadas, existen iniciativas como RareShare (23), orientada a las enfermedades raras.

BIG DATA EN REDES SOCIALES Y SALUD PÚBLICA

La salud pública es la disciplina que se encarga de la protección y mejora de la salud de la población humana. Cuando se habla de salud pública se habla de grandes poblaciones de individuos, en vez de centrarse en casos o personas concretas. La salud pública puede gestionarse a diferentes niveles en cuanto a sus tamaños de población se refiere, en función de este preciso tamaño, ya que no es lo mismo gestionar poblaciones a niveles municipales, provinciales, estatales o internacionales.

En este contexto, *Big Data* encaja muy bien dentro de lo que se refiere a la creación de datos que puedan afectar a grandes conjuntos de personas o grandes poblaciones. El propio concepto de *Big Data* nace por la generación de grandes cantidades de datos, pero estas grandes cantidades de datos pueden ser generadas de dos formas: entidades que producen grandes cantidades de datos (muchas o pocas entidades), o gran número de entidades que generan pequeñas cantidades de datos. Al final, todo se reduce a una escala en el nivel de generación de los datos: quién los crea (cuántos actores están involucrados) y qué cantidades genera cada actor.

En el entorno, por lo tanto, de salud pública, *Big Data* emerge mediante unas poblaciones que representan a gran cantidad de individuos generadores de pequeñas cantidades de datos, pero que en bloque suponen una gran cantidad de información.

Los usuarios, por tanto, son generadores de datos a pequeña escala, pero el conjunto de todos esos usuarios que generan datos crean el *Big Data*. El ejemplo más claro está precisamente en la infografía de DOMO mencionada previamente, donde los datos generados son enormes, pero son generados precisamente por diferentes personas, por un inmenso número de personas. Cambia en este contexto el modelo de generación de datos por un modelo en cierto modo distribuido y donde por lo tanto las redes sociales son una fuente de datos de información de salud pública a tener muy en cuenta, tal y como se ha analizado previamente (Evika Karamagioli, 2015; Kass-Hout & Alhinnawi, 2013).

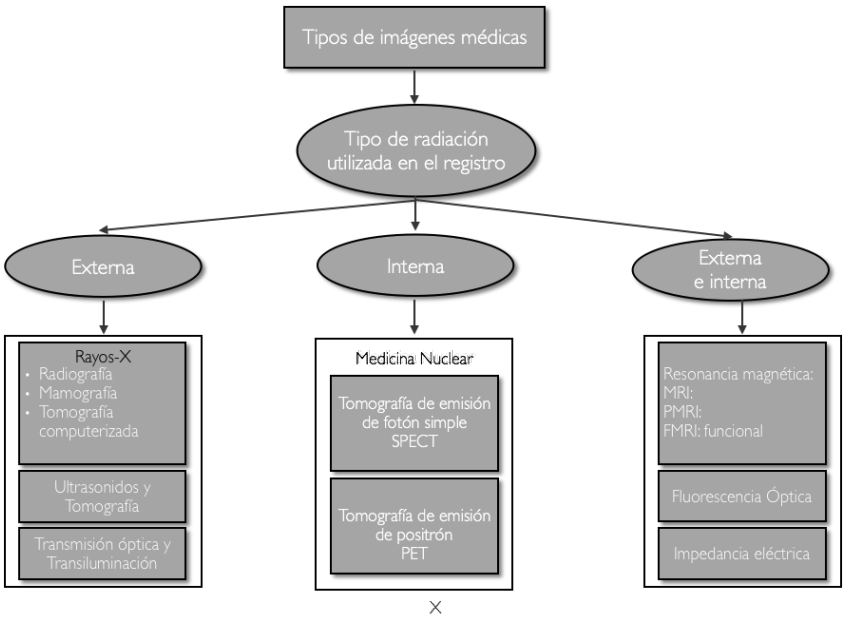
De los 500 millones de *tweets* que se generan al día, se estiman que alrededor de 1 millón están relacionados con la temática de salud. Además, los números de las redes sociales orientadas a la salud, como las descritas anteriormente, cuentan con espectaculares números en cuanto al contenido publicado, dando lugar a generación de datos plenamente orientados al dominio médico. HealthBoards.com se estima que pueda tener publicados alrededor de 4.6 millones de *posts*, con al menos 1 millón de miembros. CancerForums.net alrededor de 150.000 *posts* y Drugs-Forum alrededor de medio millón de *posts* (24).

