fceye Facultad CIENCIAS ECONÓMICAS Y EMPRESARIALES

IDENTIFICACIÓN DE PRODUCTOS DE BÚSQUEDA Y EXPERIENCIA EN LONG TAIL A PARTIR DE OPINIONES EN LÍNEA.

Caso Ciao, UK.

Trabajo Fin de Máster

En cumplimiento de los requisitos del programa
de Máster en Gestión Estratégica y Negocios Internacionales

Presentado por:

Venus A. Martínez Fortuna

Supervisor Académico: Prof. Dra. Mª del Rocío Martínez Torres
Línea de Investigación: Sistemas de información y gestión del conocimiento
Departamento: Administración de Empresas e Investigación de Mercados (Mercadeo)
Facultad: Ciencias Económicas y Empresariales (FCEYE), Universidad de Sevilla

Identificación de productos de búsqueda y experiencia en Long Tail a partir de opiniones en línea. Caso Ciao, UK.

Autor: Venus A. Martínez Fortuna
Supervisor Académico: Prof. Dr. Mª del Rocío Martínez Torres

Programa de Máster: Gestión Estratégica y Negocios Internacionales
Línea de Investigación: Sistemas de información y gestión del conocimiento
Departamento: Administración de Empresas e Investigación de Mercados (Marketing)
Facultad: Ciencias Económicas y Empresariales (FCEYE), Universidad de Sevilla

Universidad de Sevilla
Avda. Ramón y Cajal N.º 1, 41018 Sevilla

En Sevilla, a 2 de noviembre del 2018

X
Fdo.: Mª del Rocío Martínez Torres

X
Fdo.: Venus A. Martínez Fortuna

DEDICATORIA Y AGRADECIMIENTO

A mi madre Maira Fortuna por enseñarme que no existen límites más allá que los propios y a ser perseverante para cumplir mis metas. A mis tíos Peter Boersen, Manuela Fortuna de Boersen y a mis padres Tomás Martínez y Clara Guillén, por la confianza depositada en mí, su motivación constante y el apoyo incondicional.

Tabla de contenidos

INTRODUCCIÓN

"La economía exige a las empresas identificar, impactar, ofrecer mercancía, vendiéndola y reteniendo a los clientes alrededor del mundo." Josep-Francesc Valls (2017)

La Nueva Economía[1], es moldeada por la globalización y el comercio electrónico. Hoy en día, la integración efectiva en las tecnologías de la comunicación y la gestión de la información ha jugado un rol importante en el desarrollo de la estrategia empresarial y el aumento del comercio internacional (Aydin & Kilinç, 2014). Debemos entender cómo la globalización, las nuevas formas de comunicación, la información generada en las redes sociales y el comercio electrónico se interrelacionan, integrando a los usuarios alrededor del mundo en procesos internos de creación de valor (Kleemann, Voss & Rieder, 2008).

El comercio electrónico, conocido como E-commerce, se basa en cualquier transacción de compra o venta de productos o servicios en sistemas electrónicos (Vikram, 2012). Este sector industrial está creciendo sin precedentes gracias a los avances tecnológicos y el desarrollo de la sociedad de la información que facilita la creación, distribución y manipulación de los datos. Este tipo de evolución en los mercados permite ofrecer una mayor cantidad de opciones y disponibilidad, reduciendo barreras de entrada a los consumidores en diferentes partes del mundo.

En este sentido, los negocios tradicionales exigían limitar los estantes a aquellos productos que se consideran exitosos, debido al coste de oferta que demandan. Actualmente aquellas empresas presente en el entorno on-line experimentan la posibilidad de almacenar prácticamente todo. Por otro lado, el desarrollo del comercio electrónico ha permitido al consumidor empoderarse de la decisión de compra. Josep-Frances Valls (2017) explica la evolución del usuario en la participación del proceso de pre-compra y post-compra, donde puede comparar precios, opciones y marcas y compartir su reflexión sobre el producto o servicio recibido.

Phillip Nelson (1970) fue el primer autor en plantear las limitaciones de obtener información sobre la calidad de los productos y cómo afectaba el comportamiento en el mercado. A partir de este artículo, se plantea la clasificación de los productos dependiendo del momento en el que el consumidor se puede hacer una idea de la calidad de los productos en *bienes de búsqueda o bienes de experiencia*.

[1] Término adoptado por Don Tapscott (1995) en su libro. *The digital Economy: Promise and Peril in the Age of Networked Intelligence*.

A partir del concepto dado, se definen los productos de búsqueda, aquellos cuya calidad puede ser objetivamente evaluada por el consumidor antes de la compra. En el caso de los productos de experiencia, sólo después de probarlos el consumidor puede formarse una opinión sobre la calidad. Podemos concluir que el consumidor puede considerar factores que hacen al producto superior sobre otra oferta en el mercado previo a la decisión de compra en productos de búsqueda, mientras que en los productos de experiencia, sólo es posible posterior a su uso.

Para la disertación de este Trabajo Fin de Máster , utilizaremos como recurso las opiniones de los usuarios disponibles en la base de datos CIAO UK en la categoría *fashion*, con la intención de identificar si este nicho de mercado corresponde a productos de búsqueda, productos de experiencia o a ambos.

Justificación

Millones de usuarios globalmente utilizan Internet como un medio de desarrollo de las rutinas personales y empresariales básicas para el intercambio de datos, información, opiniones e intereses y como recurso operacional para satisfacer sus necesidades en la adquisición de bienes y servicios que se ofrecen y se distribuyen a través de páginas específicas de empresas facilitadoras.

En el contexto electrónico, comunicar efectivamente la información del producto es uno de los grandes retos que tienen los comercios en línea. Existe una gran cantidad de datos en la World Wide Web y por esta razón, el comprender e identificar patrones de los resultados en las búsquedas y acciones de las personas en Internet, persigue ofrecer aquella información que es relevante para el usuario y que le facilite la identificación de productos o servicios menos conocidos, influyendo en la confianza del consumidor y en la decisión de compra.

OBJETIVOS

Llegados a este punto, lo siguiente a realizar sería identificar los objetivos que guiarán este Trabajo Fin de Máster:

Objetivo General

1. Diseñar un clasificador que permita distinguir productos de experiencia y productos de búsqueda en base a las opiniones de los usuarios

Objetivos Específicos

1. Clasificar productos a partir de los comentarios en la plataforma Ciao UK

2. Analizar los productos basados en experiencia y búsqueda

Objetivo de Impacto

1. Contribuir a la comprensión de las empresas y clientes potenciales o actuales de los productos dentro de la Long Tail

Estructura del Trabajo Fin de Máster

El siguiente Trabajo Fin de Máster está organizado según el detalle explicado a continuación. En primer lugar, se introduce el marco teórico haciendo una revisión de la literatura que fundamentan los objetivos y la justificación. En consecuencia, se discute sobre el desarrollo e impacto de la globalización, el comercio electrónico y el Big Data. Posteriormente se introducen las comunidades eWOM, el contenido generado por el usuario en línea (UCG) y el uso de la minería de datos como metodología general para obtener información relevante.

Luego de explicar los antecedentes y los conceptos involucrados en este proceso de investigación, se desarrolla en detalle la metodología de investigación aplicada para cumplir con el propósito expresado en los objetivos y se desarrolla el análisis de los datos. Finalmente, se concluye el Trabajo Fin de Máster discutiendo las implicaciones y limitaciones del estudio y los planes para futuras investigaciones.

MARCO TEÓRICO

Los avances en las Tecnologías de la Información y Comunicación (ICT por sus siglas en inglés) de la mano con la globalización, han influido en la vida y el comportamiento de las personas y han abierto las posibilidades a la aplicación en diferentes áreas de la industria. Esta evolución ha transformado las formas tradicionales del marketing. Gil-Pechuán, Palacios-Marqués, Peris-Ortiz, Vendrell, & Ferri-Ramírez (2014) afirma que las compañías utilizan el Word Wide Web como base para evaluar su habilidad de competir. Actualmente, existen plataformas denominadas "Online Social Networks" o, OSNs (Redes Sociales en Línea), páginas web que permiten la interacción con otras personas en diferentes partes del mundo y facilita a los usuarios ampliar su círculo social y compartir información y experiencias personales. Las compañías han identificado estos entornos como una fuente de estrategias claves para las firmas.

El desarrollo de Internet ha impulsado el avance de la globalización definido como *el proceso cultural, económico y de información en el cual, a través de importantes avances, logran que las fronteras de los países sean menos evidentes y las relaciones entre las personas del mundo sean más cercanos*[2]. El auge de la denominada economía colaborativa, con la que millones de personas intercambian productos y servicios mediante la red, creando nuevas formas de entender la propiedad que mueven ya miles de millones de dólares en todo el mundo[3].

La red ha propiciado un gran cambio económico, revolucionando la forma de comunicarse y de entender la sociedad. Este hecho, junto a los avances en las tecnologías de la información arrastra como consecuencia, la importancia de la presencia en línea y la facilidad de encontrar un producto o servicio sin tener que recorrer largas distancias. El informe anual publicado en 2014 por SelfBank concluye los avances en las ICT facilitan a las empresas la segmentación y la elaboración de ofertas personalizadas. Por tanto, la información no es sólo una virtud de los comercios electrónicos, es un factor de éxito dentro de un modelo económico, bien conocido como E-commerce.

Las economías de escalas facilitaron las ventas a grandes cantidades de consumidores, el desarrollo de Internet repercutió de forma positiva a la disminución de costos y aumento de ventas de amplio alcance y con mejor empatía con los clientes meta. Maria Olmedilla (2016) presenta dos enunciados relevantes en el estudio de la información masiva generada por los

[2] Myro, R. (14 de julio de 2001). Globalización y crecimiento económico. *El país.* Recuperado de: https://elpais.com/diario/2001/07/14/opinion/995061608_850215.html

[3] EFECOM (16 de junio de 2015). Internet, la herramienta que contribuye al auge de la economía colaborativa. La Vanguardia. Recuperado de: https://www.lavanguardia.com/economia/20150516/54431675862/Internet-la-herramienta-que-contribuye-al-auge-de-la-economia-colaborativa.html

usuarios en línea. El primero relacionado con el diseño de un *web crawler* efectivo que permita buscar e identificar el contenido en la web generada por los usuarios y, en segundo lugar, la etapa de depuración, almacenamiento y mantenimiento de la información relevante identificada tal y como se muestra en la figura 1:

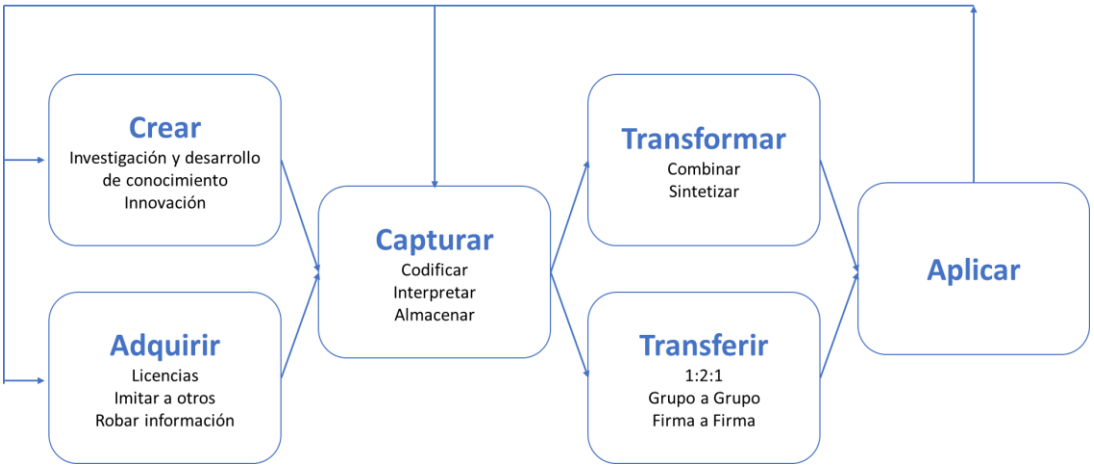

Figura 1- Ciclo de gestión del conocimiento. Fuente: Wilson, P. & Campbell, L. (2016)

Las Tecnologías de la Información y Comunicación (ICT) han transformado los métodos del marketing tradicional. Las empresas han utilizado el World Wide Web como una ventaja competitiva en un mundo globalizado. (Peris, 2014) concluye que es este nuevo entorno de competencia, la popularidad de los Online Social Networks (OSNs) han tenido un impacto rotundo y han influido en la vida de las personas. De este modo, las OSNs han crecido exponencialmente como un medio social donde grandes masas de usuarios interactúan entre ellos en su búsqueda de información para viajes, productos, organización de sus círculos sociales, etc. (Gil et al, 2014). De este modo, las personas no sólo buscan información, sino que también juegan un rol de hacer visibles su información a millones de personas.

En el año 2004, Chris Anderson introdujo el término Long Tail (cola larga) en Wired Magazine bajo el concepto de productos menos populares con menor demanda. El autor explica que el comercio electrónico ha facilitado encontrar los productos *desconocidos u olvidados* gracias a la facilidad de exhibir los productos sin costes de espacio físico, información disponible, los algoritmos de recomendación y los comentarios positivos de otros usuarios que han compartido su experiencia.

