OBS Business School

Inteligencia artificial, aprovechamiento inteligente de datos masivos usando Ingeniería del Conocimiento: Radicalización en redes sociales y consecuencias económicas de la pandemia.

José Angel Olivas, Andrés Montoro y Antonio Lorenzo
Profesores en OBS Business School.

Enero, 2023

Partner Académico:

Universitat de Barcelona

OBSbusiness.school

Autores

José Angel Olivas se licenció en Filosofía (especialidad Lógica) en 1990 (Universidad de Santiago de Compostela), obtuvo un Master en Ingeniería del Conocimiento del Dpto. de Inteligencia Artificial de la Universidad Politécnica de Madrid en 1992, y un Doctorado en Ingeniería Informática en 2000 (Universidad de Castilla–La Mancha). En 2001 fue Postdoc Visiting Scholar en el BISC (Berkeley Initiative in Soft Computing) con Lotfi A. Zadeh (creador de la Lógica Borrosa), University of California-Berkeley, USA. Desde entonces sigue colaborando activamente con el BISC y con el Centro de Inteligencia Artificial del SRI Internacional de la Universidad de Stanford (Prof. Richard Waldinger). Autor del libro "Búsqueda eficaz de información en la Web" y de más de 300 publicaciones científicas, es Doctor "Honoris causa" por la Universidad Nacional de La Plata, Buenos Aires, Argentina, desde 2020 y Académico de número de la Academia de Ciencias Sociales y Humanidades de Castilla-La Mancha, sección de Antropología, Filosofía y Pensamiento. Colaborador de la OBS desde 2015.

Andrés Montoro es Graduado en Ingeniería Informática por la Universidad de Castilla-La Mancha, máster en Ciencia de Datos e Ingeniería de Computadores en la Universidad de Granada y estudiante de doctorado en tecnologías informáticas avanzadas de la Universidad de Castilla-La Mancha, profesor en el departamento de Tecnologías y Sistemas de Información e investigador en el grupo SMILe (*Soft Management of Internet and Learning*). Su experiencia investigadora comprende trabajos en Soft Computing, Ingeniería del Conocimiento y Procesamiento del Lenguaje Natural aplicados a diferentes áreas como el discurso del odio, la radicalización y el desorden informativo.

Antonio Lorenzo es Graduado en Ingeniería en Informática por la Universidad de Castilla-La Mancha (ULCM). Desde hace 5 años es investigador y doctorando en el grupo SMILe desarrollando proyectos de analítica avanzada de datos. Con 30 años de experiencia en el sector de las TIC y las Administraciones Públicas, ha desarrollado su actividad en todas las áreas de las TIC y de la Administración Pública: coordinación y desarrollo de Sistemas de Información, administración de la infraestructura de Comunicaciones, gestión de Centro de Proceso de Datos (Grupo electrógeno, SAI's, sistemas contra incendios, seguridad física…), administración de Infraestructura de Sistemas (base de datos, sistemas operativos, almacenamiento, copias de seguridad…), elaboración y seguimiento de adquisición de contratación de bienes y servicios informáticos y atención a usuarios. Desde hace 10 años dirige el Departamento de Business Intelligence de la Junta de Comunidades de Castilla La Mancha llevando a cabo tareas de coordinación, análisis, gestión y desarrollo de cuadros de mandos para la analítica avanzada de datos, así como modelos de Inteligencia Artificial (aprendizaje automático) para el aprendizaje y la predicción.

Índice

Sección 1

Introducción

En este informe se describe y propone, dentro del ámbito de la computación, la informática y la inteligencia artificial, el uso de modelos de ingeniería del conocimiento y otras disciplinas y tecnologías más 'cognitivas' y menos 'numéricas', como el Soft Computing, el enfoque semántico/lingüístico del procesamiento del lenguaje natural, algunos aspectos de la sociología y la psicología, etc. para, entre otras cosas, hacer un mejor aprovechamiento inteligente de los datos masivos de los que se dispone en muchos contextos con el fin de diseñar sistemas más inteligentes en el sentido humano y en cuanto a sus prestaciones que muchos de los que usan únicamente técnicas numéricas aisladas.

Hay que tener en cuenta que el otro elemento central debe ser la capacidad de transferir esos modelos a la tecnología de la sociedad, el tejido empresarial y la industria, contribuyendo al desarrollo de sistemas computacionales más robustos, humanos y con mayores capacidades de anticiparse al futuro en sus diferentes formas (predicción, pronóstico, estimación, prescripción,…) que en el fondo es uno de los objetivos y obsesiones de las capacidades de inferencia que hacen a los humanos seres 'racionales'.

Esta propuesta es una continuación de la presentada en el Informe OBS: Inteligencia artificial, inteligencia computacional y análisis inteligente de datos[1], en la que se sientan las bases de lo que debe entenderse de forma rigurosa por Inteligencia Artificial.

A continuación se introduce brevemente qué se debe entender por Inteligencia Artificial y muchas características específicas que deben ser tenidas en cuenta cuando trabajamos con datos, información y conocimiento. Estos elementos no son considerados con su relevancia en la mayoría de los sistemas para el aprovechamiento de los datos masivos disponibles, lo que provoca unos resultados demasiado pobres, erróneos o irrelevantes, como vemos en muchos de los sistemas de anticipación al futuro que se publicitan, como el pronóstico electoral, la estimación de evolución de conflictos o pandemias, la evolución de factores económicos, Euribor, IPC, inflación y un largo etcétera.

Con el fin de mejorar el comportamiento y los resultados de estos sistemas, se propone (se muestra de una forma esquemática) una metodología genérica para el desarrollo de sistemas para el aprovechamiento inteligente de datos masivos combinados con la ingeniería del conocimiento del dominio de aplicación y se ejemplifica con dos casos, el primero de vigilancia en medios sociales con el fin de detectar comportamientos radicales y el segundo sobre el análisis de la influencia de la pandemia del COVID 19 en la bolsa [1].

Pueden consultarse otros ejemplos de aplicación de esta propuesta como el pronóstico electoral [2] y todos los referenciados en el mencionado anterior informe OBS sobre Inteligencia Artificial.

[1] Olivas, J. Á. (2021, 29 marzo). Informe OBS: Inteligencia artificial, inteligencia computacional y análisis inteligente de datos. OBS Business School. https://www.obsbusiness.school/actualidad/informes-de-investigacion/informe-obs-inteligencia-artificial-inteligencia-computacional-y-analisis-inteligente-de-datos

Sección 2

Inteligencia Artificial: Aprovechamiento inteligente de la combinación del conocimiento humano y los datos disponibles.

La **Inteligencia Artificial** (IA/AI –siglas en inglés–) puede ser definida como la disciplina del ámbito de la computación y los sistemas de información orientada a simular computacionalmente comportamientos humanos que pueden ser considerados como inteligentes, tanto en cuanto a actuación como a razonamiento, en el sentido de ser capaz de generar inferencias.

Pero hoy en día parece que lo que predomina es una simplificación sobre que hacer un análisis numérico básico sobre un conjunto de datos ya es IA. Calcular una regresión, aplicar un algoritmo de *clustering* o uno de clasificación basado en distancias estadísticas o en distribuciones de probabilidad, aunque puede dar muy buenos resultados para problemas muy concretos como la clasificación de imágenes o el cálculo de tendencias no cumplen la definición previa en cuanto a 'simular computacionalmente comportamientos humanos' en un sentido riguroso.

Cuando se habla de datos, deberían tenerse en cuenta algunas cosas como las que se comentan a continuación.

2.1. El uso impreciso de los términos ‘datos’, ‘información’ y ‘conocimiento’.

En el a veces mal llamado ‘mundo dirigido por datos’, es cada vez más frecuente encontrarnos con un uso impreciso e inadecuado de los términos ‘datos’, ‘información’ y ‘conocimiento’. Es habitual oír expresiones del tipo ‘análisis de información’, ‘conocimiento de datos’ y otras muchas similares, que no reflejan de forma exacta a lo que queremos referirnos.

Si consultamos el diccionario de la RAE, encontramos tres acepciones de la palabra ‘dato’:

1. Información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho. A este problema le faltan datos numéricos.

2. Documento, testimonio, fundamento.

3. Información dispuesta de manera adecuada para su tratamiento por una computadora.

La RAE nos presenta 8 acepciones muy dispersas de la palabra ‘información’, pero las más relevantes en nuestro contexto son:

1. Comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada.

2. Conocimientos comunicados o adquiridos mediante una información.

En cuanto al término conocimiento ocurre algo similar, pero podemos destacar:

1. Entendimiento, inteligencia, razón natural.
2. Noción, saber o noticia elemental de algo.
3. Saber o sabiduría.

Desde nuestro punto de vista, estas definiciones son ambiguas, imprecisas y mezclan conceptos muy relevantes en el ámbito del aprovechamiento inteligente de datos masivos.

De una forma más técnica, se suene entender como 'dato' una representación simbólica (numérica, alfabética, gráfica, …) del valor de un atributo o característica cuantitativa o cualitativa, que describe un hecho empírico, un suceso o una cualidad, atributo o característica de una entidad (persona, objeto, organización...).

La información, en cambio puede verse como la representación o visualización de un conjunto organizado de datos ordenados, distribuidos, procesados o tratados según algún criterio o método.

La asimilación inteligente de datos y/o información produce conocimiento, en cierto sentido de 'sabiduría', que nos permitirá anticiparnos a situaciones futuras, bien sea mediante predicciones, pronósticos, estimaciones, estudio de tendencias, etc.

Por último, permítase un ejemplo que represente estos tres niveles de abstracción descritos. Imaginemos un entorno hospitalario. En este contexto, los datos serían la materia prima inicial, las historias clínicas electrónicas de los pacientes, las imágenes radiológicas, los informes de urgencias, los datos de ocupación de camas o de UCI, etc. La información sería por ejemplo un gráfico de tarta que mostrase el nivel de ocupación de la UCI por meses u otro que nos permitiese visualizar el porcentaje de detección de tumores en imágenes radiológicas. Conocimiento sería, por ejemplo, el que proporcionara un sistema automático que hubiera sido capaz de extraer a partir de los datos y la información anteriores una regla del tipo: 'si el paciente pasa más de 15 días en UCI probablemente se infecte con la bacteria x'. Esto nos permitiría anticiparnos a situaciones futuras de forma automática ya que, por ejemplo, el sistema de gestión hospitalaria podría lanzar un mensaje de aviso (*early warning*) sobre esta posibilidad cuando un paciente lleve más de ese tiempo en cuidados intensivos.

Usemos de forma más precisa estos términos y no hablemos salvo en casos muy específicos de 'analizar información', se analizan los datos y se visualiza o usa la información para tomar decisiones (las personas), por ejemplo en los 'cuadros de mando' o *dashboards* y debemos avanzar hacia la obtención automática de conocimiento mediante sistemas computacionales, a partir de datos e información, como se muestra en el Ejemplo 2.

