Challenges And Opportunities In Analytic-Predictive Environments Of Big Data And Natural Language Processing For Social Network Rating Systems

J. L. Jiménez[1], I. González-Carrasco[2] and J. L. López-Cuadrado[3]

Abstract— **Social Media is playing a key role in today's society. Many of the events that are taking place in diverse human activities could be explained by the study of these data. Big Data is a relatively new parading in Computer Science that is gaining increasing interest by the scientific community. Big Data Predictive Analytics is a Big Data discipline that is mostly used to analyze what is in the huge amounts of data and then perform predictions based on such analysis using advanced mathematics and computing techniques. The study of Social Media Data involves disciplines like Natural Language Processing, by the integration of this area to academic studies, useful findings have been achieved. Social Network Rating Systems are online platforms that allow users to know about goods and services, the way in how users review and rate their experience is a field of evolving research. This paper presents a deep investigation in the state of the art of these areas to discover and analyze the current status of the research that has been developed so far by academics of diverse background.**

Keywords— **Big Data analytics, Big Data predictive, Natural Language Processing, Social network rating systems**

I. INTRODUCCIÓN

A PARTIR de la globalización de Internet 2.0 y el Internet de las cosas (IoT), la cantidad de datos que se genera día a día crece exponencialmente. La era del Big Data (BD) permite a la comunidad científica llevar a cabo diversidad de extensos estudios en casi cualquier disciplina, ya que existen ahora volúmenes de información con los que era impensable contar antes. Además, el sector de BD está proporcionando herramientas analíticas y predictivas a especialistas para el tratamiento de estos datos.

Este trabajo presenta el estado del arte de diversos componentes, el objetivo común que comparten y algunos ejemplos de los campos científicos en los que pueden aplicarse. Los tópicos a tratar en el documento son: entornos analítico-predictivos de BD, procesamiento de lenguaje natural (PLN) y Sistemas de clasificación de redes sociales (SCRS). En este trabajo, los autores tratan de responder a las siguientes preguntas:

1. ¿Cuál ha sido la experiencia de ciertos investigadores en la integración de BD con el PLN?
2. ¿Cómo puede integrarse el PLN en entornos analítico-predictivos para estudios más complejos?
3. ¿Qué técnicas ha desarrollado la comunidad científica para el análisis de la reputación a través de las calificaciones de los usuarios?

Existen diversas definiciones de BD y no hay un acuerdo sobre cómo definir algo que está creciendo tan aceleradamente cada segundo. Para los propósitos de este trabajo, los autores definen el término BD como el trabajo de reunir, organizar, limpiar y asegurar la privacidad de enormes conjuntos de datos procedentes de diversas fuentes para obtener información valiosa de tales datos. De Mauro et al. [1] han hecho un análisis más profundo de lo que el término significa basado en la definición de muchos autores. ¿Es BD una moda en informática? El futuro sigue siendo incierto, según [2] a partir de 2012 el interés y el número de estudios que se han realizado a través, o relacionados con BD ha aumentado cada año.

El BD tiene un alto potencial para emplearlo en investigación, pero primero se necesitan dos cosas: (1) pensar en cuál es el valor exacto que desea obtener de los datos, y (2) utilizar inteligentemente las herramientas apropiadas para establecer una arquitectura que eventualmente conduzca a responder a las preguntas previamente formuladas. Para efectos de este trabajo, la respuesta a la primera pregunta es la siguiente: existe un conjunto enorme de datos de medios sociales que contiene información de comentarios en línea sobre varios tipos de actividades de negocios. El valor que se espera obtener de estos datos es encontrar patrones y comportamientos de usuario que puedan ayudar a predecir indicadores, tales como: preferencias, disgustos, tendencias actuales, tendencias futuras, etc. Estos indicadores podrían ayudar a académicos o inversores a tomar decisiones sobre sus respectivas áreas. La respuesta a la segunda pregunta se discutirá en el resto del artículo.

Las técnicas de BD aumentarán su poder de cómputo cuando se combinen con otras disciplinas de ciencias de la computación. El PLN es una disciplina clave que, como se verá más adelante, funciona muy bien cuando trabaja en conjunto con el BD. Como lo indican Nadkarni et al. [3], el PLN ha estado presente en la informática desde la década de 1950 tomando prestando de diversos campos, para convertirse hoy en

[1] J. L. Jiménez, Universidad Carlos III de Madrid, Departamento de Informática, Leganés, Madrid, España, 100339395@alumnos.uc3m.es
[2] I. González-Carrasco, Universidad Carlos III de Madrid, Departamento de Informática, Leganés, Madrid, España, igcarras@inf.uc3m.es
[3] J. L. López-Cuadrado, Universidad Carlos III de Madrid, Departamento de Informática, Leganés, Madrid, España, jllopez@inf.uc3m.es

un área de gran impacto tanto para la academia como para las organizaciones.