Bing Pan & Xiang Li (2006) explica que la Long Tail se forma de la distribución de venta cuando una gran cantidad de productos especializados se venden muy poco individualmente, en

comparación con pequeña cantidad de los productos más populares genera una gran cantidad de ingresos. En resumen, la teoría de la Long Tail explica que, cuando los usuarios tienen a ampliar sus opciones, suelen dirigir su atención hacia los nichos porque satisfacen mejor los intereses más específicos y acorde con nuestros límites.

En resumen, las empresas presentes en el entorno online (iTunes, Amazon, etc.) experimentan la posibilidad de almacenar prácticamente todo, superando la disponibilidad de productos de nichos -o productos especializados- en comparación con los productos de éxito. Chris Anderson presenta el caso de la venta de un libro olvidado, en consecuencia, la recomendación de los usuarios y los algoritmos utilizados por Amazon que se tradujeron en ventas. A partir del caso presentado por Anderson, si nos remontamos a años atrás, antes del desarrollo de la ICT, podemos identificar dos variables de esta situación donde, primero las personas nunca hubieran conocido acerca del libro o, si conocieran su existencia, debido a la cantidad limitada de copias no encontrasen un ejemplar. Este fenómeno surge a partir de la proliferación de canales de comunicación en Internet que eliminan las barreras de comunicación entre ubicaciones geográficamente distantes (como se citó en Edosio, 2014).

Las compañías de comercio electrónico, los motores de búsqueda y la tecnología aplicada a la búsqueda de información, permiten a los clientes potenciales o actuales encontrar aquellos productos que van dirigidos a nicho de mercados más específicos, o productos menos conocidos para los usuarios.

En el artículo *Big Data: la revolución económica de la información* recuperado de Forbes México 2016 se señala que para el 2015 el total de información generada en Internet alcanzó 8 trillones de gigas. También señala que para final de esta década se predice que el volumen de información se habrá multiplicado 40 veces. Paola Palma (2016) defiende que el Big Data ha transformado sectores como la venta minorista, que al manipular grandes volúmenes de información permiten conocer los hábitos de consumo de las personas en tiempo real.

No es un secreto que el uso del Big Data está transformando la economía. La Organización para la Cooperación y el Desarrollo Económicos (OCDE) calcula que el valor estimado de este mercado alcanzó 17,000 millones de dólares en 2015, con un crecimiento promedio anual de 40% desde 2010[4]. Dan Ariely y Harper Collins (2010) exponen que existe una irracionalidad predecible en la toma de decisiones que como consumidores podemos controlar y contrarrestar y como vendedores poder anticipar.

[4] Palma, P. (15 de marzo de 2016). Big data: la revolución económica de la información. *Forbes.* Recuperado de: https://www.forbes.com.mx/big-data-la-revolucion-economica-la-informacion/

La aplicación del Big Data y la inteligencia artificial permite comprender el comportamiento, optimizar y generar valor en el negocio del retail, utilizando datos de una población y reduciendo al mínimo errores en la muestra. Aplicando bien el sistema de gestión de datos podemos responder a preguntas como: ¿Quién compra?, ¿Cuándo compra?, ¿Qué productos compra?, ¿En qué momentos compra?, ¿Por qué compra unos productos y no otros?, ¿Cuánto está dispuesto a pagar?, entre otras.

Como consecuencia del auge Internet, las compañías han ganado ventajas gracias a la eliminación de costes para llegar a mantener la relación con los clientes, entrar a nuevos mercados y promocionarse. Ignacio Gil-Pechuán et al. (2014) explica cómo se utilizan las plataformas OSNs, que usadas correctamente pueden atraer clientes. En el artículo *Big Data: Revolución económica de la información* escrito por Paola Palma (2016), la autora sostiene la importancia de los datos en la red, aclarando que grandes empresas transforman la información en conocimiento para tomar decisiones más acertadas como *el desarrollo de productos o mejorar procesos de negocios.* Por consiguiente, la interacción de los usuarios en Internet genera información que puede ser transformada en conocimiento y, este de igual forma puede afectar el proceso empresarial, el cual, identificado, filtrado e implementado de forma correcta, puede traer ventajas a las empresas.

Globalización e E-commerce

"El proceso de globalización es seguido por la disminución de las barreras administrativas al comercio […]." - Erdal Aydin, 2014

Globalización no es un concepto nuevo. Es mencionado una y otra vez por diferentes autores y lo relacionamos junto a la incorporación de la economía mundial. Si bien existen múltiples definiciones para la globalización, C. Passaris (2006) la define como la integración global de las economías a través del comercio y los flujos de inversión, incluyendo en su definición la producción de bienes y servicios para mejorar la competitividad internacional.

La globalización ha sido uno de los fenómenos más destacados del siglo XX que dio forma dramática a la economía mundial. El proceso de globalización no sólo integra a los países a nivel económico, también resulta en la solvencia del conocimiento, trabajo, capital y bienes. En la nueva economía global del siglo XXI, se ha transformado el panorama económico, social, educativo y político de una manera profunda e indeleble.

La disminución de las barreras administrativas al comercio, las grandes reducciones en los costos de transporte y comunicación, la fragmentación de los procesos de producción y el desarrollo de la tecnología de la información y la comunicación son consecuencias del proceso de globalización, permitiendo así, oportunidades de inversión en nuevos mercados y el acceso a nuevas materias primas y recursos (Aydin & Kilinç, 2014). Con la globalización es las organizaciones buscan una posición competitiva superior con menores costos operativos, para obtener una mayor cantidad de productos, servicios y consumidores a través de la diversificación de recursos, la creación y el desarrollo de nuevas oportunidades (de inversión) en nuevos y al acceder a nuevas materias primas y recursos.

Una amplia revisión de las bibliografías sostiene cómo el desarrollo del transporte y la ciencia de las telecomunicaciones han impulsado el proceso de globalización. J. Ibañéz (2006), en defensa del desarrollo de las telecomunicaciones, expone la idea de cómo Internet ha influido socialmente en el S. XXI. Ibañéz sustenta que, con la llegada de Internet, el conjunto interrelacionados de procesos, resultado de la transformación "global", experimentaron una expansión geográfica e intensificación en ámbitos sociales, culturales, económicos, etc. derivada del desarrollo de la velocidad de la comunicación.

Internet además de facilitar la comunicación, ha posibilitado la creación de una nueva forma de acceder a nuevos mercados a través del comercio electrónico. La adopción de Internet hace que sea más barato y más fácil para las empresas ampliar sus mercados, administrar sus operaciones

y coordinar las cadenas de valor a través de las fronteras. La reducción de los costos de las transacciones y la información, la tecnología ha reducido las fricciones del mercado y ha dado un impulso significativo al proceso de ampliación de los mercados mundiales. La adopción de las ICT fomenta la globalización al reducir el costo de las transacciones y la coordinación y al crear mercados nuevos y expandidos con economías de escala (como citado en Aydın & Kılınç, 2011).

La literatura nos muestra cómo, la Tecnología de la Información y la Comunicación (ICT) se ha convertido en uno de los elementos fundamentales de la sociedad moderna. Junto al desarrollo de las ICT, surgen los sitios de Redes Sociales en línea que permiten a los usuarios crear comunidades en la red, también conocidos como OSNs. En la última década, las OSNs se han convertido en un medio de comunicación de moda entre los usuarios de Internet (Khedo, Roushdat, Mocktoolah & Suntoo, 2012).

El crecimiento económico y el desarrollo en la nueva economía global han estado precedidos por un complejo realineamiento estructural de las corrientes de inversión, el agrupamiento de empresas comerciales, la transformación del proceso de producción y la adopción de un enfoque de marketing de nicho (Porter, 1998). La innovación juega un papel de catalizador que impulsa el crecimiento económico y es fundamental para la economía global. De este modo, las Redes Sociales en Línea han introducido una nueva forma de hacer negocios, basada en los entornos colaborativos como canales de venta, en el cual la recomendación y la viralidad cumplen un rol fundamental para el éxito empresarial: el Comercio Social (Castelló, 2011). En un principio, el término Comercio Social abordaba el contenido generado por los usuarios para la recomendación de los productos en línea, en 2006, Rubel incluyó en la definición de Comercio Social aquellas herramientas colaborativas en el comercio electrónico que permiten a los compradores obtener consejos y recomendaciones de otros usuarios que consideran “de confianza”, para así encontrar y adquirir productos y servicios.

Las OSNs surgen a partir de la globalización, las consecuencias de transformación social y repercuten sobre el comercio electrónico, abriendo camino al comercio social. Actualmente el área del comercio social incluye las herramientas en medios sociales y contenidos generados en el ámbito del comercio electrónico, incluyendo las valoraciones de los productos, recomendaciones, aplicaciones y publicidad social.

> “El comercio electrónico es la necesidad de los negocios internacionales, y viceversa. Ya sea que se trate de transacciones minoristas de negocio a cliente o de negocio a negocio, es parte importante de las transacciones comerciales” – Azamat Nogoev et al. (2011)

Amith Vikram (2012) define el comercio electrónico, comúnmente conocido como E-commerce, como cualquier transacción de compra o venta de productos o servicios en sistemas electrónicos y otras redes computacionales. El comercio electrónico ofrece múltiples beneficios a los consumidores: disponibilidad de productos, mayor cantidad de opciones y ahorro de tiempo.

Muchas publicaciones previas, enfatizan la importancia de Internet en el comercio electrónico. "Las redes sociales y las comunidades en línea ofrecen la oportunidad de compartir experiencias, sentimientos e intercambiar información. Han transformado a consumidores, sociedades y compañías con acceso generalizado a la información, comunicación y redes sociales mejoradas." (Toral et al., 2017, pp.2). El autor J. F. Valls (2017) afirma que nos hallamos en un punto donde las personas recurren a las plataformas tecnológicas para sus actividades sociales y económicas. En su disertación sostiene que actualmente la digitalización, los Big Data, la analítica, los algoritmos, la automatización y la inteligencia artificial protagonizan la humanidad. El comercio se ha ampliado globalmente, eliminando barreras de entradas y ampliando la presencia de marcas en países sin presencia física, ampliando el, el universo de marcas de los consumidores.

Gracias al Internet, el comportamiento de los consumidores se configura para acceder a mejores ofertas y mejor calidad, a partir de la información disponible para satisfacer sus necesidades y aspiraciones (Valls, 2017). El desarrollo del comercio electrónico ha permitido al consumidor empoderarse de la decisión de compra. El usuario puede comparar precios, opciones y marcas y compartir su reflexión sobre el producto o servicio recibido.

El desarrollo de la informática y las ciencias de la comunicación ha sentado una base sólida para este tipo de comercio. Las transacciones E-commerce han crecido de manera exponencial desde principios de la segunda década del siglo XXI (Castelló, 2011), emergiendo como un fenómeno empresarial cada vez más importante. En la figura no. 2 se resume el comportamiento del consumidor y las aplicaciones de esta información para proveedores de comercio electrónico.

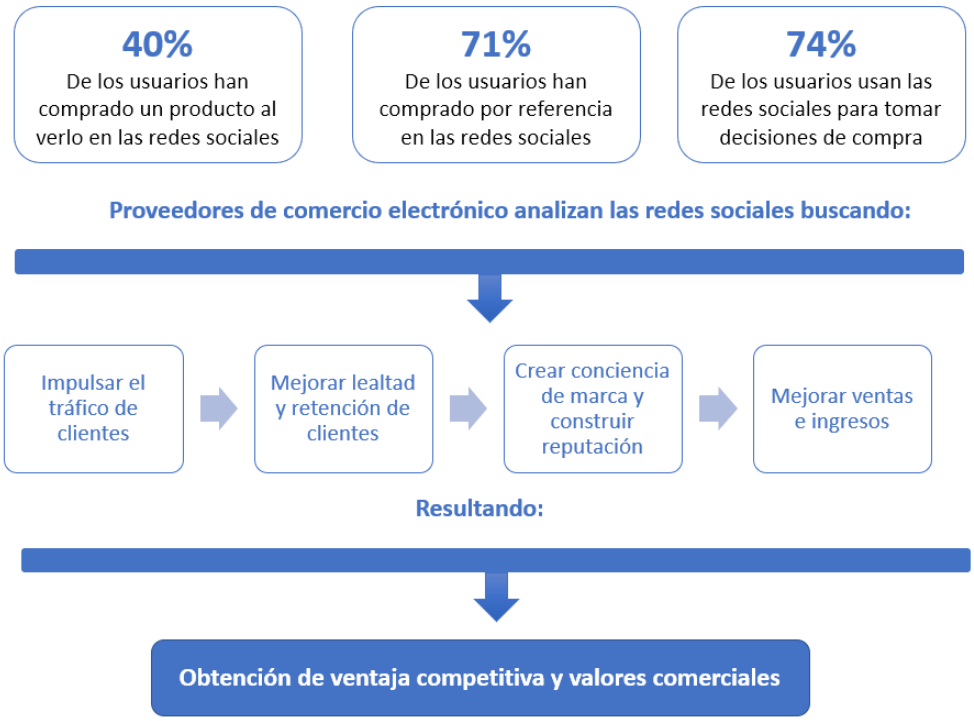

Figura 2- Uso de las OSN en E-commerce. Elaboración propia.

