UNIVERSIDAD DE GRANADA

Programa de Doctorado en Ciencias Económicas y Empresariales

Línea de investigación en Marketing y Consumo

Diseño, construcción y validación de LOGOS: una herramienta basada en el análisis de sentimientos multilingüe como apoyo a la toma de decisiones de marketing

Tesis Doctoral

Alysson Filipe Steiner Corrêa

Granada, 2019

Editor: Universidad de Granada. Tesis Doctorales
Autor: Alysson Filipe Steiner Corrêa
ISBN: 978-84-1306-368-3
URI: http://hdl.handle.net/10481/58116

UNIVERSIDAD
DE GRANADA

Diseño, construcción y validación de LOGOS: una herramienta basada en el análisis de sentimientos multilingüe como apoyo a la toma de decisiones de marketing

MEMORIA PRESENTADA POR
Alysson Filipe Steiner Corrêa

PARA OPTAR AL GRADO DE DOCTOR EN
CIENCIAS ECONÓMICAS Y EMPRESARIALES
Granada, 2019

DIRECTORES

María Isabel Viedma del Jesús

Comercialización e Investigación de Mercados

Antonio Gabriel López Herrera

Ciencias de la Computación e Inteligencia Artificial

Esta tesis doctoral ha sido realizada con el apoyo del ***Programa Observatório da Educação, de la Coordenação de Aperfeiçoamento de Pessoal de Nível Superior de la institución CAPES/Brasil*** a través de la beca Doutorado pleno no exterior.

La beca tiene como finalidad ofrecer recursos económicos para cubrir gastos de permanencia en el programa de doctorado, desplazamiento, acomodación y seguro de salud, además de un pago mensual a estudiantes brasileños de doctorado en universidades extranjeras por un periodo máximo de 48 meses. Esta beca de numero 99999.002230/2015-01 fue disfrutada de 01/07/2015 al 30/06/2019.

Dedicada a mi padre y a mi madre:
Álvaro Corrêa y Wiviane Steiner Corrêa

Y a mi hermano:
Emerson Vicente da Cruz

Agradecimientos

Esta tesis está especialmente dedicada a mis padres, D. Álvaro Corrêa y Dña. Wiviane Steiner Corrêa, porque con ella materializo un sueño que parecía utópico y distante. A mis dos hermanas Alany y Camylla y a mi hermano/mentor Emerson, por su soporte leal y entusiasta durante todos estos años.

También la dedico a mis directores Marisa y Antonio, por tratarme como más que un pupilo, como un amigo. Su apoyo, experiencia y cariño, fueron la piedra angular en la realización de este trabajo.

Finalmente, a todos aquellos que de manera directa o indirecta estuvieron a mi lado a cada paso dado, asegurándose de que no cayera en momento alguno.

Sin vosotros esto aún sería sueño.

GRACIAS
OBRIGADO

ÍNDICE

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

GLOSARIO

AI - *Artificial Intelligence*: Inteligencia Artificial. Conjunto de técnicas, algoritmos y métodos que son usados para emular ciertos procesos (normalmente informáticos) de manipulación de datos para la realización de tareas complejas.

Algoritmo: Palabra de origen árabe que hace referencia al matemático Al-Khwarizmi. Que determina el conjunto de instrucciones utilizadas en la ejecución de una tarea o resolución de un problema.

API: Interfaz de Programación de Aplicaciones, del inglés *Application Programming Interface*. Rutinas usadas por una aplicación para gestionar por lo general servicios de bajo nivel, llevados a cabo por el sistema operativo de la computadora. También usados para la interconexión de sistemas en la red, habitualmente en arquitecturas software de tipo cliente – servidor.

C: Lenguaje de programación de tipo compilado que ha sido y es muy usado por la industria del software para el desarrollo de sistemas. Es muy usado también para el desarrollo de librerías que a su vez son empleadas por otros lenguajes de programación.

Código fuente: Instrucciones escritas que integran un programa informático, y que son fácilmente editables e intercambiados y compartidos en la red.

CRM - *Customer Relationship Management*: Manejo de la Relación con el Consumidor. Sistema información automatizado que permite atender a clientes de manera personalizada. Internet es uno de los soportes tecnológicos más importantes en CRM, a la vez que uno de sus principales canales de comunicación con los clientes.

CSV - *Comma Separated Values*: Valores Separados por Comas. Forma de estructurar bases de datos utilizando comas (,) como separadores de columnas en archivos de texto

Data: Nombre que se da para cualquier información que transita en un ordenador o cualquier otro medio digital

Favoritos: Marcador de un sitio web, del inglés *bookmark*. Donde se puede guardar direcciones de sitios o cualquier otra información web preferidos

Hipervínculo: Enlace del tipo hipertexto que apunta a otro documento.

HTML: Metalenguaje informático que es usado para describir la estética, y en cierto modo el contenido, de una página web.

Interface: Interfaz o interface se trata del punto de conexión, entre los usuarios y los programas informáticos. También hace referencia al sistema, habitualmente de ventanas, en la que se confecciona una sistema informático, y que facilita la interacción con el mismo.

Login: Clave de identificación y acceso a los recursos de un ordenador o programa informático asignada a un usuario.

Medio social Web: Medio por lo cual se establecen la relaciones sociales en la web, referido fundamentalmente a las plataformas tecnológicas que lo hacen posible. Por ejemplo, Facebook, Twitter, Instagram, etc.

MongoDB: Sistema Gestor de Bases de Datos de tipo documental ampliamente usado para la confección y sobre todo el almacenamiento de grandes volúmenes de datos en contextos heterogéneos.

MySQL: Sistema Gestor de Bases de Datos conocido por ser extremadamente rápido y versátil.

Noe4J: Sistema Gestor de Bases de Datos extremadamente versátil para el modelado y trabajo de datos con estructura en red.

Open Source: Hace referencia a un programa cuyo código fuente está disponible de manera gratuita, para uso y modificación.

Paquete: Grupos de información que se transmite por una red. También hace referencia a la agrupación de librerías software, que son a su vez usadas como extensión de otros lenguajes de programación.

PHP: Lenguaje de programación de tipo interpretado que es usado para el desarrollo de páginas web.

Python: Lenguaje de programación de tipo interpretado ampliamente usado para análisis de textos y procesamiento del lenguaje natural. Hoy en día es el lenguaje de moda debido a su facilidad de uso y abundancia de librerías.

R: Lenguaje de programación de tipo interpretado ampliamente usado en el ámbito estadísticos y muy extendido en los últimos años para el análisis de datos.

Red: Conjunto de dos o más computadoras interconectadas. También hace referencia al nombre coloquial para referirse a Internet.

Shiny: Entorno de desarrollo web ligado al lenguaje de programación R. De hecho es una extensión o paquete de este lenguaje.

Servidor Web: Programa y el hardware que lo ejecuta, que maneja los dominios y páginas web, por medio de lenguajes como HML y PHP, entre otros.

SQL - *Structured Query Language:* Lenguaje de programación que permite realizar consultas a bases de datos.

RESUMEN

El nuevo panorama económico, social y tecnológico nos está llevando a una nueva forma de hacer marketing, pasando del Marketing 2.0 (centrado en el consumidor) al Marketing 3.0 (centrado en valores) (Kotler, Kartajaya, & Setiawan, 2010). La situación de inestabilidad económica de los últimos años está afectando al consumidor en sus decisiones de consumo, volviéndose a su vez mucho más escépticos respecto a las prácticas de marketing. El resultado es que hoy en día la confianza se da más en las relaciones horizontales que en las verticales, y los consumidores confían más en otros consumidores que en las empresas (Kotler, Kartajaya & Setiawan, 2010). Además, con el desarrollo de las nuevas tecnologías, Internet no solamente es un canal fundamental para los consumidores a la hora de buscar información y comparar alternativas, sino que también se ha convertido en un lugar donde buscar la experiencia y el consejo de otros consumidores, fomentando dichas relaciones horizontales. Con el desarrollo de los medios sociales, el consumidor cobra cada vez un papel más activo, convirtiéndose en un generador de contenidos y opiniones con relación a las marcas (Vela & Riera, 2013).

Estamos hablando de la emergencia de un nuevo consumidor que no solamente está más informado, sino que es más exigente, más crítico frente a los mensajes de las marcas y que le gusta que la empresa tenga en cuenta sus opiniones y ser coprotagonista del diseño y creación de los productos y servicios que va a consumir (Canales & Hernández, 2013). Además, se trata de un consumidor global, que ya no tiene reparos en comunicarse e

informarse de manera abierta, e interactuar con otros consumidores de todo el mundo. De modo que las empresas, si quieren recuperar la confianza del consumidor, es fundamental que entiendan que nos encontramos en la era de la globalización y de la participación donde el consumidor no solamente tiene un rol pasivo sino que también lleva a cabo de manera proactiva actividades de marketing, valorando cada vez más la colaboración e implicación en el proceso de creación de la marca (Kotler, Kartajaya & Setiawan, 2010). Ante este panorama, las empresas tienen un reto añadido, escuchar a sus clientes actuales y potenciales, hablen el idioma que hablen.

Este nuevo panorama del marketing está dando lugar al surgimiento de nuevas herramientas de investigación de mercados. La irrupción de Internet, y en particular el auge de los medios sociales conlleva grandes ventajas para las empresas, ya que permite la recolección de grandes cantidades de información respecto a los consumidores que utilizan estos medios para expresar sus opiniones sobre productos y servicios. Éstos comparten de manera voluntaria una gran variedad de datos, obsequiando a las empresas con información que sirve de base para nuevas ideas de productos, servicios, posicionamiento en el mercado, entre otros. Tal escenario sitúa el marketing en una nueva era, la era digital, facilitando enormemente el conocimiento del consumidor a través de la recolección de grandes volúmenes de información (Big Data).

No obstante, para que las empresas puedan ser competitivas - en esta nueva era del marketing -, no es suficiente con recopilar toda esta

información, sino que estos datos (Big Data) deben ser transformados en conocimiento que sirva para tomar mejores decisiones para el negocio (Smart Data) (Merchanteí, 2015). Según el informe de la OBS (Online Business School) en su estudio "Big Data 2016", el "65% de las empresas saben que corren el riesgo de convertirse en irrelevantes o no competitivas si no adoptan Big Data" (Maroto, 2016). A este respecto, las nuevas tecnologías aplicadas a la Minería de Datos (MD) se presentan como un mecanismo ideal para realizar tal tarea. Chen y Zimbra (2010) la definen como un soporte que aporta técnicas de extracción, clasificación, procesamiento y comprensión de opiniones manifestadas en diversas fuentes de información online y comentarios en medios sociales.

Se han desarrollado multitud de herramientas de minería de datos web, de monitoreo de medios sociales, y de minería de opiniones, las cuales son utilizadas en la actualidad en gran medida por las grandes empresas de forma estratégica para conocer a los consumidores y fomentar las relaciones de intercambio. Sin embargo, estas herramientas son la mayoría de las veces inaccesibles para las pequeñas y medianas empresas (PyMEs), y en particular para las de países en vías de desarrollo, dados sus, a veces, limitados recursos económicos y humanos. En este contexto, el objetivo de la presente tesis doctoral es diseñar y construir, en base a técnicas de minería de datos, minería de opiniones y monitoreo de medios sociales una herramienta de inteligencia de mercados, llamada `LOGOS`, que pueda ser usada de manera gratuita por las PyMEs.

Con este propósito principal, en esta tesis doctoral se han llevado a cabo cuatro aportaciones. En la primera aportación se ha realizado, tras una exhaustiva revisión de la literatura, un compendio de datasets en múltiples idiomas que han sido validados subjetivamente por la comunidad científica. Estos datasets se han utilizado como recurso de base para la segunda aportación, en la que los principales algoritmos automáticos de análisis de sentimientos han sido entrenados y comparados en su faceta de analizar textos según idioma. Con estos resultados aportamos la tercera contribución, el diseño y construcción de LOGOS, una herramienta de Análisis de Sentimientos multilingüe, la cual ha sido validada en la última aportación, realizando aplicaciones con datos reales de medios sociales para dos empresas de carácter multinacional.

Los resultados han puesto de manifiesto el potencial de la herramienta en el contexto de este tipo de análisis y han desvelado futuras acciones de mejora para su posterior evolución. Al tratarse de una herramienta de código abierto, destaca por su flexibilidad y facilidad de uso, por lo que las PyMEs podrán adaptarla en función de sus necesidades.

En definitiva, el gran reto del marketing en el futuro no es solo conocer bien a los nuevos consumidores sino ganarse la confianza de un consumidor muy bien informado, muy influyente y cada vez más escéptico en relación con las prácticas llevadas a cabo por las empresas. Con el desarrollo de LOGOS esperamos ayudar a que las PyMEs, gracias a sus posibilidades de análisis, puedan acercarse más a sus consumidores, ganarse su respeto y

confianza, y de este modo favorecer su supervivencia en los mercados globales y competitivos actuales.

ABSTRACT

The new economic, social and technological scenario is leading us to a new way of dealing with marketing, passing from Marketing 2.0 (consumer focused) to Marketing 3.0 (values focused) (Kotler, Kartajaya & Setiawan, 2010). The economic instability of recent years is affecting the consumer in their purchasing decisions, making them much more sceptical about marketing practices. The result is that, nowadays, consumer's trust is given more in a horizontal way than in a vertical way, and consumers rely more on each other than on companies (Kotler, Kartajaya & Setiawan, 2010). In addition, with the development of new technologies, Internet is not only a fundamental channel for consumers to look for information and to compare alternatives, but it has also become a place to search for the advice, and experience of other consumers, promoting these horizontal relationships. With the development of social media, consumers are increasingly taking on a more active role, becoming a generator of content and opinions regarding brands (Vela & Riera, 2013).

We are describing the needs of a new and more up to date consumer, more demanding and more critical against the messages of the brands and who likes the company to take into account their opinions and act as a co-producer of the design and creation of the products and services they will consume (Canales & Hernández, 2013). Furthermore, it is a global consumer, who has no hesitance about openly communicate to get informed and interact with other consumers around the world. In that way, if companies want to regain consumer's trust, it is essential that they

understand we live in the era of globalization, where the consumer not only has a passive role but also proactively performs marketing activities, valuing more their participation with the brand creation process (Kotler, Kartajaya & Setiawan, 2010). Given this overview, companies have an added challenge; to listen to their current and potential customers, and speak the same language as them.

This new marketing prospect is leading to the emergence of new market research tools. The establishment of the Internet, and in particular the rise of social media entails great advantages for companies since it allows the collection of large amounts of information regarding consumers who use the media to express their opinions about products and services. Consumers voluntarily share a large variety of data, giving companies information that serves as the basis for new ideas for products, services, market positioning, etc. Such a scenario places marketing in a new era, the digital era, greatly increasing consumer knowledge through the collection of massive volumes of information (Big Data).