Actualmente el comercio en línea constituye uno de los mayores activos para las grandes corporaciones, de hecho, muchas de las más exitosas startups están basados en los comentarios en línea vertidos por los usuarios pertenecientes a determinada red social. Esas expresiones y la posterior interacción entre usuarios se están convirtiendo en un factor clave para encontrar un consenso común sobre un tema determinado a través de millones de registros en datos no estructurados. Qi et al. [4] establecen que las revisiones en línea podrían ser la fuente de ideas para que los fabricantes diseñen nuevos productos que sean más adecuados para los clientes. En la Fig. 1 se presenta un esquema de los temas tratados en este artículo y su relación. En la sección II del artículo, se desarrollan estos conceptos.

Figura 1. Tópicos cubiertos en este artículo.

El objetivo del presente trabajo es determinar el estado actual de la técnica e identificar futuras líneas de investigación. Asimismo, profundizar en los desafíos que plantea BD para la investigación en ciencia informática y proponer algunos enfoques para abordar estos desafíos. Las contribuciones son:

• Una completa revisión de las tecnologías relacionadas con el BD, el PLN y los SCRS, haciendo hincapié en los aspectos en que convergen.

• Una comparación de las técnicas que se han desarrollado en las disciplinas mencionadas anteriormente, incluyendo parámetros de integridad, escalabilidad y disponibilidad.

• Una discusión sobre los desafíos actuales y las principales oportunidades que surgen en la confluencia de estas tecnologías.

Para reunir los trabajos a los que se hace referencia en este trabajo, los autores tomaron en consideración en primer lugar, aquellos estudios que trataron cada tema con un diseño de investigación exploratoria. Cabe mencionar que todos los criterios de búsqueda siempre tomaron en consideración: título, resumen y palabras clave. La búsqueda se realizó buscando material que contenga los temas Análisis Predictivo de BD, PLN y SCRS, pero los resultados casi nulos llevaron a la decisión de ampliar la búsqueda.

Posteriormente, la búsqueda estuvo abierta a la consideración de sistemas BD Analytics, BD o SCRS, junto con el PLN, y al hacer varias combinaciones de estos criterios es como se recopiló el material principal para este trabajo. De acuerdo con la clasificación establecida por Paré et al. [40] los autores desarrollaron una revisión descriptiva con el objetivo de descubrir el estado actual de la técnica para proponer una serie de lineamientos en las brechas descubiertas en los trabajos revisados.

El resto del artículo está organizado de la siguiente manera. En la Sección 2 se realiza la revisión de las tecnologías. En la sección 3 se presentan estudios de casos en los que las tecnologías han convergido total o parcialmente. La sección 4 compara los casos revisados en la sección 3, también discute sobre los desafíos que surgen y las oportunidades que existen para hacer frente a estos desafíos. Finalmente, las conclusiones se resumen en la sección 5.

II. PANORAMA GENERAL

2.1 Big Data

Dado el vasto volumen de información que se genera actualmente en medios digitales, el enfoque de bases de datos relacionales para el almacenamiento tuvo que evolucionar, dando lugar a la creación del paradigma Big data. ¿Por qué es BD tan importante y por qué los académicos deberían prestar más atención al impacto global de BD?

En [5] se afirma que, debido a su enorme potencial para generar valor comercial, se está convirtiendo en "el centro de la investigación académica y corporativa". Se puede derivar que BD es una disciplina de la Informática que, a diferencia de otras tecnologías que están más vinculadas a los procesos de negocio, puede ser de gran utilidad tanto para proyectos de investigación [6] como para proyectos empresariales.

2.1.1 Big Data Analítico

Industrias y organizaciones de todo el mundo están obteniendo información valiosa de las cantidades masivas de la información que tienen a través de la aplicación de técnicas de alto nivel de BD. Estas técnicas se conocen comúnmente como Big Data Analytics (BDA) y consisten en un conjunto de algoritmos, estadísticas avanzadas y análisis aplicados. Iqbal et al. [7] lo definen como: "(BDA) se refiere a las técnicas utilizadas para examinar y procesar BD de modo que los patrones subyacentes ocultos se revelan, las relaciones se identifican y otros conocimientos sobre el contexto de la aplicación bajo investigación están expuestos".