La generación de estos datos por parte de los usuarios puede ser de especial relevancia dentro de la salud pública. Un claro ejemplo de la importancia de la relación existente entre *Big Data*, redes sociales y salud pública es precisamente la última convocatoria del programa en retos sociales del programa europeo H2020, tal y como se vio reflejado en el programa de trabajo para el año 2016-2017 con el tópico «*Big Data supporting Public Health policies* (SC1-PM-18-2016)» (25).

Pero la importancia de esta temática no finaliza solo en esta convocatoria de la unión europea. Otro ejemplo claro de la importancia de esta temática es la famosa conferencia internacional de la WWW en cuya edición de 2013 en Rio de Janeiro se organizó un *workshop* sobre redes sociales, *Big Data* y salud pública (26). El mismo año, se publicó un número especial de la revista *Journal of Biomedical Informatics* sobre la información biomédica en entornos de redes sociales (Rodríguez-González, Mayer, & Fernández-Breis, 2013) y más recientemente se ha organizado un *special track* sobre análisis de datos médicos y datos sociales en la conferencia IEEE *Computer Based Medical Systems* (CBMS) (27). Son muchas las iniciativas que apuestan por esta línea de investigación. Existen diferentes estudios e investigaciones ya llevadas a cabo tanto a nivel nacional como internacional sobre el análisis de datos a gran escala provenientes de fuentes de información de tipo social. Algunos de estas investigaciones incluyen por ejemplo la investigación llevada a cabo por la escuela universitaria de enfermería de Indiana y ChaCha (28), una red social de preguntas-respuestas con el objetivo de analizar mediante técnicas de procesamiento de lenguaje natural y herramientas de exploración de datos en tiempo real aquellas preguntas y respuestas relacionadas con salud y bienestar. Otras investigaciones se han centrado en el enlace entre los datos extraídos de las redes sociales y los registros médicos publicados en estas redes (Padrez *et al*., 2015) o para directamente analizar la información pública de las redes sociales con el objetivo de poder identificar posibles brotes de epidemias o condiciones patológicas (Asamoah, Sharda, & Kumarasamy, 2015; Capurro *et al*., 2014; Nambisan, Luo, Kapoor, Patrick, & Cisler, 2015; Paul & Dredze, 2011; Xie *et al*., 2013) tal y como hacen ya plataformas existentes como HealthMap (29) mediante el análisis masivo de información de redes sociales y noticias, o como hizo en su momento GoogleFlu (30) mediante el análisis y geolocalización de búsquedas relacionadas con la gripe. Otros estudios se han centrado por ejemplo en el análisis de grupos de Facebook en cuanto a elementos de bienestar y salud como es la alimentación saludable (Leis *et al*., 2013),grupos relacionados con el cáncer de mama (Bender, Jimenez-Marroquin, & Jadad, 2011) o estudios sobre el uso de Facebook para contestar a preguntas de medicina en la página de una conocida revista de medicina (Rodríguez-González, Menasalvas-Ruiz, & Pujadas, 2016) entre muchos otros estudios de la misma índole. Algunos proyectos europeos como TrendMiner (31) incluían en sus casos de uso precisamente el análisis de las redes sociales en el entorno de salud, en este caso concretamente para la búsqueda de interacciones entre fármacos (Segura-Bedmar, Martínez, Revert, & Moreno-Schneider, 2015).

El crecimiento tanto de las redes sociales de propósito general (Facebook, Twitter, Instagram, etc.) como las de uso específico (entre las que destacan las analizadas previamente) ha dado lugar a un *boom* en lo que se refiere a la generación de información relacionada con la salud. Las redes sociales de uso específico, en las que se realiza un intercambio de la información relacionada con determinadas patologías y sus tratamientos entre pacientes, son una fuente muy importante de datos que puedan servir para extraer datos de gran utilidad como efectos adversos, nuevas opciones de tratamiento o incluso para poder detectar casos de errores en prescripciones clínicas entre otra información interesante a extraer.