Big Data

"La recolección, almacenamiento, manejo y análisis en tiempo real de volúmenes gigantescos de información originó lo que se conoce como Big Data" Forbes, 2016

Tim Berners-Lee propone en 1991 las directrices para la creación y uso de una red interconectada a nivel mundial disponible en cualquier momento que termina convirtiéndose en el Word Wide Web. Sin embargo, la primera definición de término *Big Data* se originó para 1997 y fue presentada por dos investigadores de la NASA, Michael Cox y David Ellsworth (Lógica & Magdalena, 2015) a partir del incremento masivo en los volúmenes de datos, derivados de los servicios en internet.

Según la última edición del informe de *We Are Social y Hootsuite*, en el 2018 el número de usuarios en Internet alcanza los 4.000 millones especificando que, en España el 85% de la ciudadanía está conectada[5]. Las actividades que realizan los usuarios a través de Internet son la

[5] BBVA (2018). Los cuatro consejos definitivos para proteger tus datos en Internet. *La Vanguardia*. Recuperado de: https://www.lavanguardia.com/tecnologia/20180523/443666265462/cuatro-consejos-proteger-datos-Internet-brl.html

búsqueda de información, entretenimiento, comunicación, acceso a contenidos audiovisuales y uso de redes sociales.

En el reporte *Cisco Visual Networking Index 2017*, la empresa Cisco estima que para el 2020, habrá 11,600 millones de dispositivos conectados, incluyendo la comunicación máquina a máquina, superando a la población mundial actual de 7,800 millones de personas.

En el Word Wide Web existen dominios que sirven de almacén de información en bases de datos centralizadas para luego ser distribuidas. José Riquelme, Roberto Ruiz & Karina Gilbert (2006) describen la revolución digital y su impacto en la información digitalizada, logrando capturar, procesar, almacenar, distribuir, y transmitir con una facilidad sin precedentes. De este modo, el recolectar información se convierte en asunto sencillo, ahora el problema radica en la identificación de información relevante (Riquelme et al., 2006).

El papel de la tecnología de la información y las comunicaciones en la nueva economía ha sido fundamental (Passaris, 2006). Hoy en día se aplican metodologías de análisis inteligente de datos, que nos permite extraer valores de análisis predictivo y de comportamiento, que pueden utilizarse para detectar las tendencias del mercado. Sin embargo, el acceso a grandes volúmenes de información limita la capacidad humana de analizar los datos manualmente y crea la necesidad de automatizar el proceso, utilizando algoritmos sofisticados para filtrar la información útil (Riquelme et al., 2006).

Beye y Laney (2012) definen el Big Data como activos de información caracterizados por su volumen, velocidad y variedad, que requieren soluciones innovadoras y eficientes para su procesamiento para mejorar el conocimiento y la toma de decisiones en las organizaciones.

El manejo de información a través del Big Data permite a las empresas aprovechar los datos generados por los usuarios en línea[6]. Estos datos adquieren valor al ser procesados para identificar patrones, correlaciones e interacciones de manera que se conviertan en conocimiento para desarrollar mejores productos y servicios, mejorar procesos y tomar decisiones más acertadas (Krumm, Davies & Narayanaswami, 2008). Esta interacción ha transformado la economía al orientar la toma de decisiones empresariales enfocadas a lo que el cliente quiere. Por consiguiente, el proceso de extracción de conocimiento debe ser eficiente y cercano al tiempo real, porque almacenar todos los datos observados es casi inviable. En ese sentido, las técnicas computacionales avanzadas están explotando el potencial de la tecnología para capturar

[6] Recuperado del artículo *Big Data: la revolución económica de la información* (Forbes, 2016)

y analizar cantidades tan grandes de datos de Internet en formas cada vez más poderosas (Eynon 2013).

Al utilizar la instrumentación adecuada para la recopilación de datos, es posible aprovechar la información que proviene del contenido generado por el usuario, como clics, tweets, opiniones de los usuarios, ofertas de subasta, elecciones del consumidor o intercambios de redes sociales (Chang et al. 2014). De este modo, la evolución en los métodos asistidos por computadora está cambiando la forma en que se realiza la investigación de las Ciencias Sociales y el procesamiento de datos (Demchenko, et al. 2013). En la figura no. 3 podemos observar la representación de una base de datos a partir de los métodos computacionales y la participación de los usuarios en línea.

El Big Data trata de las herramientas de captura, búsqueda, descubrimiento y análisis que ayudan a obtener conocimiento de datos no estructurados. Esta tendencia brinda la posibilidad de hacer que muchos espacios sociales sean cuantificables, por lo que se pueden estudiar siguiendo un enfoque cuantitativo (Boyd & Crawford 2012). De tal manera que los negocios de comercio electrónico pueden obtener una mejor comprensión del comportamiento del consumidor a partir del del intercambio electrónico de datos (EDI), y se puede aplicar para mejorar el servicio al cliente y guiar la estrategia comercial (Edosio, 2014).

Mediante la deducción estadística, las simulaciones y la prospectiva, que requieren billones de ecuaciones lineales, se pueden anticipar los patrones de consumo. Un ejemplo práctico es Amazon, que utiliza un software especial para analizar cookies e identificar patrones en el hábito de compra de los consumidores, con la intención de brindar ofertas personalizadas, anuncios y descuentos para tal consumidor (Mosavi & Vaezipour, 2013). De igual forma, existe una gran cantidad de información compartida entre las comunidades de eWOM (Olmedilla, 2016).

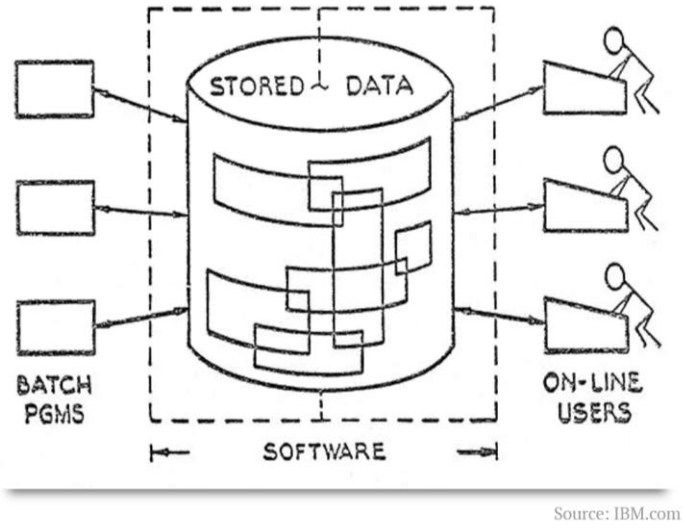

Figura 3- Representación de un Sistema de Base de Datos. Fuente: Camps et al. (2014)

A finales de la década de 1980 y principios de los 90, la popularidad de los sistemas de planificación de recursos empresariales (ERP) ofrecían la posibilidad de coordinarse e integrarse entre los departamentos de la empresa. La influencia de la información trajo un nuevo problema en la gestión de los datos. La dificultad de mantenimiento y aplicaciones de la información comenzaron a emerger las plataformas de *bussiness intelligence* o inteligencia empresarial (BI). A partir de este momento podemos mencionar la preocupación por el Knowledge Discovery in Databases o Descubrimiento del conocimiento en bases de datos (KDD) acuñado en 1989 que es utilizado para referirse a todo el proceso de extracción de conocimiento útil que queremos descubrir a partir de una base de datos.

Boca en boca (WOM) y Boca en boca electrónico (eWOM)

"Si tiene un cliente insatisfecho en Internet, no se lo dice a sus 6 amigos, le cuenta a sus 6,000 amigos" -Jeff Bezos, presidente de Amazon (2013).

En la literatura, Ramón y Morin (2009) definen las redes sociales como un espacio relacional por actores sociales diferenciados que buscan establecer entre sí distintos procesos ya sea de cooperación, amistad, negociación, subordinación, solidaridad, entre otros. La construcción de redes sociales nos permite acceder a nuevas informaciones más allá de nuestro espectro de conocimiento y/o experiencia.

En 1966 los autores Katz y Lazarsfeld describieron el término Boca en Boca o Word of Mouth (WOM) como el intercambio de información de marketing entre los consumidores, que desempeña un rol en el cambio de comportamiento y actitudes hacia los productos y servicios.

En 1967 Arndt caracterizó el término WOM como la "comunicación oral de persona a persona entre un receptor y un comunicador que el receptor percibe como no comercial, con respecto a una marca, producto o servicio". Junto a esta definición, los autores Katz y Lazarsfeld (1966) sugieren que ese intercambio de información debe desempeñar un papel fundamental en el comportamiento y la actitud del receptor frente al bien o servicio.

Los usuarios con frecuencia utilizan WOM como fuente de información sobre marcas, productos y servicios de las organizaciones para tomar decisiones de compra. Esta información puede ser positiva o negativa[7]; en el caso de las opiniones positivas, los clientes relatan experiencias agradables, vívidas e innovadoras, recomendaciones a otros e incluso una exhibición llamativa. En un WOM negativo se incluyen comportamientos como denigración del producto, relatando experiencias desagradables, rumores y quejas privadas. A lo largo de este trabajo, WOM será utilizado para referirse a aquellas comunicaciones informales entre personas, sobre evaluaciones de bienes y servicios.

WOM puede influir positivamente en las decisiones o negativamente, por esta razón, las empresas, mercadólogos, editores, etc. prestan atención sobre lo que se dice de su producto debido a que la exposición de comentarios negativos se asocia con baja probabilidad de comprar un producto, mientras que los positivos se vinculan a una alta probabilidad de compra (Duan, Gu y Whinston, 2008).

Si bien el boca a boca es una de las formas más antiguas de transmitir información (Dellarocas, 2003), el auge y la difusión de Internet ha llevado a la aparición de una nueva forma de boca a boca (WOM): boca a boca electrónica (eWOM), considerada uno de los medios informales más influyentes entre los consumidores, las empresas y la población en general (Huete-Alcocer, 2017).

Henning-Thurau et al. (2010) se refiere al término "boca en boca" electrónico (eWOM), a como el intercambio de información a través de Internet, sobre un producto o servicio de una organización. Los consumidores con frecuencia utilizan el boca a boca como fuente de información sobre marcas, productos y servicios de las organizaciones para tomar decisiones de compra. El hecho de recibir información de una fuente ajena a la compañía produce más

[7] Arndt J. Role of product-related conversations in the diffusion of a new product. Journal of Marketing Research. 1967 Aug;4(3):291–295. doi: 10.2307/3149462.

confianza en el consumidor que el marketing directo que puedan realizar las organizaciones tales como panfletos, publicidad, promociones, etc.

Independientemente de la forma de WOM o eWOM, su finalidad recae en el intercambio de información sobre las experiencias con diversos productos y servicios. Ver tabla 1 para una mejor comprensión entre ambos estilos de WOM:

	WOM	eWOM
Credibilidad	Conoce a la persona que comunica (Credibilidad positiva)	Anonimato entre los interlocutores (Credibilidad negativa)
Privacidad	Conversación privada, interpersonal y en tiempo real	Puede ser leída o vista por cualquiera en cualquier momento
Rapidez de difusión	Ambos interlocutores están presentes y los mensajes son compartidos de forma lenta	Los mensajes se pueden compartir por Internet y transportada en cualquier momento
Asequibilidad	Menos asequible	Fácilmente asequible

Tabla 1 - Diferencias entre WOM y el eWOM. Fuente: Nuria Huete-Alcocer (2017)

Las comunidades eWOM a menudo tienen un fuerte impacto en los juicios de productos porque la información recibida es más accesible (Olmedilla, 2016). Las opiniones publicadas incluyen una gran variedad de productos y servicios, y se han convertido en parte del proceso de toma de decisiones para los consumidores (Khedo et al., 2012) y con las características de que las conversaciones en el eWOM se pueden producir de forma anónima.

Hoy en día los usuarios se han convertido en entusiastas creadores de información gracias al boom de las redes sociales. El contenido generado por los usuarios o User Generated Content (UGC) se utiliza para definir aquellos datos creados voluntariamente por personas, usualmente en la web. Estos datos son utilizados por empresas, mercadólogos, agencias, entre otros, como una herramienta efectiva para atraer a los consumidores (iab, 2015).

El contenido generado por los clientes dentro de eWOM se pueden intercambiar a través de una variedad de comunidades en línea como foros de discusión, sistemas de tablones de anuncios electrónicos, grupos de noticias, blogs y sitios de revisión (Goldsmith & Horowitz, 2006). En contraste con el WOM tradicional, las conversaciones son visibles para el resto de los consumidores (Toral et al., 2014).

La transformación del cliente pasivo a activo que desea participar en todos los procesos de producción y el desarrollo de redes sociales, están cambiando la visión de la producción misma, forzando a las organizaciones a crear un vínculo con el mercado e interactuar, así como también a ser abiertos y cooperativos con los clientes y otras partes interesadas en los procesos de producción. Los sitios de recomendación en línea (eWOM) utilizan SNS (Social Network Sites) como medio de comunicación doble con tus consumidores potenciales. En este sentido, la correcta aplicación del eWOM en SNS permiten crear un lazo de lealtad, crear una reputación online, conocer la satisfacción de los clientes (Gil-Pechuán et al, 2014).