Debido al alto valor económico que tiene para las organizaciones y a la poderosa capacidad de analizar datos gigantescos, hoy en día BDA está siendo utilizado en sectores como la gestión de la cadena de suministro [9], mercadotecnia [10] y el proceso de toma de decisiones [11], sólo por mencionar algunos. Las redes sociales son un gran recurso para aplicar BDA [8], por ejemplo: comprender las preferencias del usuario, saber las tendencias diarias, comprender el comportamiento de los usuarios con afinidades relacionadas, analizar los nuevos hábitos de la población, etc. En el futuro cercano, posteriores investigaciones y regulaciones oficiales tendrán que establecer los límites de BDA [12].

2.1.2 Big Data Predictivo

Comúnmente conocido "Big Data Analytics Predictive" (BDP), constituye un nivel superior al de BDA en el sentido de que el primero puede hacer predicciones a partir de los resultados del análisis de datos. BDP está ayudando a las

organizaciones a lograr algo que siempre han estado buscando: tomar mejores decisiones. Lo cual se está logrando mediante el procesamiento de grandes cantidades de datos a través de diversas técnicas [13] con el propósito de hallar patrones ocultos en tales datos y descubrir correlaciones desconocidas.

Las habilidades requeridas para desarrollar un estudio del tipo BDP son diversas, entre estas se encuentran: aprendizaje máquina, análisis estadístico y cuantitativo [14], minería de datos [15], visualización de datos, creatividad y resolución de problemas y habilidades de programación. BDP no es un proyecto experimental o una tecnología emergente, es una realidad con éxito probado en diferentes campos, que van desde Recursos Humanos [14] a ciencias de la salud [15], campo en el que los autores estiman que BDP tendrá amplia aplicación en el corto plazo gracias a la gran cantidad de datos que se almacenan generan díariamente.

2.2 Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN) siempre ha estado unido a la relación humano-computador, ya sea porque necesitan "hablarse" o “entenderse” entre ambos [16]. Se trata de una disciplina de la informática siempre en evolución y que se vuelve cada vez más compleja debido a las mayores exigencias del sector.

"El PLN comenzó en la década de 1950 como la intersección de la inteligencia artificial y la lingüística" [4]. Desde entonces diversas teorías han ayudado a la evolución del PLN, como la notación Backus-Naur Form (BNF) y herramientas como analizadores léxico y sintáctico. El estado actual del arte revela que el estudio del PLN está involucrado con la conjunción de técnicas como: autómatas de estado finito y de árbol, gramática libre del contexto y etiquetado de parte del discurso [17], por mencionar algunos.

Una de las áreas donde el PLN está colaborando y siendo más útil en la ciencia es en Medicina y Biomedicina. Una breve investigación del término PLN en localizadores de bases de datos científicas conduce a encontrar numerosos artículos publicados en revistas no relacionadas a ciencias de la computación, nos dice sobre la enorme participación del PLN en estos campos. Otras áreas de aplicación se pueden encontrar por ejemplo en los sistemas de información geográfica (GIS) [18] o entrevistas motivacionales [19].

2.3 Sistemas de clasificación de redes sociales

Web 2.0 o Web Social es un término que involucra a los sitios web que facilitan la interoperabilidad y la colaboración entre sus usuarios. Existen diversos tipos de redes sociales donde los usuarios pueden interactuar o compartir sus experiencias de diversas formas. Aprovechando las posibilidades ilimitadas de Web 2.0 surgieron varias *startups* para desarrollar plataformas que permitieran a los usuarios evaluar y comentar los productos (libros, coches, celulares, tabletas, etc.) y servicios (hoteles, viajes, vuelos, etc.).

Los usuarios de estas redes comúnmente dan una calificación a "algo" y de igual manera pueden hablar de su experiencia. También expresan sus sentimientos sobre la base del grado de satisfacción que obtuvieron. La relación que existe entre lo que el usuario expresa textualmente y la calificación no siempre es clara: se puede expresar textualmente que su experiencia fue mala a terrible y en una escala de 1 a 5 estrellas, pueden dar un 4 o 5, También puede suceder la manera inversa

La forma en que una comunidad de usuarios construye la reputación de una empresa a través de un SCRS, puede desempeñar un papel clave en la decisión de un cliente potencial para comprar ese bien o contratar ese servicio. En [20] se describen los escenarios donde la reputación en los SCRS es injusta debido a la colaboración de varios usuarios, este trabajo también propone métodos para descubrir tales patrones. El grado de reputación de los usuarios es una manera de recompensar a alguien por su actividad y ayudar a otros miembros de la comunidad, lo cual aumenta su popularidad. Existen estudios sobre cómo los SCRS ayudan a construir reputaciones sostenibles o clasificar a un usuario en [21] y [22].

III. ESTUDIOS DE CASO

Diversos estudios han estado relacionados con algunas de las tecnologías tratadas en este documento, todos con un objetivo diferente al de este trabajo. Esta sección cubre algunos de los esfuerzos más notables desarrollados hasta ahora en cada uno de estos campos para abordar los objetivos propios de este trabajo. Dado que BD es relativamente una nueva tecnología, el plazo que se consideró para los proyectos que implican directamente a BD o que habla de este, fue para el período 2012-2016.