2.2. La deificación actual de los ‘datos’.

Desde nuestro punto de vista, asistimos diariamente a un constante proceso de deificación, casi religioso, de ensalzar excesivamente el valor y el potencial de los datos. No debemos perder de vista que, como hemos detallado en otras contribuciones académicas, los datos son en su mayoría representaciones parciales e incompletas de determinados aspectos de la realidad, intrínsecamente sesgados (por el propio diseño de la estructura de su adquisición o de dónde residen) y necesariamente con diversas cargas de ruido, imprecisión, incertidumbre, errores y, una de las cosas más relevantes actualmente, intencionalidad, buena o mala.

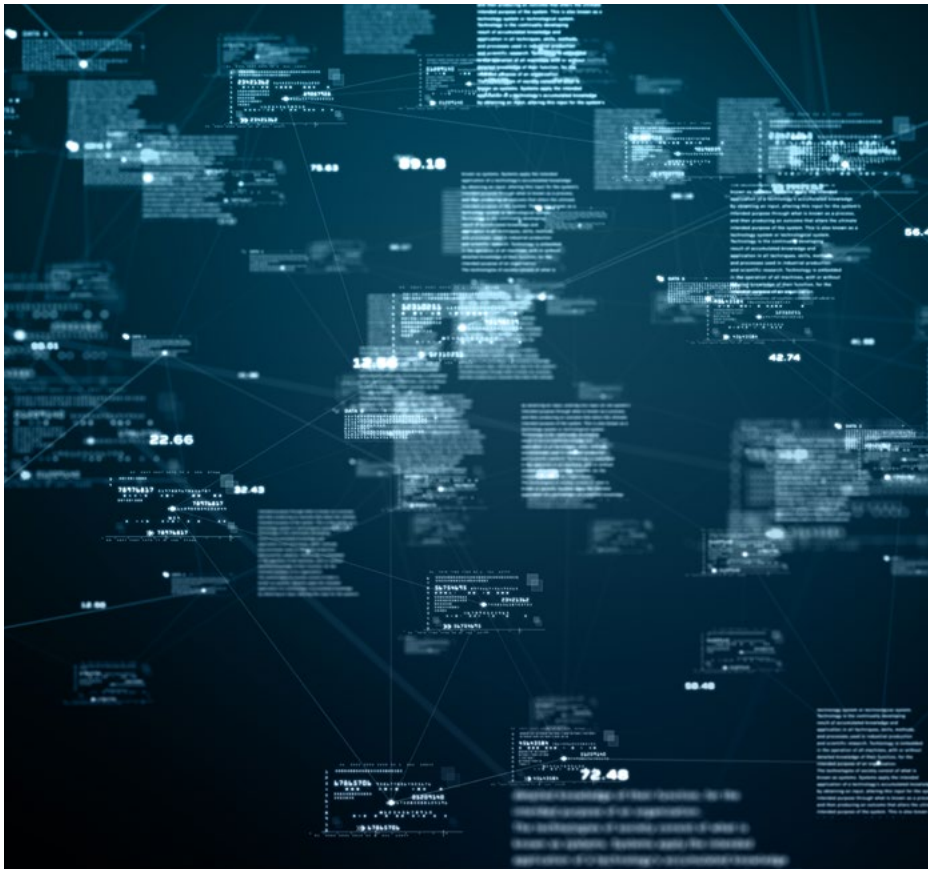
Las técnicas y herramientas encaminadas a su aprovechamiento solo permiten en general como entrada datos numéricos, con cierta relación de orden (total o parcial), estructurados y normalizables. En este contexto podríamos decir que los datos no estructurados y de otros tipos no existen. Por ejemplo, si queremos aprovechar el contenido de un conjunto de tweets o de imágenes radiológicas (en principio datos no estructurados o semi estructurados), lo primero que se hace es normalizarlos y estructurarlos, por ejemplo, convirtiendo cada tweet en un vector que almacena la frecuencia de aparición de cada término presente en dicho micromensaje o cada imagen en un vector numérico sobre el color del contenido de cada pixel.

Todo esto provoca una especie de 'efecto embudo', dónde en su parte ancha está la realidad, con sus matices, imprecisiones, riqueza de detalles y dimensiones, y por la estrecha su representación numérico-estructurada (única entrada factible a las diversas técnicas y herramientas), reduciendo drásticamente esta riqueza. Y la consecuente pérdida de representatividad.

Por otra parte, en las técnicas y herramientas se produce el efecto que solemos denominar 'la caja del matemático', que consiste en que cuanto más robusta es la técnica y más propiedades importantes cumple (consistencia, completitud, convergencia…), más pequeña es la caja, menos matices del mundo representa. Y viceversa, modelos menos 'robustos' pueden ser capaces de considerar más matices de la realidad. Por ejemplo, la lógica o matemática clásicas manejan dos cuantificadores, el 'para todo' y el 'existe' de una forma muy robusta, pero son incapaces de formalizar los matices de otros cuantificadores como 'la mayoría', 'demasiados', etcétera, que otros modelos formales menos robustos sí son capaces de manipular. Además, siguiendo el ejemplo, estos cuantificadores pueden ser contextuales. Por ejemplo 'la mayoría' no significa lo mismo (como posible abstracción porcentual) si decimos a nuestros alumnos 'la mayoría vais a sacar una buena nota en esta asignatura' que si decimos en un contexto de urbanismo 'en 2026, la mayoría de los coches que veamos circulando serán híbridos o eléctricos'. Y claramente en ninguno de los casos pensamos en 'la mitad más uno'.

Por lo tanto, cuando alguien argumenta que algo es objetivo o irrefutable porque se apoya en datos, el argumento es falaz por definición. Ni hacienda tiene unos datos perfectos ni completos de todos nosotros. Por ejemplo, en esta dura pandemia que nos azota, se están tomando decisiones falaces en base a supuestos datos objetivos. Nada más lejos de la realidad: ¿Cómo se cuantifica cómo se amontona la gente en un determinado medio de transporte?, ¿cuántas de las personas que conocemos que se han infectado recientemente están incluidas de forma completa en las bases de datos?... En lo más duro de la pandemia, en diversos ámbitos se adelantaba o retrasaba el envío de algunos datos con el fin de alterar la percepción política de la situación, en cuanto a mejora o empeoramiento. Y no hablemos de cuando un político o gestor habla de 'aplanar la curva'… Por favor, seamos serios desde un punto de vista ético y científico y no veamos los datos como una especie de 'dios pagano' que nos proporcionará la verdad absoluta. Nada más lejos de la realidad.

2.3. La ilusión del análisis de datos ‘no estructurados’.



Habitualmente se suele distinguir entre datos estructurados, semi-estructurados y no-estructurados. Los primeros son aquellos que están organizados de una forma clara en cuanto a campos, registros y relaciones, con formatos prediseñados. Habitualmente los valores para un determinado campo son números en un intervalo y con un formato específico, o etiquetas lingüísticas dentro de unos valores posibles o en lo que técnicamente puede denominarse ‘lenguaje regimentado’ (en el sentido Fregeano de la ‘Conceptografía’).

Los datos no-estructurados son aquellos que no tienen estas características. No están sujetos a formatos predefinidos, como puede ser un video o imagen abiertos o un texto libre. Cuando hay algún tipo de homogeneidad, limitación u organización suelen denominarse semi-estructurados.

Hoy en día es habitual que se hable de análisis de datos estructurados y no-estructurados, pero es una afirmación incierta, imprecisa e inadecuada en varios aspectos. La palabra ‘análisis’ nos evoca el tratamiento numérico o estadístico básico, lo cual es evidentemente muy restrictivo e inicial. Desde nuestro punto de vista debería hablarse de ‘aprovechamiento’ que es un concepto más

amplio, que va más allá del procesamiento numérico y que evoca un fin útil y no meramente un proceso descriptivo.

Actualmente no hay herramientas que puedan manipular directamente datos no-estructurados. Y las técnicas disponibles tampoco de una forma completa. Por ejemplo, es posible aplicar técnicas generales de Procesamiento de Lenguaje Natural a un documento o conjunto de documentos, pero para poder trabajar con ellos, debemos preprocesarlos, es decir estructurarlos, convertir cada documento en un vector dónde cada componente, representa la frecuencia (inversa, TF-IDF) de aparición de cada término presente en dicho documento y probablemente se han eliminado los signos de puntuación, se han desechado las palabras 'vacías' (*stop words*), como artículos, pronombres, etc. y se han reducido muchos términos a su raíz léxica (*stemming* o lematización), verbos, plurales, géneros…

Otro ejemplo podría ser el procesamiento inteligente de imágenes, que podríamos considerar datos semi-estructurados, ya que al tratar una colección todas suelen tener un formato similar.

Esta conversión imprescindible de datos no-estructurados a estructurados implica una inevitable pérdida de representatividad, de semántica en el caso de texto, lo que aunque facilita su procesamiento computacional, aleja aún más los datos de la fidelidad en su descripción de una determinada realidad.

Es por ello que se debe ser muy cuidadoso en este proceso de preprocesado y transformación de no-estructurados a estructurados, con el fin de minimizar estas pérdidas de representatividad y así maximizar las posibilidades de aprovechamiento de los mismos. Por otra parte, se debe seguir investigando en el desarrollo de modelos, técnicas y herramientas que nos acerquen al tratamiento real de datos no-estructurados.

Supongamos que estamos tratando de aprovechar un hilo de Twitter y nos encontramos con algo irónico sobre un político como 'este es muy listo, no creo que fuera mejor si ganara' y lo preprocesamos para manipularlo computacionalmente, quedaría algo como 'ser… listo… creer… ser… mejor… si… ganar…', esta conversión llevaría a una evidente y sustancial pérdida en la semántica inicial del mensaje. En el caso no de un microtexto, sino de un texto más grande, esta pérdida puede ser sustancial y condicionar los resultados de las técnicas que se usen, sobre todo si éstas son lexicográficas basadas en las apariciones de los diferentes términos.

Lo primero que se suele hacer con los datos no-estructurados es estructurarlos…

2.4. El auge de diferentes tipos y herramientas de almacenamiento y gestión de datos.

Los que ya llevamos algunas décadas vinculados al aprovechamiento inteligente de datos masivos, recordamos que hace no tantos años cuando nos referíamos a un 'especialista en bases de datos' pensábamos en alguien con conocimiento en bases de datos SQL, Microsoft (Access, Foxpro, Excel…), *ORACLE*, *Dbase* (*Clipper*…) y quizá alguna otra herramienta específica y los correspondientes SGBD (Sistemas Gestores de Bases de Datos) era alguien que dominaba la tecnología relativa al almacenamiento de datos en sistemas computacionales.