FIGURA 1
TAXONOMÍA DE DIFERENTES MODALIDADES DE IMAGEN MÉDICA SEGÚN EL TIPO DE RADIACIÓN UTILIZADA EN SU REGISTRO



Fuente: Adaptada de (Dhawan, 2013)

Las redes sociales de profesionales permiten también, además de extraer información relacionada con la descrita para el caso de redes sociales de tipo paciente-paciente, la extracción y análisis de diagnósticos de forma colaborativa.

Finalmente, las redes sociales de propósito general son junto con las anteriores una fuente de información distribuida geográficamente de gran calidad. Los datos a extraer de estas redes permitirán aplicar políticas de salud pública.

IMAGEN MÉDICA

La imagen médica constituye hoy una de las principales fuentes de información utilizadas por el médico para el diagnóstico y terapia. Se pueden considerar una ventana del cuerpo humano, que, de una manera prácticamente inocua, permite la caracterización de diferentes enfermedades, facilitando su diagnóstico y tratamiento. Como consecuencia, todas las tecnologías relacionadas con la imagen médica han evolucionado notablemente en las últimas décadas.

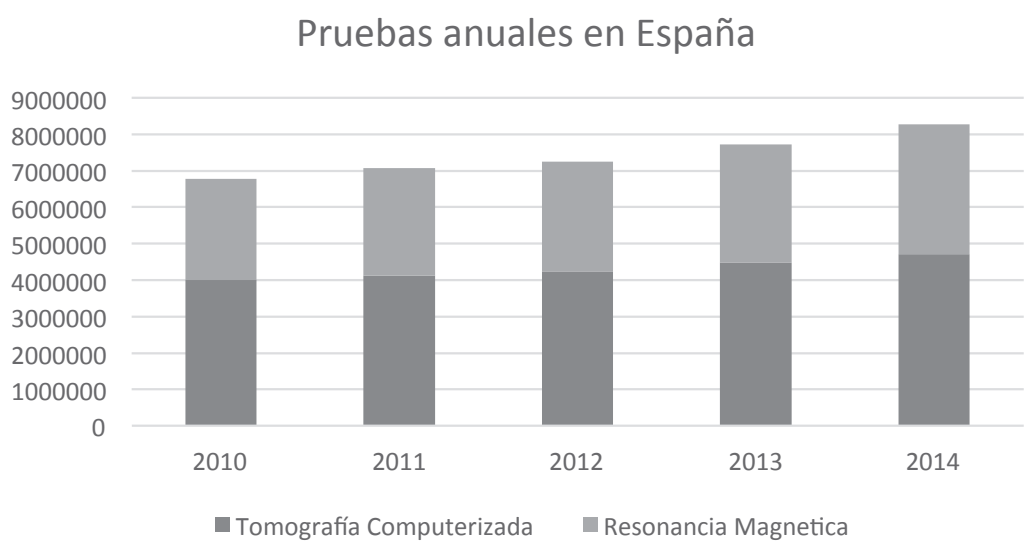
En este artículo nos vamos a referir principalmente a 4 de las V's que, desde el punto de vista de *Big Data*, caracterizan a la imagen médica: variabilidad, volumen, velocidad de generación y sobre todo el valor que se extrae de ellas. La variabilidad viene dada principalmente por el gran número de modalidades de imagen médica, como queda representado en la Figura 1.

En cuanto al volumen, por un lado, hay que considerar la cantidad de imágenes que se generan diariamente en cada centro asistencial y ese número extrapolarlo a nivel regional, nacional y supra-nacional. La Figura 2 muestra, a modo ilustrativo, la evolución de este volumen de datos para dos de las modalidades más utilizadas, Resonancia Magnética y Tomografía Computerizada, desde el año 2010 al 2014 en España. Como se puede apreciar, la suma de este número de pruebas en el año 2014 superó las 8.000.000 pruebas, exactamente fueron 8.278.465.

Ahora bien, para realmente entender el volumen de datos que es preciso almacenar y procesar en el ámbito de la imagen médica, se debe tener presente que una imagen 2-D (Figura 3) de tamaño medio, está formada por más de un millón de unidades de información (pixeles); y que la mayoría de las imágenes médicas son 3-D e incluso 4-D. En este sentido, existen estudios que consideran que el grado de complejidad del procesado de las imágenes médicas, estimado en número de píxeles es equiparable a la complejidad del análisis de datos genómicos.