However, for companies to be competitive - in this new era of marketing - it is not enough to collect all this information. This data (Big Data) should be transformed into knowledge that serves to take better decisions to the business (Smart Data) (Merchanteí, 2015). According to the OBS report (Online Business School) on its study "Big Data 2016", "65% of companies know that they are in risk of becoming irrelevant or non-competitive if they do not adopt Big Data" (Maroto, 2016). In this sense, the new technologies applied to Data Mining (MD) are introduced as an ideal mechanism to

perform such a task. Chen and Zimbra (2010) define it as support that provides techniques for extracting, classifying, processing and understanding opinions expressed in a variety of online information sources and comments on social media.

A large amount of web data mining, social media monitoring, and opinion mining tools have been developed, which are currently largely and strategically used by big companies to get to know consumers and foster exchange relationships. However, these tools are mostly inaccessible for small and medium-sized enterprises (SMEs), and in particular for those in developing countries, due to their, sometimes, limited economic and human resources.

In this context, the aim of this doctoral thesis is to design and build, based on data mining techniques, opinion mining and social media monitoring, a business intelligence tool, called LOGOS, that can be used for free by the SMEs. With this main purpose, four contributions have been made in this doctoral thesis. In the first contribution, a survey of datasets in multiple languages that were subjectively validated by the scientific community has been carried out after an exhaustive review of the literature. These datasets were used as a base resource for the second contribution, in which the main automatic sentiment analysis algorithms were trained and compared in their role of analyzing texts according to language. With these results we move forward to the third contribution, the design and construction of LOGOS, a multilingual Sentiment Analysis tool, which was validated in the

last contribution, making real data applications of social networks for two multinational companies.

The results brought into light the potential of the tool in the context of this type of analysis and exposed future improvement actions for its subsequent evolution. As per being an open-source tool, it stands out for its flexibility, and straightforward, so that SMEs can adapt it according to their needs.

In short meaning, the big challenge of marketing in the future is not only to know new consumers very well but as well to gain the trust of a very knowledgeable, influential and increasingly sceptical consumer in relation to the practices carried out by companies. With the establishment and development of LOGOS, we hope to help SMEs, acknowledging its analysis possibilities, to get closer to their consumers, gain their respect, and trust. Therefore, supporting their survival in the current global and competitive markets.

RESUMO

O novo cenário econômico, social e tecnológico nos direciona à uma nova maneira de fazer marketing. Passamos do Marketing 2.0 (com foco no consumidor) ao Marketing 3.0 (centrado na criação de valor) (Kotler, Kartajaya & Setiawan, 2010). Existe muitas variáveis afetam e mudam o comportamento do consumidor. Neste estudo, destacamos a situação de instabilidade econômica dos últimos anos, que vem afetando o consumidor e suas decisões de compra, tornando-o muito mais cético em relação às práticas de marketing. O resultado é que hoje em dia, a confiança se estabelece mais nas relações horizontais do que nas verticais, de modo que os consumidores confiam mais em outros consumidores do que nas próprias empresas (Kotler, Kartajaya & Setiawan, 2010). Além disso, com os avanços das novas tecnologias, a Internet deixa de ser apenas um canal fundamental para os consumidores quando se trata de buscar informações e comparar alternativas, e passa a ser também um local onde buscam conselhos e experiências de outros consumidores, promovendo essas relações horizontais. Com o desenvolvimento das mídias sociais, os consumidores assumem um papel cada vez mais ativo, como atores que produzem conteúdo e opinam sobre as marcas (Vela & Riera, 2013).

Estamos falando do surgimento de um novo consumidor que ademais de ser mais informado, mais exigente e muito mais crítico em relação às mensagens das marcas, valoriza as empresas quem levem em consideração suas opiniões e que possa ser co-participante no design e na criação dos produtos e serviços que irá consumir (Canales & Hernández, 2013). Além

disso, se trata de um consumidor global, que não tem medo de se informar, se comunicar abertamente e interagir com outros consumidores do mundo inteiro. Portanto, se as empresas desejam conquistar a confiança deste novo consumidor, é essencial que entendam que estamos na era da globalização e da participação. Neste contexto, o consumidor não apenas desempenha um papel passivo, mas também realiza ações de marketing de forma proativa, valorizando cada vez mais a colaboração e o envolvimento no processo de criação das marcas (Kotler, Kartajaya & Setiawan, 2010). Diante desse cenário, as empresas têm um desafio adicional, ouvir aos seus atuais e potenciais clientes, independentemente do idioma que falem.

Esse novo panorama de marketing impulsiona o surgimento de novas ferramentas de pesquisa de mercado. A Internet, e em particular a ascensão das mídias sociais, trazem grandes vantagens para as empresas, pois permitem a coleta de informações sobre os consumidores que utilizam cada vez mais essas mídias para expressar suas opiniões e recomendações. Eles compartilham de forma voluntária uma ampla variedade de dados, fornecendo às empresas informações que servem de base para novas ideias de produtos, serviços, posicionamento de mercado, entre outras. Essa realidade posiciona o marketing em uma nova era digital, facilitando substancialmente o conhecimento do consumidor por meio da coleta de grandes volumes de informações (Big Data).

No entanto, para que as empresas sejam competitivas - nesta nova era do marketing -, não basta apenas coletar toda essa informação. É necessário que esses dados (Big Data) sejam transformados em

conhecimento que sirva para tomar melhores decisões de negócios (Smart Data) (Merchanteí, 2015). De acordo com o relatório da OBS (Online Business School) em seu estudo "Big Data 2016", "65% das empresas sabem que correm o risco de se tornarem irrelevantes ou não competitivas se não tiverem em conta o manuseio e o tratamento do Big Data" (Maroto, 2016). Nesse sentido, as novas tecnologias aplicadas à Mineração de Dados (MD) são apresentadas como um mecanismo ideal para executar essa tarefa. Chen e Zimbra (2010) a definem como um suporte que fornece técnicas para extrair, classificar, processar e entender opiniões expressas em várias fontes de informações on-line e comentários das mídias sociais.

Existem diversas ferramentas para mineração de dados na web, monitoramento de mídias sociais e mineração de opiniões. Atualmente, essas ferramentas são amplamente utilizadas de maneira estratégica por grandes empresas para atender aos consumidores e estreitar relações. No entanto, essas ferramentas muitas vezes são inacessíveis, principalmente para pequenas e médias empresas (PMEs) e, particularmente, para empresas de países em desenvolvimento devido aos recursos econômicos e humanos às vezes limitados.

Nesse contexto, o objetivo desta tese de doutorado é projetar e construir, com base em técnicas de mineração de dados, mineração de opiniões e monitoramento de mídias sociais, uma ferramenta de inteligência de mercados chamada LOGOS, que possa ser usada gratuitamente pelas PMEs. Com esse objetivo principal, esta obra aporta quatro contribuições principais. Na primeira contribuição, após uma ampla

e minuciosa revisão da literatura, foi elaborado um compêndio de *datasets* de sentimentos em múltiplos idiomas e validados subjetivamente pela comunidade científica. Esse conjunto de dados foi utilizado como recurso base para a segunda contribuição, na qual os principais algoritmos de análise automática de sentimentos foram treinados e comparados com base em distintos idiomas. Com esses resultados, realizamos a terceira contribuição, o design e a construção da LOGOS, uma ferramenta de Análise de Sentimentos multilíngue que foi validada com a última contribuição, através de aplicações práticas com dados reais provenientes de mídias sociais de duas empresas multinacionais.

Os resultados revelaram o potencial da ferramenta no contexto de análise de mercado e potenciais ações de melhoria para futuras evoluções do sistema. Por se tratar de uma ferramenta de código aberto, LOGOS se destaca por seu aspecto flexível e facilidade de uso, de modo que as PMEs possam adaptá-la de acordo com suas necessidades.

Em suma, o grande desafio do marketing no futuro não é apenas conhecer bem os novos consumidores, mas ganhar a confiança de um consumidor experiente, influente e cada vez mais cético em relação às práticas de marketing realizadas pelas empresas. Neste cenário, o uso da ferramenta LOGOS oferece diversas possibilidades de análises. Se espera com isso, ajudar as PMEs a se aproximarem dos consumidores, ganhando seu respeito e confiança, o que favorece a sobrevivência dessas empresas nos mercados globais e competitivos de hoje.

ESTRUCTURA DE LA MEMORIA

Esta tesis doctoral está organizada en los siguientes nueve capítulos:

- **Capítulo 1.** Abarca los fundamentos introductorios y justificación de esta tesis. En este primer capítulo se describe brevemente el trabajo previo del que parte la tesis y se abordan de forma general los elementos de base que articulan este trabajo: el concepto Big Data y el uso de los medios sociales por parte de las empresas.

- **Capítulo 2.** Constituye el marco conceptual en el que se apoya esta tesis. En este segundo capítulo se realiza una exhaustiva revisión bibliográfica de los distintos enfoques, recursos y tareas relacionadas con el descubrimiento de conocimiento en bases de datos, la minería de datos/opiniones y el análisis de sentimientos a través de algoritmos de aprendizaje automático y las herramientas de monitoreo de medios sociales. Finalmente se detallan los principales trabajos de investigación llevados a cabo en el ámbito objeto de estudio, así como las principales aplicaciones al marketing de la Minería de Opiniones y el Análisis de Sentimientos para enmarcar la línea de desarrollo de este trabajo.

- **Capítulo 3.** Se describen el objetivo principal de esta tesis y los objetivos específicos derivados de cada una de las aportaciones desarrolladas. En segundo lugar, se detallan los aspectos

metodológicos de la investigación. Concretamente, se abordan los diferentes métodos aplicados en la ejecución de las diferentes aportaciones propuestas para dar respuesta a las cuestiones de investigación planteadas en esta tesis.

- **Capítulo 4.** Se describe la **Aportación 1,** en la que se realiza una exhaustiva revisión de la literatura para elaborar un compendio de recursos de base (*datasets*) relacionados con el Análisis de Sentimientos automáticos en múltiples idiomas y validados científicamente. Tales recursos fueron fundamentales para la realización de los análisis de la aportación siguiente.

- **Capítulo 5.** Este capítulo corresponde a la **Aportación 2** en la que se trabaja la experimentación con métodos automáticos de análisis de sentimientos. Se prueban diferentes algoritmos automáticos con diferentes grupos de datos (*datasets obtenidos de la aportación 1*) de sentimientos para obtener sus niveles de bondad en relación con precisión y rapidez para diferentes idiomas. Este estudio tuvo como base los índices de clasificación presentados en las tareas de Análisis de Sentimientos para determinar cuál algoritmo – según el idioma - debe ser utilizado por la herramienta web ideada y desarrollada a continuación. Los resultados se analizan estadísticamente para comprobar la robustez de los métodos.

- **Capítulo 6.** Este capítulo corresponde a la **Aportación 3** donde se diseña, desarrolla e implementa el sistema LOGOS: un sistema web de inteligencia empresarial que, aplicando técnicas de Minería de Datos y Análisis de Sentimientos sobre datos multilingües procedentes de medios sociales, sirva como soporte de decisiones estratégicas de marketing para las empresas. En el capítulo se detalla su construcción e implementación en términos de diseño y de funcionalidad. Se describen todas las fases del proceso de creación y ampliación del sistema, así como los matices técnicos pertinentes a la programación informática de LOGOS.

- **Capítulo 7.** Este capítulo corresponde a la **Aportación 4** donde se lleva a cabo la etapa de validación de LOGOS por medio de cuatro aplicaciones del sistema. Concretamente, se realiza un análisis y una interpretación de los datos arrojados por la herramienta a modo de ejemplo de cómo este conocimiento puede ser utilizado por las empresas en su día a día. Las aplicaciones se realizaron utilizando los mensajes descargados del medio de micro mensajes Twitter relacionados a dos empresas. Fueron generados informes con base a esos grupos de datos que sirven como material de apoyo a las decisiones de marketing para estas empresas.

- **Capítulo 8.** En este capítulo se presenta el repositorio, que se ha elaborado para poner a disposición de la comunidad todas

las aportaciones de la presente memoria doctoral, así como otra información asociada.

- **Capítulo 9.** En este capítulo final, se exponen las principales conclusiones y reflexiones relativas a esta tesis doctoral en función de los resultados obtenidos en las diferentes aportaciones. Se presentan y se discuten también las limitaciones relacionadas, así como las posibles futuras líneas de actuación derivadas de este trabajo.

CAPÍTULO 1: INTRODUCCIÓN

1.1. TRABAJO PREVIO

1.2. JUSTIFICACIÓN DE LA TESIS

Este capítulo introductorio se centra en describir los fundamentos y aspectos más relevantes que explican y justifican la realización de esta tesis, centrada principalmente en el fenómeno de los medios sociales y su influencia en el ámbito de las empresas.

1.1. TRABAJO PREVIO

La presente tesis ha sido desarrollada dentro del marco del Programa de Doctorado en Ciencias Económicas y Empresariales de la Universidad de Granada, dentro de la línea de investigación en Marketing y Consumo.

Concretamente, este trabajo busca dar continuidad al estudio precedente titulado ***DANDO ALAS A Red Bull: Una aplicación de la Minería de Opiniones, para conocer qué piensan y de qué hablan los seguidores de la marca en Twitter***, llevado a cabo como Trabajo Fin de Máster (TFM) en el *Máster Universitario en Marketing y Comportamiento del Consumidor* por la Universidad de Granada (UGR) y la Universidad de Jaén (UJA) (Steiner-Correa, 2017).

El Trabajo Fin de Máster desarrollado tuvo como principal objetivo la extracción, filtrado y clasificación de mensajes cortos provenientes de la red de micro mensajes Twitter. Concretamente, se llevó a cabo una aplicación para evaluar las opiniones manifestadas por seguidores de la marca de bebidas energéticas Red Bull a través de los tuits generados en esta red en el periodo de análisis considerado. Los resultados permitieron detectar los ítems de interés del público de la marca, así como los sentimientos

expresados sobre productos y servicios a través de un análisis de la polaridad (positivo, negativo o neutro) de los comentarios, comparando para ello clasificaciones realizadas por algoritmos automáticos vs percepciones subjetivas (medidas a través de un cuestionario). Los resultados obtenidos pusieron de manifiesto cómo la correcta extracción, almacenaje y procesamiento de la información presente en los medios sociales resulta útil como base de conocimiento para que las empresas entiendan cada vez más al detalle a sus clientes, sus comportamientos y expectativas hacia la marca y que, partir de este conocimiento, puedan tomar decisiones de mercado cada vez más acertadas.