Esto ha cambiado mucho, sobre todo de una forma acelerada en los últimos años. Cuando se comenzó a hablar del 'Big Data' era frecuente que se hablase de datos SQL y NO SQL, desde mi punto de vista de una forma muy imprecisa y desenfocada, describiendo los primeros como aquellos cuyo volumen no crece o lo hace poco a poco, las necesidades de proceso se pueden asumir en un único servidor y no hay picos de uso del sistema por parte de los usuarios más allá de los previstos, y los NO SQL como aquellos en los que el volumen crece muy rápidamente en momentos puntuales, las necesidades de proceso no se pueden prever y hay picos de uso del sistema por parte de los usuarios en múltiples ocasiones.

Se comenzaron a especializar los sistemas de almacenamiento y gestión de los datos, que desde una visión simplificada podríamos dividir en estos cuatro grupos:

1. **Bases de datos documentales**, como *Couchbase*, *MarkLogic* o *mongoDB* pensadas para trabajar con documentos de diferentes formatos y tamaños.

2. **Bases de datos gráficas**, como *Neo4j* o *InfiniteGraph* orientadas a manejar datos de temas con características que puedan ser asimiladas a estructuras de grafos, como sistemas planetarios o galaxias, estructuras químicas o internet y los medios sociales, que en el fondo no deja de ser un gran conjunto de nodos interconectados.

3. **Bases de datos de columna ancha**, como *redis*, *amazonDynamoDB* o *riak*, que gestionan bases de datos NoSQL y organizan el almacenamiento de datos en columnas flexibles que pueden repartirse entre varios servidores o nodos, utilizando un mapeo multidimensional para referenciar los datos por columna, fila o marca de tiempo, es decir, los nombres y el formato de las columnas pueden variar de una fila a otra en la misma tabla.

4. **Bases de datos Key-value**, como *accumulo*, *Hipertable*, *Cassandra*, *amazonSimpleDB* o la más conocida *ApacheHBASE*, que son bases de datos sencillas que utilizan una matriz asociativa (piense en un mapa o un diccionario) como modelo de datos fundamental, en el que cada clave se asocia a uno y sólo un valor en una colección. Esta relación se denomina par clave-valor. Siguen la filosofía del famoso paradigma *Map-reduce*, que puede ser considerado uno de los pilares de Big data desde sus inicios en Google a principios de este siglo.

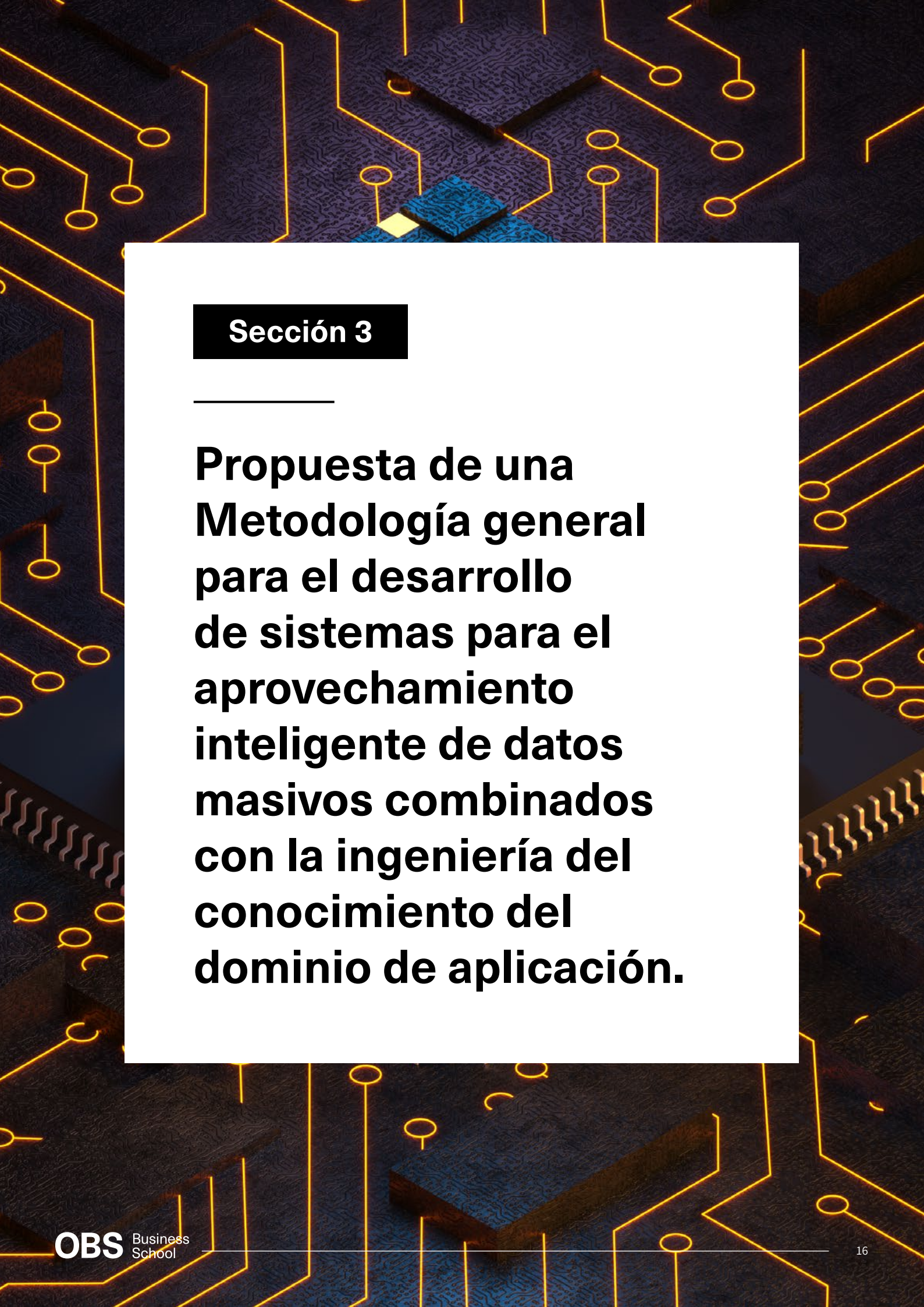
Cada vez más, la tendencia es almacenar y procesar datos en la 'nube', con herramientas como *amazonWebServices*, Windows Azure o *Cloudera* para *Hadoop* y en particular destacaría *Google Colab*, que permite a cualquier usuario escribir y ejecutar un programa *Python* en el navegador y es especialmente adecuado para tareas de aprendizaje automático (descubrimiento de comportamientos regulares -patrones- en datos), análisis de datos y educación.

Todo esto nos lleva, entre otras cosas, al concepto de 'data lake' o lago de datos, que trataremos posteriormente con más detalle.

Como vemos, el perfil de un especialista en bases de datos ha cambiado radicalmente en los últimos años y cada vez es más inabarcable en su totalidad.

Todo esto nos avoca al conocido **concepto de '**data lake**'**, que podemos sintetizar como ese conjunto heterogéneo de datos almacenados en bases de datos muy diferentes, como las que se han mencionado, que pueden estar siendo modificadas simultáneamente desde diferentes lugares (por ejemplo la base de datos de clientes de una compañía puede estar siendo manipulada simultáneamente desde el departamento de marketing, el de facturación, el de mantenimiento y el de ventas).

Por ello, para abordar problemas complejos no suele ser suficiente la aplicación aislada de algún procedimiento, algoritmo o método de análisis de datos numéricos, que es lo que se suele hacer en muchas ocasiones. Estos mecanismos sólo admiten como entrada conjuntos de datos estructurados, numéricos y normalizables, que dista mucho el contenido de un lago de datos unido a todo el conocimiento que puede haber sobre el tema en cuestión. Por lo tanto, en casos reales complejos es necesario además de hacer un análisis numérico de los datos disponibles (con todos los problemas y matices comentados) combinarlo con la ingeniería del conocimiento del entorno, que nos guie en dicho aprovechamiento de datos masivos.



Sección 3

Propuesta de una Metodología general para el desarrollo de sistemas para el aprovechamiento inteligente de datos masivos combinados con la ingeniería del conocimiento del dominio de aplicación.

Se presenta en este informe un esquema de metodología que se está desarrollando en profundidad en diferentes trabajos académicos (tesis doctorales) y científicos relacionados con los autores. La metodología es genérica y común para todas las posibles tareas, y consta esencialmente de las siguientes fases:

1. Estudio de viabilidad de la propuesta, usando técnicas adecuadas, como por ejemplo el *Test de Slagel*.

2. Descripción detallada de la propuesta, en cuanto a elementos de partida disponibles (datos, expertos…), objetivos generales y elementos a alcanzar, diferenciando claramente si lo que se persigue es clasificación, segmentación, predicción, pronóstico, prescripción, estimación, prospección y cualquier otra forma de anticipación al futuro.

3. Determinación del alcance y límites, qué cosas se van a tratar dentro de ese dominio y cuáles no.

4. Adquisición del conocimiento del dominio: A partir de expertos, datos… usando técnicas adecuadas para tal fin, como entrevistas (estructuradas y no estructuradas), emparrillados (*repertory grids*), Discusión focalizada, Análisis de casos tipo, Simulación de escenarios hacia delante, Método de los incidentes o decisiones críticas, Método de los objetivos, Método de la división del dominio, Método de la reclasificación o descomposición de metas, Método del "*Teachback*", Método del "*Role*" inverso, Método de las veinte preguntas, Método del escenario incompleto, Técnicas de entrevista en grupo, como por ejemplo "*Brainstorming*", Toma de decisiones por consenso, Método Delphi, Técnica del grupo nominal. (Similar al método Crawford Slip), etcétera.

5. Representación del conocimiento: Elaboración de la Taxonomía del problema a abordar, Ontologías de los elementos del dominio, determinación de la base de conocimiento (conjunto de reglas de inferencia a aplicar para extraer las nuevas conclusiones) y otras muchas cosas que se pueden necesitar.

6. Aplicación de Modelos de inferencia y razonamiento aproximado combinados con el aprovechamiento y análisis de los datos disponibles (*machine learning*…).

7. Evaluación de los modelos aplicados y en su caso del sistema final (Validación, verificación, usabilidad, utilidad, rendimiento…).

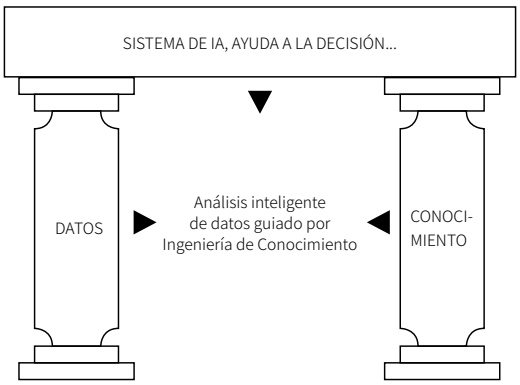
8. Documentación completa del sistema o los programas y algoritmos usados y/o desarrollados y de los resultados obtenidos y su evaluación.