Guo et al. [23] desarrollaron un sistema que permite trabajar con un enorme conjunto de datos con opiniones de TripAdvisor, ellos le dan mayor prevalencia al uso de "dimensiones" que les permiten trabajar directamente con la satisfacción de los visitantes específicos. Liu et al. [24] también han estudiado el comportamiento de la satisfacción del usuario a través del análisis de los comentarios de TripAdvisor. En su trabajo usan BD para obtener información sobre estos datos para hoteles en China, haciendo una comparación entre turistas nacionales y extranjeros de hasta ocho idiomas diferentes que no sean el chino mandarín. Sus hallazgos conducen a conclusiones interesantes sobre lo que se encuentra detrás de estas críticas.

Auto considerado como el primer estudio "para analizar el rol de las emociones en las críticas en línea de clientes", Felbermayr y Nanopoulos [25] se ocupan de dos aspectos importantes en tales críticas. Primero, consideraron el papel de las emociones y cómo éstas pueden influir en una crítica, para varios productos. Segundo, investigaron en conjuntos de datos donde los críticos también son votados por otros críticos y cómo este factor constituye un factor cualitativo a tomar en cuenta. Los mercadólogos emplean este estudio para analizar el efecto de las críticas en línea.

Las críticas en línea también han servido como una manera de proteger a los niños del uso de juguetes peligrosos a través del estudio de expresiones en redes sociales. Winkler et al. [26] realizaron una serie de experimentos para: (i) caracterizar cómo el contenido de los comentarios expresa una preocupación directa de seguridad, (ii) identificar reseñas que no hablan directamente sobre cuestiones de seguridad, sino que hablan de otros riesgos potenciales y (iii) diferenciar aquellas que no hablan de seguridad en lo absoluto o expresan confianza en el producto.

Rahman et al. [27] demostraron cómo a través del uso de BDA y Aprendizaje Automático, se puede predecir el futuro consumo de energía y cuánto se deberá generar en próximos años. Un estudio no relacionado a BD que trata a fondo el problema de la identificación de comportamiento irregular en los SCRS, se encuentra en [28]. En cuanto a la integración de BD y PNL, Nesi et al. [29] presentan una arquitectura que permite la integración de BD, combinando las funciones Mapreduce e integrándolas con la aplicación GATE (una plataforma fuente NLP).

Con un enfoque diferente a BD, Agerri et al. [30] proponen un sistema en el que utilizan varias máquinas virtuales para ejecutar sus propios módulos de PLN: las herramientas IXA. Ellos emplean el máximo poder de cómputo que puedan obtener de las máquinas virtuales para dividir sus necesidades de procesamiento de texto, ejecutadas por cada máquina que cuenta con las herramientas IXA.

En investigación con BD, se emplean varios modelos, pero hay algunos tienen mayor relevancia que el resto, entre estos están: las máquinas de vectores de soporte (SVM) y la asignación latente de Dirichlet (LDA) constituyen un referente en esta área. Gross y Murthy [31] demuestran en su estudio que el uso de LDA fue mejor que SVM, aunque eso depende en la naturaleza del material fuente para los estudios de PLN.

Hasta ahora se ha hablado sobre el uso de BD para diversos estudios, pero no todos emplean esta técnica. Khanaferov et al. [32] desarrollaron un sistema para extraer datos de Twitter con el propósito de reunir datos sobre salud y obesidad. En su investigación utilizan el PLN para la minería de textos de los tweets en inglés. El método que utilizan para la capa de datos es "una base de datos relacional OLTP", donde se ejecutan las operaciones analíticas de minería de datos en línea.

Gudivada et al. [33] han hecho un estudio a fondo sobre la importancia y la necesidad de integrar más estudios que incluyan PLN y BD. Comienzan hablando de las tareas y aplicaciones de PLN para posteriormente presentar algunas fuentes de datos para la investigación de PLN. Respecto a estudios que hablen de BDA se tiene por ejemplo que, en [34], Marine-Roig y Anton presentan una metodología utilizada para extraer inteligencias comerciales en contenido generado por los usuarios, tales como blogs turísticos o reseñas de viajes, ellos destacan "la utilidad de BDA para dar soporte a destinos inteligentes".

En [35] Yin et al. llevaron a cabo tres estudios para analizar los efectos de las críticas en línea de varios usuarios, el tercero de estos estudios emplea Yahoo! Shopping para explorar "los efectos de las emociones discretas sobre la utilidad de la crítica en un entorno real".