1.2. JUSTIFICACIÓN DE LA TESIS

Caracterizadas por proporcionar medios que estimulan la creación colaborativa de contenido (Lévy, 2004) y fomentan interacción social (Kaplan & Haenlein, 2010), los medios sociales permiten a los internautas expresar opiniones y criticar o recomendar productos y/o servicios (Hunt, 2010), generando una especie de "economía de reputación" (Anderson, 2006). Así pues, estos medios pueden actuar como un barómetro midiendo cómo una organización es vista y sentida por sus clientes. Se trata de un espacio donde el consumidor gana voz y el diálogo con las empresas pasa a ser directo y constante (Dourado, 2010), ofreciendo a la vez información relevante a las organizaciones que apuestan por un enfoque relacional (Kotler & Armstrong, 2008).

En tal escenario, se hace indispensable el seguimiento constante de los medios sociales en busca de ideas importantes que aporten beneficios a ambas partes, empresas y clientes. En ese sentido, el abanico de recursos que aporta la Minería de Datos (MD) (Liu 2015) y sus variantes, se presentan como recursos capaces de transformar estos datos en conocimiento competitivo para las empresas (Chen & Zimbra, 2010).

1.2.1. EL FENÓMENO DE LOS MEDIOS SOCIALES

Tras la evolución a la Web 2.0 (O'Reilly, 2007), las páginas de Internet dejaron de ser independientes en términos de contenidos y servicios, pasando a constituir entornos que reúnen y comparten diversos recursos con sus usuarios. Este fenómeno, amplió y facilitó el proceso de interacción entre los internautas creando una red humana conectada a nivel global (Berners-Lee, Fischetti, & Foreword By- Dertouzos, 2000). Esta realidad permitió la interacción entre los subscriptores y cedió paso a la creación de una red global humana, donde la información se crea y se difunde también por sus usuarios (UGC, *User Generated Content)*, dando paso a la denominada inteligencia colectiva (O'Reilly, 2007).

Lo que antes era limitado a personas de alto nivel técnico, pasa a ser accesible a cualquier usuario de la red. Semejante escenario, creó un ambiente óptimo para el florecimiento de distintas redes que, ahora más que nunca, conectan no sólo ordenadores sino también personas con sentimientos, pensamientos e intenciones.

Tenzer, Ferro y Palacios (2009) argumentan que una red social corresponde en su sentido más amplio a un punto de encuentro de organizaciones, asociaciones e individuos que utilizan la interacción para compartir contenido, generar respuestas o analizar problemas de forma colaborativa. Además, es evidente el deseo manifestado por los consumidores actuales de participar activamente de la producción y consumo de la información, **co-creando y coparticipando** junto a las marcas, con sugerencias y opiniones sobre los productos de su interés estimulando la generación de valor. Esto hace que éstos asuman el papel de productores y consumidores a la vez, es decir, pasan de ser consumidores a "prosumidores[1]". Este fenómeno resultado de los cambios sociales, económicos y tecnológicos, enmarca un cambio en el comportamiento del consumidor que surge tras la aparición de la Web 2.O, los Medios Sociales y los avances de las nuevas Tecnologías de la Comunicación, conceptos que enmarcan el **Marketing 3.0** (Kotler et al., 2010)**.**

Actualmente, es extenso el abanico de medios sociales. *Facebook*, por ejemplo, permite la publicación de fotos y vídeos además de la creación de grupos, eventos, páginas personales y también comerciales. A pesar de ser un medio social con enfoque más personal, viene adaptándose

[1] El concepto de "prosumer" (prosumidor) fue introducido por Alvin Toffler en los 80 y se refiere al consumidor/a que abandona la faceta pasiva para convertirse en generador de contenidos y creador de ideas y opiniones que ejercen influencia a la comunidad de compradores de una marca o un producto.

fuertemente para uso comercial y profesional. A su vez, *YouTube* es un medio social que tiene el enfoque principal en compartir vídeos, sin embargo, también ofrece hoy en día servicios de "*streaming"* de música. Además de permitir que los usuarios suban vídeos, también permite a sus usuarios que comenten e interactúen con los contenidos en su plataforma. Por otra parte, *Twitter* es un medio del tipo micro-mensajes que se destaca por su dinamismo ya que permite la publicación de mensajes con un máximo de 280 caracteres en tiempo real, lo que promueve su aspecto conciso y dinámico en la red.

Por otro lado, existen medios más segmentados que tratan de temas específicos como, por ejemplo, *MySpace* que por su enfoque musical tiene éxito entre los usuarios que se dedican a la música como cantantes, grupos musicales y aficionados, entre otros. *LinkedIn*, tiene un enfoque bastante profesional. Este medio permite la creación de un Currículo Vitae online y permite a sus usuarios compartir experiencias profesionales, buscar trabajo e incluso la captación de oportunidades de negocios. *Instagram* y *Pinterest* aparecen para atender a los más adeptos de las imágenes, fotos y videos cortos tipo Gif. Estas dos redes atienden a aficionados y profesionales como tatuadores, artistas, arquitectos, diseñadores, fotógrafos, decoradores, personas relacionadas con el mundo de la moda, entre otros. También existen redes aún más especializadas como *Mendeley*, *ResearchGate* y *Academia*, que tratan de conectar personas interesadas en la investigación científica como profesores, académicos, investigadores e incluso aficionados a la ciencia.

Cada uno de estos medios traen consigo aspectos distintivos, no obstante, conservan el enfoque de red social. Dentro de esta variedad, esta tesis doctoral se centra fundamentalmente en el medio social del tipo micro-mensaje Twitter, debido principalmente a su aspecto dinámico, de comunicación inmediata y directa entre las empresas y sus consumidores.

1.2.2. El MEDIO DE MICRO-MENSAJES TWITTER

Creado en 2007, Twitter ocupa actualmente el puesto de mayor relevancia en redes de micro-mensajes. La empresa cuenta con más de 328 millones de usuarios activos al mes que generan más de 500 millones de tuits al día en más de 40 idiomas. Un *tuit* publicado puede ser replicado por otros usuarios en su muro de mensajes o también ser marcado como favorito por los mismos (Twitter, 2019).

Para conocer bien el mundo de Twitter es necesario entender algunos términos y expresiones que de manera frecuente son utilizados por sus usuarios:

- **Usuarios/Tuiteros**: personas que utilizan el medio social Twitter y que son identificadas en la red por medio del uso del símbolo @ + (su nombre en la red), por ejemplo, `@tesisdoctoral2019`.
- **Tuit**: se trata de un mensaje corto de hasta 280 caracteres que publica el usuario.
- **Tuitear:** verbo que se utiliza cuando se publica un mensaje en Twitter.

- **Retuit:** palabra que se utiliza para identificar al tuit que ha sido replicado por otros usuarios de la red.
- **Retuitear:** verbo que se utiliza cuando un usuario replica/reenvía un tuit publicado por otro usuario a sus seguidores en la red. En otras palabras, es la acción de hacer un retuit.
- **Seguidor**: persona que "sigue" a otra. Ser seguidor significa recibir en su muro de mensajes todos los Tuits que genera la persona a la cual sigues. En este sentido, la popularidad de un usuario puede ser conocida por el número de seguidores que este tiene.
- **Etiqueta (Hashtag)**: término que se utiliza para identificar o localizar a temas de interés. Su forma sintáctica es: # + nombre (sin espacios), por ejemplo `#tesisdoctoral2019`. De este modo se puede hacer referencia o "etiquetar" en la red a cualquier tema, frase o acontecimiento.
- **Tendencias (Trending topics)**: término utilizado para denominar un fenómeno de gran alcance/importancia en Twitter. Un tema puede llegar a ocupar un lugar en los trending topics cuando un gran número de usuarios comenta sobre él. En general se definen a través de #hashtags.
- **Mención**: término utilizado cuando un usuario menciona a otro usuario(s) en un tuit utilizando el @ + nombre del usuario mencionado. De esta manera, el usuario mencionado, recibe una notificación informativa sobre la referida mención.

Relacionado a las distintas funciones que puede tener Twitter, Carvalho (2010) destaca algunas como: a) conversación; b) participación,

principalmente en los flujos de información de carácter periodístico e informativo; c) actualización, especialmente en momentos donde es necesaria agilidad en la diseminación de la información; y d) difusión para proporcionar a las organizaciones un soporte mediático. Estas características facilitan al internauta abandonar la faceta pasiva para convertirse en generador activo de contenidos, aportando ideas y opiniones sobre productos y servicios que hayan experimentado, influenciando a la comunidad de compradores de una marca o producto.

Ante lo expuesto, debido a la importancia del tema para las organizaciones y ante la necesidad de trabajos académicos respecto a la utilización de Twitter en el ámbito de las empresas, esta tesis se centra en estos aspectos y contribuye al desarrollo de esta línea de investigación en el ámbito corporativo. A continuación, se profundiza en la relación de Twitter y las empresas y en la importancia de esta red social para las mismas.

1.2.3. TWITTER Y LAS EMPRESAS

Hoy día, los usuarios de los medios sociales comparten de manera voluntaria gran variedad de datos que resultan muy útiles para las transacciones comerciales. Estas manifestaciones ganan cada vez más protagonismo y destacan como valiosa fuente de recomendaciones, revisiones y opiniones (Anderson, 2006). De este modo, los usuarios obsequian a las empresas con información, la cual, bien recolectada y

clasificada correctamente, sirve como fundamentación para nuevas ideas de productos, servicios o posicionamiento de la marca, entre otros.

En este sentido, las empresas empezaron a adoptar Twitter como un ambiente idóneo para las prácticas del marketing relacional, buscando fortalecer las relaciones con los clientes a través del conocimiento de sus intereses, potenciando su propuesta de valor y el nivel de lealtad de las relaciones. Dourado (2010) comenta que los medios sociales suponen una importante transformación, puesto que eliminan al intermediario en las comunicaciones entre empresas y clientes, lo que proporciona una relación más estrecha y facilita la tarea de identificar las necesidades y deseos de los individuos. Tal y como apunta González-Fernández-Villavicencio (2015, p.6) "*la clave está en ser capaces de establecer un diálogo continuo y mantenido en el tiempo. Indiscutiblemente, hoy día las redes sociales son los canales de comunicación con más posibilidades*".

Por lo tanto, los medios sociales se han constituido como un soporte imprescindible para el desarrollo de las estrategias de marketing (Culnan, McHugh, & Zubillaga, 2010). No obstante, aunque cada vez son más las empresas que hacen uso de estos medios, en particular Twitter, como herramienta de comunicación de marketing, hoy día, el uso de los medios sociales de manera estructurada y estratégica es predominante en empresas de gran tamaño, debido a su disponibilidad de recursos humanos y fundamentalmente económicos. En contrapartida, en las pequeñas y medianas empresas (PyMEs), el escenario es bastante distinto y los medios sociales todavía son una herramienta en proceso de expansión en lo que se refiere a su utilización a nivel estratégico (López, Morales, & Cavero, 2018;

Marolt, Zimmermann, & Pucihar, 2018; Olvera-Lobo, Castillo-Rodríguez, & Gutiérrez-Artacho, 2018; Pérez, Carreras, & Bustamante, 2018). Debido quizás a su tamaño, las PyMEs cuentan con recursos económicos más limitados y suelen tener más dificultad a la hora de acceder a recursos tecnológicos que ayuden al manejo y análisis de la gran cantidad de información y conocimiento (explícito e implícito) disponible en los medios sociales. Si bien es cierto, que existen recursos informáticos a disposición pública que permiten el monitoreo de medios sociales y el análisis de los datos subyacentes, hasta nuestro conocimiento no existe una solución funcional flexible que facilite, al público en general y a las PyMEs en particular, la puesta en marcha y la ejecución de análisis de datos de medios sociales de una manera fácil y adaptable a las necesidades particulares de cada empresa, a bajo o nulo costo. Es por ello, que todavía se requiere de trabajos académicos como los que aborda la presente tesis doctoral.

1.2.4. SEGUIMIENTO DE MEDIOS SOCIALES – DEL *BIG DATA AL SMART DATA*

Debido a los competentes mercados actuales, incorporar estrategias de negocio exitosas es algo crucial en el escenario corporativo. Para ello, saber más sobre los clientes se hace indispensable. En la búsqueda de este "elixir", es imprescindible hoy día el uso activo y el seguimiento constante de los medios sociales a partir de la recolección, almacenamiento y procesamiento de datos. Para ello, la Minería de Datos (MD) se presenta como técnica ideal, capaz de recabar, almacenar y procesar grandes cantidades de información *(Big Data),* transformando datos aparentemente irrelevantes en *insights (Smart Data)* que sirvan para tomar

decisiones relevantes para el negocio (Merchanteí, 2015). Según el informe de la OBS (Online Business School) de su estudio titulado “Big Data 2016”, el 65% de las empresas saben que corren el riesgo de convertirse en irrelevantes o no competitivas si no adoptan las habilidades de manejo del *Big Data* por medio de la Minería de Datos (Maroto, 2016).

Chen & Zimbra (2010) definen la Minería de Datos como una herramienta que aporta técnicas de extracción, clasificación, procesamiento y comprensión de opiniones manifestadas en diversas fuentes de información online, medios sociales y otros mecanismos.

Algunas de las derivaciones de la Minería de Datos (MD) son la Minería de Opiniones (MO) y el Análisis de Sentimientos (AS), que aplicados de forma subjetiva o a través de algoritmos automáticos, permiten identificar la polaridad de los mensajes y clasificar opiniones a partir de un determinado conjunto de datos, ayudando a las empresas a comprender qué piensan y qué sienten los consumidores (Pang & Lee, 2008). En esta misma línea, Liu (2015) entiende la Minería de Opiniones y el Análisis de Sentimientos como un estudio automático de opiniones, expresiones, emociones, evaluaciones y otros aspectos que el usuario pueda mostrar en sus mensajes de texto. Finalmente, Larose (2014, p.10) define estas técnicas como “el proceso de descubrir significativas y nuevas correlaciones, patrones y tendencias por tamizado, a través de grandes cantidades de datos almacenados en los repositorios, utilizando tecnologías de reconocimiento de patrones, así como las técnicas estadísticas y matemáticas”.

En este contexto, esta tesis doctoral pretende proporcionar aporte técnico y científico a las técnicas actuales de monitoreo o seguimiento de los medios sociales. Concretamente, LOGOS emerge como un sistema concebido para atender a las necesidades latentes de las PyMEs. Para ello, se desarrolla una metodología y un sistema web gratuito de apoyo a decisiones que se conecte a medios sociales y realice tareas de Minería de Datos y Análisis de Sentimientos en distintos idiomas a través de algoritmos automáticos, para reunir e interpretar información vital sobre los clientes, las empresas y sus competidores con el objetivo de apoyar a las decisiones comerciarles. El nombre LOGOS surge inspirado en la palabra griega (λόγος) que significa “la palabra meditada, reflexionada o razonada”, que también puede ser entendida como: "inteligencia", "pensamiento" y "sentido".