El uso de las SNS proporciona un mayor y mejor acceso a mercados, información, tecnología y otros recursos que favorecen las posibilidades de supervivencia, crecimiento y éxito en general (Gulati et al. 2000). Los consumidores pueden intercambiar información eWOM a través de diversos tipos de plataformas como son los blogs, micro blogs, foros de discusión, sitios web de revisión, sitios web de compras, comunidades de consumidores virtuales y sitios web de medios sociales (como citado en Olmedilla, 2016).

La Asociación de Investigación en los Medios de Comunicación (AIMC) confirma que 77% de los usuarios en España ha consultado opiniones de otras personas en Internet, 48% han confiado en la opinión de los usuarios sobre un producto o servicio y confían en estos, mientras un 43% ha comentado su propia opinión sobre un producto o servicio en línea. Lo anterior demuestra la importancia de la presencia en la web de las empresas en OSNs para sustentar sus estrategias[8].

E-commerce y comunidades eWOM

"Los medio WOM e eWOM influyen tanto en las empresas como en los consumidores, ahora que se han convertido en una de las fuentes de información más influyentes para la toma de decisiones y procesos de fabricación." -Huete-Alcocer (2017)

El comercio social9 es una forma de comercio mediado por las redes sociales que involucra la convergencia entre los entornos tradicionales y en línea (Bai, Yao & Dou, 2015; Chen & Shen, 2015; Shanmugam et al., 2016) y se manifiesta en las plataformas eWOM. Mahmood Hajli (2015) defiende que las plataformas eWOM como Medios Sociales en Línea (OSNs) permiten acceder a opiniones de clientes, recomendaciones, panel de discusión y redacción e incluso incluir

[8] AIMC (6 de marzo de 2018). Resultados de la 20ª Encuesta a Usuarios de Internet, Navegantes en la Red. *AIMC. Recuperado* de https://www.aimc.es/blog/internauta-espanol-alto-grado-confianza-la-compra-online-esta-continuamente-conectado/

[9] Término usado por Araceli Castelló (2011) en su artículo: La venta online a través de medios sociales: el social commerce.

calificaciones de una opinión. En este entorno, los clientes tienen acceso a conocimientos y experiencias que les ayuda a comprender mejor su propósito de compra y tomar decisiones más informadas y precisas (como citado en Chen, A., Lu, Y., Wang, B., 2017).

Duan et al. (2008) sostienen que las comunidades eWOM influyen directamente en los clientes y crean interés con eficacia y flexibilidad a pesar de las fronteras geográficas. En el mismo contexto, el crowdsourcing en el comercio social puede verse como una integración de usuarios en los procesos internos que crean valor.

Como ejemplo práctico, el conocimiento de las personas en las comunidades eWOM sobre el uso de productos o servicios, permiten a aquellos poco conocidos (Long Tail), competir con los productos de éxito. Sin embargo, Standifird (2001) considera que las plataformas eWOM puede inhibir el fenómeno de Long Tail, promoviendo la venta de productos populares con altas calificaciones. Por lo tanto, la parte principal de la distribución de ventas se vuelve más gruesa generando eventos de alta frecuencia en la cola corta.

Por otro lado, Park y Lee (2009) sostienen que las opiniones de los usuarios pertenecen a una categoría de producto y el efecto del eWOM depende del producto en sí. A partir de esta teoría, dependiendo la naturaleza de los atributos específicos, estos pueden clasificarse generalmente como productos de búsqueda o productos de experiencia (Cui, Lui & Guo 2012). Ver figura 4:

Figura 4- Beneficios y usos del UGC. Fuente: iab (2015)

Long Tail

El Long Tail explica cómo nuestra economía y cultura está cambiando de los mercados masivos a millones de nichos. Narra el efecto de las tecnologías, facilitando a los consumidores encontrar y comprar productos especializados gracias al "efecto infinito del espacio de almacenamiento" y los nuevos mecanismos de distribución, rompen el cuello de botella de la venta al por menor de los mercados tradicionales.

Bajo un contexto general, los autores Benghozi y Benhamou (2010) definen el Long Tail desde un punto teórico, haciendo referencia al principio de la distribución estadística que enfatiza el peso considerable de eventos infrecuentes o menores en comparación con eventos frecuentes o mayores. La Long Tail se aplica a una curva de distribución, que a menudo forma leyes de poder y, por lo tanto, son distribuciones de Long Tail en el sentido estadístico (Olmedilla, 2016).

Alan Lew (2008) afirma que, a pesar de que los productos de Long Tail proporcionan ventas individuales razonablemente bajas, pueden obtener ganancias al proporcionar una mayor diversidad de productos en conjunto. Por tanto, el concepto Long Tail describe la estructura y el éxito de las actividades basadas en Internet al representar un nuevo enfoque para la comercialización y venta de productos que no existían antes del advenimiento de Internet (Lew 2008).

En el contexto ¨en línea¨ de venta minorista se venden productos menos “populares” que aquellos comercios tradicionales, generando una nueva tendencia en la economía. Dentro de su descripción sostenía que las recomendaciones creadas por los usuarios en plataformas impulsan la personalización y, por tanto, el incremento del consumo de productos de Long Tail de nicho de productos de menor venta. Por otro lado, Anderson sostiene que las recomendaciones basadas en contenidos generados por el usuario guían a los consumidores a especialización, satisfaciendo sus necesidades con productos de menor venta. En esta línea, surge una nueva tendencia en la economía: los productos de nichos tendrán una mayor demanda en comparación con los comercios de minoristas tradicionales. En la figura 4 se ilustra un ejemplo de curva de Cola Larga o Long Tail:

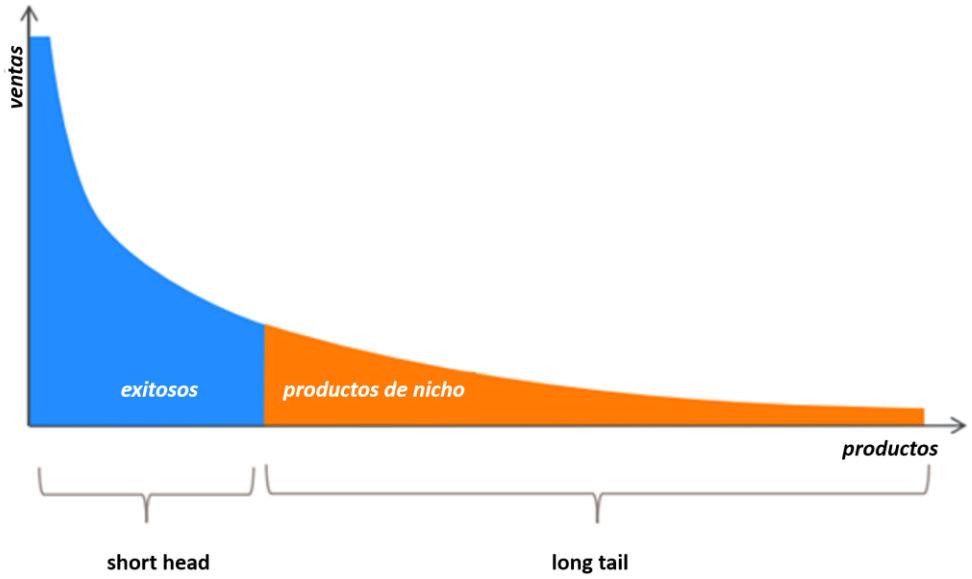

Figura 5- Ilustración del Long Tail. Fuente: María Olmedilla (2006)

En la parte naranja del gráfico de ventas anterior específicamente, muestra una curva de demanda estándar que podría aplicarse a cualquier industria según el nivel de ventas y tipo de producto. En este gráfico el eje vertical representa "ventas" y el eje horizontal "tipos de productos". La sección azul representa los productos estrella, protagonistas y la sección naranja son los productos no populares, o nichos.

Los eventos de Long Tail rara vez ocurren de forma individual, pero el número agregado total de eventos de baja frecuencia que se encuentran en la cola puede ser superior al grupo de eventos de alta frecuencia en la cabeza corta (Lew 2008). En resumen, la teoría de la Long Tail explica que, cuando los usuarios tienen a ampliar sus opciones, suelen dirigir su atención hacia los nichos porque satisfacen mejor los intereses más específicos y acorde con nuestros límites. Tal y como sostiene María Olmedila (2016), la proliferación de canales en línea, los consumidores encontrarán más fácil buscar y descubrir esos productos poco conocidos ya que Internet ha eliminado muchas de las barreras de comunicación entre ubicaciones geográficamente distantes.

Chris Anderson (2008) delineó dos ideas diferentes pero relacionadas. El primero es que las ventas de productos por debajo de un cierto volumen aumentan proporcionalmente a las de los productos que superan un cierto volumen, ya que las limitaciones físicas y de costos de la selección desaparecen. La segunda idea es que los canales en línea están cambiando la forma de la curva de demanda ya que las personas están interesadas en explorar esos productos de nicho, que es más probable que estén orientados a sus intereses particulares que en aquellos orientados al mercado masivo." En este sentido, Alan Lew (2008) argumenta cómo el enfoque económico

de Internet proporciona una vía de éxito en el mercado de Long Tail. Este libera a muchas empresas de los factores de ubicación tradicionales, proporcionando un medio económico para que individuos y empresas lleguen a clientes potenciales. Si los costos de almacenamiento y distribución son altos, solo los productos más rentables se pueden vender a un precio competitivo.

El concepto de Long Tail describe la estructura y el éxito de las actividades basadas en Internet, al representar un nuevo enfoque para el mercadeo y los productos de nicho. Este fenómeno muestra la capacidad de obtener ganancias al proporcionar una mayor diversidad de productos en conjunto, opuesto a la idea de ofrecer una reducida variedad de productos que se venden mucho (Lew, 2008).

La Long Tail es el resultado de cómo nuestra economía y cultura está cambiando de los mercados masivos a millones de nichos. Las tecnologías facilitan a los consumidores encontrar y comprar productos especializados gracias al "efecto infinito del espacio de almacenamiento", junto a los nuevos mecanismos de distribución, rompen el cuello de botella de la venta al por menor de los mercados tradicionales.

Productos de búsqueda y de experiencia

"[…] el valor incremental de los nuevos medios de comunicación será el suministro de información en un formato más accesible, menos costoso y más personalizable. Esperamos que esto reduzca los costos de búsqueda aumente los beneficios esperados a través de la nueva información"- Klein (1998)

A medida que las economías mundiales se vuelven más interdependientes y el desarrollo de la web nos abre puertas a nuevos mercados, la oferta de productos se expande. Uno de los mayores desafíos para los usuarios es que no podemos inspeccionar todos los productos ofertados para luego seleccionar cuál es el que mejor se adapta a nuestras necesidades (Nakayama et al., 2010).

Nelson (1970) y otros autores posteriores, siendo Derby & Karni (1973) de los más influyentes en el área, identificaron la necesidad de entender el compartimiento de pre-compra, compra y post-compra y cómo puede ayudar a comprender el Long Tail. Estos autores fueron quienes propusieron el modelo "SEC" correspondiente a las siglas en inglés de los productos de "Búsqueda, Experiencia, Confianza" (Search, Experience, Credence) utilizado por economistas y mercadólogos. Es importante entender la diferencia de cada uno de estos productos como consecuencia de la evolución de las ICT al convertir Internet en un mecanismo de búsqueda de información rápida y confiable. Los bienes de confianza no serán abordados en este Trabajo Fin

de Máster, pero se explicará con la intención de ofrecer al lector una mejor comprensión de los tipos de productos del "SEC".

Según el modelo SEC, los productos se categorizan según el grado de evaluación de la calidad del producto antes o después de la compra (experiencia de uso). Los productos de búsqueda son aquellos cuya calidad puede ser objetivamente evaluada por el consumidor antes de la compra, permitiendo al cliente poder comparar con mayor facilidad y como consecuencia, están más relacionados a sustitución y competencia por precio al poder buscar alternativas que ofrezcan mayores ventajas. En el caso de los productos de experiencia, sólo después de probarlos, puede el consumidor formarse una opinión sobre la calidad, las características principales y los beneficios del producto (Nelson 1970). Los productos de confianza son aquellos que, incluso tras su compra o uso, el consumidor no puede realizar una evaluación de su calidad. Un ejemplo son los suplementos dietéticos (Derby & Karni 1973).