Una interesante línea de investigación en el campo de las críticas en línea es el aspecto de la reputación de los revisores y la influencia que estos pueden tener entre otros usuarios. Zhu et al. [36] han explorado el dataset Yelp! para encontrar aspectos de relevancia en este sentido. Ellos encontraron que a pesar de que un crítico de alto prestigio pueda influir en los futuros clientes, la utilidad del comentario es algo que los clientes tienen en cuenta en el proceso de toma de decisiones. Un enfoque de la utilidad de las críticas en línea mediante el uso de redes neuronales, se encuentra en [37].

El trabajo desarrollado por Salehan y Kim [38] acerca de la utilidad de las críticas en línea, se basa en la minería de sentimientos, por lo tanto, centran su modelo basado en 5 variables cuya medición ha llevado, entre otras conclusiones, a que el sentimiento positivo en el título de la revisión atrae a más lectores y que "las revisiones más largas son más propensas a atraer lectores y ser percibidas como útiles". Esto demuestra que las nuevas críticas tienen pocas posibilidades de ser votadas más útiles cuando para el mismo producto o servicio, hay otras revisiones que han sido más votadas en el pasado. Ngo-Ye y Sinha [39] apoyan esta tesis a través de su investigación, donde desarrollaron un modelo que basado en las dimensiones RFM del revisor, ayudan a identificar rápidamente nuevas críticas que sean de utilidad

IV. DISCUSIÓN

En este apartado los autores compararon los trabajos citados en este documento que hayan evaluado algún modelo existente, propuesto uno nuevo o desarrollado un paquete tecnológico. La Tabla I concentra los indicadores más significativos para los propósitos de este estudio: Big Data, Procesamiento de Lenguaje Natural y Sistemas de clasificación de redes sociales, fueron el enfoque clave durante la investigación. La última columna indica el modelo, teoría o sistema que da soporte a cada investigación.

TABLA I
COMPARACIÓN DE LOS MÉTODOS ESTUDIADOS

A	BD	BDA	BDP	PLN	CGU	OCU	PMU
[3]	✗	✗	✗	✓	✗	✗	SVM, HMM, CRF, NG
[4]	✓	✓	✓	✓	✓	✓	SVM, CA
[6]	✗	✓	✗	✗	✗	✗	CC
[9]	✗	✓	✗	✗	✗	✗	SCA
[11]	✓	✓	✗	✗	✓	✗	B-DAD
[15]	✗	✗	✓	✗	✗	✗	H/M
[17]	✗	✗	✗	✓	✗	✗	FS
[19]	✗	✗	✗	✓	✗	✗	DSF, RNN
[20]	✗	✗	✗	✗	✓	✓	CD
[22]	✗	✗	✗	✗	✗	✓	IRUA
[23]	✓	✗	✗	✗	✓	✓	LDA
[24]	✓	✗	✓	✗	✓	✓	AT
[25]	✓	✗	✗	✗	✓	✓	NRC, GALC
[26]	✗	✗	✗	✗	✓	✓	TM, SA
[27]	✓	✗	✓	✗	✗	✗	ML, ANN
[28]	✗	✗	✗	✗	✗	✓	CFRS
[29]	✓	✗	✗	✓	✗	✗	H/M
[30]	✓	✗	✗	✓	✗	✗	MTPC
[31]	✓	✗	✗	✓	✓	✗	LDA
[33]	✓	✗	✗	✓	✓	✗	BPP
[34]	✓	✓	✗	✗	✓	✓	LIWC
[35]	✗	✗	✗	✗	✓	✓	LIWC
[36]	✗	✗	✗	✗	✓	✓	ELM
[37]	✗	✗	✗	✗	✗	✓	HPNN
[38]	✓	✗	✗	✗	✓	✓	SM
[39]	✗	✗	✗	✗	✓	✓	RFM, BOW, TRM

Explicación de abreviaturas:
A: Autores
ANN: Red neuronal artificial
AT: Prueba de Anova
BD: Relacionado a/con Big Data
BDA: Big Data Analytics
BDP: Big Data Predictive
BOW: Bolsa de palabras
BPP: Paradigma del proceso por lotes
CA: Análisis Conjunto
CC: Cómputo en la nube
CD: Detección de colusión
CFRS: Sistemas de recomendación de filtrado colaborativo
CGU: Contenido generado por el usuario
CRF: Campos aleatorios condicionales
DSF: Modelo de frase discreta
ELM: Modelo de probabilidad de elaboración
FS: Tecnología de estados finitos
GALC: Codificador de etiqueta del afecto de Ginebra
H / M: Hadoop y algoritmos MapReduce
HMM: Modelos de Markov ocultos
HPNN: Modelo de predicción útil utilizando una red neuronal
IRUA: Algoritmo iterativo para la clasificación de la calidad del objeto y la reputación del usuario
LDA: Asignación latente de Dirichlet
LIWC: Petición lingüística y recuento de palabras
ML: Aprendizaje máquina
MTPC: Computación Paralela Multi-Hilos
NG: N-Grams
PLN: Procesamiento de lenguaje natural
PMU: Principales Métodos Utilizados
NRC: Consejo Nacional de Investigación de Canadá
OCU: Relativo a opiniones o calificaciones del usuario
RFM: Dimensiones de frescura, frecuencia y valor monetario
RNN: Red Neural Recursiva
SA: Análisis de sentimientos
SCA: Análisis de la Cadena de Suministros
SM: Minería de sentimientos
SVM: Máquina de Vector de Soporte
TM: Minería de textos
TRM: Modelos de regresión textual