Esta motivación encuentra apoyo en la escasez de herramientas de monitoreo de medios sociales 100% gratuitas, que hagan tareas de Minería de datos y Análisis de Sentimientos (ej. extracción, procesamiento, visualización y clasificación de la información) en múltiples idiomas a través de una interfaz sencilla, que sea de código libre y que además, estén pensadas con un enfoque técnico y de diseño desde la perspectiva del marketing.

Bajo estas premisas, son cuatro las principales aportaciones de esta tesis (ver Figura 1):

a) La primera aportación trata de identificar, describir y catalogar un conjunto amplio de *datasets* de mensajes cortos, en siete idiomas distintos (inglés, español, portugués, árabe, alemán, italiano y

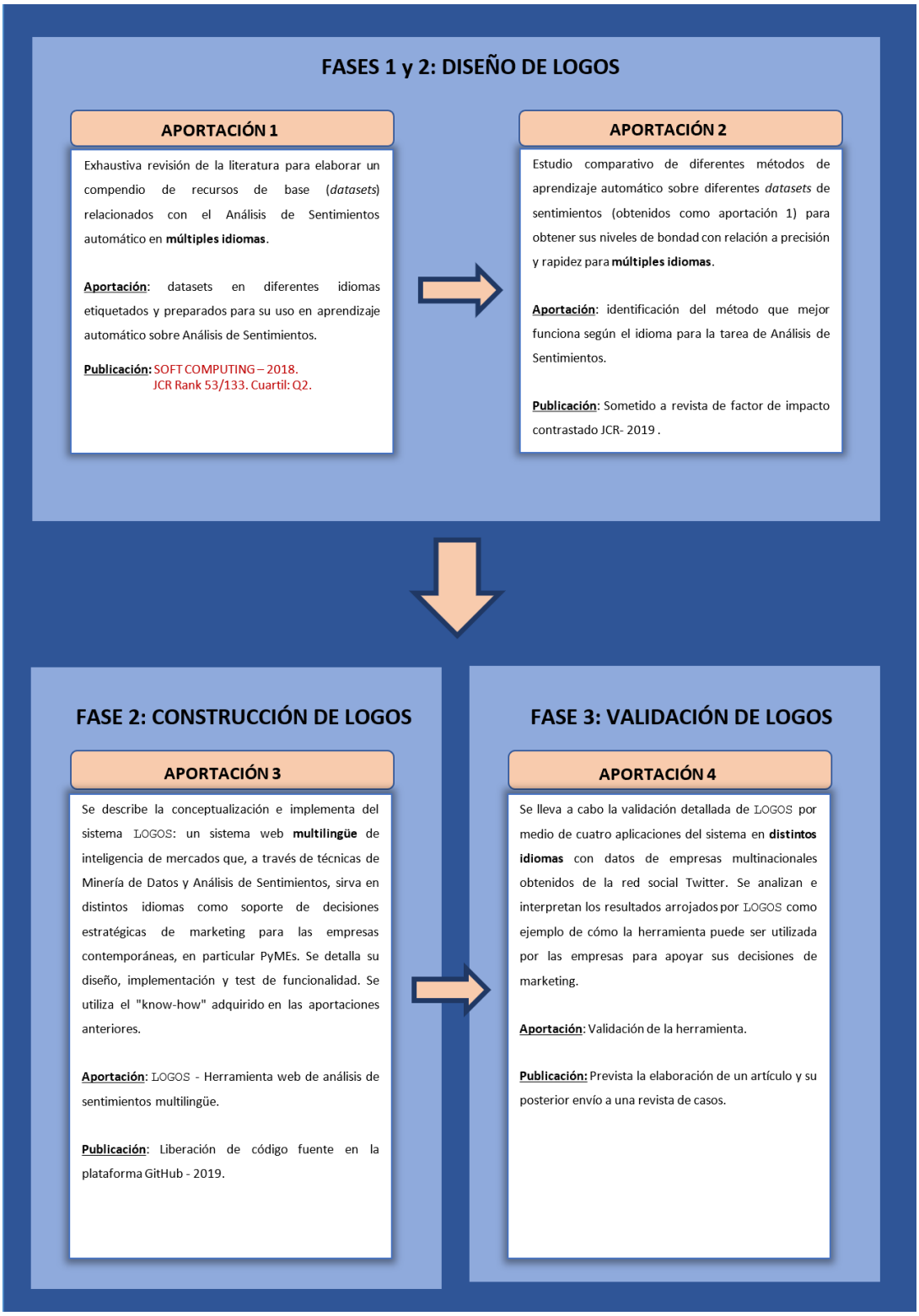
francés), clasificados de manera subjetiva y validados por la comunidad científica para las tareas de Análisis de Sentimientos automáticos. Este compendio de recursos, derivado de una extensa y exhaustiva revisión de la literatura, sirvió como pilar de los análisis comparativos llevados a cabo en la segunda contribución de esta tesis.

b) Con la segunda aportación se realiza una comparativa entre cuatro algoritmos frecuentemente utilizados en la literatura en las tareas de AS, frente a los distintos *datasets* recopilados en la primera aportación. Estos análisis fueron llevados a cabo para determinar el grado de bondad de los indicadores de clasificación de cada algoritmo. Este paso previo y medular, fue el que propició elegir – según el idioma - el algoritmo de mejor rendimiento para las tareas de clasificación realizadas por LOGOS, sistema creado a continuación.

c) En la tercera aportación se lleva a cabo el desarrollo propiamente dicho de LOGOS. Se trata de un sistema web de inteligencia empresarial aplicado a medios sociales, que utiliza la Minería de Datos y el Análisis de Sentimientos para generar distintos informes que sirvan de apoyo a las decisiones de marketing de las PyMEs. Se plantean y se ejecuta cada paso de las etapas de diseño, programación y validación del sistema, los cuales fueron pensados y construidos bajo el enfoque de un sistema de apoyo a decisiones de marketing.

d) El cuarto estudio pone en valor a los tres estudios precedentes, ya que consiste en cuatro aplicaciones de la herramienta LOGOS sobre dos empresas multinacionales (Nike y Samsung) con datos multilingües. Este último estudio, además de corroborar las bases teóricas utilizadas en esta tesis, es el que afianza a LOGOS como un sistema de soporte de decisiones de marketing eficaz, práctico, intuitivo, ágil, intuitivo y gratuito, para las PyMEs.

Figura 1. Fases y principales aportaciones de la tesis.



Fuente: Elaboración propia.

CAPÍTULO 2: MARCO CONCEPTUAL DE LA TESIS

2.1. INTRODUCCIÓN

2.2. DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS (KDD)

2.3. LOS DATOS

2.4. MINERÍA DE DATOS

2.5. MINERÍA DE OPINIONES Y ANÁLISIS DE SENTIMIENTOS

2.6. HERRAMIENTAS DE MONITOREO DE MEDIOS SOCIALES WEB

2.7. APLICACIONES Y TRABAJOS RELACIONADOS

2.1. INTRODUCCIÓN

Los avances en la tecnología de la información en los últimos años han cambiado la forma de pensar y de aplicar el marketing y en consecuencia la manera en cómo las organizaciones manejan información sobre sus clientes. Este nuevo escenario, aliado a los avances de las tecnologías de gestión de la información y dotado de un gran volumen de datos, proporciona tanto nuevas oportunidades como nuevos retos para las empresas que desean diferenciarse de sus competidores.

En este sentido, son muchas las organizaciones que entienden que la información contenida en estos datos sirve como soporte clave para la toma de decisiones de marketing. Sin embargo, gran parte de este valioso conocimiento está oculto entre los datos y necesita ser desvelado (Parenteau et al., 2016).

Por otro lado, la volatibilidad actual de los clientes, resultado de un mercado competitivo, es intensa y diversa. Esta volatibilidad exige que las compañías tomen decisiones más asertivas, minimicen los errores y se preocupen de fidelizar a sus clientes, manteniendo relaciones más duraderas a través de estrategias de negocios cada vez más a medida de su público. Dado que los especialistas en marketing interactúan con los clientes a través de muchos puntos de contacto, incluido un extenso abanico de distintas medios sociales, existen muchas oportunidades para escuchar y adquirir información sobre ellos, de manera que permita el desarrollo de estrategias eficaces (Ciribeli & Paiva, 2011; Lahuerta-Otero & Cordero-Gutiérrez, 2016; Lim, Chen, & Chen, 2013).

En este sentido, información que ayude a las organizaciones en temas relacionados con la segmentación de mercado, la comprensión de las preferencias del cliente, la atención y satisfacción del cliente o la recopilación de datos de la competencia entre otros, puede ser conseguida a través de los medios sociales (Bijmolt et al., 2010). El análisis del Big Data ha sido adoptado como una tecnología disruptiva que ha remodelado la inteligencia empresarial. El *Business Intelligence* (BI) – traducido al español como Inteligencia Empresarial - se presenta como un enfoque holístico que se basa de estos datos para obtener información comercial que apoye a los procesos de toma de decisiones. El término *Business Intelligence* fue presentado por Howard Dresner del Grupo Gartner en 1989 para describir un conjunto de conceptos, métodos y procesos que buscan mejorar la toma de decisiones empresariales mediante el uso de sistemas de soporte (Carvalho, 2010; Müller, Linders, & Pires, 2010; Solomon Negash & Gray, 2008; Solomun Negash, 2004; Watson & Wixom, 2007). Debido a la creciente importancia del análisis de información dirigido a los procesos de toma de decisiones a nivel estratégico, táctico y operativo, la aplicación del BI se ha vuelto cada vez más popular en la comunidad empresarial (Ferreira, Pedrosa, & Bernardino, 2017). Prueba de ello es que, el BI viene adquiriendo un rol fundamental junto a las empresas que buscan mejorar su rendimiento y establecer ventajas competitivas (Cordero-Guzmán & Rodríguez-López, 2017; Fan, Lau, & Zhao, 2015). El BI ofrece respuesta a las necesidades actuales de acceso correcto, rápido y fácil a información relevante mediante el uso intensivo de las tecnologías de información (TI). De este

modo, los gestores pueden tomar decisiones mejor fundamentadas y consecuentemente más acertadas en los más diversificados contextos organizacionales (Herschel, 2019; Mariani, Baggio, Fuchs, & Höepken, 2018).

Para Ferreira, Pedrosa y Bernardino (2017), las organizaciones aún tienen poco conocimiento de sus clientes, así como sobre los procesos más adecuados que proporcionen un ajuste perfecto entre las necesidades de su público y lo que se les ofrece como empresa. Para estos autores, el BI trae a las organizaciones la solución a este problema, ya que permite la obtención de conocimiento sustancial sobre los clientes a partir de plataformas sociales, así como de los sistemas de comercio electrónico.

Las soluciones ofrecidas por el BI son diversas, como, por ejemplo, el descubrimiento de patrones de compra, la mejora de la gestión de las relaciones con el cliente, la posibilidad de una mejor administración del stock, el soporte a las acciones de marketing, entre otros. En la práctica, estas soluciones se traducen en aplicaciones que están relacionadas con la identificación y segmentación de perfiles de clientes (Abbasoglu, Gedk, & Ferhatosmanoglu, 2014), la gestión de reputación y la ontología (términos y que se usan para describir y representar un cierto dominio) de productos (Lau, Li, & Liao, 2014; Petasis, Spiliotopoulos, Tsirakis, & Tsantilas, 2014; Wei Di, Neel Sundaresan, Robinson Piramuthu, 2014), los sistemas de análisis y recomendación de marketing (Cai et al., 2014; Fresno, 2014; Lu, Ba, Huang, & Feng, 2013; Moen, Havro, & Bjering, 2017), las estrategias de precios y el análisis de la competencia (Ingenbleek & van der Lans, 2013;

Liozu & Hinterhuber, 2013), la publicidad basada en la ubicación geográfica y el análisis dinámico de los medios web (Bruwer & Johnson, 2010; Fan et al., 2015; Fong, Fang, & Luo, 2014).

En la misma línea, los investigadores Lim, Chen y Chen (2013) definen el *Business Intelligence* como tecnologías, sistemas, prácticas y aplicaciones que analizan datos comerciales para ayudar a las empresas a comprender mejor su negocio y el mercado. Conocimiento que puede ser utilizado para mejorar productos y servicios, lograr una mejor eficiencia operativa además de fomentar las relaciones con los clientes. Para Foley y Guillermette (2010) y Watson (2010), el BI debe ser entendido como un proceso analítico apoyado por la tecnología que reúne y transforma datos fragmentados de la empresa y del mercado en información y conocimiento sobre objetivos, oportunidades y posiciones de una organización. Vale mencionar que no se trata solo de herramientas software y sistemas, sino de todo el proceso de administración de datos que comprenden aplicaciones, infraestructuras, herramientas y prácticas que permiten mejorar el acceso y el análisis de la información que sirvan de apoyo en la toma de decisiones gerenciales (Carl Anderson, 2019; Elbashir, Collier, Sutton, Davern, & Leech, 2013).

En definitiva, el BI se constituye de un conjunto de tecnologías tales como extracción y almacenaje de datos, procesamiento analítico en línea (OLAP), sistemas de soporte de decisiones (del inglés, *Decision Support*, DS), cuadro de mando integral, entre otros, para mejorar el flujo de trabajo y el proceso de toma de decisiones tanto a nivel táctico

como estratégico dentro y fuera de las empresas (Carl Anderson, 2019; Lee & Longo, 2016).

En este sentido, las empresas contemporáneas han integrado los medios sociales para alimentar a estos sistemas de inteligencia fomentando la obtención de datos respecto a sus clientes. La posibilidad de interactuar con los clientes en plataformas abiertas y acceder a su información previamente disponible genera oportunidades de diálogos de alto valor para las organizaciones (Weiler, Matt, & Hess, 2019). Esta nueva realidad de "creación de conversaciones" puede desarrollar relaciones más significativas y potencialmente más rentables con los clientes (Rodríguez & Santamaría, 2012).