Podemos concluir que, en cuanto a los productos de búsqueda, el consumidor puede considerar factores que hacen al producto superior sobre otra oferta en el mercado previo a la decisión de compra, mientras que en los productos de experiencia no es posible sin uso previo. Por último y en contraste con los bienes descritos anteriormente, tenemos los bienes de confianza, aquellos cuyas cualidades y beneficios no podrán ser percibidos por los consumidores, incluso después de comprarlos. Sintetizando las definiciones anteriores Valerie Zeithaml (1981) representa los atributos de los bienes y su facilidad de evaluación a través de la figura no.3:

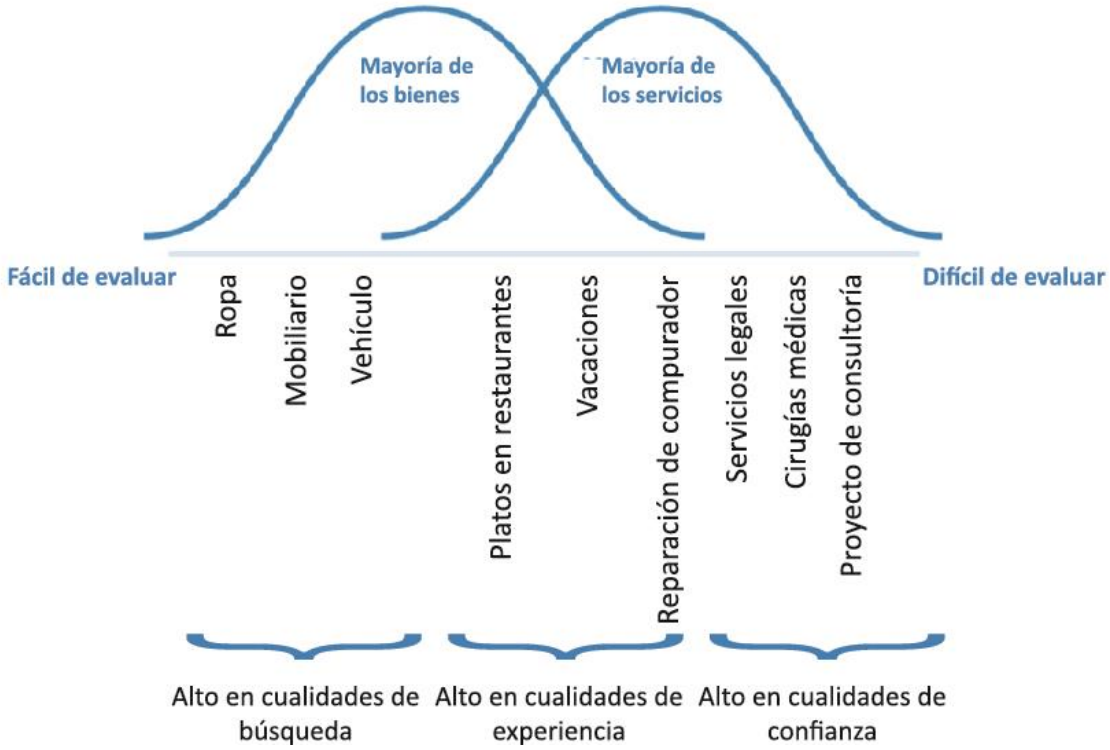

Figura 6 Modelo SEC. Fuente: Valerie Zeithaml (1981)

Huang, Lurie y Mitra (2009) argumentan que Internet sirve como una importante fuente de información para los productos de búsqueda y experiencia, enfatizando que la web proporciona un canal útil para dar a conocer y propagar información sobre la calidad y poder "experimentar" el producto antes de comprarlo.

RECOPILACIÓN DE INFORMACIÓN

Los clientes recopilan información en el proceso de compra de un producto para ayudarles a tomar una decisión. Las personas pueden solicitar opiniones de sus amigos o consultar los comentarios y recomendaciones de otros clientes en línea (como mencionado en Chen, 2017). La búsqueda de información durante el proceso de compra es similar a la compra de escaparates, donde se transfiere información relacionada de un producto al cliente potencial.

Base de datos y búsqueda en redes de información

"Hoy en día, prácticamente todo se hace electrónicamente; las personas intercambian información a través de Internet y participan en la compra y venta a través de este medio." - Assuncoa et al. 2013

Internet y las redes sociales se han convertido en el medio natural de las operaciones básicas de las empresas y la rutina de las personas, convirtiendo estas plataformas en una base para el intercambio de datos, información, confidencias, emociones, necesidad de, opiniones e intereses (Valls, 2017). La Web se ha convertido en el medio más dinámico y estimulante para encontrar y recuperar información de las bases de datos automatizadas. Funciona como una plataforma donde podemos buscar y localizar texto con facilidad en cualquier ordenador que esté conectado a la red, y transferir o grabar la información[10].

Las "bases de datos" son el método preferido para el almacenamiento estructurado de datos donde, utilizando diferentes tipos de motores de búsqueda o Internet Search Engine (ISE), permite a los usuarios encontrar información con costes inferiores a los sistemas tradicionales (como citado en Camps et al., 2014) y en tiempo récord.

Los motores de búsqueda o mecanismos de búsqueda (*search engine*) son programas que permiten buscar dentro de una base de datos web (Omedilla, M., 2016). Un motor de búsqueda tiene por finalidad proporcionar resultados relevantes para el usuario a partir de la solicitud de información que emiten los usuarios, visitando las páginas Web y realizando la indexación (Stark, 2001).

Actualmente los motores de búsqueda se clasifican en tres categorías principales: motores de búsqueda temática, también conocidos como directorios o catálogos; *crawlers* o motores de

[10] Esteve Fernández. Institut Universitari de Salut Pública de Catalunya. Campus de Bellvitge. Universitat de Barcelona. Ctra. Feixa Llarga, s/n. 08907 L'Hospitalet (Barcelona)

búsqueda por palabras claves y sistemas basados en el c*ontent-routing* o enrutamiento de contenido. Uno de los elementos más importantes de un motor de búsqueda es el crawler, de acuerdo con Najork, Gollapudi & Panigrahy (2009), un crawler visita una o más URL, descarga las páginas web asociadas, extrae cualquier hipervínculo que contenga y continúa recursivamente descargando la web. En su proceso básico busca, adquiere, indexa y mantiene páginas que representan un segmento estrecho de Web en lugar de rastrear toda la web.

Estar bien posicionado entre los resultados de búsqueda de clientes potenciales es importante. Por lo tanto, la visibilidad y la correcta la clasificación de los productos en la Long Tail en medios digitales logran una mejor efectividad a la hora de identificar los productos (Gil- Pechuán et al., 2014). El UGC es una forma efectiva de aumentar la indexación de los motores de búsqueda, especialmente en relación con una campaña de Optimización de los Resultados de Búsqueda o Search Engine Optimization (SEO) social.

El éxito y popularidad de un motor de búsqueda están condicionados por su capacidad de producir los resultados más relevantes a partir de atributos específicos y su capacidad de filtrar la información fiable para los usuarios. Según los datos de la consultora NetMarket Share de mazo del 2017, en el segmento de motores de búsqueda, Google tiene una supremacía global. Su popularidad surge a partir de su facilidad para encontrar los resultados de forma más relevante para los usuarios que aquellos buscadores que existían en el momento de su creación.

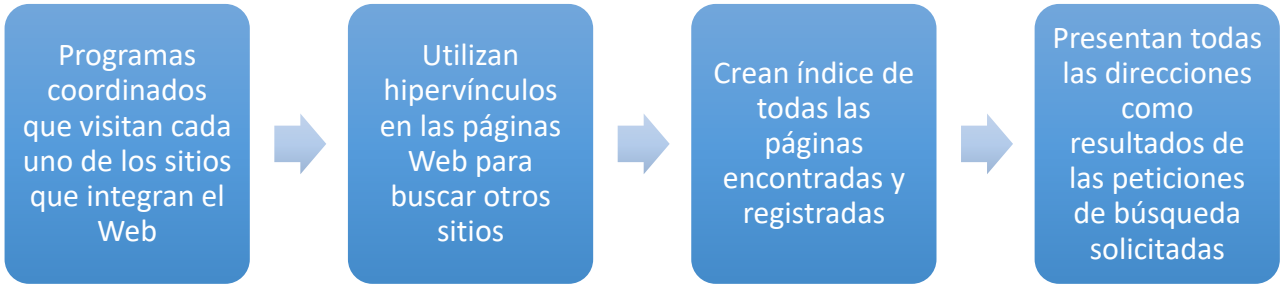

Figura 7 Proceso de búsqueda en base de datos. Fuente: Elaboración propia

Huang et al. (2009) investigan las diferencias en los patrones de búsqueda de los consumidores entre bienes experiencia y búsqueda en línea. Si bien los usuarios son capaces de utilizar motores de búsqueda en la base de datos de las plataformas eWOM (Olmedilla, 2016), al usuario exponerse a las evaluaciones de productos a partir de la experiencia de los consumidores, la búsqueda y la conducta de compra se ve afectadas. Por esta razón, los ISE plantean la necesidad de optimizar los resultados de búsquedas correspondiente a las palabras claves que identifica el usuario.

Los consumidores hacen uso de las herramientas de búsqueda y descubrimiento en Internet, como los motores de recomendación, y estos se relacionan directamente con un aumento en la participación de productos especializados. Nicholas Carroll (2010) expone que la utilización de un SEO permite alcanzar posiciones de alto rango en Internet, a través de motores de búsqueda y programación, marketing o conocimiento del contenido. Desde el punto de vista de los medios y las empresas, permite controlar la exposición del consumidor al comentario que busca para la toma de decisiones.

Minería de datos

La Web como repositorio de información, desafía a quienes buscan extraer información útil debido a la variedad de estilos y formatos de escritura, la falta de calidad de muchos documentos, el amplio espectro de asuntos y demás. Este escenario ha requerido del desarrollo de algoritmos y métodos específicos que permitan extraer información útil (Mendoza, 2011).

Los datos disponibles aceleran la innovación, mejoran la toma de decisiones y requieren estrategias de acción en tiempo real, por esta razón, es vital una correcta selección e integración de datos, y uso de que proporcionen un conocimiento de los clientes para adelantarse a sus necesidades y aspiraciones (Valls, 2017). De este modo, las empresas globales que mejor manejan los datos, las analíticas y los algoritmos, son aquellas que registran un mayor crecimiento económico, como Google, Facebook, y otras.

Los datos objeto de tratamiento son cada vez más numerosos y variados, y se analizan mediante métodos no convencionales para disponer de todo tipo de información útil y conocer qué podría atraer a los clientes. Para lograr este fin, es necesario una etapa de análisis que permita cuantificar los conocimientos, descubrir relaciones entre los valores, innovar y solucionar problemas (Valencoso, 2016, López, 2012 & Valls, 2017).

La minería de textos abarca una amplia variedad de técnicas para la agrupación, búsqueda y recuperación, visualización de información, etc. e incluye técnicas estadísticas, lingüísticas y de aprendizaje automático (Hashimi et al., 2015). Una de las principales ventajas de la minería de texto es que permite trabajar automáticamente con grandes volúmenes de material, como se requiere en el contexto en línea, debido a la gran cantidad de información disponible en Internet (Camps et al., 2014), extrayendo información con a partir de datos no estructurados.

Gracias a la minería de datos en la Web, podemos extraer información en Internet para mejorar la precisión de las recomendaciones de búsqueda de interés, organizar grandes volúmenes de información e utilizar máquinas de aprendizajes para detectar tendencias y por ende, la categorización de documentos (Cacheda, Fernández & Huete, 2011)

Al proceso de extracción de datos modelados de páginas web en Internet se le conoce como *scraping.* Fidel Cacheda Seijo et al. (2011) sustentan que este término se debe por el proceso de extracción de un "fragmento" de información específico. Un scraping efectivo permite filtrar e identificar el texto relacionado con la descripción.

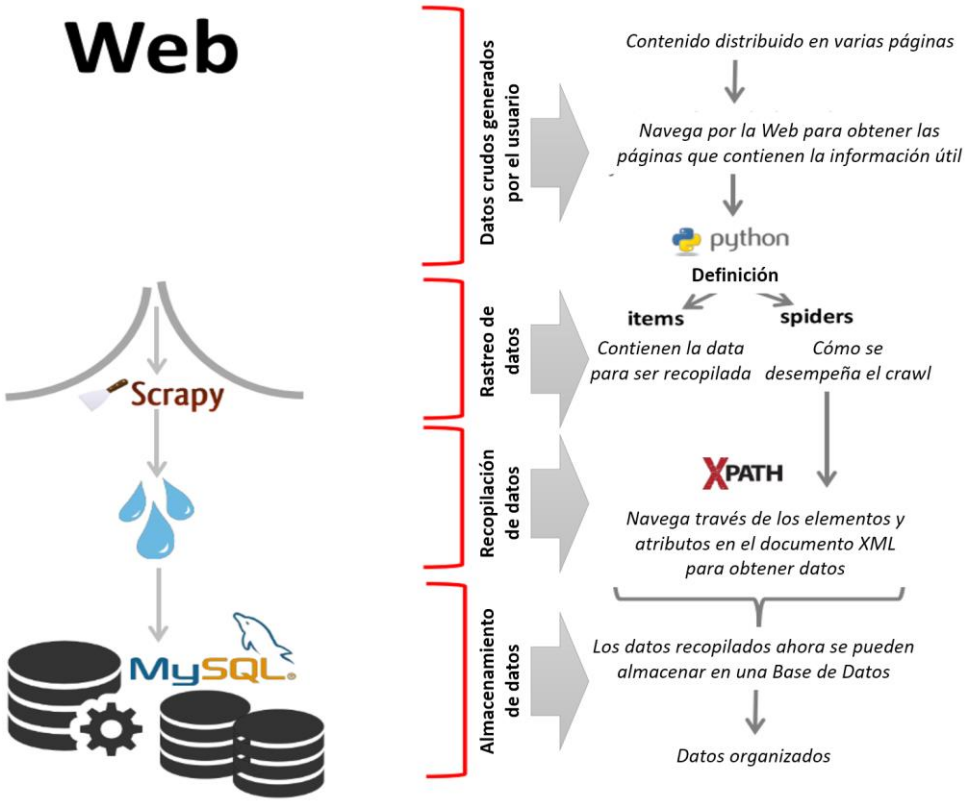

Figura 8- Proceso de recopilación de datos. Fuente: Maria Olmedilla (2016)

En la publicación *Text Mining: Techniques, Applications and Issues* (2016), los autores Ramzan Talib et al., describen el proceso genérico de minería de texto bajo los siguientes pasos (ver figura 9):

- Recopilación de datos no estructurados de la base de datos
- Se realizan operaciones de preprocesamiento y limpieza para detectar y eliminar anomalías, eliminando los stop-words, identificando la raíz de las palabras, indexando datos y asegurando la captura de la información esencial.
- Procesamiento automático para operaciones de procesamiento para auditar y limpiar la información.