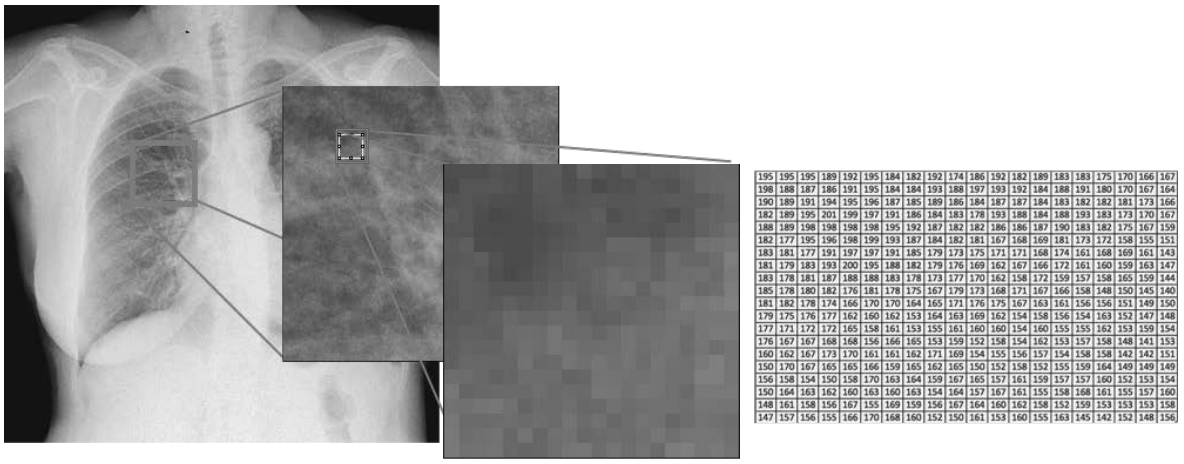
Respecto a la velocidad de generación de los datos, que a su vez está muy relacionada con el volumen, un buen indicador es la experiencia personal de cualquier persona que acude a un servicio de urgencias, pero para dar algunos datos objetivos, citaremos los datos correspondientes a un hospital de la red pública de la Comunidad de Madrid del año 2013 (33) (último año del que se dispone de datos). Los valores medios diarios de Ecografías, Tomografía Computerizada, Resonancia Magnética y Radiografías registradas fueron de 11, 85, 53, 600, respectivamente. Un aspecto muy importante a tener en cuenta es que para explotar la información

FIGURA 2
EVOLUCIÓN DEL NÚMERO DE RESONANCIAS MAGNÉTICAS Y TOMOGRAFÍAS COMPUTERIZADAS REALIZADAS EN ESPAÑA DESDE EL AÑO 2010 HASTA EL 2014 (32)



Fuente: Elaboración propia

FIGURA 3
IMAGEN MÉDICA 2-D, CON DIFERENTES NIVELES DE AMPLIACIÓN DE UNA ZONA Y REPRESENTACIÓN DE LOS VALORES ASOCIADOS A LOS PÍXELES DE DICHA ZONA



Fuente: Elaboración propia

de este volumen de datos a una velocidad equiparable a la que se producen, son necesarias técnicas y herramientas que permitan un procesado y análisis de las imágenes de forma automática o cuasi-automática. Solo disponiendo de estas técnicas, realmente se podría extraer todo el valor asociado a la imagen médica.

Por lo tanto, uno de los grandes retos es el desarrollo de nuevos algoritmos y metodologías para el procesado, análisis e interpretación del enorme volumen de datos que este tipo de imágenes representan, con objeto de ayudar a los profesionales sanitarios a explotar toda la información contenida en las mismas. Dado que la extracción de la información de estas imágenes depende de muchos factores como, por ejemplo: modalidades, condiciones de registro, dispositivos, etc., no es posible disponer de procedimientos generales para localizar y/o identificar objetos tales como estructuras anatómicas o lesiones, sino que es preciso desarrollar métodos particulares para cada tipo de imagen y/o patología.