Los resultados llevan a hallazgos interesantes: primero se destaca que varios de los trabajos utilizados en la primera sección de este artículo, hablan en su mayoría sobre el paradigma BD y las mejoras que ofrecen al sector comercial. En el campo del PLN, existen más documentos técnicos y mayor consenso entre los autores ya que, como se mencionó, se trata de un área de la Informática con seis décadas de existencia. En la era del BD, se puede afirmar que la investigación conjunta de PLN y las Ciencias Médicas tiene un largo camino por recorrer en los próximos años para el estudio de millones de expedientes médicos, lo cual llevará a beneficios significantes para los pacientes de diversas áreas.

Dado que las áreas de estudio son demasiado grandes para poderlas conjuntar en una sola investigación, pocos han sido los trabajos publicados que logren abarcar todas las áreas científicas tratadas en este documento. Como se puede ver en la Tabla I, sólo el trabajo desarrollado en [4] puede considerarse como una investigación que tiene todas estas áreas: BD, PLN y SCRS. Esto es debido en parte a que BD es relativamente una nueva tecnología y el plazo que se consideró para los proyectos que implican directamente a BD o que habla de este, fue para el período 2012-2016.

El área de aplicación de este tipo de investigaciones puede ser aplicada favorablemente en áreas comerciales que emplean BD, pero también podrían ser utilizadas en otros ámbitos como: política, ciencias sociales, turismo, salud, detección de fraude, etc.

Otro aspecto a destacar es la investigación desarrollada hasta ahora en BDP. Algunos autores han escrito sobre la prospectiva para esta área, pero además de [4] sólo [24] y [27] han hecho la investigación de un estudio de caso real con resultados prácticos. De los tres trabajos mencionados, sólo [4] y [24] trabajan con datos de medios sociales, siendo el CGU un área con alto potencial para la investigación, se puede afirmar que existe aún mucho trabajo por hacer.

V. CONCLUSIONES Y TRABAJO FUTURO

Después de revisar los resultados de la Tabla I, se tiene constancia que hay pocos trabajos realizados en proyectos que combinen todas las áreas en que este documento se enfoca para el periodo analizado (2012-2016). Sin embargo, dado que BD es una nueva línea de investigación, la publicación de nuevos estudios en el futuro cercano es una realidad. No obstante, BD sigue siendo considerado por algunos expertos de la industria como una moda temporal pero, en cuanto a las publicaciones revisadas, BD abre un nuevo horizonte a los académicos y profesionales para los desarrollos tecnológicos del mañana.

Este documento habla principalmente del trabajo con datos procedentes de fuentes de medios sociales, pero eso es sólo una parte de todo lo que se entiende como BD. Los resultados de [27] demuestran que otros tipos de datos pueden tratarse eficazmente en entornos analítico-predictivos de BD. Dado que este tipo de estudios no implican el conocimiento del PLN, se aconseja una sólida base en estadística y ML para desarrollar sistemas inteligentes de pronóstico.

Muchos de los autores en los artículos presentados han involucrado diversas técnicas de TM para la recolección de datos y luego realizar sus experimentos. TM es un área de importancia para la informática, y por ello las grandes empresas de redes sociales como Facebook o Tripadvisor podrían publicar conjuntos de datos abiertos para fines académicos, al igual que Yelp lo ha estado haciendo durante los últimos años. Encriptar y cifrar la identidad de los usuarios pueden ser algunas soluciones para abrir sus datos, ya que aún es difícil llevar a cabo una investigación académica basada únicamente en TM y viejos datasets publicados en sitios como Amazon S3.

Con las nuevas tecnologías de generación de datos como el Internet de las Cosas y la telefonía móvil 5G, las redes sociales no serán la principal fuente de BD, y fenómenos sociales como el consumo de energía, las nuevas enfermedades o el humor social podrían ser mejor analizados y predichos con la conjunción de estas tecnologías. Finalmente se recomienda que, para llevar a cabo investigación en esta área, los integrantes del equipo cuenten con perfiles que cubran áreas tanto computacionales como matemáticas.