Todas estas tareas se apoyan y extienden la combinación de la Ingeniería del conocimiento y el aprendizaje automático en el sentido propuesto en foros como el congreso AAAI-MAKE (Stanford), *AAAI Spring Symposium on Combining Machine Learning with Knowledge Engineering*[2], y que refleja la siguiente figura:

[2] AAAI-MAKE.2023: Challenges Requiring the Combination of Machine Learning and Knowledge Engineering. https://www.aaai-make.info/

Figura 01 ⟶ SISTEMAS DE IA GUIADOS POR IC

Fuente: Elaboración propia



En el sentido de que ambas columnas pueden interactuar para generar y/o comprobar/refutar hipótesis, proponer modelos, patrones, etc. en ambos sentidos.

A continuación se muestran dos ejemplos, el primero sobre cómo puede ser aplicada para la detección y clasificación de determinados contenidos y mensajes de redes sociales, y un segundo para el aprovechamiento de datos masivos sobre el covid-19 y pandemias anteriores similares unidos al conocimiento sobre el contexto de estas epidemias, estableciendo así modelos predictivos y conclusiones más cercanas a la realidad que con el uso de únicamente análisis estadístico numérico.

Sección 4

Ejemplo 1: Detección de comportamiento radical en redes sociales.

Internet y en concreto las redes sociales se han convertido en un nuevo caldo de cultivo para los procesos de radicalización que antes no existía. Para identificar y comprender el comportamiento radical en los medios sociales, se aborda desde la perspectiva mostrada centrándonos en describir, analizar y proponer métodos para la adquisición, representación y utilización del conocimiento sobre este fenómeno además de los datos disponibles y el análisis de los propios mensajes de las redes.

Como se ha presentado, una taxonomía es una representación jerárquica del conocimiento que permite captar los conceptos clave y las relaciones relevantes entre entidades en un determinado dominio. Esta conceptualización puede permitirnos caracterizar el perfil radical en los medios sociales. Se ha utilizado un método híbrido de adquisición de conocimientos para construir esta taxonomía. En primer lugar, recurriendo a expertos en la materia (un policía especializado y un profesor universitario del Instituto Europeo e Internacional de Derecho Penal de la Universidad de Castilla-La Mancha), y después analizando la literatura de referencia sobre el comportamiento radical y algunos marcos jurídicos penales. El resultado de la conceptualización de los conocimientos adquiridos es una taxonomía que modela el comportamiento radical, de la que una parte puede verse en la figura 2. Se pueden distinguir tres partes principales.

Figura 02 ⟶ PARTE DE LA TAXONOMÍA SOBRE EL COMPORTAMIENTO RADICAL.

Fuente: Elaboración propia

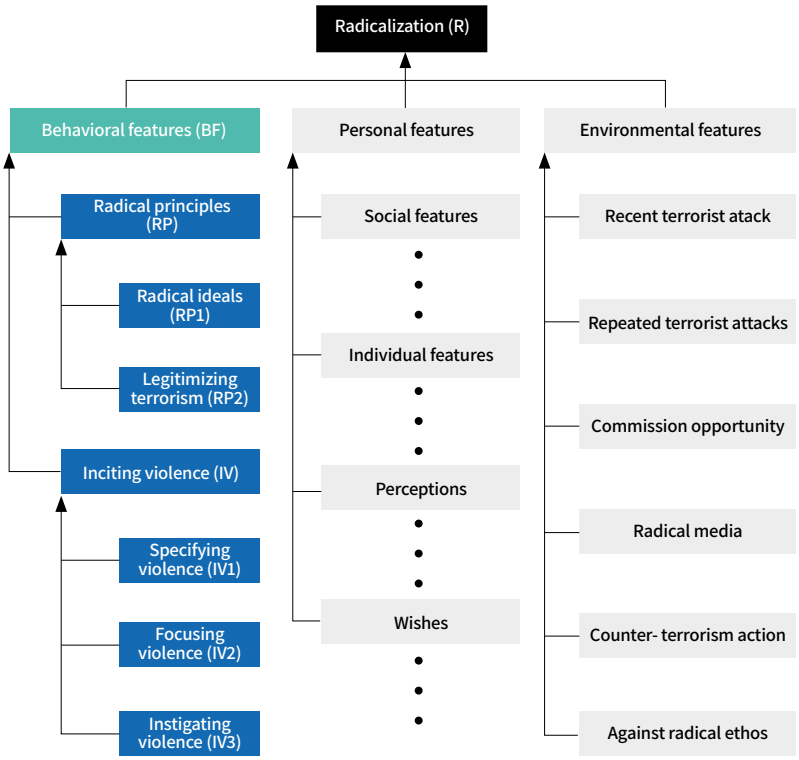
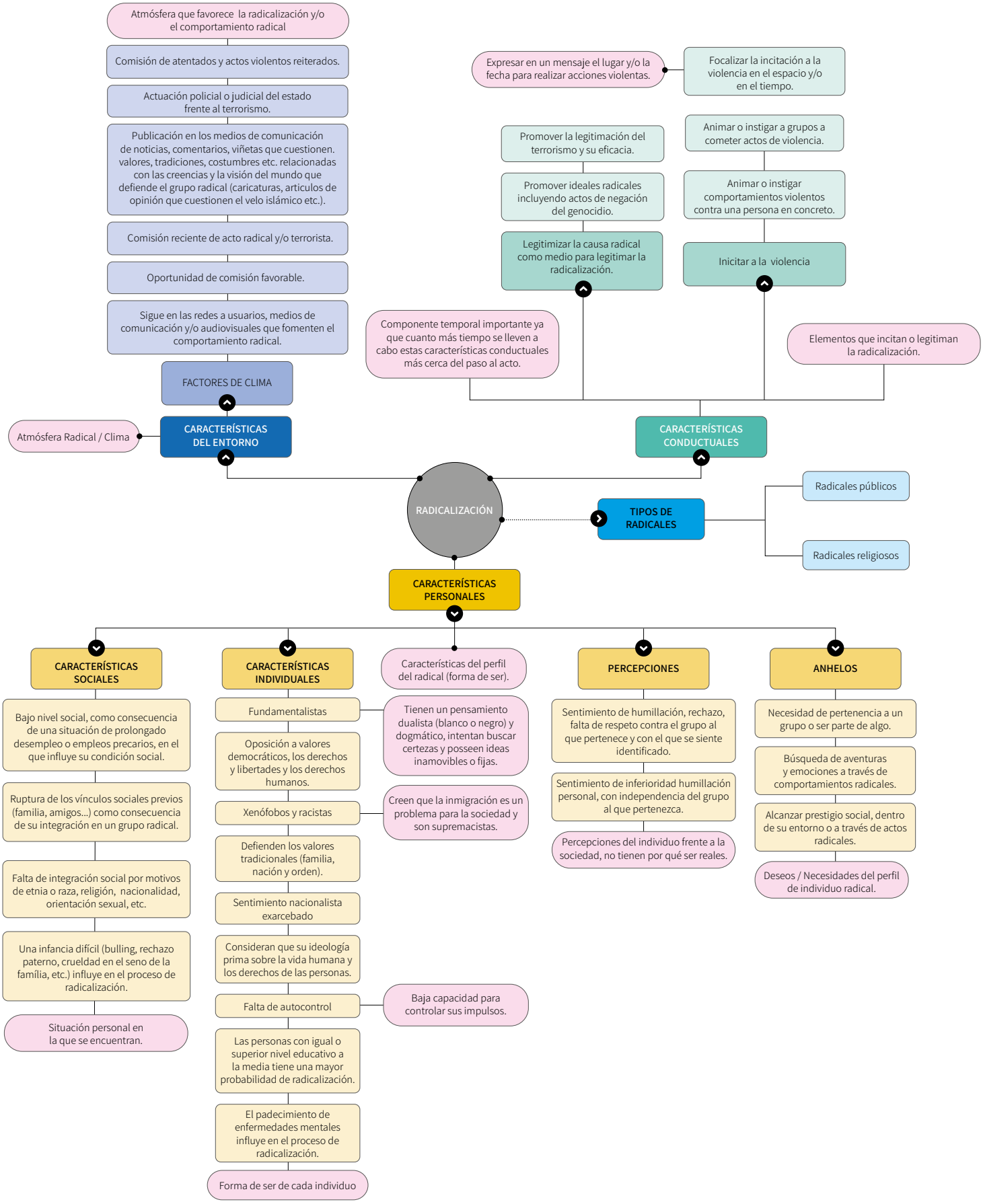


Figura 03 ⟶ TAXONOMÍA COMPLETA SOBRE EL COMPORTAMIENTO RADICAL

Fuente: Elaboración propia



- Rasgos conductuales: conductas que incitan y/o legitiman el comportamiento radical.
- Características personales: se refieren a la forma en que el individuo está inmerso en el proceso de radicalización, destacando sus características sociales e individuales y sus deseos y anhelos.
- Características ambientales: se refieren al clima que favorece el proceso de radicalización y la transición al acto terrorista.

De esta taxonomía se pueden extraer conclusiones parciales desde los nodos hoja hasta su núcleo. La taxonomía está compuesta por variables lingüísticas [3] y a partir de ella se ha implementado una base de conocimiento con reglas borrosas [3]. Debido al gran tamaño de la taxonomía, se ha seleccionado una pequeña parte de la misma, concretamente los rasgos de comportamiento (véase la Fig. 2), por lo que la base de conocimiento se implementa sobre este subconjunto de la taxonomía. A continuación se describe brevemente el significado de cada una de las etiquetas que constituyen las características de comportamiento:

- Principios Radicales (RP):
  - » Ideales Radicales (RP1): promueven creencias radicales.
  - » Legitimación del terrorismo (RP2): promueven la legitimación del terrorismo y su eficacia.
- Incitación a la violencia (IV):
  - » Especificar la violencia (IV1): Centrar la incitación a la violencia en el espacio y/o tiempo. Es decir, expresar en un post de la red social el lugar y/o fecha para las acciones violentas.
  - » Focalizar la violencia (IV2): Alentar o instigar un comportamiento violento contra una persona o subgrupo en particular.
  - » Instigar a la violencia (IV3): Alentar o instigar actos de violencia.

    (Esta etiqueta es la generalización de las anteriores).

La base de conocimientos consiste en una base de datos que contiene conjuntos de términos utilizados para describir cada una de las clases de la taxonomía y conjuntos de etiquetas para cada concepto representadas por números borrosos (*term-sets* compuestos por conjuntos borrosos normalizados y convexos) [X] que definen la semántica dentro de un dominio predefinido de cada una de las etiquetas de la taxonomía, y una base de reglas compuesta a partir de los términos lingüísticos definidos en la base de datos y que sigue la forma de entradas múltiples-salida única:

$$\textbf{Si } X_1 \textbf{ es } A_1 \textbf{ y } \ldots \textbf{ y } X_n \textbf{ es } A_n$$
$$\textbf{entonces Y es B}$$

(Ejemplo: Si la temperatura es media y la humedad es alta entonces la velocidad del ventilador debe ser alta)

El motor de inferencia utiliza la extensión del *Modus Ponens* clásico, llamado *Modus Ponens Generalizado* propuesto por Zadeh [4]. Dependiendo de la red social de la que se extrae el mensaje a analizar, se utilizan unos metadatos u otros como entrada al motor de inferencia. Por ejemplo, en el caso de Twitter, las entradas del motor de inferencia, además de la etiqueta lingüística, son el número de seguidores, el número de *likes* y el número de *retweets*. Cada una de estas variables se representa con una función de pertenencia dividida también en diferentes particiones borrosas según el conocimiento adquirido.