- El análisis de patrones es implementado por el *Management Information Systems* o Sistema de Información de Gestión (MIS).
- La información procesada en los pasos anteriores se utiliza para extraer información valiosa y relevante para la toma de decisiones efectiva y oportuna y el análisis de tendencias, convirtiéndose en conocimiento.

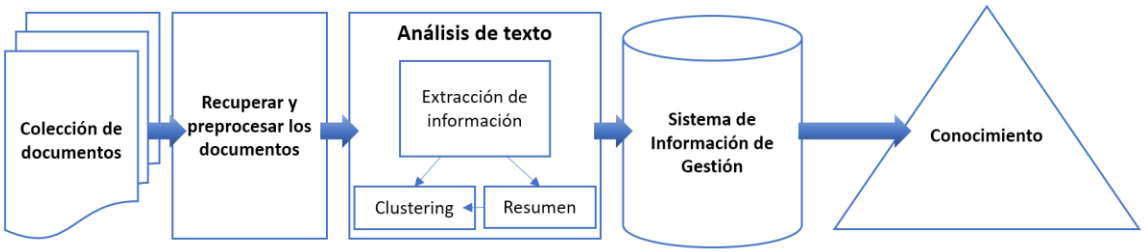

Figura 9- Transformación de datos a conocimiento. Fuente: Ramzan Talib et al. (2016)

Pregunta de investigación

Considerando que este estudio se ha basado en el análisis de contenido generado por el usuario. Este contenido está presente y es distribuido en las plataformas en línea a disposición del público.

Este trabajo pretende se busca identificar Atributos Únicos dentro de tres clasificaciones: "experiencia", "búsqueda" y "mix", en un análisis comparativo de opiniones en línea sobre productos "Moda". Continuando con la investigación, proponemos la siguiente pregunta de investigación:

> *Pregunta de investigación:* ¿Es posible identificar si el nicho de mercado pertenece a la categoría de productos de experiencia, de búsqueda o mixta a través de Atributos Únicos utilizando opiniones generadas por los usuarios en plataformas eWOM?

Los Atributos Únicos hacen referencia a propiedades distintivas asociadas a una clasificación determinada. Para una mejor comprensión, se espera que los Atributos Únicos predigan si el comentario describe un producto de "experiencia", "búsqueda" o "mixto", basándose en la frecuencia que el atributo aparezca en la opinión compartida. De esta forma, si bien los términos de "experiencia" hace referencia a palabras subjetivas, estas estarán presentes en comentarios subjetivos frecuentemente y, los términos de "búsqueda", que van relacionados a palabras que denoten objetividad, aparecerán con mayor frecuencia dentro de comentarios objetivos (Toral et al., 2017).

METODOLOGÍA

Dado que la literatura sobre el Long Tail no distingue entre los productos de búsqueda y experiencia basada en los nichos de producto, este Trabajo Fin de Máster se centra en el diseño de un clasificador que permita discernir entre productos de experiencia y productos de búsqueda en base a las opiniones de los usuarios para comprender mejor el fenómeno de la Long Tail. Para este fin, la metodología se divide en dos partes. En el primer apartado se explica la captura de datos y posteriormente, cómo se han analizado.

Captura de datos

Los comentarios utilizados para el desarrollo de este trabajo fueron facilitados por el grupo de investigación SEJ-548 *Big Data & Business Intelligence in Social Media* de la Universidad de Sevilla, recopilados en el período abril-mayo del año 2015.

La recopilación de datos para obtener el conjunto de opiniones y comentarios que se utilizó en este documento se basó en acceso al registro de la base de datos del sitio web Ciao UK. Ciao ha sido una de las comunidades en línea más grandes del mundo (alrededor de 1.3 millones de miembros) que revisa críticamente y califica millones de productos y servicios (alrededor de 7 millones de revisiones en 1.4 millones de productos) en beneficio de otros consumidores. Está disponible de forma gratuita para los consumidores en versiones locales en los principales mercados de Europa occidental (Toral et al., 2017).

Siguiendo el marco teórico sobre los datos generados por el usuario pueden obtener información relevante a partir de las redes sociales. Para extraer los datos se utilizó el lenguaje de programación Python versión 2.7 y se combinó con el spider de código abierto de rastreo web *Scrapy,* disponible en http://www.scarpy.org. Para rastrear la información en Scrapy se programó diferentes clases llamadas *Spiders* que permitió definir cómo se iba a rastrear la información en el sitio web de Ciao UK.

Los comentarios recopilados fueron de diversos nichos sin discriminar los comentarios por subcategorías de productos. Para este Trabajo Fin de Máster se utilizó un total de 959 comentarios pertenecientes al nicho *fashion* o moda y los comentarios fueron recopilados en formato *.cvs* y posteriormente convertidos en formato Excel del paquete Microsoft Office. Los datos obtenidos de las opiniones son: autor, descripción del artículo y cuerpo del comentario. Los comentarios estructurados por los usuarios abarcan prendas de vestir y accesorios de ambos sexos como carteras, bufandas, pantalones, camisas, entre otros.

Análisis de datos

Para el análisis de nuestra base de datos se implementaron diferentes técnicas para extraer la información relevante. Como primer paso se realizó el preprocesamiento de datos que implicó de manera sucesiva remover la puntuación, convertir todas las letras en minúscula, eliminar los *stop-words,* lematizar las palabras y convertir las palabras derivadas a su raíz. Las stop-words se definen como aquellas palabras comunes los documentos de texto y que no contienen (o muy poca) información útil para encasillar o clasificar las diferentes clases de documentos. Ejemplos de stop-words en inglés tenemos "have, who, is, if…". Remover estas palabras es útil cuando trabajamos con frecuencias de palabras o valores normalizados (Raschka, 2015).

En la mayoría de los casos los productos ofrecidos en línea fallan al comunicar las características propias del nicho, por lo que necesario identificar sus características únicas. Al recolectar información investigaciones previas en el área de estudio apoyan el uso de un análisis de varianza (ANOVA por sus siglas en inglés) para encontrar diferencias significativas entre los documentos recopilados (Toral et al., 2017). Por esta razón, posterior al preprocesamiento de datos se realizó un análisis ANOVA para encontrar diferencias significativas entre cada clase con respecto a las otras dos categorías denominados *Unique Attributes* o Atributos Únicos.

A continuación, se procedió al *feature generation* utilizando 340 comentarios seleccionados al azar dentro de la base de datos (Hu, Zhang, Wu & Zeng, 2017). Esta etapa se dividió en dos fases: (1) creación de un *bag of words* o listado de palabras y (2) la identificación de los grupos de palabras (frases) que identificaran las categorías de comentarios de búsqueda, experiencia y mix. En la primera fase se creó un documento donde se clasificó palabras de búsqueda y de experiencia con relación a su significado dentro los comentarios, que se utilizó en la fase dos como plantilla de referencia para la segunda fase del feature generation. La segunda fase y finalidad de este proceso consistió en la división de los 340 comentarios en 3 partes iguales, en la que cada grupo de palabras se asignó a una categoría manualmente: búsqueda, experiencia y mix.

Finalmente se desarrolló el *feature selection*, este proceso permitió seleccionar un subconjunto de palabras claves esparcidas *en* el documento resultante del feature generation. Para este caso se utilizó el algoritmo *TF-IDF*, un valor normalizado que equilibra la frecuencia de las palabras con su rareza a lo largo de la base de datos y cuya intención es medir la importancia en un documento o corpus de documentos, para descartar palabras frecuentes y poco relevantes. Si bien existen términos muy frecuentes que aparecen en la mayoría de los documentos, esta alta frecuencia no tiene valor discriminativo. El valor TF-IDF permite descartar esas palabras frecuentes

globalmente, enfatizando el valor de las palabras frecuentes localmente en u subconjunto de documentos de la base de datos.

El *time frequency* o frecuencia del documento tf_{ik} del término $_i$ en el documento es definido por la cantidad de veces que $_i$ ocurre en $_k$, mientras que el *inverse document frequency* o frecuencia inversa de documentos mide la importancia de un término específico por su relevancia dentro del documento calculada al dividir el número total de documentos por el número de documentos que contienen el término y luego obtener el logaritmo de ese cociente. El *TF-IDF* se calcula:

$$tf_{ik} * idf_i = \frac{f_{ik}}{\sum_{i=1}^{t} f_{ij}} * \log \frac{N}{n_k}$$

Donde N: número total de documentos en el corpus.

El rendimiento del algoritmo de clasificación depende de las características seleccionadas de las palabras claves. Al realizar el TF-IDF en cada categoría, se buscó el peso de relevancia de las palabras para ser consideradas como Atributos Únicos. Para maximizar la precisión del clasificador, posterior al TF-IDF se utilizó el método de clasificación supervisada de los *k-nearest neighbors* o k vecinos más cercanos (*k-NN*) en el lenguaje de programación Python, que pertenece a la subcategoría de los modelos no paramétricos descrita como *Instance-based learning* o métodos basados en instancia porque comparara datos memorizados en la memoria del conjunto de entrenamiento con las nuevas instancias de problemas (Witten & Frank, 2005)**.** El método k-NN estima la probabilidad de que un elemento *x* pertenezca a cada clase C_j. Esta función de densidad, expresada como:

$$F(\frac{x}{C_j})$$

Para k-NN se asignó cada documento a la clase mayoritaria de sus vecinos k más cercanos, donde k es un parámetro. En función de la distancia métrica seleccionada, el algoritmo k-NN encontró en el conjunto de datos las muestras *k* más similares (cercanas) y asignó la categoría por mayoría de cercanía de sus *k* vecinos. La distancia utilizada es una generalización de la distancia Euclidean y Manhattan que se define como:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ri} - x_{rj})^2}$$

Posteriormente se utilizó el *F1 score* o valor F como medida de desempeño del test, basado en la precisión (correcta identificación) y el *recall* o exhaustividad (encontrar todos los valores positivos dentro de la muestra) (Raschka, 2015). El valor resultante es una media de los valores de precisión y exhaustividad, de tal forma que:

$$F_1 = 2 * \frac{Precisión * Exhaustividad}{Precisión + Exhaustividad}$$

Donde:

$$Precisión = \frac{\{documentos\ relevantes \cap documentos\ recuperados\}}{\{documentos\ recuperados\}}$$

$$Exhaustividad = \frac{\{documentos\ relevantes \cap documentos\ recuperados\}}{\{documentos\ relevantes\}}$$

Es indispensable estimar el desempeño del modelo con nuevos datos. En este sentido, como siguiente paso se ejecutó la técnica de validación cruzada *k-fold cross-validation*, una técnica de remuestreo sin remplazo que persigue validar los modelos creados en proyectos de inteligencia artificial (Raschka, 2015). En esta investigación se utiliza la técnica con k = 10, por esta razón se explica el modelo con el uso de esta variable. El proceso consiste en dividir los datos de entrenamiento en k pliegues, utilizando k-1 pliegues para el entrenamiento y 1 como dato de prueba para evaluar si funciona el modelo. Este paso se repite k cantidad de veces hasta que cada división se pruebe y al terminar, se estima el rendimiento de precisión o error (*variable E_a)* de cada prueba de pliegue para calcular el desempeño medio del modelo. La siguiente figura sintetiza el proceso utilizado:

Figura 10 Recuperado de Randal Raschka (2015)

La figura 10 muestra la interacción de los pliegues de entrenamiento junto al pliegue de prueba para cada iteración. La media de esas iteraciones representa la precisión del procedimiento.

El clasificador fue entrenado (input) y probado utilizando los valores dentro las tres categorías de los valores de los Atributos Únicos, TF y TD-IDF, que posteriormente fueron contrastados utilizando el algoritmo k-NN y el F1 score como métrica de comparación.

RESULTADOS

Para este trabajo de investigación, la metodología propuesta se aplicó a la base de datos recopilada de la plataforma Ciao UK que constó de 959 comentarios en inglés correspondientes al nicho *fashion.*

En el preprocesamiento de datos se eliminaron los signos de puntuación (el punto, la coma, el punto y coma, las comillas, los signos de interrogación, entre otros), se utilizaron listas predefinidas de stop-words y también se identificaron manualmente aquellas palabras que se consideraban stop-words para este caso para su eliminación, se lematizaron las palabras y se convirtieron a su raíz aquellas palabras derivadas. Palabras como nombres de famosos, acrónimos de marcas, letras que representaban el nombre de un modelo, etc. fueron conservadas por su relevancia en los comentarios.