Por ello, las herramientas análisis de medios sociales deben ser manejadas de manera acertada por las organizaciones. Es función del marketing encontrar y gestionar la aplicación de estos recursos sociales de manera correcta para entender verdaderamente las necesidades y preferencias reales de sus mercados (Olbrich & Holsing, 2011). A este respecto, la Minería de Datos (MD) es parte de un proceso automático en el que se combinan el descubrimiento de conocimiento a través de la extracción de patrones y el análisis de datos (*KDD, Knowledge Discovery in Databases*), y contribuye a la consecución de conocimiento estratégico para que organizaciones puedan decidir de manera más acertada sobre sus acciones de marketing (Herschel, 2019). En términos prácticos, las soluciones de MD facilitan la extracción y clasificación de la información disponible en los medios sociales por lo que están ganando terreno,

consolidándose como herramienta de soporte para las decisiones de marketing (Johny & Scholar, 2017; Mohanty & Das, 2017). Berry y Linoff (2011) apuntan que cuanto más precisos y refinados sean los procesos de MD, más precisos y fiables serán los datos recogidos respecto a los consumidores y en consecuencia, más adecuadas serán las respuestas ofrecidas a ellos. En definitiva, las herramientas de Minería de Datos, cuando son bien aplicadas, proporcionan a los especialistas de marketing el tipo de conocimiento apropiado para tomar decisiones fundamentadas y más acertadas. Este conocimiento real del cliente, combinado con la tecnología interactiva disponible al día de hoy, facilita la gestión de relaciones de éxito entre empresa y consumidor (Bharti & Lecturer, 2016; Johny & Scholar, 2017; Mohanty & Das, 2017; Vijaya & Sivasankar, 2017).

Sobre esta base, el propósito de esta esta tesis doctoral es construir un sistema Web de Inteligencia empresarial gratuito y multilingüe, que se sirva de técnicas de Minería de Datos, Minería de Opiniones y Análisis de Sentimientos para extraer y transformar los datos de la red social Twitter en conocimiento competitivo de apoyo a las decisiones de marketing.

2.2. DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS (KDD)

Como ha quedado patente en el epígrafe anterior, la MD es parte de un proceso automático de descubrimiento de conocimiento a través de la extracción de patrones y de análisis de datos, llamado KDD (*Knowledge Discovery in Databases*). El KDD a su vez, abarca las tareas de

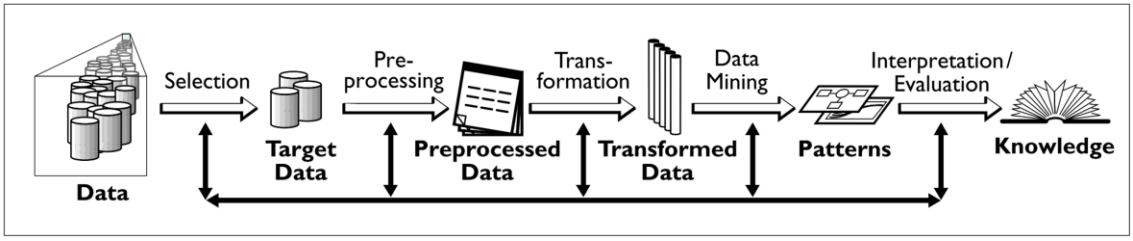
preprocesamiento, de MD y de presentación de resultados (Agrawal & Srikant, 1994; Chen, Han, & Yu, 1996). A pesar de que algunos autores consideran el KDD y la Minería de Datos como sinónimos (Han & Kamber, 2006; Wang, 2009), la gran mayoría de autores los entienden como procesos distintos. En este sentido, el KDD sería responsable por abarcar todo el proceso de descubrimiento del conocimiento, siendo la MD una de las actividades de este proceso que debe ser iterativo y escalonado (Azevedo, 2017; Chaurasia & Pal, 2014; Deepashri & Kamath, 2017; García-Peñalvo & Conde-González, 2017; Jiang, Kumar, Subrahmanian, & Faloutsos, 2017; Moreno & Lara, 2017). La MD como etapa del KDD, es responsable de la selección de los métodos que serán utilizados para localizar patrones, formar representaciones y adecuar ajustes de parámetros de algoritmos para la tarea deseada (Fayyad, Piatetsky-Shapiro, & Smyth, 1996a). Para Fayyad, Piatetsky-Shapiro y Smyth, autores pioneros en esta área de investigación, el proceso de descubrimiento de conocimiento en base de datos busca solucionar el problema de la llamada "**era de la información**": la sobrecarga de datos.

Al día de hoy investigaciones más recientes siguen corroborando esta información y comparten otras definiciones del KDD como un proceso no trivial de identificación de patrones que sean válidos, nuevos (previamente desconocidos), potencialmente útiles y comprensibles para mejorar el entendimiento de un problema o procedimiento de toma de decisiones (Azevedo, 2017; Mishra & Kumar, 2017).

Uno de los objetivos centrales del KDD es hacer comprensible a los analistas datos aparentemente sin sentido. El factor de comprensión de los datos está relacionado con la intuitividad de su representación. Por ejemplo: el registro de un servidor Web no es una representación comprensible. Sin embargo, datos estadísticos extraídos de este registro, tales como totales de acceso o clasificación de los accesos realizados, proporcionan información en un formato más intuitivo y comprensible. En este sentido se suele ver el KDD como un proceso interactivo, iterativo, cognitivo y exploratorio (Jiang et al., 2017).

Concretamente, el proceso del KDD se constituye principalmente de 7 pasos (Figura 2) (Deepashri & Kamath, 2017; Fayyad et al., 1996a; Sharma, Sharma, Sharma, & Shrivatava, 2014):

Figura 2. Visión general de las etapas del proceso de KDD.



Fuente: Fayyad, Piatetsky-Shapiro y Smyth (1996b).

1. **Definición del tipo de conocimiento a descubrir**: presupone una comprensión del dominio de la aplicación, así como del tipo de decisión que tal conocimiento puede contribuir a mejorar.

2. **Selección**: creación de un conjunto de datos objetivo: seleccionar un conjunto o subconjunto de datos sobre los que realizar el descubrimiento.

3. **Limpieza y preprocesamiento de datos**: operaciones tales como eliminación de ruidos cuando sea necesario, recopilación de la información necesaria para modelar o estimar el ruido, elegir estrategias para manipular campos de datos ausentes o *missing data* o formatear datos para adecuarlos a la herramienta de minería.

4. **Transformación, reducción de datos y proyección**: localización de características útiles para representar a los datos dependiendo del objetivo de la tarea, con el propósito de reducir el número de variables y/o instancias a ser consideradas para el conjunto de datos, así como el enriquecimiento semántico de las informaciones.

5. **Minería de datos:** seleccionar los métodos a utilizar para localizar patrones en los datos, seguido de la búsqueda efectiva por patrones de interés y el mejor ajuste de los parámetros del algoritmo para la tarea en cuestión.

6. **Interpretación/Evaluación:** interpretación de los patrones extraídos, con un posible retorno a los pasos 1-6 para posterior interpretación (paso 7).

7. **Implementación del conocimiento descubierto**: incorporar este conocimiento como parte del sistema, documentarlo y reportarlo a las partes interesadas.

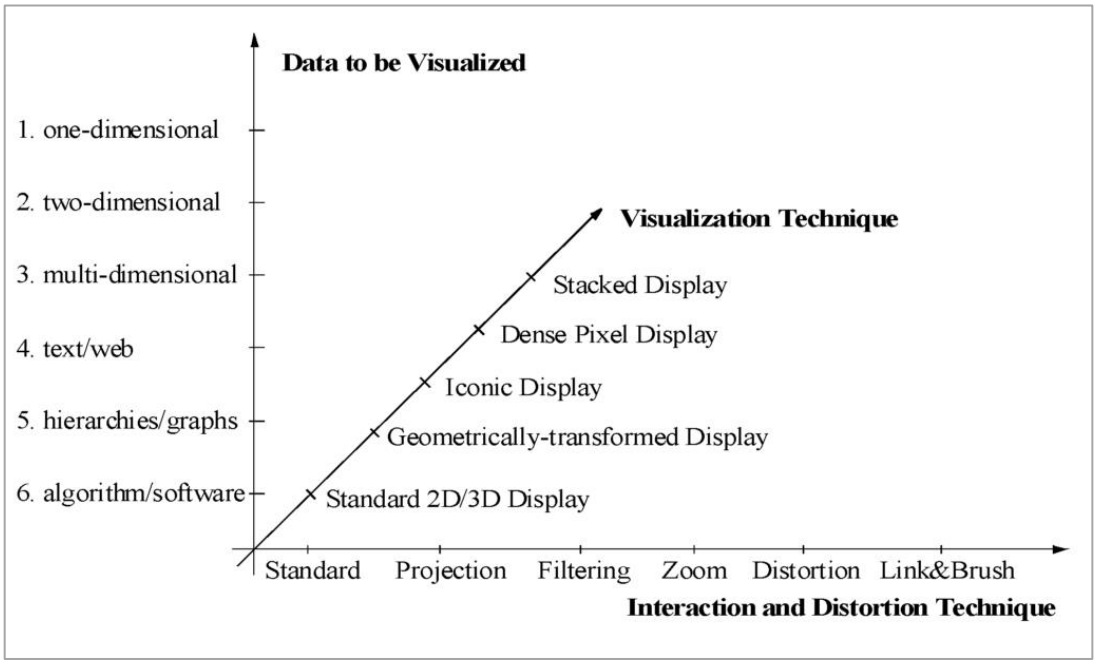
2.3. LOS DATOS

Conocer los tipos de datos con los que se va a trabajar es requisito fundamental para que se pueda escoger el/los métodos(s) adecuado(s) para los procesos de Minería de Datos. De manera general los datos pueden categorizarse fundamentalmente en dos tipos: cuantitativos o cualitativos. Los cuantitativos son representados por valores numéricos, pudiendo ser discretos y continuos. Los datos cualitativos a su vez contienen valores nominales y ordinales o categóricos.

Conocer la cualidad de los datos antes de aplicar las técnicas de Minería de Datos es de extremada importancia para la obtención de resultados consistentes. En este sentido, uno de los primeros pasos es obtener una visualización general que ayude a comprender la forma de los datos y que viabilice la elección adecuada de la técnica de MD a utilizar.

Los autores Bandaru, Ng y Deb (2017), en su artículo de revisión sobre técnicas de visualización, destacan la importancia de estos procedimientos y presentan los diferentes enfoques relacionados al tema. En la Figura 3, el autor Keim (2002) representa la evolución de estas técnicas según el tipo de dato a ser tratado.

Figura 3. Técnicas de visualización y su evolución.



Fuente: Keim (2002).

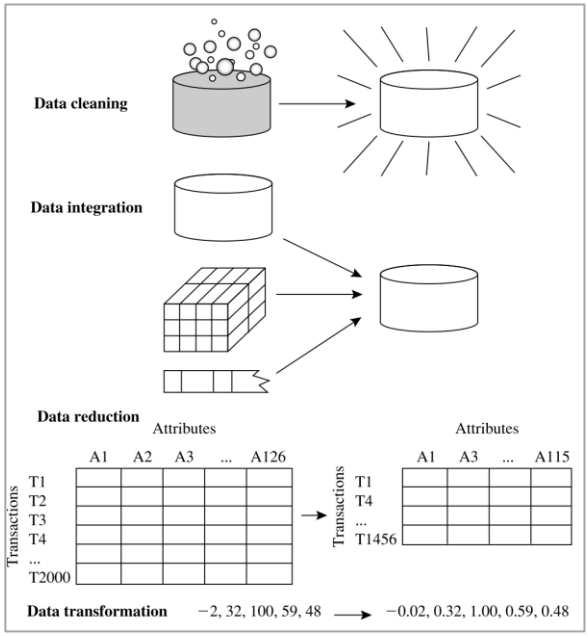
Una vez se obtiene una perspectiva inicial definida de los datos, es necesario explotarlos para, además de adquirir más conocimiento sobre ellos, encontrar valores que puedan venir a comprometer su calidad como, por ejemplo, los valores en blanco o nulos o las variables duplicadas, entre otros.

Según vayan siendo encontrados y corregidos esos posibles problemas, la comprensión sobre los datos se irá ampliando, así como su calidad. Para Olson y Delen (2008), este proceso de preparación de los datos es tan importante, que este paso puede representar hasta un 50% del trabajo en la mayoría de los proyectos de Minería de Datos. Para otros autores como

McCue (2007), esta parte del proceso gana aún más protagonismo llegando a representar hasta un 80% del proyecto.

El proceso correspondiente a la preparación de los datos para su posterior minería se llama preprocesamiento de datos que según los autores Han, Kamber y Jian (2012) se compone de 4 tareas principales: limpieza, integración, transformación y reducción de los datos, las cuales detallamos a continuación (Figura 4).

Figura 4. Formas de preprocesamiento de datos.



Fuente: Han et al. (2012).

Limpieza de los datos: es frecuente encontrar inconsistencias en los datos como, por ejemplo, registros incompletos o valores no válidos entre otros. Esta etapa trata de eliminar/minimizar estos problemas de modo que no influyan en los resultados finales.

Las técnicas utilizadas en esta etapa van desde la eliminación de registros, que puedan presentar algún problema, y la atribución de valores, hasta la aplicación de técnicas de agrupamiento para ayudar en el descubrimiento de valores adecuados.

Integración de los datos: debido a que los datos pueden provenir de fuentes diversas como bases de datos, archivos de texto, hojas de cálculo, almacenes de datos, vídeos, imágenes, entre otras, es necesaria la integración de éstos en un repositorio único y consistente. Este requisito exige un análisis profundo de los datos a fin de identificar las redundancias, dependencias entre variables y valores conflictivos como, por ejemplo, categorías diferentes para los mismos valores, reglas distintas para los mismos datos, entre otros.

Transformación de los datos: esta etapa es muy importante ya que algunos algoritmos consideran valores numéricos y otros valores categóricos. Para esos casos, se deben transformar los valores numéricos en categóricos o categóricos en numéricos donde, por ejemplo, los valores brutos de un atributo numérico (edad) se reemplazan por etiquetas de intervalo (0-10, 11-20, etc.) o etiquetas conceptuales (joven, adulto, senior). Algunas de las diversas técnicas aplicadas en este paso son, a) el suavizado *(smoothing)*, que remueve los valores erróneos también llamados ruidos, que puedan presentar los datos; b) el agrupamiento *(aggregation)*, que reúne valores en espectros sumarios; c) la generalización

(generalization), que convierte a los valores muy específicos en valores más genéricos; d) la normalización (*normalization*), que introduce las variables en una misma escala; y e) la creación de nuevos atributos generados a partir de otros ya existentes.

Reducción de los datos: en la Minería de Datos se maneja un gran volumen de información. En algunos casos este volumen puede ser tan grande que hace realmente complejo el proceso de minería. Por este motivo, las técnicas de reducción de datos deben ser aplicadas para que el volumen de datos original pueda ser menor sin perder su representatividad. Este significativo paso permite que los algoritmos de aprendizaje se ejecuten con más eficiencia manteniendo resultados de calidad. Algunas de las estrategias adoptadas en esta etapa son la creación de estructuras óptimas para los datos, como los cubos de datos *(data cube)*, la selección de subconjuntos de atributos (*attribute subset selection*), la reducción de la dimensionalidad *(dimensionality reduction)* entre otras (Han et al., 2012; Smith, 2002).

Una vez entendidos y preprocesados los datos, se pasa a la siguiente etapa, la Minería de Datos.