Una vez que se ha extraído de cada imagen particular la información deseada, el siguiente paso con-

siste en asociar este conocimiento a las imágenes (anotación semántica de imágenes), de forma que, una vez almacenadas en grandes bases de datos, sea posible llevar a cabo búsquedas semánticas de las mismas o, en otras palabras, búsquedas por conceptos. La anotación consiste en un conjunto de palabras capaces de describir la información contenida en la imagen. Tradicionalmente, la anotación de las imágenes se ha llevado a cabo manualmente por personal especializado. Sin embargo, este proceso tiene algunas desventajas, tales como el coste de tiempo y la subjetividad del operador. Aunque el mismo operador delimite y anote la misma imagen en momentos diferentes, la anotación no será necesariamente la misma. El enfoque alternativo es la anotación automática o semiautomática realizada por un algoritmo. La mayor parte de las aproximaciones que se encuentran en la literatura se basan en técnicas de aprendizaje automático (Menasalvas & Gonzalo 2016).

Un problema que se ha detectado para poder pasar a la aplicación clínica de los estudios más teóricos es la necesidad de disponer de grandes bases de datos que permitan de una forma objetiva y rigurosa determinar y validar cuales son los mejores algoritmos a utilizar para cada aplicación. En este sentido, es de gran interés, las competiciones propuestas en los últimos años (34), en las cuales se han proporcionan datos anotados de casos reales de los que se conoce la información realidad-terreno, y se ha establecido un protocolo de validación común. De este modo, se asegura que los resultados proporcionan por diferentes algoritmos son técnicamente comparables y fiables para su aplicación clínica.

CONCLUSIONES

La aplicación del paradigma de *Big Data* al entorno de la salud supondrá una mejora de magnitud aún no predecible en la calidad de la atención a los pacientes, así como en la prevención, diagnóstico y tratamiento de enfermedades, unido a una reducción notable en los costes de sanidad. Para alcanzar estos logros es fundamental la integración de todos los datos procedentes de muy diferentes fuentes, así como el desarrollo de nuevas tecnologías que permitan la explotación de dichos datos. No obstante, el verdadero valor de *Big Data* en salud se conseguirá solo si los diferentes actores implicados en el proceso (administraciones públicas, empresas privadas, hospitales, médicos, centros de investigación, universidades, …) se comprometen en este proyecto de forma conjunta para llevar el ámbito sanitario a una nueva era. Esto solo se podrá llevar a cabo en el marco de un ecosistema de *Big Data* en salud en el que se integre conjuntamente con la tecnología, políticas adecuadas sobre privacidad y confidencialidad, infraestructuras y una cultura de uso compartido de los datos. Todo ello conlleva una serie de retos que hay que afrontar desde diferentes perspectivas y grado de profundidad.

NOTAS

[1] Del inglés, Electronic Health Record.
[2] Informe Índice SEIS http://www.seis.es/html/INDICE%202013%20DEFINITIVO%20ver%202%20SECURED.pdf
[3] Del inglés, Internet of Things.
[4] Del inglés, Magnetic Resonance Imaging.
[5] Del inglés, Computed Tomography.
[6] https://www.domo.com/blog/data-never-sleeps-4-0/
[7] https://www.amnhealthcare.com/uploadedFiles/MainSite/Content/Healthcare_Industry_Insights/Industry_Research/final.pdf
[8] Doximity https://www.doximity.com/
[9] Sermo http://www.sermo.com/
[10] Figure 1 https://figure1.com/
[11] SharePractice https://sharepractice.com/
[12] WeMedUp http://www.wemedup.com/
[13] Doc2Doc http://doc2doc.bmj.com/
[14] Ippok http://www.ippok.com/
[15] DoctorDice http://www.doctordice.com/
[16] MedCenter http://www.medcenter.com/
[17] Medicalia http://medicalia.ning.com/
[18] PatientsLikeMe https://www.patientslikeme.com/
[19] Tudiabetes.org http://www.tudiabetes.org/
[20] Stupidcancer http://stupidcancer.org/
[21] Curetogether http://curetogether.com/
[22] Aorana http://www.aorana.com/
[23] RareShare http://www.rareshare.org/
[24] Estadísticas de post publicados en temas de salud. http://www.naccho.org/topics/infrastructure/informatics/resources/upload/dredze-naccho-webinar.pdf
[25] Informe Big Data supporting Public Health policies https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/2442-sc1-pm-18-2016.html
[26] Conferencia internacional WWW de 2013. http://www2013.org/2013/04/25/social-networks-and-big-data-meet-public-health/
[27] Social data and medical data analytics special track. IEEE Computer-Based Medical Systems 2017 (CBMS 2017). http://midas.ctb.upm.es/sdma
[28] IU School of Nursing and ChaCha partner to conduct interdisciplinary Big-Data research about health and wellness http://news.iupui.edu/releases/2015/02/News%20Release.shtml
[29] HealthMap http://www.healthmap.org/en/
[30] GoogleFlu https://www.google.org/flutrends/about/
[31] TrendMiner http://www.trendminer-project.eu/
[32] http://stats.oecd.org/Index.aspx?DataSetCode=HEALTH_HCQI
[33] http://www.madrid.org/cs/Satellite?cid=1142475426064&language=es&pagename=HospitalPuertaHierroMaja%2FPage%2FHPHM_contenidoFinal
[34] http://www.miccai.org