Ahora la tarea consiste en asignar automáticamente los nodos hoja de la taxonomía (RP1, RP2, IV1, IV2, IV3) a un mensaje de las redes sociales (por ejemplo, un mensaje de Twitter). Hay varios problemas siendo el principal la falta de datos. Como partimos de un conocimiento nuevo, no hay datos etiquetados disponibles.

Haremos uso de una categorización semántica de las etiquetas lingüísticas de la taxonomía utilizando una ontología de dominio para cada tipo de perfil radical. Una ontología es una especificación explícita de una conceptualización en un dominio en una forma compartida y acordada [5]. La ontología propuesta está compuesta por una serie de temas que modelan las características de un tipo específico de radical y un conjunto de reglas para relacionar estos temas con las categorías o etiquetas de la taxonomía. Para construir la ontología, se ha seguido el siguiente proceso:

1. Selección de dominios. Para probar el prototipo, se ha seleccionado el dominio de la extrema derecha (y la supremacista) en España.

2. Búsqueda y captura de datos del dominio. Es difícil encontrar mensajes radicales en las redes sociales, ya que éstas son proactivas a la hora de eliminar los mensajes radicales y de odio. Por ello, hemos buscado otras alternativas para la extracción de datos, concretamente, libros y foros supremacistas y nacionalsocialistas (por ejemplo el foro S*tormfront* en español).

3. Análisis exploratorio y preprocesamiento de los textos extraídos. Utilizando técnicas típicas del Procesamiento del Lenguaje Natural.

4. Extracción de temas y términos relevantes. La extracción de temas y términos se ha llevado a cabo mediante cuatro técnicas:

    a. La primera, utilizando la frecuencia de términos - frecuencia inversa de documentos (TF-IDF) con los datos del dominio.

    b. La segunda, utilizando uno de los algoritmos más usados para estos fines, el *Latent Dirichlet Allocation* (LDA) [6] con hilos extraídos de foros supremacistas españoles. Para seleccionar el número de temas, hemos utilizado el modelo de coherencia, que mide el grado de similitud semántica entre las palabras con la puntuación más alta en cada tema.

    c. La tercera forma de extraer temas y términos del dominio ha sido utilizando el algoritmo de Gradiente Integrado. Para aplicar este método hemos etiquetado manualmente, a partir del nombre

de los hilos de los foros supremacistas, un conjunto de mensajes que distinguen cuatro clases (homófobo, xenófobo, antisemita y supremacista). Una vez etiquetados los datos, implementamos y entrenamos una red neuronal LSTM para obtener los gradientes integrados y, a partir de ellos, los términos principales de cada mensaje.

d. Y por último, para agrupar cada uno de los términos y temas de forma coherente, hemos utilizado el conocimiento adquirido del dominio.

La ontología se compone de una lista de temas definidos en los pasos anteriores con un total de 12 temas. Algunos de los temas más representativos son los siguientes:

- **Tema 1:** cualquier concepto referido a la población judía y su religión.
- **Tema 2:** cualquier concepto referido al régimen actual (entendido como forma de gobierno).
- **Tema 3:** conceptos referidos a la feminización de la sociedad y al colectivo LGTBI.
- **Tema 4:** conceptos referidos al genocidio y su negación.
- **Tema 5:** conceptos que tengan que ver con la raza y su perfeccionamiento.

Cada uno de los temas está compuesto por una lista de conceptos y estos, a su vez, contienen una lista de otros conceptos asociados que pueden ser similares, como sinónimos o términos relacionados con el concepto padre extraídos de la red semántica *ConceptNet*[3] (https://conceptnet.io/), o también pueden ser conceptos cercanos, aquellos términos que suelen aparecer junto al concepto padre. La Fig. 4 muestra la arquitectura de la ontología.

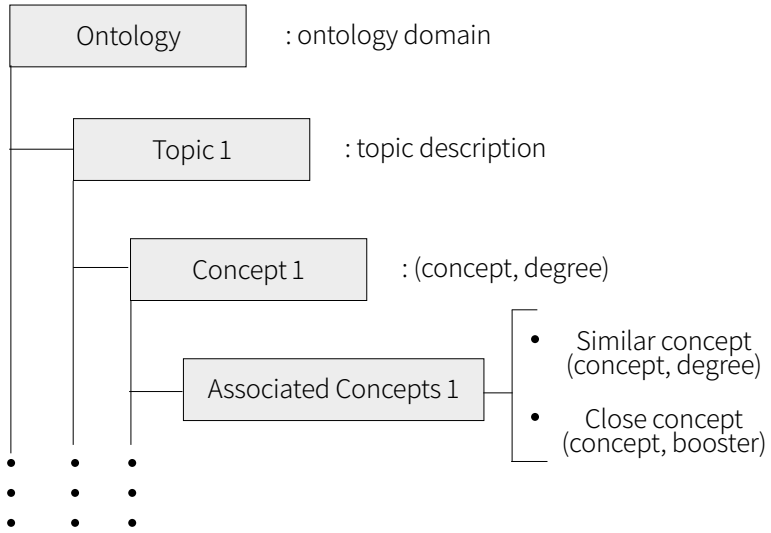
Cada concepto y sus conceptos similares tienen un grado de relación (pertenencia) con cada tema. Los conceptos pertenecientes a cada tema extraído en el proceso de creación de la ontología, tienen el mayor grado de pertenencia al tema al que pertenecen. El grado de pertenencia de los

[3] ConceptNet. https://conceptnet.io/

conceptos similares asociados a cada concepto principal es proporcionado por *ConceptNet*. Los términos cercanos actúan como reforzadores de los conceptos principales y similares.

Figura 04 ⟶ ESQUEMA DE LA ONTOLOGÍA

Fuente: ConceptNet



Ejemplo del cálculo del grado de radicalización de un usuario de Twitter basado en el contenido de su publicación.

El nivel de radicalización de usuario se puede observar a través de las características de comportamiento de la taxonomía. Los metadatos del tweet y del usuario son los siguientes:

- Tweet: "La verdad antes que la paz y siempre defenderé lo que es bueno y verdadero. Está claro que no me equivoco, las consecuencias de señalar al sionismo y a ciertos estratos de esa raza como los que dominan el mundo son evidentes. Lo sigo diciendo y lo volvería a decir mil veces más"[4].
- Seguidores: 9737
- *Retweets*: 35
- Me gusta: 238

Una vez obtenido el mensaje y sus metadatos, el proceso de evaluación del potencial radical del usuario en función de sus características de comportamiento es el siguiente:

- Paso 1: cotejar el mensaje con la ontología para obtener los temas si existen. En este caso, hay dos conceptos principales, el primero el sionismo (t_1^1) que pertenece al tema 1 con $\mu(t_1^1)= 1$ porque es un

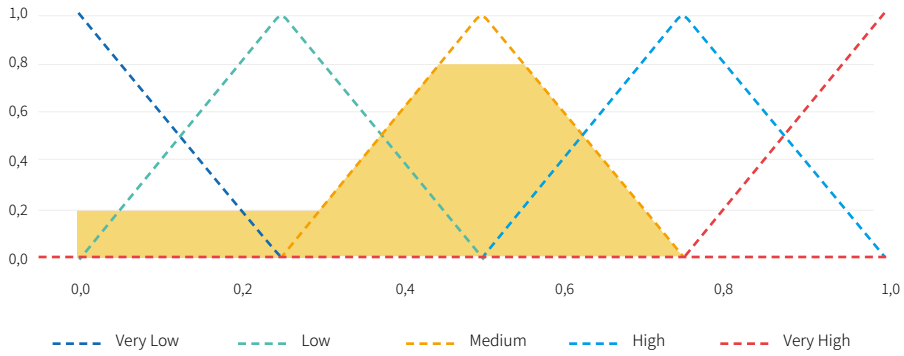
[4] El tuit original en español publicado por un conocido radical ha sido extraído mediante Wayback Machine https://web.archive.org/web/20210217085548/ https://twitter.com/isabelmperalta

concepto extraído del proceso de extracción de términos[5]. El segundo es raza (t_1^5) con $\mu(t_1^5)=1$. El (t_1^1) tiene tres conceptos cercanos: raza, dominio y mundo. Y (t_1^5) tiene cuatro conceptos cercanos: Sionismo, estrato, dominar y mundo.

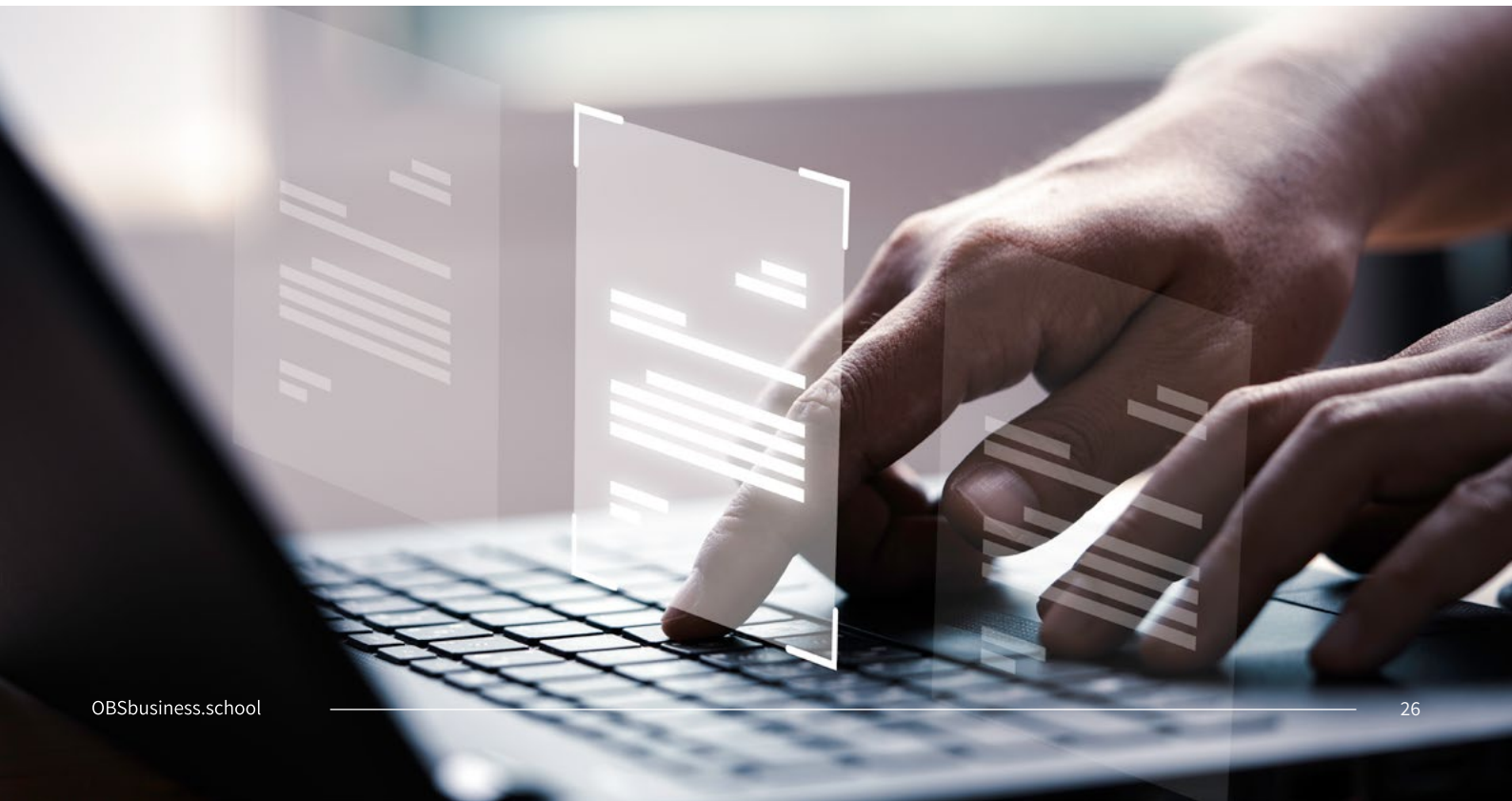
- Paso 2: emparejar con las reglas de la ontología. La regla que se activa es: *si Tema 1 y Tema 5 entonces RP1*.
- Paso 3: ahora se conoce la etiqueta de la taxonomía a la que pertenece el mensaje, por lo que es el momento de lanzar el proceso de inferencia borrosa teniendo en cuenta la etiqueta y los metadatos.