REFERENCIAS

[1] A. De Mauro, M. Greco, M. Grimaldi, "What is Big Data? A Consensual Definition and a Review of Key Research Topics", In: AIP Conference Proceedings, pp. 97-104, 2015.

[2] H. Özköse, E. Sertac, C. Gencer, "Yesterday, Today and Tomorrow of Big Data", Procedia-Social and Behavioral Sciences, vol. 195, pp. 1042–1050, 2015.

[3] P. Nadkarni, L. Ohno-Machado, W. Chapman, "Natural language processing: an introduction", Journal of the American Medical Informatics Association, vol. 18, pp. 544-551, 2011.

[4] J. Qi, Z. Zhang, S. Jeon, Y. Zhou, "Mining Customer Requirements from Online Reviews: A Product Improvement Perspective", Information and Management, vol. 53, no. 8, pp. 951-963, 2016.

[5] S. Fosso, S. Akter, A. Edwards, G. Chopin, D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study", International Journal of Production Economics, vol. 165, pp. 234–246, 2015.

[6] J. Belaud, S. Negny, F. Dupros, D. Michéa, B. Vautrin, "Collaborative simulation and scientific big data analysis: Illustration for sustainability in natural hazards management and chemical process engineering", Computers in Industry, vol. 65, pp. 521–535, 2014.

[7] R. Iqbal, F. Doctor, B. More, S. Mahmud, U. Yousuf, "Big Data analytics: Computational intelligence techniques and application áreas", International Journal of Information Management, pp. 1-16, 2016.

[8] W. Tan, B. Blake, I. Saleh, S. Dustdar, "Social-Network-Sourced Big Data Analytics", IEEE Computer Society, 2013.

[9] G. Wang, A. Gunasekaran, E. Ngai, T. Papadopoulos, "Big data analytics in logistics and supply chain management: Certain investigations for research and applications", International Journal of Production Economics, vol. 176, pp. 98–110, 2016.

[10] S. Erevelles, N. Fukawa, L. Swayne, "Big Data consumer analytics and the transformation of marketing", Journal of Business Research, vol. 69, pp. 897–904, 2016.

[11] N. Elgendy, A. Elragal, "Big Data Analytics In Support of the Decision Making Process", Procedia Computer Science, vol. 100, pp. 1071–1084, 2016.

[12] L. Baruh, M. Popescu, "Big data analytics and the limits of privacy self-management", United Kingdom: SAGE, 2015.

[13] R. Perrons, D. McAuley, "The case for 'n«all': Why the Big Data revolution will probably happen differently in the mining sector", Resources Policy, vol. 46, pp. 234-238, 2015.

[14] N. Shah, Z. Irani, A. Sharif, "Big data in an HR context: Exploring organizational change readiness, employee attitudes and behaviors", Journal of Business Research, vol. 70, pp. 366-378, 2017.

[15] N. Saravana, T. Eswari, P. Sampath, S. Lavanya, "Predictive Methodology for Diabetic Data Analysis in Big Data", Procedia Computer Science, vol. 50, pp. 203–208, 2015.

[16] X. Yue, G. Di, Y. Yu, W. Wang, H. Shi, "Analysis of the Combination of Natural Language Processing and Search Engine Technology", Procedia Engineering, vol. 29, pp. 1636–1639, 2012.

[17] A. Maletti, "Survey: Finite-state technology in natural language processing", Theoretical Computer Science, pp. 1-16, 2016.

[18] D. Calì, A. Condorelli, S. Papa, M. Rata, L. Zagarella, "Improving intelligence through use of Natural Language Processing. A comparison between NLP interfaces and traditional visual GIS interfaces", Procedia Computer Science, vol. 5, pp. 920–925, 2011.

[19] M. Tanana, K. Hallgren, Z. Imel, D. Atkins, V. Srikumar, "A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing", Journal of Substance Abuse Treatment, vol. 65, pp. 43–50, 2016.

[20] M. Allahbakhsh, A. Ignjatovic, B. Benatallah, S. Beheshti, N. Foo, E. Bertino, "Representation and querying of unfair evaluations in social rating systems", Computers & Security, vol. 41, pp. 68-88, 2014.

[21] M. Ekmekci, "Sustainable reputations with rating systems", Journal of Economic Theory, vol. 146, pp. 479–503, 2011.

[22] X. Liu, Q. Guo, L. Hou, C. Cheng, J. Liu, "Ranking online quality and reputation via the user activity", Physica A, vol. 436, pp. 629–636, 2015.