BIBLIOGRAFÍA

AGGARWAL, C. C., HAN, J., WANG, J., & YU, P. S. (2003). «A Framework for Clustering Evolving Data Streams». In *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29* (pp. 81–92). Berlin, Germany:

VLDB Endowment. Retrieved from http://dl.acm.org/citation.cfm?id=1315451.1315460

AGUILAR-RUIZ, J. S., & GAMA, J. (2005). «Data Streams». *Journal of Universal Computer Science*, *11*(8), 1349–1352.

ASAMOAH, D., SHARDA, R., & KUMARASAMY, A. T. (2015). «Can Social Media Support Public Health? Demonstrating Disease Surveillance using Big Data Analytics». *AMCIS 2015 Proceedings*. Retrieved from http://aisel.aisnet.org/amcis2015/HealthIS/GeneralPresentations/12

BASEL KAYYALI, DAVID KNOTT, & STEVE VAN KUIKEN. (2013). «The big-data revolution in US health care: Accelerating value and innovation». In *McKinsey Company*. McKinsey Company. Retrieved from http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care

BENDER, J. L., JIMENEZ-MARROQUIN, M.-C., & JADAD, A. R. (2011). «Seeking Support on Facebook: A Content Analysis of Breast Cancer Groups». *Journal of Medical Internet Research*, *13*(1). https://doi.org/10.2196/jmir.1560

BONNIE FELDMAN, ELLEN M. MARTIN, & TOBI SKOTNES. (2012). «Big Data in Healthcare Hype and Hope». Retrieved from http://www.west-info.eu/files/big-data-in-healthcare.pdf

CAPURRO, D., COLE, K., ECHAVARRÍA, M. I., JOE, J., NEOGI, T., & TURNER, A. M. (2014). «The Use of Social Networking Sites for Public Health Practice and Research: A Systematic Review». *Journal of Medical Internet Research*, *16*(3), e79. https://doi.org/10.2196/jmir.2679

CUNNINGHAM, H. (2002). «GATE, a General Architecture for Text Engineering». *Computers and the Humanities*, *36*(2), 223–254. https://doi.org/10.1023/A:1014348124664

DEY, N., KARÁÇ A, W. B. A., CHAKRABORTY, S., BANERJEE, S., SALEM, M. A., & AZAR, A. T. (2015). «Image mining framework and techniques: a review». *International Journal of Image Mining*, *1*(1), 45-64.

DHAWAN, A. P. (2013). «Medical Image Analysis», *Volumen 31 de IEEE Press Series on Biomedical Engineering*, John Wiley & Sons

DOMINGOS, P., & HULTEN, G. (2000). «Mining High-speed Data Streams». In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 71–80). New York, NY, USA: ACM. https://doi.org/10.1145/347090.347107

EVIKA KARAMAGIOLI. (2015). «Social media as a big public health data source: review of the international bibliography». *PeerJ Preprint*.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., & SMYTH, P. (1996). «From data mining to knowledge discovery: an overview». In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 1–34). Menlo Park, CA, USA: American Association for Artificial Intelligence. Retrieved from http://dl.acm.org/citation.cfm?id=257938.257942

FERNÁNDEZ BAIZÁN, C., MENASALVAS RUIZ, E., MARBÁN GALLEGO, Ó., & PEÑA SANCHEZ, J. M. (2001). «Minimal Decision Rules Based on the A Priori Algorithm». *International Journal of Applied Mathematics and Computer Science*, *11*(3), 671–704.