La etapa del feature selection fue un proceso manual que se dividió en dos partes: primero creó un *bag of words* donde se enlistó 122 palabras de búsqueda y 127 palabras de experiencia a partir de los primeros 60 comentarios de la base de datos y posteriormente, se eligió 347 comentarios los que se dividieron individualmente en fragmentos con un número de palabras similar y en función a la longitud del comentario. A continuación, se creó una lista de Excel con 3 hojas de trabajo nombradas de acuerdo con la categoría correspondiente: búsqueda, experiencia y mix. A cada fragmento se le designó a una hoja de Excel (según clasificación) a partir de los atributos que le dieran significado. El bag of words sirvió como plantilla de referencia en caso de tener dudas durante el proceso de clasificación.

Los términos o palabras incluidos en el bag of words representan candidatos potenciales a ser Atributos Únicos de productos de búsqueda, experiencia o mix. Para redefinir la selección de los atributos se condujo la prueba estadística ANOVA que analizó la diferencia de las palabras entre las tres categorías posibles. La diferencia nula de las palabras en relación con las tres posibilidades de categorización representa una igualdad entre las etiquitas por lo que se anula la exclusividad, de lo contrario una distinción significativa de una palabra en una de las categorías frente a las otras dos representa un atributo único designado para esa clasificación. Un total de 664 Atributos Únicos fueron identificados a través de esta prueba estadística.

Para obtener los valores de TF y TF-IDF se utilizó la hoja de Excel resultante del feature generation con los fragmentos de comentarios asignados a cada categoría de productos, sin embargo, se identificó un desequilibrio en la cantidad de comentarios pertenecientes a cada categoría. Se determinó la necesidad de aumentar el data set para erradicar este desbalance con la intención de regularizar y prevenir el desajusto del peso de las categorías (Wong, Gatt & McDonnell, 2016).

El aumento de la data set consistió en repetir fragmentos aleatorios en las categorías que lo exigieran hasta que tuvieron una cantidad balanceada de celdas en cada clase (659 comentarios por cada categoría). Se identificaron un total de 4,192 palabras con su TF y su TF-IDF correspondiente utilizando un algoritmo en Python.

Como es señalado en la sección metodología se utilizó el clasificador k-NN. Como datos de entrada, se seleccionaron las primeras 2,000 palabras que resultaron del feature selection (TF, TF-IDF). La muestra de palabras seleccionada se dividió en 80% *training* y 20% *test.* Con la parte de training se entrenó el clasificador k-NN con k=1, que ha dado la siguiente matriz de confusión para la parte de test:

Etiqueta de destino	Etiqueta predicha	Conteo
Experiencia	Experiencia	92
Experiencia	Mix	13
Experiencia	Búsqueda	29
Mix	Experiencia	4
Mix	Mix	146
Mix	Búsqueda	0
Búsqueda	Experiencia	14
Búsqueda	Mix	0
Búsqueda	Búsqueda	115

Tabla 2- Matriz de confusión

La matriz de confusión resultante nos permitió visualizar el desempeño del algoritmo k-NN. En las predicciones se observó una razonable distinción entre cada una de las categorías.

Con el fin de determinar la eficiencia de los clasificadores, se utilizó una estrategia de validación cruzada k-fold. Este proceso consistió en dividir los datos de entrenamiento en k pliegues, en este caso k=10, utilizando 9 de estos pliegues para el entrenamiento y 1 como dato de prueba para evaluar si funcionó el modelo. Este paso se repitió las k veces (10), el rendimiento del clasificador se midió con el pliegue que se queda fuera del training set y el valor de rendimiento final se tomó de la media de los valores de cada repetición. Como métrica de desempeño de la prueba se usó la validación cruzada F1 score (combinación de precisión y recall) que, a partir del 10 fold-cross que respaldado de evidencias teóricas obtienen el mejor resultado (toral et al., 2017).

Cada clasificador para el proceso de entrenamiento y prueba tomaron como entrada las características de un conjunto de datos e hicieron predicciones dentro de los conjuntos de clases. El primer clasificador utilizó el bag of words con el mayor TF, el segundo también utilizó el bag of

words pero con mayor TF-IDF y por último, el k-NN utilizó los 664 Atributos Únicos resultantes de la prueba estadística ANOVA. La tabla no. Tal resume la precisión de los tres clasificadores considerando el input y detallando el valor medio del F1 score:

	Cantidad de Bag of words	***F1 score (Precisión)***
TF	2,000	0.87 (+/- 0.05)
TF-IDF	2,000	0.87 (+/- 0.05)
Atributos Únicos	664	0.90 (+/- 0.04)

Tabla 3- Aplicación del clasificador 1-NN a diferentes bags of words

Como última prueba y sólo con intención de validación, se utilizó el algoritmo Clasificador Logístico en Python en lugar de k-NN que obtuvo como resultado:

Accuracy: 0.84 (+/- 0.04)

Los mejores resultados se obtuvieron a partir de los Atributos Únicos con una puntuación F1 de 0.90 de 664 términos. Por otro lado, los valores TF y TF-IDF a pesar de utilizar un bag of words mayor tuvieron un resultado más bajo comparado con los Atributos Únicos. A partir de este principio podemos concluir que los Atributos Únicos son el mejor conjunto de predictores de clases logrando asociar cada conjunto de palabras de forma más eficiente.

Para determinar el valor de k en k-NN en F1-score, se cambió el valor de k del clasificador k-NN, concluyendo que k=1 sería más eficiente que valores más altos (ver figura no.11).

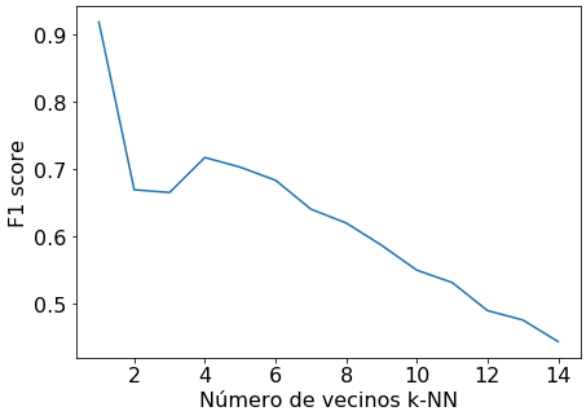

Figura 11- Uso del F score para determinar el mejor valor k en k-NN

Como último paso y con la intención de determinar la categoría a la que pertenece el nicho fashion se utilizó el clasificador con el total de los 959 comentarios pertenecientes a nuestra base de datos original, esta vez clasificando cada comentario de forma individual a partir de los Atributos Únicos. Finalmente, el clasificador k-NN con k=1 se determinó que:

	Conteo
Búsqueda	99
Experiencia	815
Mix	42

Tabla 4. Resultado de la clasificación de los comentarios en la base de datos

En la tabla 4 podemos observar los resultados del clasificador a partir de los Atributos Únicos con k-NN donde k=1 que, de los 959 comentarios, un 84% de los comentarios pertenecen a experiencia, un 10% a atributos de y el 6% restante a mix. En resumen, los productos dentro de la categoría fashion pertenecen a la categoría de productos de experiencia.

DISCUSIONES E IMPLICACIONES

"La tecnología, en forma de escucha social, llega a los usuarios en tiempo real y tiene rápida acción que puede llevar a una monetización mejorada". -iab, 2015

La nueva economía se compone de fuerzas interactivas que incluyen la globalización, la liberalización del comercio y la revolución de la tecnología de la información y las comunicaciones. En el mismo contexto, a largo de esta literatura se respalda cómo la reestructuración de la sociedad económica en la comercialización y el desarrollo de productos, centrado en pequeños segmentos de mercados minoritarios en el medio On-line, pone a disposición miles de ofertas de bienes y servicios.

A través de plataformas digitales los usuarios buscan información ante la necesidad de tomar decisiones. Por consiguiente, el manejo de los datos es uno de los retos más importantes desde el auge de Internet. Es necesario prestar una atención rigurosa a la evidencia y los datos, reconociendo el conocimiento como un activo institucional. Siguiendo este pensamiento, la identificación de información relevante en la Long Tail puede ofrecer nuevas formas para que los académicos y profesionales estudien el comportamiento del consumidor (Olmedilla, 2016) e identifiquen nuevas tendencias. Su comprensión en las Ciencias Sociales brinda la posibilidad de que los minoristas utilicen la disponibilidad de productos especializados como fuente de ventaja competitiva frente a la competencia.

La creación y entrenamiento de un clasificador que permita identificar productos de búsqueda y experiencia en los nichos de Long Tail a partir de las opiniones en línea, visto desde la perspectiva de la evolución del conocimiento en las industrias, impulsa la innovación y la especialización del Long Tail para el crecimiento global del mercado de nichos. Recordamos que la innovación sirve como catalizador que impulsa el motor del crecimiento económico debe reconocerse como un postulado fundamental de la nueva economía global.

A lo largo del documento se revisó comentarios en Internet generados por los usuarios donde compartían sus opiniones y percepciones sobre productos pertenecientes al nicho de mercado fashion. Un problema que tienen los investigadores al recopilar información es la identificación de las palabras utilizadas por los usuarios para describir su experiencia pueden ser diversas y evolucionar con el tiempo y, por otro lado, determinar la singularidad de los términos puede convertirse en un desafío.

Los Atributos Únicos de nichos de mercado pueden ser recopilados a partir de la percepción y satisfacción de los clientes a través de los comentarios en línea. El análisis de textos abiertos

requiere técnicas basadas en la minería de datos no estructuradas (Raschka, 2015). Por esta razón en enfoque de ese estudio se concentró en aplicar una metodología cuantitativa enfocada en el análisis comparativo entre un conjunto de clases (búsqueda, experiencia y otros) para determinar qué buscan los usuarios como atributo en el momento de buscar información.

Con el fin de identificar el mejor método para la clasificación de los nichos de mercado a partir de las publicaciones de los usuarios, se utilizaron 3 tipos de valores diferentes (TF, TF-IDF y Atributos Únicos). A partir de una muestra para cada valor se entrenó el clasificador haciendo uso del 80% de los datos para entrenamiento y el 20% restante de ensayo probando su precisión a través del k-NN donde k=1 y un f score de 10. A partir de esta prueba se determinó que los Atributos Únicos son más discriminantes que el TF y el TF-IDF, mostrando los términos más identificativos de cada clase y logrando una mayor precisión de segregación.

Por último, se realizó la clasificación de todo el conjunto de datos basada en el método de los Atributos Únicos. Los resultados del clasificador en la categoría fashion concluyeron en un gran margen de atributos basados en la experiencia (84%), con una presencia minoritaria de comentarios de búsqueda (10%) y mix (6%).

Contribución de la investigación

Este Trabajo Fin de Máster contribuye en la investigación de la identificación de atributos de los productos de la cola larga según la categoría de los nichos basados en el SEC. Hoy en día la Web brinda una amplia plataforma de experimentación para las Ciencias Sociales (Toral et al., 2017) y Empresariales. Las investigaciones se deberían enfocar en estudiar el comportamiento de los usuarios en plataformas sociales y erradicar la carencia de estudios relacionados a los contenidos generados con el usuario.

Esta investigación abordó la recuperación de información desde una perspectiva interdisciplinaria, entrelazando la ciencia de la computación, la estadística, la minería de datos, el machine learning y las Ciencias Sociales. También se definió una metodología basada en el entrenamiento de tres clasificadores para obtener un análisis de palabras claves para las diferentes categorías de producto dentro de la Long Tail seguido de un análisis comparativo para probar cómo los Atributos Únicos dentro de contenido generados por los usuarios pueden sintetizar la búsqueda de información relevante para su aplicación posterior en actividades empresariales.

Al encontrar un método que permita distinguir Atributos Únicos para diferentes categorías de productos y nicho de mercado, tiene un impacto significativo en la estrategia de marketing, SEO,

diseño de productos y comunicación empresarial. Es de obligación para los minoristas en línea encontrar y explotar el potencial de estos atributos que pueden servir de línea guía para obtener distinción entre la competencia.

Por último, el diseño propuesto en la metodología de la minería de datos puede complementar la literatura existente proponiendo alternativas de búsqueda de información más allá de lo tradicional. El uso de Internet y las OSNs permite recopilar la información al instante, identificar tendencias y el comportamiento del mercado a través del análisis de percepción del consumidor.

Limitaciones y propuesta de investigación a futuro

La información recopilada para el desarrollo de este trabajo se limitó a comentarios de productos en el sitio web Ciao UK, pero puede ser aplicado en otras plataformas eWOM a nivel global, incluyendo el uso de servicios. Un análisis posterior y más extenso puede ser dirigido a la identificación de otras categorías de productos a partir de comentarios presentes en otras páginas web para personalizar las campañas de marketing, perfeccionar la experiencia de búsqueda y eficientizar el proceso de producción y venta de las empresas.

Con relación al caso de estudio, es importante destacar que existen dos limitaciones importantes en este estudio (1) las APIs sólo proporcionan una muestra de la información existente en las plataformas sociales y (2) los comentarios provenientes de *bots* o cuentas falsas o inactivas que pueden comprometer la confiablidad del comentario. Para evitar recopilar información poco confiable, se podría comprobar del uso de algoritmos que permitan la segmentación de los comentarios a partir de atributos como la reputación, el historial de validaciones, entre otros.