[23] Y. Guo, S. Barnes, Q. Jia, "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation", Tourism Management, vol. 59, pp. 467–483, 2017.

[24] Y. Liu, T. Teichert, M. Rossi, H. Li, F. Hu, "Big data for big insights: Investigating language-specific drivers of hotel satisfaction with 412,784 user-generated reviews", Tourism Management, vol. 59, pp. 554–563, 2017.

[25] A. Felbermayr, A. Nanopoulos, "The Role of Emotions for the Perceived Usefulness in Online Customer Reviews", Journal of Interactive Marketing, vol. 36, pp. 60–76, 2016.

[26] M. Winkler, A. Abrahams, R. Gruss, J. Ehsani, "Toy safety surveillance from online reviews. Decision Support Systems", vol. 90, pp. 23–32, 2016.

[27] M. Rahman, A. Esmailpour, J. Zhao, "Machine Learning with Big Data An Efficient Electricity Generation Forecasting System", Big Data Research, vol. 5, pp. 9–15, 2016.

[28] Z. Yang, Z. Cai, X. Guan, "Estimating user behavior toward detecting anomalous ratings in rating systems", Knowledge-Based Systems, vol. 111, pp. 144–158, 2016.

[29] P. Nesi, G. Pantaleo, G. Sanesi, "A Hadoop based platform for natural language processing of web pages and documents", Journal of Visual Languages and Computing, vol. 31, pp. 130–138, 2015.

[30] R. Agerri, X. Artola, Z. Beloki, G. Rigau, A. Soroa, "Big data for Natural Language Processing: A streaming approach", Knowledge-Based Systems, vol. 79, pp. 36–42, 2015.

[31] A. Gross, D. Murthy, "Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing", Neural Networks, vol. 58, pp. 38–49, 2014.

[32] D. Khanaferov, C. Luc, T. Wang, "Social Network Data Mining Using Natural Language Processing and Density Based Clustering", IEEE International Conference on Semantic Computing, pp. 250-251, 2014.

[33] V. Gudivada, D. Rao, V. Raghavan, "Big Data Driven Natural Language Processing Research and Applications", Handbook of Statistics, vol. 33, pp. 203-238, 2015.

[34] E. Marine-Roig, S. Anton, "Tourism analytics with massive user-generated content: A case study of Barcelona", Journal of Destination Marketing & Management, vol. 4, pp. 162–172, 2015.

[35] D. Yin, S. Bond, H. Zhang, "Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews", MIS Quarterly, vol. 38, no. 2, pp. 539-560, 2014.

[36] L. Zhu, G. Yin, W. He, "Is this opinion leader's review useful? Peripheral cues for online review helpfulness", Journal of Electronic Commerce Research, vol. 15, no. 4, pp. 267-280, 2014.

[37] S. Lee, J. Choeh, "Predicting the helpfulness of online reviews using multilayer perceptron neural networks", Expert Systems with Applications, vol. 41, pp. 3041–3046, 2014.

[38] M. Salehan, D. Kim, "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics", Decision Support Systems, vol. 81, pp. 30-40, 2016.

[39] T. Ngo-Ye, A. Sinha, "The influence of reviewer engagement characteristics on online review helpfulness: A text regression model", Decision Support Systems, vol. 61, pp. 47-58, 2014.

[40] G. Paré, M. Trudel, M. Jaana, S. Kitsiou, "Synthesizing information systems knowledge: A typology of literature reviews", Information & Management, vol. 52, pp. 183-199, 2015.

José Luis Jiménez Márquez es estudiante de Doctorado en el departamento de Informática en la Universidad Carlos III de Madrid. Obtuvo el Grado de Maestro en Ciencias Computacionales y el de Licenciado en Informática, ambos por el Instituto Tecnológico de Orizaba, México. Email: 100339395@alumnos.uc3m.es.

Israel González Carrasco es un Profesor Visitante en el departamento de Informática en la Universidad Carlos III de Madrid. Es Ingeniero Informático y Doctor en Ciencia y Tecnología en Informática por la Universidad Carlos III de Madrid. Ha publicado diversos artículos en publicaciones internacionales. Ha estado involucrado en varios proyectos internacionales y es miembro de la junta de revisión de varias revistas internacionales. Email: igcarras@inf.uc3m.es.

José Luis López Cuadrado es un Profesor Visitante en el departamento de Informática en la Universidad Carlos III de Madrid. Es Ingeniero Informático y Doctor en Ciencia y Tecnología en Informática por la Universidad Carlos III de Madrid. Su investigación se centra en la web, las redes neuronales artificiales, la mejora de procesos y la ingeniería de software. También es coautor de varios artículos en revistas de impacto internacional, conferencias nacionales e internacionales. Email: jllopez@inf.uc3m.es.