FERRUCCI, D., & LALLY, A. (2004). UIMA: «An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment». *Nat. Lang. Eng.*, *10*(3–4), 327–348. https://doi.org/10.1017/S1351324904003523

GABER, M. M., KRISHNASWAMY, S., & ZASLAVSKY, A. (2005). «On-board Mining of Data Streams in Sensor Networks». In *Advanced Methods for Knowledge Discovery from Complex Data* (pp. 307–335). Springer London. https://doi.org/10.1007/1-84628-284-5_12

IAN H. WITTEN, EIBE FRANK, & MARK A. HALL. (2011). *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition* (3 edition). Burlington, MA: Morgan Kaufmann.

KASS-HOUT, T. A., & ALHINNAWI, H. (2013). «Social media in public health». *British Medical Bulletin*, *108*(1), 5–24. https://doi.org/10.1093/bmb/ldt028

LEIS, Á., MAYER, M. Á., TORRES NIÑO, J., RODRÍGUEZ-GONZÁLEZ, A., SUELVES, J. M., & ARMAYONES, M. (2013). «Grupos sobre alimentación saludable en Facebook: características y contenidos». *Gaceta Sanitaria*, *27*(4), 355–357. https://doi.org/10.1016/j.gaceta.2012.12.010

LOPER, E., & BIRD, S. (2002). NLTK: «The Natural Language Toolkit». In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1* (pp. 63–70). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1118108.1118117

MENASALVAS, E. & GONZALO-MARTÍN, C., (2016). Machine Learning for Health Informatics, Volume 9605 of the series Lecture Notes in Computer Science pp 221-242.

NAMBISAN, P., LUO, Z., KAPOOR, A., PATRICK, T. B., & CISLER, R. A. (2015). «Social Media, Big Data, and Public Health Informatics: Ruminating Behavior of Depression Revealed through Twitter». In *2015 48th Hawaii International Conference on System Sciences (HICSS)* (pp. 2906–2913). https://doi.org/10.1109/HICSS.2015.351

PADREZ, K. A., UNGAR, L., SCHWARTZ, H. A., SMITH, R. J., HILL, S., ANTANAVICIUS, T., … MERCHANT, R. M. (2015). «Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department». *BMJ Quality & Safety*, bmjqs-2015-004489. https://doi.org/10.1136/bmjqs-2015-004489

PAUL, M. J., & DREDZE, M. (2011). «You Are What You Tweet: Analyzing Twitter for Public Health». In *Fifth International AAAI Conference on Weblogs and Social Media*. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880

RODRÍGUEZ-GONZÁLEZ, A., MAYER, M. A., & FERNÁNDEZ-BREIS, J. T. (2013). «Biomedical information through the implementation of social media environments». *Journal of Biomedical Informatics*, *46*(6), 955–956. https://doi.org/10.1016/j.jbi.2013.10.006

RODRÍGUEZ-GONZÁLEZ, A., RUIZ, E. M., & PUJADAS, M. A. M. (2016). «Automatic extraction and identification of users' responses in Facebook medical quizzes». *Computer Methods and Programs in Biomedicine*, *127*, 197–203. https://doi.org/10.1016/j.cmpb.2015.12.025

SEGURA-BEDMAR, I., MARTÍNEZ, P., REVERT, R., & MORENO-SCHNEIDER, J. (2015). «Exploring Spanish health social media for detecting drug effects». *BMC Medical Informatics and Decision Making*, *15*(Suppl 2), S6. https://doi.org/10.1186/1472-6947-15-S2-S6

SEIFERT, S., KELM, M., MOELLER, M., MUKHERJEE, S., CAVALLARO, A., HUBER, M., & COMANICIU, D. (2010, March). «Semantic annotation of medical images». In *SPIE medical imaging* (pp. 762808-762808). International Society for Optics and Photonics.

XIE, Y., CHEN, Z., CHENG, Y., ZHANG, K., AGRAWAL, A., LIAO, W.-K., & CHOUDHARY, A. (2013). «Detecting and Tracking Disease Outbreaks by Mining Social Media Data». In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (pp. 2958–2960). Beijing, China: AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=2540128.2540556

ZHANG, D., ISLAM, M. M., & LU, G. (2012). «A review on automatic image annotation techniques». *Pattern Recognition*, *45*(1), 346-362.