Por otro lado, se puede profundizar la investigación bajo un enfoque relacionado al explorar el comportamiento del consumidor en los diferentes nichos de mercado y sus efectos. En este sentido, se puede probar el uso este clasificador para mejorar el SEO y el CRM a través del CGU. DE esta forma, evaluar las opciones para la optimización la comunicación, publicidad y el facilitar la búsqueda de información al cliente. De igual manera, la clasificación de los comentarios puede aportar en el desarrollo de productos y servicios de la empresa de forma más rápida y eficiente, para erradicar posibles fallos en la detección de tendencias y/o tiempo de entrega.

Conclusiones

Luego de una revisión exhaustiva de diferentes autores y la prueba de los tres diferentes métodos utilizados para la clasificación de productos (TF, TF-IDF y Atributos Únicos) de experiencia, búsqueda o mix en base de las opiniones de los usuarios en la plataforma Ciao UK en una de las

categorías de productos de Long Tail, podemos afirmar que los mejores resultados de clasificación se obtuvieron a partir de los Atributos Únicos, con una puntuación F1 de 0.90 a partir de 664 términos. Los valores TF y TF-IDF a pesar de utilizar un bag of words mayor que los Atributos Únicos tuvieron un resultado inferior. En resumen, determinamos que los Atributos Únicos son el mejor conjunto de predictores de clases, logrando asociar cada conjunto de palabras de forma más eficiente.

Por otro lado, al analizar los comentarios en la plataforma de Ciao UK en la categoría fashion en long tail, se mostró en evidencia que el 84 % se basan en la experiencia, el 10 % en búsqueda y el restante mix , lo que demuestra que los usuarios realizan sus compras en base de las opiniones basadas en la experiencia de los demás usuarios más que en las especificaciones técnicas que puedan ser proporcionadas por la empresa. De esta forma, se pueden aplicar estos hallazgos en plataformas eWOM para mejorar el marketing de los productos de venta de long tail basado en la vivencia del cliente.

Este trabajo propone un enfoque cuantitativo para la obtención de Atributos Únicos dentro de las clasificaciones de nichos de mercado utilizando técnicas de minería de datos. Los hallazgos revelan que es posible obtener un conjunto de atributos para cada clase que sirvan de predictores para cada categoría a través de comentarios compartidos en línea. A través de palabras claves se pueden discriminar aquellas características que buscan los clientes o clientes potenciales dentro de una categoría de producto para su aplicación en la estrategia empresarial y de marketing. Internet facilita a las empresas informar a sus clientes de productos y servicios, por lo que es vital para las empresas prestar suma importancia en hacer llegar información relevante al cliente que permita la identificación de factores claves y orgánicos, destacando su posición sobre los productos estrellas.

BIBLIOGRAFÍA

AIMC (6 de marzo de 2018). Resultados de la 20ª Encuesta a Usuarios de Internet, Navegantes en la Red. AIMC. Recuperado de https://www.aimc.es/blog/internauta-espanol-alto-grado-confianza-la-compra-online-esta-continuamente-conectado/

Anderson, C. (2006). The Rise and Fall of the Hit. *Wired Magazine* 14(7). Recuperado de: https://www.wired.com/2006/07/longtail/

Ariely, D., & HarperCollins (2010). *Predictably Irrational, Revised and Expanded Edition: The Hidden Forces That Shape Our Decisions*. Londres: HarperCollingsPublishers.

Aydin, E., & Kilinç, B. (2014). The Relationship between Globalization and E-Commerce: Turkish Case. *Procedia - Social and Behavioral Sciences*, pp. 150 1267 – 1276.

BBVA (2018). Los cuatro consejos definitivos para proteger tus datos en Internet. La Vanguardia. Recuperado de: https://www.lavanguardia.com/tecnologia/20180523/443666265462/cuatro-consejos-proteger-datos-Internet-brl.html

Benghozi, P.J. & Benhamou, F. (2010). The long tail: Myth or reality? *International Journal of Arts Management* 12(3), pp. 43-53.

Buttle, F.A. (1998). Word of mouth: Understanding and managing referral marketing. *Journal of Strategic Marketing* 6(3), pp. 241-254.

Cacheda, F., Fernandez Luna, J. M., Huete Guadiz, J. F. (2011), *Recuperación de información. Un enfoque práctico y multidisciplinar*, Córdoba: RA-MA Editorial.

Camps, R., Casillas, L. A., Costal, D., Ginestá, M., Martín, C. & Pérez, O. (2014). Base de datos (Formación de Posgrado). Recuperado de: https://www.researchgate.net/publication/43668137_Bases_de_datos

Carroll, N. (2010). Search Engine Optimization. P*acific-Sociology* 6, pp. 4613-4629.

Castelló, A. (2011). La venta online a través de medios sociales: el social commerce. *Revista académica del Foro Iberoamericano sobre Estrategias de Comunicación* 4(1), pp. 83-104. recuperado de:

https://www.researchgate.net/publication/298783941_La_venta_online_a_traves_de_medios_sociales_el_social_commerce

Chen, A., Lu, Y., Wang, B. (2017). Customers' purchase decision-making process in social commerce: A social learning perspective. *International Journal of Information Management* 37 (6): 627-638.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. Management science, 49(10), pp. 1407-1424.

Duan, W., Gu, B., Whinston, A. B. (2008). Do online reviews matter? — An empirical investigation of panel data. *Decision Support Systems*. 45(4), pp. 1007–1016. doi: 10.1016/j.dss.2008.04.001.

Edosio, U. (2014). Big Data Analytics and its Application in E-Commerce. (Artículo). Recuperado de: https://www.researchgate.net/profile/Uyoyo_Edosio/publication/264129339_Big_Data_Analytics_and_its_Application_in_E-commerce/links/53cf8ef30cf2f7e53cf811e0/Big-Data-Analytics-and-its-Application-in-E-commerce.pdf.

EFECOM (16 de junio de 2015). Internet, la herramienta que contribuye al auge de la economía colaborativa. La Vanguardia. Recuperado de: https://www.lavanguardia.com/economia/20150516/54431675862/Internet-la-herramienta-que-contribuye-al-auge-de-la-economia-colaborativa.html

Evolution and Development of E-commerce Market and E-Cash. The International Conference on Measurement and Control Engineering 2nd. Estados Unidos.).

Gil-Pechuán, I., Palacios-Marques, D., Peris-Ortiz, M., Peris Vendrell, E. & Ferri-Ramirez, C. (2014). *Strategies in E-Business: Positioning and Social Networking in Online Markets.* New York: Springer.

Goldsmith, R. E. & Horowitz, D. (2006). Measuring motivations for online opinion seeking. *Journal of interactive advertising* 6(2), pp. 2-14.

Goldsmith, Ronald E., and David Horowitz. 2006. "Measuring motivations for online opinion seeking." *Journal of interactive advertising* 6 (2): 2-14.

Hennig-Thurau, T., Skiera, B., Malthouse, E., Friege, C., Gensler, A., & Lobschat, L. (2010), The Impact of New Media on Customer Relationships. *Journal of Service Research* 13(3), pp. 311-330.

Hu, X., Zhang, C., Wu, M. & Zeng, Y. (2017). *Research on Long Tail Recommendation Algorithm.* Conf. Series: Materials Science and Engineering (pp. 1-6). Beijing: School of Information and Communication Engineering, Beijing University of Posts and Telecommunications.

Huang, P., N.H. Lurie, & S. Mitra (2009) Searching for experience on the Web: An empirical examination of consumer behavior for search and experience goods. *Journal of Marketing* 73(2), pp. 55–69.

Huete-Alcocer, N. (2017). A Literature Review of Word of Mouth and Electronic Word of Mouth: Implications for Consumer Behavior. *Frontiers in Psycholo*gy, 8, 1256. Recuperado de: http://doi.org/10.3389/fpsyg.2017.01256

Khedo, K. K., Roushdat, S. M., Mocktoolah, S. & Suntoo, R. (2012). "Online Social Networking as a Tool to Enhance Learning in the Mauritian Education System". *Journal of Emerging Trends in Computing and Information Sciences* 3(6): 907-912.

Kleemann, F., Voß, G.G., and Rieder, K. (2008). Un(der)paid Innovators: The Commercial Utilization of Consumer Work through Crowdsourcing. *Science, Technology & Innovation Studies,* 4(1), pp. 5-26.

Koutris, N. (2009). *Online Information Search for Experience Goods*: An Empirical Investigation of the Product Information Effects on Consumers (Tesis Doctoral). Recuperado de: https://thesis.eur.nl/pub/6241/Koutris,%20Nikolaos%20325791.doc

Krumm, J., Davies, N. & Narayanaswami, C. (2008). "User-Generated Contect". *IEEE Pervasive Computing*, 7(4):10-11.

Laband, D. N. (1991). An Objective Measure of Search Versus Experience Goods. *Economic Inquiry*, 29(3), pp. 497–509.

Lew, A. (2008.). Long tail tourism: New geographies for marketing niche tourism products. *Journal of Travel & Tourism Marketing* 25(3-4), pp. 409-419.

Lógica, B. & Magdalena, R. (2015). Using Big Data in the Academic Environment. 7th International Conference, The Economies of Balkan and Eastern Europe Countries in the changed world, EBEEC. Congreso llenado en Pitesti, Romania. Recuperado de: https://www.sciencedirect.com/science/article/pii/S2212567115017128

Myro, R. (14 de julio de 2001). Globalización y crecimiento económico. El país. Recuperado de: https://elpais.com/diario/2001/07/14/opinion/995061608_850215.html

Najork, M., Gollapudi, S. & Panigrahy, R. (2009). *Less is more: sampling the neighborhood graph makes SALSA better and faster.* In 2nd ACM Intl. Conference on Web Search and Data Mining, pp. 242-251.

Nakayama, M., Sutcliffe, N. & Wan, Y. (2010). Has the Web transformed experience goods into search goods*?. Electronic Markets* 20(3-4), pp. 251-262. doi: 10.1007/s12525-010-0041-z

Nelson, P. (1970). Information and Consumer Behavior. *Journal of Political Economy* 78(2), pp. 311-329.

Nogoev, A., Menon, M., Samadi, B., Mohseni, S. & Yazdanifard, R., (octubre de 2011). *The*

Raschka, S., (2015), *Python Machine Learning,* UK: Packt Publishing.

Palma, P. (15 de marzo de 2016). Big data: la revolución económica de la información. Forbes. Recuperado de: https://www.forbes.com.mx/big-data-la-revolucion-economica-la-informacion/

Rubels, S. (2005). Social Commerce. *Social commerce today.* Recuperado de: http://socialcommercetoday.com/steve-rubels-original-2005-social-commerce-post/

Riquelme, J., Ruiz, R., Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. *Revista Iberoamericana de Inteligencia Artificial* 10(29), pp. 11-18. Recuperado de: https://www.researchgate.net/publication/28140441_Mineria_de_Datos_Conceptos_y_Tendencias [accessed Aug 24 2018].

Rubel, S. (2005). 2006 Trends to Watch Part II: Social Commerce. *Micro Persuasion*,

Standifird, S.S (2001). Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings. *Journal of management* 27(3), pp. 279-295.

Stark, N. S. (2011). *Motores de búsqueda en Internet* (Trabajo Fin de Investigación). Recuperado de http://www.unlu.edu.ar/~tyr/tyr/TYR-motor/stark-motor.pdf).

Stauss, B. (2000) Using new media for customer interaction: a challenge for relationship marketing. In T. Hennig-Thurau & U. Hansen (eds) Relationship Marketing. Berlin: Springer, pp. 233–253.

Sundaram, D.S., Mitra, K., & Webster, C. (1998). Word of-Mouth Communications: A Motivational Analysis. *Advances in Consumer Research, 25*, 527–531.

Talib, R, Hanif, M. K., Ayesha, S. & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *(IJACSA) International Journal of Advanced Computer Science and Applications,* 7(11), pp. 414-418.

Tapscott, D. (1995). *The digital Economy: Promise and Peril in the Age of Networked Intelligence*. US: McGraw-Hill.

Toral, S., Martínez-Torres, M. R., Gonzalez-Rodriguez, M. R. (2017). "Identification of the Unique Attributes of Tourist Destinations from Online Reviews." *Journal of Travel Research,* 57 (7): 908-919.

Valls, J. F., (2017), *Big Data: atrapando al consumidor,* Barcelona: Profit Editorial.

Vara, D. & Cuza, I., (2012). The search experience credence product classification paradigm in the eyes of the electronic consumer. *Management & Marketing Challenges for the Knowledge Society, 7*(3), pp. 449-464.

Vikram, A. (2012). E-commerce: Opportunities and Challenges. En A. Vikram (presidencia), Conference: National Conference, Department of Management Studies of Koshys Institute of Management Studies. Congreso llevado en Bangalore, India. Recuperado de: https://www.researchgate.net/publication/273455693_E-commerce_Opportunities_and_Challenges_ISBN978-81910530-3-6

Weinberger, M.G. & Dillon, W.R. (1980). The effect of unfavorable product rating information. *Advances in Consumer Research,7*(1), pp. 528-32.

Wilson, P. & Campbell, L. (2016). Developing a knowledge management policy for ISO 9001:2015. *Journal of knowledge management* 20(4), pp. 829-844.

Wong, S.C., Gatt, A., Stamatescu, V., & McDonnell, M.D. (2016). Understanding Data Augmentation for Classification: When to Warp? *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1-6.

Zeithaml, V. (1981). *Marketing of Services*. McGraw-Hill Education.