Figura 05 ⟶ NIVEL DE RADICALIZACIÓN BASADO EN LAS CARACTERÍSTICAS DEL MENSAJE ESTUDIADO, DE SU EMISOR Y EL ENTORNO Y EN EL APRENDIZAJE A PARTIR DE OTROS MENSAJES

Fuente: Elaboración propia



La Fig. 5 muestra el resultado del proceso de inferencia y se puede concluir que el usuario analizado tiene un grado medio de radicalización con confirmación alta.



Sección 5

Ejemplo 2: Aprovechamiento inteligente de los datos de la influencia de la covid-19 en los mercados de valores.

Tradicionalmente la predicción en entornos de datos numéricos se ha llevado a cabo mediante técnicas estadísticas. Las técnicas estadísticas ofrecen buenos resultados cuando hay una proyección de los datos actuales al futuro, es decir hay tendencias claras. En últimos años se han usado también las técnicas de aprendizaje automático. Esto ha sido posible debido a la gran cantidad de datos que se dispone, así como al aumento de la capacidad de computación que han experimentado los sistemas informáticos. Algunos algoritmos de aprendizaje automático permiten extraer información de datos no estructurados e inferir automáticamente reglas a partir de datos. No obstante, estas técnicas, en la predicción, ofrecen resultados limitados debido a que solo ofrecen buenos resultados cuando el caso a predecir es del tipo de casos con el que ha sido entrenado. Tanto las técnicas estadísticas como las técnicas de aprendizaje automático se basan casi exclusivamente en el análisis de los datos históricos. Ambas han dado buenos resultados en algunos ámbitos de los seguros, la banca o la medicina.

Dentro del ámbito de la predicción hay otro tipo de eventos en los que no solamente es suficiente el analizar los datos históricos, es necesario aportar conocimiento extra, específico del evento a predecir, que determina el resultado. Los resultados deportivos, electorales o bursátiles suelen ser de este tipo de eventos. El equipo de fútbol del FC Barcelona (FCB) es el único en la historia de la Copa de Europa de Fútbol que ha llegado a dos semifinales con una ventaja de tres goles en el primer partido y perdió las dos semifinales en el segundo partido. Ocurrió en la semifinal de 2018 contra la Roma, en el primer partido el FCB ganó 3-0 y en el segundo partido perdió por 3 goles. Se clasificó la Roma para la final. También sucedió en mayo de 2019, en el partido Barcelona-Liverpool, ganó el Barcelona 3-0, y en el segundo partido el Liverpool ganó por cuatro goles. Desde un punto de vista estadístico el FCB tenía una alta probabilidad de clasificarse. Desde el punto de vista del aprendizaje automático, si se hubieran analizado todos los partidos previos al de Copa de Europa en los cuales un equipo marca tres goles en el partido de ida, el FCB se hubiera clasificado para la final. No obstante, el FCB, no se clasificó ninguno de los dos años. Como se propone en este informe, en este tipo de eventos se necesita conocimiento extra, habitualmente suministrado por humanos conocedores o expertos en el dominio, que indique entre otras cosas los factores clave, específicos del evento a predecir, para mejorar la predicción respecto a las técnicas estadísticas y técnicas de aprendizaje automático.

Para este ejemplo se aplicará la metodología propuesta, basada en este caso principalmente en el conocimiento experto, que permite extraer conclusiones de cuánto tiempo tardarán las Bolsas de Valores en cambiar la tendencia decreciente y recuperar los valores anteriores a las bajadas como consecuencia de la pandemia del covid-19.

Las Bolsas de valores permiten introducir órdenes y negociar la compra y venta de valores. Desde el punto de vista de las empresas, la Bolsa de valores es una vía de financiación con el objetivo de crecer y expandirse. La empresa obtiene liquidez como consecuencia de la venta de acciones que servirá para hacer mejoras en la empresa: abrir nuevas sedes, contratar más personal, comprar maquinaria, invertir en investigación, etc. Desde el punto de vista del inversor, la Bolsa de valores sirve para obtener rentabilidad. La rentabilidad se obtiene comprando acciones cuando tienen un precio bajo para venderlas cuando tienen un precio alto. El inversor también puede obtener ganancias como consecuencia de los dividendos. La crisis sanitaria de la covid-19 en todo el mundo produjo una crisis económica mundial. La covid-19 es altamente contagiosa. Para evitar la expansión del virus, los Gobiernos impusieron medidas de confinamiento y limitaciones a la movilidad de la población, provocando que la economía mundial se paralizara: la producción y elaboración de productos, las exportaciones/importaciones de bienes y servicios, transacciones financieras, etc.

Diversos autores han estudiado cómo afectaron las crisis sanitarias a las bolsas de valores. A continuación, se muestran unos ejemplos. Dastkhan y sus colaboradores [7] estudian cómo afecto la epidemia del SARS-CoV de 2003 a los mercados de valores de China, Japón, Taiwan, Singapur y Hong Kong, desde los 5 años anteriores y posteriores a la epidemia, utilizando la técnica estadística de cointegración suave y variable en el tiempo. Noy, I. y su equipo [8] estudian, de forma retroactiva, los efectos económicos del SARS-CoV, indicando

que la cadena de suministro en China no se vio afectada, y el movimiento transfronterizo de mercancías continuó sin perturbaciones significativas. En cambio, el turismo fue muy afectado, entre marzo y abril de 2003 el total de llegada de visitantes cayó en un 63% y se cancelaron más del 45% de sus vuelos programados en abril de 2003. Se recuperaron los niveles económicos pre epidémicos a finales de julio de 2003. Gormsen y sus colaboradores [9] han estudiado los datos agregados del mercado de valores y mercado de futuros sobre los dividendos para cuantificar cómo evolucionan las expectativas de los inversores sobre el crecimiento económico con la crisis del SARS-CoV-2. Albulescu [10] desarrolla un estudio que relaciona la crisis del covid-19 con el precio del petróleo y la incertidumbre económica. Fan, [11] hace un estudio del coste en la economía mundial de la crisis de covid-19. Determina que el coste de una pandemia moderadamente grave es del 1% de los ingresos mundiales y de una pandemia grave del 4%-5%. Minh [12] intenta predecir el precio de las acciones. Para ello utiliza el análisis técnico bursátil, que da buenos resultados en general, pero no tiene en cuenta los acontecimientos inesperados. Para mejorarlo utiliza el Procesamiento del Lenguaje Natural (PLN) de las noticias económicas. Complementa el sistema realizando un "análisis de sentimiento" de cada palabra. El método propuesto necesita gran cantidad de tiempo de entrenamiento, así como muchos recursos computacionales.

La metodología utilizada se basa en la propuesta anteriormente, y se va a aplicar para determinar qué factores influyeron en recuperación los principales mercados de valores tras la pandemia de la covid-19, en concreto:

I. Determinar por qué el evento a predecir es complejo.

II. Técnicas comunes de predicción.

III. Análisis de los datos históricos y actuales.

IV. Extracción de conocimiento: Factores claves.

V. Resultados.

I. A finales del s. XX, y en especial durante el s. XXI, como consecuencia de acuerdos comerciales entre países se produjo una reducción de las barreras comerciales, creándose una interdependencia económica mundial, provocando un aumento del volumen y de la variedad de las transacciones internacionales de bienes y servicios, así como de los flujos internacionales de capitales. A las Bolsas de valores les influyen multitud de noticias o eventos, unos planificados (indicadores macroeconómicos, resultados empresariales, publicaciones de índices de referencia, …) y otros no planificados (desastres naturales, relaciones comerciales, guerras entre países…). Los eventos, tanto los planificados como los no planificados, crean un punto de inflexión en las Bolsas de valores cambiando de forma abrupta su tendencia. Por ejemplo, el S&P500 llevaba 10 años en tendencia creciente, y con motivo del covid-19 ha cambiado su tendencia a decreciente (en 20 días perdió el 30% de su valor).

II. La predicción de índices bursátiles usa técnicas especificas basadas en la estadística. Se representan con velas japonesas y se analizan mediante el análisis técnico. El objetivo del análisis técnico es la interpretación de los gráficos bursátiles y aplicar lo que ha sucedido en situaciones pasadas a situaciones futuras. La base principal del análisis técnico es que el mercado sigue patrones que se repiten, asociándole una tendencia o cambio de esta. Cuando se repite ese patrón, se supone que en el futuro el mercado bursátil se comportará de similar forma [13]. Sin ánimo de ser exhaustivos, los dos tipos de situaciones en las que se determina un cambio de tendencia son: Hombro-cabeza-hombro (cambio de tendencia a decreciente) / Hombro-cabeza-hombro invertido (cambio de tendencia a creciente) y Doble suelo (cambio de tendencia a creciente) / Doble techo (cambio de tendencia a decreciente). El inconveniente del análisis técnico es que no se tienen en cuenta otros factores exógenos que pueden hacer variar la tendencia de las Bolsas.