2.4. MINERÍA DE DATOS

Desde el campo de la Inteligencia Artificial (AI), en la década de los 60 emergen las técnicas de extracción de datos (Fayyad et al., 1996a, 1996b).

El crecimiento exponencial de las bases de datos ha creado la necesidad de desarrollar tecnologías que utilicen la información y el conocimiento de forma inteligente. Por lo tanto, con el paso de los años, la MD se ha convertido en un área de investigación muy importante que ha ido ampliando cada vez más su abanico de posibilidades por medio de la implementación de las más diversas aplicaciones (Bandaru et al., 2017; Femina & Sudheep, 2015). Debido a su carácter versátil, la MD es considerada multidisciplinar ya que varía y se adapta según el área de actuación (Moreno & Lara, 2017).

Para Han et al. (2012), este fenómeno se explica debido al hecho de que la Minería de Datos ha incorporado muchas técnicas provenientes de otros dominios, como la estadística, el aprendizaje automático, el reconocimiento de patrones, los sistemas de bases de datos, la recuperación de información, la visualización, los algoritmos y la computación de alto rendimiento, entre otros (Figura 5).

Figura 5. Técnicas adoptadas por la Minería de Datos.

Fuente: Adaptación desde Han et al. (2012).

Los mismos autores la definen desde un enfoque más estadístico como el análisis de grandes conjuntos de datos con el fin de encontrar relaciones inesperadas y de resumir los datos a una forma que sean útiles y comprensibles.En la misma línea, Cabena, Hadjinian, Stadler, Verhees y Zanasi (1998), la definieron como un área interdisciplinar que reúne técnicas de aprendizaje automático, reconocimientos de patrones, estadísticas, banco de datos y visualizaciones, para conseguir extraer informaciones de grandes bases de datos.

Finalmente, los precursores del concepto, Fayyad, Piatetsky-Shapiro y Smyth (1996a) la definieron desde el enfoque del aprendizaje automático como un paso en el proceso de descubrimiento del conocimiento, que consiste en la realización de análisis de datos por medio de algoritmos que, bajo ciertas limitaciones computacionales, producen un conjunto de patrones de los datos.

Aunque las definiciones encontradas sobre la Minería de Datos dejan a entender que el proceso de extracción del conocimiento se da de manera principalmente automática, el análisis humano de los datos es un recurso imprescindible.

Como se ha mencionado anteriormente y, según la literatura, la Minería de Datos abarca una extensa diversidad de aplicaciones (Chattamvelli, 2015a; García-Peñalvo & Conde-González, 2017; Han et al., 2012; Porter, Nisbet, Miner, & Miner, 2018; Witten, Frank, Hall, & Pal, 2016). Las más comunes son:

Descripción: Esta aplicación describe los datos permitiendo una potencial interpretación de los resultados obtenidos. Esta tarea suele ser utilizada de manera híbrida con otras técnicas de explotación de datos como la caracterización o la discriminación de datos. Esto se debe a la necesidad de comprobar la influencia de ciertas variables en el resultado final obtenido.

Clasificación: La clasificación es una de las tareas más utilizadas puesto que se centra en identificar a qué clase pertenece un determinado registro. La tarea de clasificación puede ser utilizada para diversos fines como: determinar cuándo una operación de tarjeta de crédito es fraudulenta, identificar en una escuela qué clase es la más indicada para un determinado estudiante, diagnosticar dónde una determinada enfermedad puede estar presente, identificar cuándo una persona puede representar alguna amenaza de seguridad y la polarización de mensajes de texto, entre otras. Se utiliza, por ejemplo, para obtener respuestas a preguntas relacionadas a los riesgos de conceder créditos a determinados clientes o también sobre el estado de una enfermedad de un paciente según su analítica.

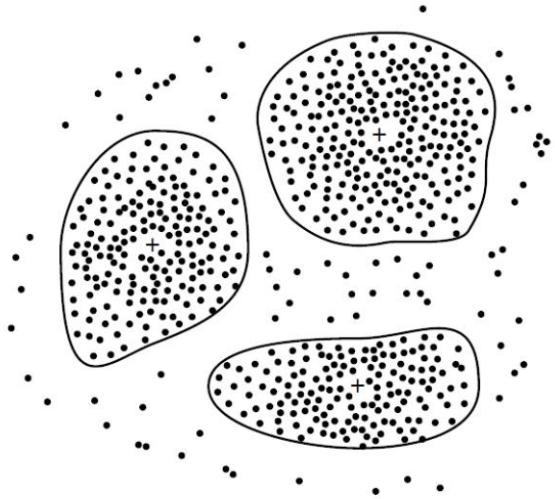
Estimación o Regresión: Esta tarea tiene similitud con la de clasificación, aunque es más usada cuando el registro es de valor numérico y no categórico. De este modo, se puede estimar el valor de una determinada variable por medio de las demás. En este sentido, por ejemplo, se podrían estimar valores mensuales de gastos para diferentes consumidores teniendo como base el historial de sus

hábitos de compras e indicar cuál será el valor de gasto por un nuevo consumidor. Esta tarea también puede ser usada, por ejemplo, para estimar el gasto de una familia de tres personas en determinados meses del año, las ventas de una tienda para el mes siguiente, o también la tensión arterial ideal de un paciente en base a su edad, sexo y masa corporal.

Predicción: Esta tarea también coincide en algunos aspectos con las tareas de clasificación y estimación, aunque se centra en descubrir el valor futuro de un determinado atributo. Es comúnmente utilizada para predecir los valores de las acciones en la banca o, por ejemplo, el porcentaje de incremento del tráfico de datos. Algunos métodos de clasificación y regresión, con sus debidos matices, también pueden ser usados en la predicción.

Agrupación: la tarea de agrupaciones permite identificar y aproximar los registros que tienen características similares. Una agrupación o *clúster*, consiste en una colección de registros similares entre sí, pero diferentes de otros registros pertenecientes a otras agrupaciones. Esta tarea difiere de la tarea de clasificación puesto que no necesita que los registros estén categorizados previamente (aprendizaje no-supervisado). Además, el agrupamiento no busca clasificar, estimar o predecir alguna variable. Lo que procura es identificar a los grupos de datos que sean de alguna manera similares entre sí (Figura 6).

Figura 6. Tarea de agrupaciones de registros em tres *clústeres*.



Fuente: Han et al. (2012).

Algunos ejemplos de aplicaciones de esta tarea son la segmentación de mercados para un determinado nicho o también para reducir a un conjunto de atributos similares los registros con centenas de atributos. Contesta preguntas como: ¿Cuáles son los diferentes tipos de clientes que compran en mi tienda?

En la literatura, son diversas las aplicaciones de *clúster,* como por ejemplo, las encuestas de mercado, el reconocimiento de patrones, el procesamiento de imágenes, el análisis de datos, la segmentación de mercado, la taxonomía de plantas y animales, las investigaciones geográficas, la clasificación de documentos Web, la detección de fraudes, entre otras (Cutting, Karger, Pedersen, & Tukey, 2017). La tarea de Asociación por ejemplo, permite identificar qué atributos se relacionan entre sí. Se considera una de las tareas más conocidas puesto que presenta

buenos resultados, sobre todo en los análisis relacionados con los "carritos de la compra" (*market basket*) identificando qué productos son preferidos por los consumidores en los procesos de venta cruzada. Responde preguntas como: ¿si un cliente de un supermercado compra papitas para bebe también compra pizza?

Según el tipo de dato que se vaya a analizar, existe una ramificación de la Minería de Datos. Por ejemplo, si el dato en cuestión es textual, se habla de minería de textos, si el dato procede de blogs o de web se habla de minería web. En esta tesis tratamos con la minería de textos aplicando técnicas de Minerías de Opiniones y Análisis de Sentimientos.

2.5. MINERÍA DE OPINIONES Y ANÁLISIS DE SENTIMIENTOS

El Análisis de Sentimientos (AS), también llamado de Minería de Opiniones (MO), es un recurso que a su vez deriva de la Minería de Datos. Este campo de estudio es el que analiza las opiniones, sentimientos, evaluaciones, valoraciones, actitudes y emociones de las personas hacia entidades, productos, servicios, organizaciones, individuos, asuntos, eventos, temas y sus atributos (Liu, 2017; Zimbra, Abbasi, Zeng, & Chen, 2018).

En esta subsección son descritas y detalladas las características, similitudes, aplicaciones y enfoques de las técnicas de Minería de Opiniones y del Análisis de Sentimientos.

2.5.1. DIFERENCIAS SEMÁNTICAS

En el universo del monitoreo y análisis de medios sociales, es común el aforismo y una falta de consenso sobre las terminologías. Consumidores y profesionales acaban por considerar como similares conceptos como *"monitoreo de marca", "monitoreo de redes", "monitoreo de reputación", "análisis de influencia de marketing", "minería de conversación", o "inteligencia del consumidor on line"* (Zabin & Jefferies, 2008) .

En esta línea, las técnicas de extracción, tratamiento de opiniones, sentimientos y subjetividad en el texto pueden considerarse como Minería de Opiniones, Análisis de Sentimientos y/o Análisis de la Subjetividad. Sin embargo, también se podrían considerar como Extracción de Opiniones, del inglés, *Opinion extraction* (Benkhelifa & Laallam, 2018). Estos múltiples términos reflejan las diferencias existentes tanto en sus usos originales como en los usos que se han desarrollado posteriormente en la literatura técnica y científica sobre los mismos.

Históricamente, el termino Minería de Opiniones aparece por primera vez en un artículo de Dave et al. (2003) titulado "*Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*" en el "*12th international conference on World Wide Web*". Debido al gran alcance del evento, el término ganó popularidad dentro de las comunidades científicas relacionadas con la búsqueda y recuperación de información en la Web. Los autores manifestaron que la herramienta ideal de Minería de Opiniones sería capaz de procesar un conjunto de datos en base a un elemento determinado, generando una lista de atributos de este como

calidad, precio entre otras características, junto con su polaridad (positiva, neutra o negativa). A partir de este trabajo, la mayoría de las investigaciones posteriores que se ajustaban a esta descripción fueron identificadas como Minería de Opiniones (Liu, 2007).

El término Análisis de Sentimientos es anterior y su historia es algo similar al de Minería de Opiniones. El término "sentimiento", utilizado para hacer referencia a los procesos de análisis automáticos de evaluación de texto con indicación predictiva de juicio (positiva, neutra o negativa), surge por primera vez en los trabajos de Das y Chen (2001) y Tong (2001). Posteriormente en el evento de la *"Association for Computational Linguistics"* (ACL) 2001 y en la *"Conference on Empirical Methods in Natural Language Processing"* (EMNLP) del mismo año, se presentaron trabajos que trataban de dar continuidad a esta línea de investigación (Pang, Lee, & Vaithyanathan, 2002; Turney, 2002). Complementariamente, Nasukawa y Yi (2003) publicaron su trabajo titulado, "*Sentiment analysis: Capturing favorability using natural language processing"*. En el mismo año Yi, Nasukawa, Bunescu y Niblack (2003) publicaron su investigación llamada "*Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques".* Estos hechos en conjunto pueden explicar la popularidad del término "Análisis de Sentimientos" entre las comunidades con enfoque en el procesamiento y clasificación del lenguaje natural.

En la literatura se encuentran una cantidad considerable de artículos que hacen mención al "Análisis de Sentimientos" para tratar de solucionar

tareas de clasificaciones de polaridad (positiva, neutro o negativa) (Alaei, Becken, & Stantic, 2019b; Anandarajan, Hill, & Nolan, 2018a; Bhadane, Dalal, & Doshi, 2015; Flekova, Preoţiuc-Pietro, & Ruppert, 2015; Hussein, 2018; Peng, Ma, Poria, Li, & Cambria, 2019; Ribeiro, Araújo, Gonçalves, André Gonçalves, & Benevenuto, 2016; L. Wang, Niu, & Yu, 2019; L. Zhang & Liu, 2017). Sin embargo, como también sucede con el término “Minería de Opiniones”, muchos autores consideran “Análisis de Sentimientos” de forma más amplia, entendido como tratamiento computacional de opinión, sentimiento y subjetividad en el texto. (Hussein, 2018). En definitiva, desde esta visión más amplia, "Análisis de Sentimientos" y "Minería de Opiniones” comparten el mismo campo de estudio.

Aunque la Minería de Opiniones y el Análisis de Sentimientos estén íntimamente relacionados y aparezcan solapados frecuentemente en la literatura, algunos autores marcan las diferencias de cada técnica otorgando a la Minería de Opiniones la misión de extraer y analizar las opiniones de la gente acerca de una entidad, producto o servicio, mientras que el Análisis de Sentimientos se encarga de identificar la polaridad o en qué sentido (positivo, neutro o negativo) versan estas opiniones (Tsytsarau & Palpanas, 2012). Por ejemplo, en la sentencia “Esta marca hace muy buenas galletas y además son muy baratas”, el Análisis de Sentimientos se ocuparía de decir si es una sentencia positiva, negativa o neutra (en este caso positiva), mientas la Minería de Opiniones se ocuparía de decir cuál es la opinión contenida en la sentencia, en este caso, que son galletas buenas y baratas.

2.5.2. CLASIFICACIÓN Y APLICACIONES

Tanto la MO como la AS tratan de abordar el problema de la clasificación (de opiniones y sentimientos) desde tres niveles específicos: documento, frase y característica o aspecto.

- **Documento:** La clasificación a nivel de documento identifica el sentimiento de este como un todo y se encarga de clasificarlo como positivo o negativo.

- **Frase:** La clasificación a nivel de frase identifica el sentimiento a un nivel más específico, determinando polaridad para cada frase de un texto.

- **Característica o aspecto:** La clasificación a nivel de característica identifica los sentimientos relacionados a aspectos específicos en base a un producto, servicio o entidad (Chen & Zimbra, 2010; Liu, 2012, 2017; Wilson, Wiebe, & Hoffmann, 2009).

Por otro lado, las aplicaciones del Análisis de Sentimientos se dividen en seis grandes grupos (Liu, 2017; Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma, 2015):

Clasificación de sentimiento o polaridad de sentimiento, que consiste en clasificar la información fundamentalmente en tres categorías: positiva, negativa o neutra (Liu, 2017; Rushdi Saleh, Martín-Valdivia, Montejo-Ráez, & Ureña-López, 2011; L. C. Yu, Wu, Chang, & Chu, 2013). Las categorías también pueden ser representadas de diversas

maneras, como escalas numéricas [-1,0,1] indicando negativo, neutro y positivo respectivamente, o [0 – 5], donde cero indica máxima negatividad y 5 máxima positividad (S.-T. Li & Tsai, 2013; Liu, 2017; Martín-Valdivia, Martínez-Cámara, Perea-Ortega, & Ureña-López, 2013).