III. Desde 1980 ha habido multitud de epidemias y pandemias que han afectado a los índices bursátiles. Quitando la epidemia de covid-19 en 2020, las epidemias y pandemias en los últimos 30 años más importantes han sido: En 1981 el Sida, en 2003 el SARS, Gripe aviar y el dengue en 2006, en 2009 la Gripe A, el Cólera en 2010, el MERS en 2013, el Ébola en 2014 y el Zika en 2016. Las primeras semanas, cuando se declara la epidemia la Bolsas bajan, pero si se analiza el impacto a 1 mes, 3 meses y 6 meses de estas epidemias y pandemias en la Bolsas se observa que de media del *MSCI World Index* ha subido un +3,08% en tres meses y un +8,50% en 6 meses[6].

Respecto a las epidemias y pandemias del siglo XXI, en la siguiente tabla se muestra por número de casos de personas infectadas, número de muertes y país más afectado[7] [9]:

Año	Virus	Período	Principales zonas afectadas	Infectados	Fallecidos
2003	SARS (SARS-Cov)	2002-2003	China	8.000	800 (10%)
2006	Gripe Aviar (H5N1)	2005-2006	Malasia, Indonesia, Singapur	200	100 (50%)
2009	Gripe A (H1N1) (gripe porcina)	2009-2010	México, Argentina	+700.000	+18.000 (3%)
2010	Cólera	2009-2010	Camerún, Chad, Níger y Nigeria	80.000	4.000 (5%)
2013	MERS	2013-2014	Arabia Saudita, Egipto, Omán, Qatar	2.000	700 (35%)
2014	Ébola	2014-2016	República Democrática Congo, Guinea, Liberia y Sierra Leona	28.000	14.000 (50%)
2016	Zika	2015-2016	Brasil, Argentina, Colombia.	1.000.000	10.000 (10%)
2019	SARS-CoV-2	2019-2020	China, Irán, Italia, España, EEUU	200.000	10.000 (5%)

[6] https://www.marketwatch.com/story/heres-how-the-stock-market-has-performed-during-past-viral-outbreaks-as-chinas-coronavirus-spreads-2020-01-22

[7] Organización Mundial de la Salud: https://www.who.int/emergencies/mers-cov/en/, https://www.who.int/health-topics/ebola/, https://www.who.int/csr/don/archive/disease/severe_acute_respiratory_syndrome/en/

Hasta la fecha, la influencia en los mercados de valores de las epidemias y pandemias anteriores al covid-19 no ha sido muy grande porque los países a los que afectaron no tenían una gran influencia económica mundial o porque los países afectados tenían un número de infectados bajo. La covid-19, hasta el 22 de febrero de 2020, había estado contenida en Asia (China, Corea del Sur, Japón, Singapur, Tailandia, Taiwan…), fecha en la que los mercados bursátiles de los países occidentales tenían una tendencia claramente alcista. El 20 de marzo de 2020 la OMS declara en nivel 6 (pandemia en expansión) de la covid-19. La caída media, desde el lunes 24 de febrero hasta el 19 de marzo de 2020, de las Bolsas de valores más importantes fue:

S&P500 (EE.UU.)	Dow Jones (EE.UU.)	MCSI Index (Global)	Hang Seng (China)	Nikkei (Japón)	Euro Stoxx (Europa)	DAX (Alem.)	Ibex (España)	FSTE MIB (Italia)
-29,55%	-33,33%	-32,66%	-18,29%	-34,24%	-41,77%	-37,47%	-41,05%	-43,88%

IV. Se ha desarrollado un cuadro de mandos con más de 15 fuentes de datos para hacer una analítica avanzada de los datos de la covid-19 relacionándolo con los índices bursátiles. El cuadro de mandos contiene las siguientes características:

Filtros: Continentes, países, si el país pertenece al grupo G20, si el país pertenece a la UE27, si el país está en el hemisferio sur o norte, las estaciones del año en el hemisferio sur y las estaciones del año en el hemisferio norte, el año, el mes, las olas de covid-19 y las principales variantes.

KPIs: Población, Contagiados, Defunciones, personas con el calendario completo de vacunación, % infectados respecto a la población, % fallecidos respecto a los infectados y % vacunados con el calendario completo respecto a la población.

Se han desarrollado cuatro vistas en el cuadro de mandos:

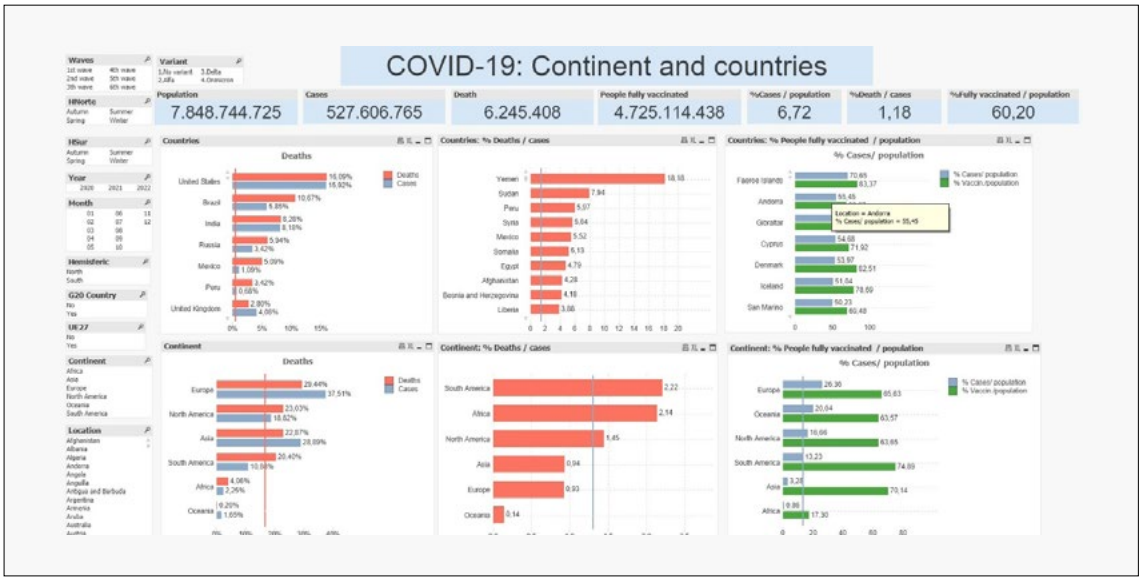
a. Vista de "Evolución del covid-19". Se muestra el número de personas infectadas por mes y año, el número de personas fallecidas por mes y año y el número de personas vacunadas con el calendario completo por mes y año. Además, se muestra la cantidad de personas contagiadas y fallecidas por año, así como por oleadas de covid-19

b. Vista de "covid-19 en continentes y países". Número de personas contagiadas y fallecidas por continentes y países. % fallecidos respecto a los infectados por continentes y países. % de personas vacunadas con la pauta completa respecto a continentes y países.

Figura 06 →

EJEMPLO DE VISTA DE COVID-19 EN CONTINENTES Y PAÍSES.

Fuente:Elaboración propia

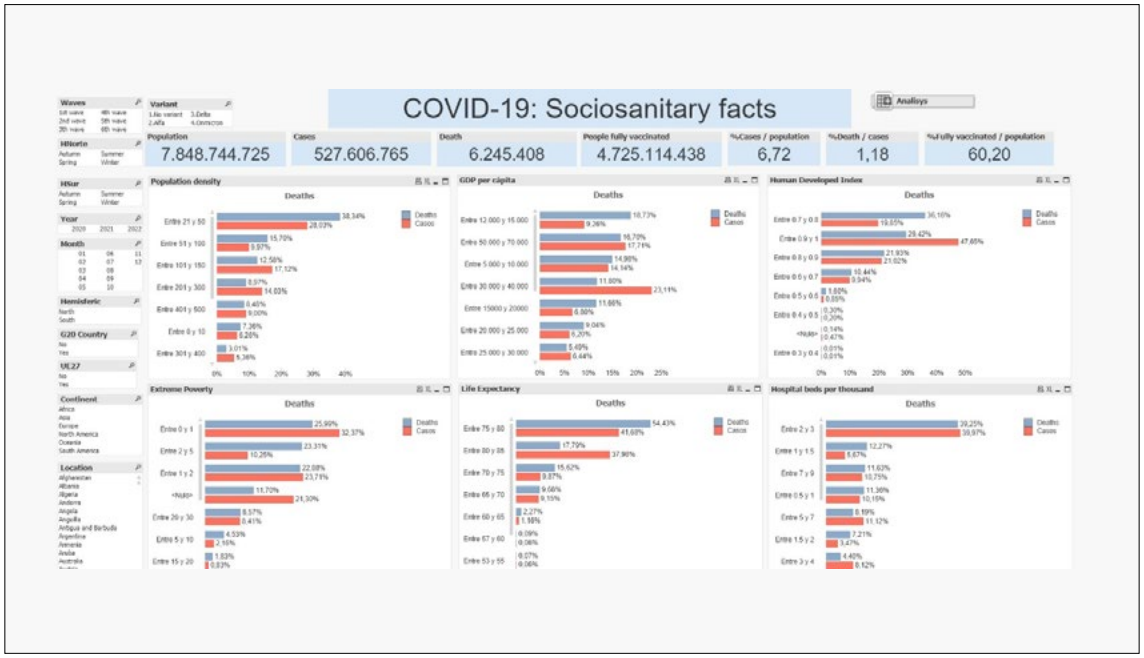


c. Vista de "covid-19 y factores sociales y de salud": Densidad de población, PIB per cápita, Índice de Desarrollo Humano, Pobreza Extrema, Esperanza de Vida y Camas Hospitalarias por mil personas, Inversión económica para la covid-19.

Figura 07 →

EJEMPLO DE VISTA DE COVID-19 Y FACTORES SOCIALES Y DE SALUD.

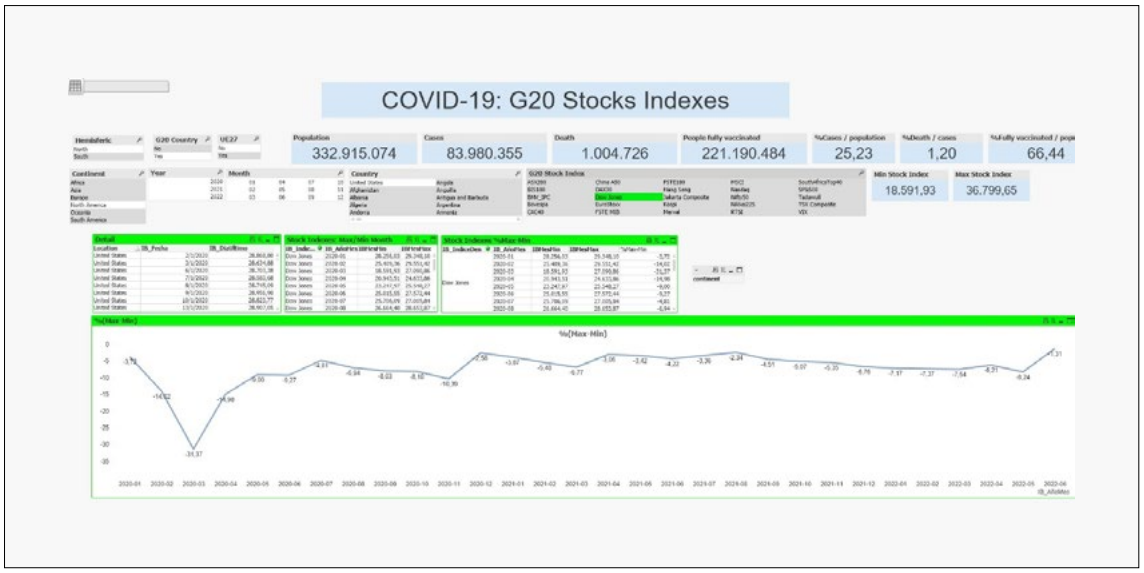
Fuente: Statista



d. Vista de "índices bursátiles". Se muestra, para los índices bursátiles más importante de los países del G20, la evolución desde enero de 2020 hasta diciembre de 2021.

Figura 08 → EJEMPLO DE VISTA DEL COVID-19 E ÍNDICE DOWJONES

Fuente: Statista



V. Del análisis de los datos del cuadro de mandos extrae que los tres factores específicos que más influyen en la recuperación de los niveles pre-pandemia son: la gestión política (adoptar desde el primer momento medidas de aislamiento y confinamiento de la población, uso de mascarilla, vacunación…), la gestión económica (establecer ayudas y subvenciones para evitar la bancarrota de empresas y el aumento de pobreza de familias) y la gestión sanitaria (disponer de los recursos humanos y materiales suficientes para atender a la población hospitalizada).

Teniendo en cuenta que la OMS decreta en el primer trimestre de 2020 el nivel 6 (pandemia en expansión) de la covid-19, se van a indicar las condiciones que determinan la recuperación de los niveles pre-pandemia covid-19 de los principales índices bursátiles:

- Los países con un PIB per cápita alto, inversión económica alta para mitigar los efectos de la covid-19 (EEUU) o la gestión sanitaria de la covid-19 fue buena o muy buena, es decir, pocos fallecidos en primer/segundo trimestre de 2020 (Corea del Sur, China, Argentina), la recuperación de sus principales índices bursátiles fue a mediados de 2020.

- Si no, Si la gestión sanitaria fue baja o muy baja.

    a. Si la gestión política fue alta (India, Japón, Turquía, Alemania) entonces la recuperación de los principales índices bursátiles fue a finales de 2020.

    b. Si no, si la covid-19 afectó muy gravemente a un sector fundamental del país, como el turismo (Italia y Francia), la recuperación de los principales índices fue a principios de 2021.

    c. Si no, recuperación de los principales índices bursátiles fue a principios de 2021 (México, Canadá).

- Si no, Si la gestión política fue baja o muy baja, no tomando medidas de aislamiento y confinamiento de la población (UK, Brasil) entonces la recuperación fue a principios de 2021.
- Si no, Si la gestión económica fue baja o mal gestionada (Rusia y Australia) entonces la recuperación de los índices bursátiles fue a mediados de 2021.

Sección 6

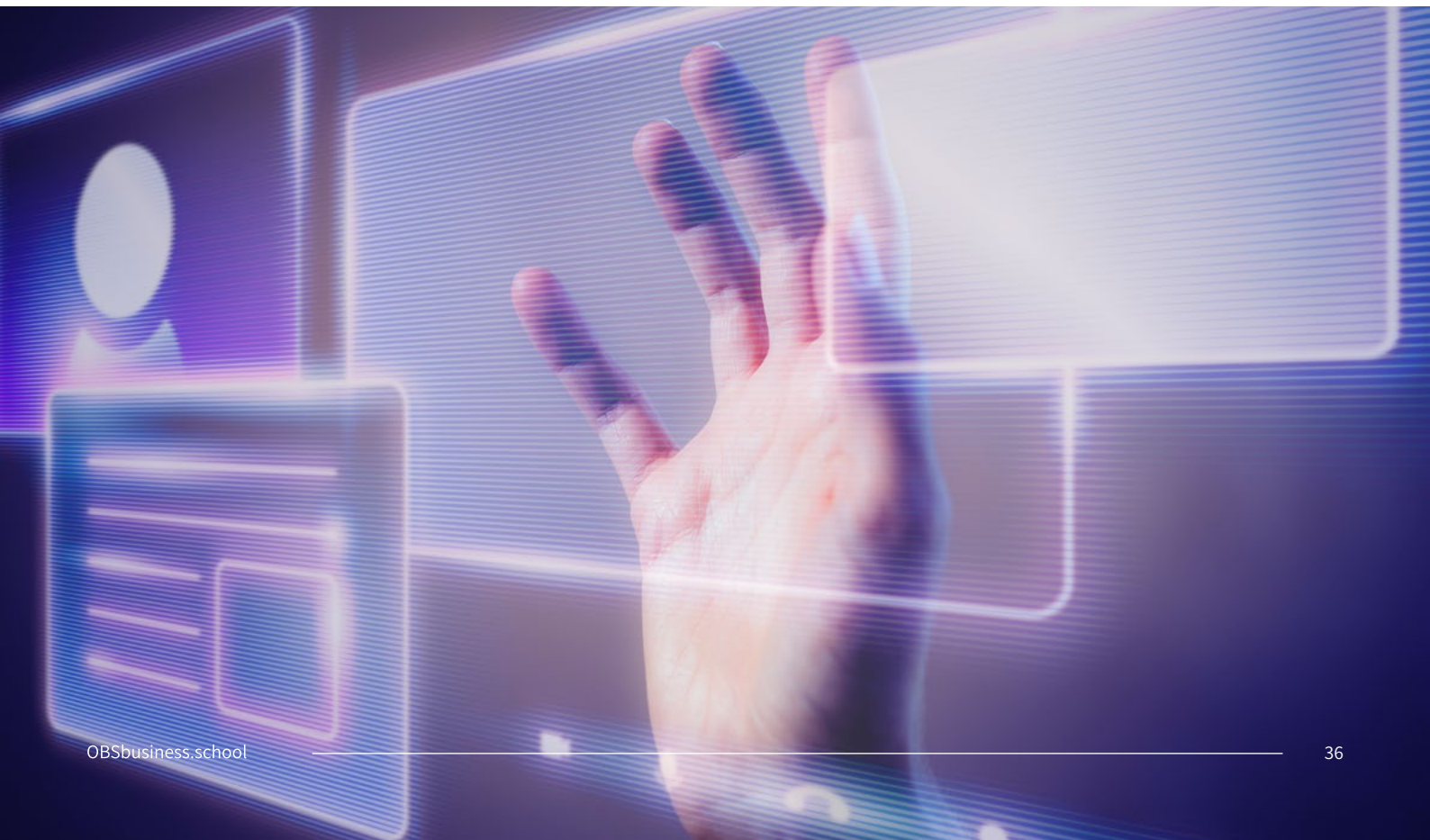
Conclusiones

En este informe se ha descrito y propuesto el uso de modelos de ingeniería del conocimiento y otras disciplinas y tecnologías para llevar a cabo un mejor aprovechamiento inteligente de los datos masivos de los que se dispone hoy en día en muchos contextos, principalmente los relacionados con el Big Data con el fin de diseñar sistemas computacionales más inteligentes en el sentido humano y en cuanto a sus prestaciones que muchos de los que usan únicamente técnicas numéricas aisladas y que son los que hoy en día se asocian principalmente con el término 'inteligencia artificial'.

Se ha introducido brevemente qué se debe entender por Inteligencia Artificial y muchas características específicas que deben ser tenidas en cuenta cuando trabajamos con datos, información y conocimiento. Se han presentado diversas críticas y comentarios acerca de la 'deificación' actual de los datos masivos y de cómo ha evolucionado el concepto de bases de datos hasta el actual de '*data lake*'.

Con el fin de mejorar el comportamiento y los resultados de estos sistemas, se ha propuesto esquemáticamente una metodología genérica para el desarrollo de este tipo de sistemas que combinan el aprovechamiento inteligente de datos masivos con la ingeniería del conocimiento del dominio de aplicación.

Finalmente, se han descrito dos casos de ejemplo, el primero sobre la observación y vigilancia en redes sociales con el fin de detectar comportamientos radicales y el segundo sobre la influencia de la pandemia del COVID 19 en la bolsa, teniendo en cuenta diferentes fuentes de datos tanto de la pandemia actual como de otras anteriores, así como otros datos contextuales y se han combinado con otras fuentes de conocimiento sobre el dominio como puede ser el de los expertos humanos.



Referencias bibliográficas

1. Lorenzo Sánchez A., Olivas J. A. (2020). Intelligent data analysis of the influence of COVID-19 on the stock market using Case Based Reasoning. J. Comput. Sci. Technol. 20(2): 10.

2. Lorenzo Sánchez A., Olivas J. A. (2019). A Case Study of Forecasting Elections Results: Beyond Prediction based on Business Intelligence. J. Comput. Sci. Technol. 19(2): 14.

3. Olivas, J. A. (2002). La Lógica Borrosa y sus aplicaciones. BOLE.TIC, nº 24, pp. 21 - 28.

4. Zadeh, L. A. (1987). Fuzzy Sets and Applications (Selected Papers, edited by R. R. Yager, S. Ovchinnikov, R. M. Tong, H. T. Nguyen), John Wiley, Nueva York, 1987.

5. Gruber T. R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), pp. 199–220.

6. Blei D. M., Ng A. Y., and Jordan M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3(4), pp. 993–1022.

7. Dastkhan H., Gharneh N. S. (2018). How the ownership structures cause epidemics in financial markets: A network-based simulation model. Physica A: Statistical Mechanics and its Applications 492, pp. 324-342.

8. Noy I., Shields S. (2019). The 2003 Severe Acute Respiratory Syndrome Epidemic: A Retroactive Examination of Economic Costs.

9. Gormsen N. J., Koijen R. S. (2020). Coronavirus: Impact on Stock Prices and Growth Expectations. University of Chicago, Becker Friedman Institute for Economics Working Paper, (2020-22).

10. Albulescu, C. (2020). Coronavirus and oil price crash: A note. arXiv preprint arXiv:2003.06184.

11. Fan V. Y., Jamison D. T., Summers L. H. (2016). The inclusive cost of pandemic influenza risk (No. w22137). National Bureau of Economic Research.

12. Minh D. L., Sadeghi-Niaraki A., Huy H. D., Min K., Moon H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. IEEE Access 6, pp. 55392-55404.

13. John J. Murphy (1999). Technical Analysis of the Financial Markets. New York Institute of Finance, New York.

OBS Business School

School of **Business Administration & Leadership**

School of **Innovation & Technology Management**

Planeta Formación y Universidades