Clasificación de la subjetividad, que consiste principalmente en identificar una oración como subjetiva u objetiva. Una sentencia objetiva se relaciona con los hechos reales y suele ser más fácil de clasificar. Una sentencia subjetiva tiende a expresar otros tipos de información como una creencia, valoración o sensación individual y personal directamente relacionada con sus experiencias anteriores. Algunos autores ven esta tarea como paso previo a la clasificación de sentimientos. De modo que una buena clasificación subjetiva, potencia los resultados de la clasificación de sentimientos automática (Barbosa & Feng, 2010; Esuli & Sebastiani, 2006; Liu, 2017; Montoyo, Martínez-Barco, & Balahur, 2012; Sarvabhotla, Pingali, & Varma, 2011b).

Resumen de la opinión, que consiste principalmente en la identificación y extracción de los principales atributos y sentimientos respecto a una entidad que están presentes dentro de uno o varios documentos (Wang et al. (2013). De este modo, esta tarea trata la detección de opiniones tanto en un documento como también en varios, buscando relaciones, características y/o vínculos entre ellos

(Beineke, Hastie, Manning, & Vaithyanathan, 2004; Li et al., 2018; Liu, 2017; Ma, Sun, Lin, & Ren, 2018; Pang & Lee, 2004).

Recuperación de opinión que, a través de aspectos como relevancia y tipo de consulta, recuperan expresiones/opiniones en los documentos. Por lo general esta técnica se aplica para la elaboración de rankings de documentos (Kim, Song, & Rim, 2016; Lee, Song, Lee, Han, & Rim, 2012; Luo, Osborne, & Wang, 2015; Peleja & Lisboa, 2015).

Sarcasmo e ironía, que trata de la detección de sentencias que lleven características de esta naturaleza. Según algunos autores es la tarea más difícil dentro del AS (Bauwelinck, Jacobs, Hoste, & Lefever, 2019; Delia Irazú Hernańdez Farías, Patti, & Rosso, 2016; Delia Irazú Hernández Farías, Patti, & Rosso, 2018; Poria, Cambria, Hazarika, & Vij, 2016; S. Zhang, Zhang, Chan, & Rosso, 2019).

Detección de emociones, que trata de identificar las distintas emociones contenidas en los textos como alegría, rabia, disgusto, irrelevancia, tristeza y otros.

Esta aplicación en particular se ha vuelto más popular en los últimos años debido a su gran potencial de aplicación en áreas como marketing, ciencias políticas, psicología, inteligencia artificial, entre otras (George, Barathi Ganesh, Anand Kumar, & Soman, 2018; Montoyo et al., 2012; Ortigosa, Martín, & Carro, 2014; Recupero, Dragoni, Buscaldi, Alam, & Cambria, 2018).

Otros análisis, como la detección del género del autor del documento (*genre or authorship detection)* (Gómez-Adorno, Sidorov, Pinto, Vilariño, & Gelbukh, 2016; Hangya & Farkas, 2016; Montesi & Navarrete, 2008; Sanchez-Perez, Markov, Gómez-Adorno, & Sidorov, 2017; Savoy, 2012; Seki, Kando, & Aono, 2009), o tareas como la detección de contenidos con el objetivo de distorsionar la opinión pública sobre una entidad, conocido como detección de opiniones en spam (*opinion spam detection)* (Li, Ott, Cardie, & Hovy, 2014; Ott, Choi, Cardie, & Hancock, 2011; Ren & Ji, 2017; Xie, Wang, Lin, & Yu, 2012).

Uno de los mayores desafíos de estas técnicas está relacionado con la subjetividad encontrada en los textos además de aspectos relativos a ironía y sarcasmo, errores gramaticales, abreviaciones y expresiones coloquiales que dificultan aún más una correcta clasificación (Balog, Mishne, & Rijke, 2006; Jindal & Liu, 2006).

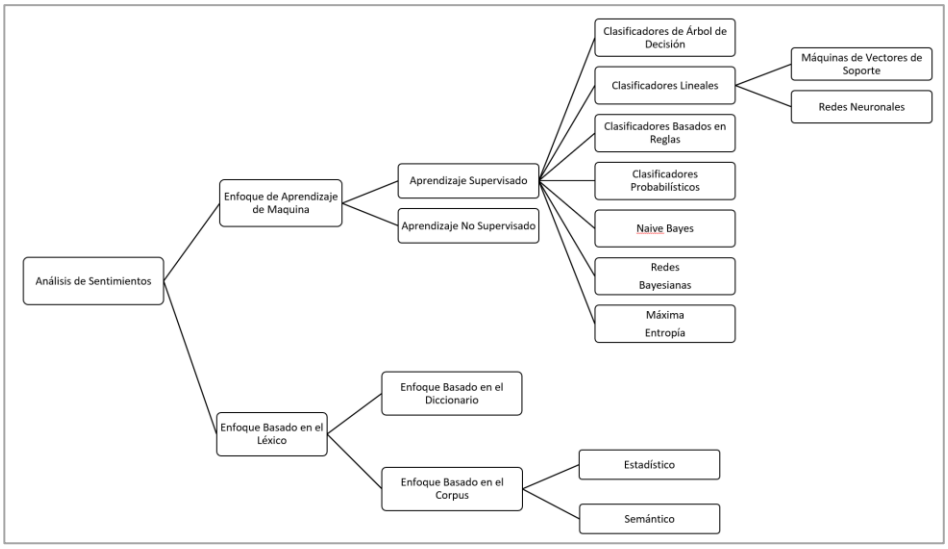
Por este motivo, las técnicas de procesamiento del lenguaje natural (PLN), o derivado del inglés (NLP) *Natural Language Processing*, son un recurso indispensable para lograr buenos resultados en la Minería de Opiniones. Perteneciente al área de las ciencias de la computación, inteligencia artificial y lingüística, el PLN se ocupa de los procesos y estudios relacionados con las interacciones entre los computadores y el lenguaje humano. Quizás el punto central de este recurso sea diseñar mecanismos para que humanos y maquinas puedan comunicarse entre sí de manera eficaz (Chowdhury, 2003).

La realización de tareas importantes relativas al PLN como la reducción del texto a palabras raíz (*steaming*) o filtrado de palabras vacías (*stop words*), entre otras, ayudan a optimizar la información y potencian los procesos de clasificación ya que facilitan el entendimiento de la información por los ordenadores.

2.5.3. ENFOQUES DEL ANÁLISIS DE SENTIMIENTOS

Son dos los enfoques principales del Análisis de Sentimientos, el basado en léxico (*lexicón-based)*, que se basa en conjuntos de palabras preestablecidos y el aprendizaje automático (*machine learning)* que se basa en el aprendizaje automático (Figura 7) (Medhat, Hassan, & Korashy, 2014). Estos dos enfoques suelen ser usados separadamente, aunque también pueden ser aplicados en conjunto de manera híbrida (Medhat et al., 2014). A continuación, se discuten cada uno de ellos.

Figura 7. Técnicas de clasificación de sentimientos.



Fuente: Adaptación de Medhat et al. (2014).

2.5.3.1. ENFOQUE BASADO EN EL LÉXICO - LEXICÓN-BASED

Este enfoque hace uso de diccionarios semánticos, también denominados léxicos de sentimientos u opiniones, que contienen conjuntos de palabras previamente rotuladas para la clasificación. De este modo, cada palabra recibe una indicación de su polaridad (positiva, negativa o neutra). Además, los léxicos pueden atender a un dominio específico con palabras relacionadas a éste (Riloff & Wiebe, 2003; Silva, Carvalho, & Sarmento, 2012), o de tipo más general como por ejemplo el proyecto *SentiWordNet*[2] presentado por Esuli & Sebastiani (2006).

Para el idioma inglés, el diccionario *SentiWordNet* y el léxico *Sentistrength2011* son ampliamente utilizados en la literatura (Alharbi & de Doncker, 2019; Anandarajan, Hill, & Nolan, 2018b; Maipradit, Hata, & Matsumoto, 2019; Quijote, Zamoras, & Ceniza, 2019; Rathore, Arjaria, Khandelwal, Thorat, & Kulkarni, 2019). Para el idioma portugués, el *OP-Lexicon* (Souza, Vieira, Chishman, & Alves, 2011) es el léxico de referencia y para el idioma español el recurso más utilizado es el *iSOL*, creado a partir de la lista de palabras del profesor Bing Liu (Bing Liu's Opinion Lexicon) (Hu & Liu, 2004) y ampliamente utilizado por Molina-González, Martínez-Cámara, Martín-Valdivia y Perea-Ortega, (2013). Para el idioma italiano se destacan el léxico *ItalWordNet* (Maks et al., 2014; Roventini, Alonge,

[2] http://sentiwordnet.isti.cnr.it/ (acceso el 07/08/2019).

Calzolari, Magnini, & Bertagna, 2000) y el más reciente *polarITA* (Hernández Farías, Laganà, Patti, & Bosco, 2017). Para el idioma alemán el léxico de referencia es el *GermanPolarityClues* presentado y ampliamente utilizado en los trabajos de (Maks et al., 2014; Sidarenka & Stede, 2016). Finalmente para el idioma francés se destaca el léxico *WOF for French* (Benoît Sagot & Fišer, 2008) creado en el año 2008 y actualizado nuevamente en el año 2012 (Benoit Sagot, Fišer, & others, 2012).

Los léxicos en los idiomas inglés, español u portugués, pueden ser descargados desde la página web creada como repositorio del conocimiento de esta tesis por medio del enlace http://hipatia.ugr.es/steiner/index.php/sentiment-lexicons/.

En relación con la construcción/compilación de los diccionarios de sentimientos, ésta puede realizarse de dos maneras. De forma manual, opción poco utilizada debido a la gran cantidad de tiempo que demanda, o automática por medio de dos diferentes métodos: en base a un diccionario o en base a un corpus. Estos métodos también pueden ser utilizados de forma simultánea para conseguir mejores resultados. Comentamos cada uno de ellos:

- **Con base en un diccionario:**

Este método parte de un pequeño grupo de palabras recogidas manualmente cuya polaridad es conocida. A continuación se buscan los sinónimos y antónimos de las palabras del grupo inicial utilizando corpus de

sinónimos como *WordNET*[3] (Miller, 1995) o *Thesaurus* (Mohammad, Dunne, & Dorr, 2009). Las nuevas palabras se añaden al grupo inicial ampliándolo exponencialmente. El proceso se repite hasta que no se encuentren nuevas palabras. Una vez concluida esta etapa, se revisa manualmente la clasificación para eliminar los errores (Hu & Liu, 2004; Kamps, Marx, Mokken, & Rijke, 2004; Kim & Hovy, 2004).

- **Con base en un corpus:**

Este segundo método se aplica cuando hay la necesidad de encontrar la orientación o polaridad de palabras en un contexto especifico. Para realizar esta tarea, se utilizan patrones sintácticos conjuntamente con una lista de palabras semilla relacionadas al dominio u objeto del estudio (Hatzivassiloglou & McKeown, 1997).

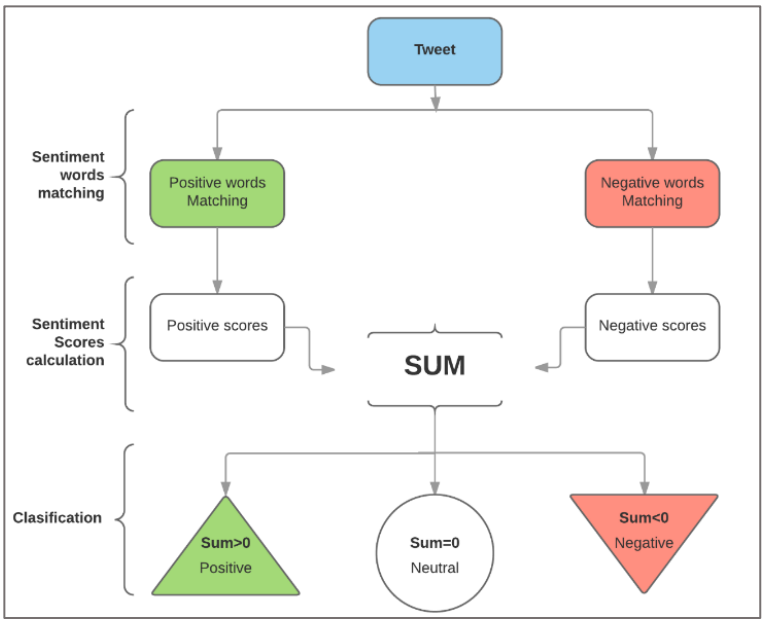
Una vez construidos y validados los léxicos, éstos son utilizados en las tareas de clasificación a través de diferentes métodos. El más común y ampliamente utilizado en la literatura (Hu & Liu, 2004; Liu, Hu, & Cheng, 2005), se basa en el cálculo de la diferencia del número de palabras consideradas positivas y del número de palabras consideradas negativas. Para esto, es vital disponer de un amplio conjunto de palabras validado. Cuantas más palabras compongan el léxico, más precisa será la clasificación.

[3] https://wordnet.princeton.edu/ (acceso el 07/08/2019).

El clasificador pasa por cada sentencia y busca en el léxico la polaridad de cada palabra que ésta contiene, calculando el número de palabras positivas y negativas en el texto.

Concretamente cuando ocurre el emparejamiento, el clasificador le asigna a la palabra el número que le corresponde, siendo 1 si la palabra se considera positiva, -1 si se considera negativa, y 0 si se considera neutra. Cuando no ocurre el emparejamiento la palabra no recibe puntuación alguna y no influye en los resultados, de ahí que se considere vital que los léxicos sean lo más amplios posibles (Figura 8).

Figura 8. Clasificador léxico de sentimientos.



Fuente: Elaboración propia a partir de M. Hu & Liu (2004).

Al final, se comparan las puntuaciones obtenidas para cada clase de sentimiento, atribuyendo al documento la clase con la mayor puntuación. Matemáticamente eso se suele resolver con base en la resta de

puntuaciones positivas menos las puntuaciones negativas. Si el resultado es mayor que 0, la sentencia se clasifica como positiva y si el resultado es menor que 0, se clasifica como negativa y si el resultado es igual a 0, se clasifica como neutra (Hu & Liu, 2004; Liu et al., 2005).

2.5.3.2. ENFOQUE BASADO EN EL APRENDIZAJE AUTOMÁTICO

El enfoque de aprendizaje automático es una técnica de clasificación de sentimientos perteneciente al campo de la inteligencia artificial (IA), centrada en desarrollar recursos computacionales competentes en adquirir conocimientos de forma automática. Los algoritmos de aprendizaje automático trabajan sobre el principio de inferencia llamado "inducción". Consiste en obtener conclusiones genéricas a partir de un conjunto de ejemplos también llamados corpus o *datasets*. Los algoritmos toman decisiones en base a soluciones consideradas adecuadas, utilizadas en problemas anteriores (Mitchell, 1997; Weiss & Kulikowski, 1991).

El aprendizaje automático inductivo se divide principalmente en dos tipos: supervisado y no-supervisado. Sin embargo, también existen otros métodos menos comunes como los de aprendizaje automático semi-supervisado y aprendizaje activo.

En relación con su representación, los sistemas de aprendizaje también pueden ser clasificados como simbólicos y no-simbólicos. Los simbólicos representan el conocimiento de manera que se facilita la lectura e interpretación de la información. Algunos ejemplos de los métodos

simbólicos son árboles de decisión y conjuntos de reglas (Kubat, Bratko, & Michalski, 1998; Michalski, 1983). Por otro lado, los métodos no-simbólicos, también llamados de caja-negra, desarrollan representaciones propias internas del conocimiento que no son fácilmente interpretables, como por ejemplo las redes neurales artificiales, K-NN del inglés (*K-Nearest Neighbors*), Redes bayesianas y SVM (Support Vector Machine).

A continuación, se describen las principales características de los diferentes métodos de aprendizaje automático citados anteriormente.

A. Aprendizaje automático supervisado

En este método, el algoritmo genera un modelo que puede ser se predicción, clasificación, agrupamiento entre otros, a partir de ejemplos contenidos en grupos de datos llamados *datasets* o corpus de entrenamiento. Es decir, el aprendizaje supervisado parte de ejemplos por los que se conoce a priori la clase o conjunto al que pertenece. De este modo, busca proporcionar al algoritmo la capacidad de producir respuestas correctas para nuevas entradas en base a conceptos aprendidos anteriormente. En otras palabras, el algoritmo utiliza la información contenida en grupos de datos (dataset/corpus), para aprender sobre un concepto y predecir de manera más acertada (Haykin, 2004). Estos grupos de datos son discutidos en profundidad en el Capítulo 4 de esta tesis doctoral.

Para testar las predicciones obtenidas por el clasificador, en general el dataset se divide en dos grupos: dataset de entrenamiento y dataset de prueba. El primer grupo permite el aprendizaje del concepto por el algoritmo. El segundo se utiliza para medir la efectividad en la clasificación del concepto aprendido. Subsecuentemente se crea un modelo de clasificación; cuando éste presenta una baja tasa de acierto, ocurre un sub-ajuste o *underfitting*. Ya cuando la tasa de acierto es bastante alta se produce un sobre-ajuste del modelo u *overfitting* (Monard & Baranauskas, 2003).

Las fases del proceso de aprendizaje supervisado son: (a) selección de atributos que constituye en la selección de un subconjunto de atributos; (b) medida de proximidad para medir el grado de semejanza entre dos objetos analizados; (c) criterio de agrupamiento para agrupar los objetos según sus altas o bajas similitudes, (d) algoritmo de agrupamiento, donde se elige el algoritmo que será utilizado; (e) verificación de los resultados e (f) interpretación de los resultados (Han et al., 2012). Estas fases también pueden ser aplicadas a los procesos de aprendizaje no-supervisado que detallaremos a continuación.

B. Aprendizaje automático no-supervisado

En el aprendizaje automático no-supervisado no se utilizan los *datasets* de entrenamiento para que el algoritmo aprenda sobre un concepto. Por lo tanto, los ejemplos no están previamente rotulados de modo que se

determina una medida de calidad para que el algoritmo agrupe las entradas que le son sometidas. Este método se suele utilizar cuando se buscan patrones o medidas de categorización que faciliten la comprensión de la información. Se intentan descubrir similitudes y diferencias entre los patrones existentes que permitan derivar en conclusiones útiles en relación a los datos (Souto, Lorena, Delbem, & Carvalho, 2003).

Concretamente, se puede decir que los métodos no-supervisados son aplicados como paso previo a los métodos supervisados con el fin de encontrar patrones y realizar agrupaciones en muestras de datos no estructurados. Subsecuentemente se aplica el método supervisado para la rotulación/clasificación de estas agrupaciones.

Las fases del proceso de aprendizaje no-supervisado son: (a) selección de atributos, (b) medida de proximidad, (c) criterio de agrupamiento, (d) algoritmo de agrupamiento, (e) verificación de los resultados e (f) interpretación de los resultados.

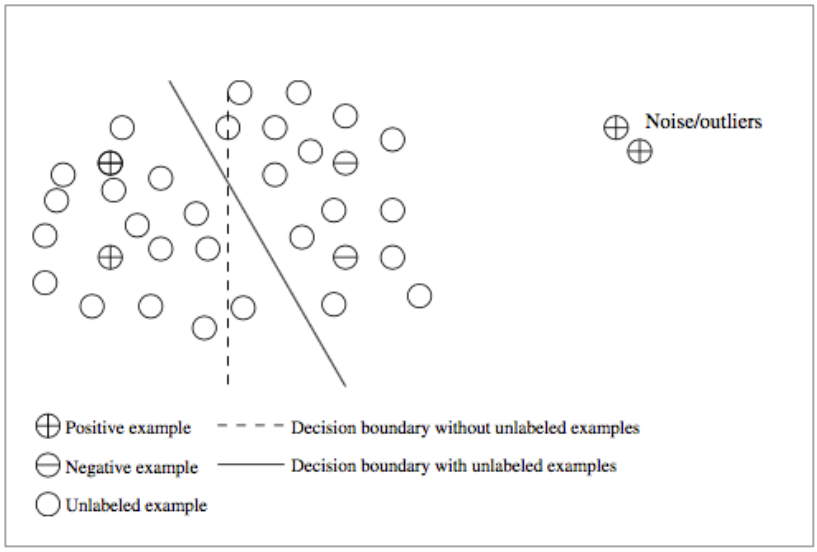
C. Aprendizaje automático semi-supervisado

Se trata de una clase de técnicas que utiliza ejemplos rotulados y no rotulados para generar el modelo de aprendizaje. En este enfoque, los ejemplos etiquetados se utilizan para aprender modelos y conceptos de clases (ej. pos/neg) y los ejemplos no etiquetados se utilizan para perfeccionar los límites entre las clases (Han et al., 2012).

Para un problema de dos clases, podemos pensar en el conjunto de ejemplos que pertenecen a la clase de los ejemplos considerados positivos y los que pertenecen a la otra clase de los ejemplos considerados negativos.

Como se observa en la Figura 9, cuando consideramos también los ejemplos no rotulados, la clasificación tiende a ser más precisa. De modo que – como vemos en el ejemplo – una entrada que antes era de clase negativa cambia de clase y pasa a ser neutra. En el otro caso, se observa como una entrada de clase positiva pasa a ser de clase negativa.

Figura 9. Aprendizaje semi-supervisado.



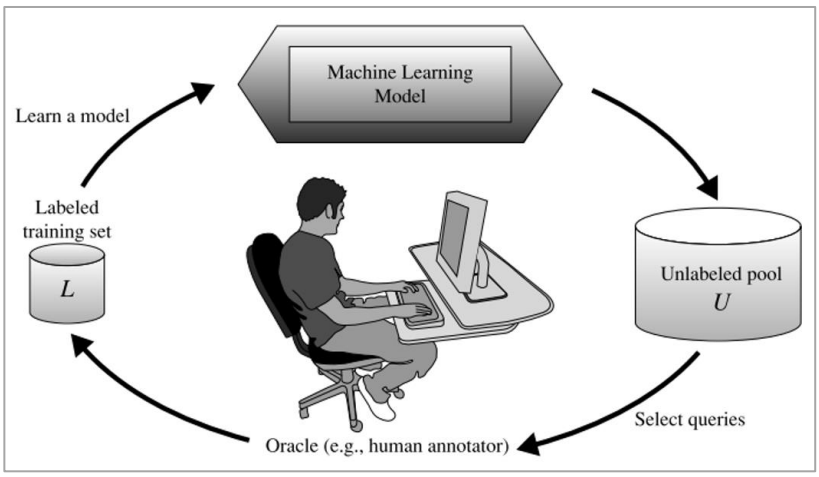
Fuente: Han et al. (2012).

D. Aprendizaje maquina activo

El enfoque de aprendizaje automático activo recurre a usuarios como parte fundamental del proceso de aprendizaje.

Se suele aplicar utilizando a un experto de una determinada área en la clasificación de los datos. El objetivo es ampliar la calidad del modelo mediante el conocimiento humano (Figura 10) (Han et al., 2012). Sin embargo, este tipo de aprendizaje es poco utilizado actualmente debido al gran volumen de datos que se maneja.

Figura 10. El ciclo de aprendizaje activo con base en grupos.



Fuente: Han et al. (2012).

2.5.4. MÉTODOS CLÁSICOS DE APRENDIZAJE AUTOMÁTICO

A continuación, se describen los cuatro métodos de aprendizaje automático más frecuentes en la literatura respecto al Análisis de Sentimientos (Liu, 2012; Tsytsarau & Palpanas, 2012).

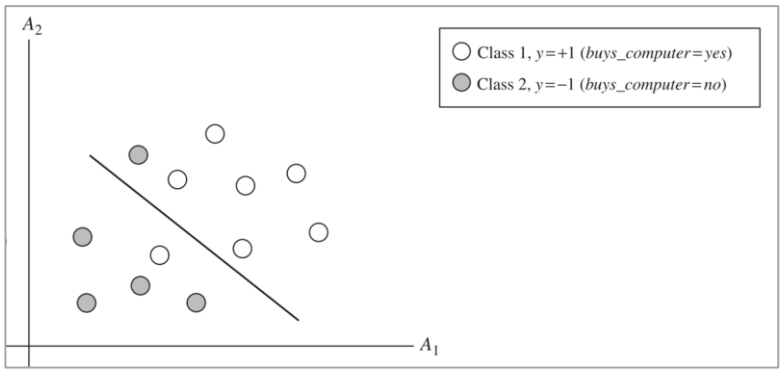
Estos son: *SVM (Support Vector Machines), Naïve Bayes, Decision Trees, Random Forest, Redes neuronales y Deep-Learning*.

2.5.4.1. SVM (SUPPORT VECTOR MACHINES)

El método SVM fue desarrollado inicialmente por AT&T Bell Laboratories, dirigidos por Vapnik y su equipo (Vapnik, Golowich, & Smola, 1997). Inicialmente aplicado al mundo de la industria y posteriormente llevado a las esferas más comunes, en un corto período de tiempo este método se convirtió en unos de los mejores sistemas disponibles para las tareas de reconocimiento de objetos (Bernhard Schölkopf & Burges, 1999).

Las SVM tienen base en la teoría de aprendizaje estadística que establece una serie de principios que deben ser seguidos para la obtención de clasificadores capaces de predecir correctamente nuevos datos a parte de un aprendizaje precedente (Vapnik & Chervonenkis, 2015). Concretamente consiste en representar cada documento (o texto en el contexto de esta tesis) a través de un punto o vector en espacio n-dimensional y trazar un hiperplano que separe de manera óptima las clases (mínimo dos) en cuestión (Pilászy, 2005). La correcta separación de hiperplano maximiza el margen de acierto en la clasificación de los datos de entrenamiento (Figura 11).

Figura 11. Hiperplano Support Vector Machines



Fuente: Han et al. (2012).

El vector con márgenes rígidos define fronteras directas a partir de datos linealmente separables. Sea T un conjunto de entrenamiento con n datos $X_i \in X$ y sus respectivos rótulos e $i \in Y$, en que X constituye el espacio de los datos e $Y = \{-1, +1\}$. T es linealmente separable si es posible separar los datos de las clases +1 y −1 por un hiperplano (B Schölkopf & Smola, 2002).

Debido a sus excelentes resultados, en este trabajo utilizamos la versión SVM^{light} (Joachims, 1999) ampliamente utilizada en la literatura para gestionar tareas de clasificación binarias (Esuli & Sebastiani, 2010; Pang & Lee, 2008).

2.5.4.2. NAÏVE BAYES (REDES BAYESIANAS)

El método Naïve Bayes desarrollado por Thomas Bayes en el siglo XVIII (Han, Kamber, & Jian, 2012, 350), es uno de los métodos más utilizados para clasificar textos ya que requiere baja capacidad de procesamiento hardware y ofrece resultados rápidos y efectivos.

Este método solo requiere una pequeña cantidad de datos de entrenamiento para estimar las medias y varianzas en la tarea de clasificación. Básicamente asume que los elementos en el conjunto de datos son independientes unos de otros y sus ocurrencias en diferentes conjuntos de datos indican su relevancia para ciertos atributos (Langley, 1992).

Por ejemplo, una fruta puede ser considerada como una manzana si es roja, redonda y tiene alrededor de 4 pulgadas de diámetro. De este modo, Naïve Bayes considera todas estas propiedades que contribuyen de forma independiente a la probabilidad de que esta fruta sea una manzana.

En la literatura Naïve Bayes ha sido utilizado en diferentes tareas como categorización de documentos, detección de spam en correo electrónico, clasificación de correo electrónico por prioridad y también en la detección de contenido sexual explícito (Han et al., 2012).

Esto solo es posible debido a sus diferentes variaciones como:

Multinomial Naïve Bayes utilizado cuando el número de ocurrencia de una palabra influye significativamente en la clasificación del texto (Pan et al., 2018; Sarkar & Bhowmick, 2018).

Binarized Multinomial Naïve Bayes utilizado cuando el número de ocurrencia de una palabra no influye significativamente en la clasificación, basta con que la palabra esté o no presente (Gupte, Joshi, Gadgul, & Kadam, 2014; Bo Pang & Lee, 2008).

Bernoulli Naïve Bayes utilizado cuando la ausencia de alguna palabra influye en el problema de la clasificación (Bhuta, Doshi, Doshi, & Narvekar, 2014; Pang & Lee, 2008; Wilson et al., 2009).

2.5.4.3. ÁRBOL DE DECISIÓN (DECISION TREE)

El árbol de decisión (*Decision Tree*) pertenece a la familia de algoritmos inductivos Top Down (*Top Down Induction of Decision Trees*). El método representa a los resultados gráficamente en forma de árbol, donde cada nodo interno indica una prueba realizada sobre un atributo. Las conexiones entre los nodos representan el resultado del test superior y las hojas indican a qué clase pertenece el registro en cuestión (Quinlan, 1986).

Las implementaciones de esta técnica de clasificación tienden a ser pequeñas y utilizan algoritmos estándares como el ID3 que selecciona los atributos para dividir los datos mediante el uso de la información de ganancia. El método C4.5, utilizado por Li & Jain (1998), se basa en el método de podas derivado del inglés "*pruning".* Este reduce el tamaño del árbol eliminando las partes que no tienen gran relevancia en la calificación (Quinlan, 1986).

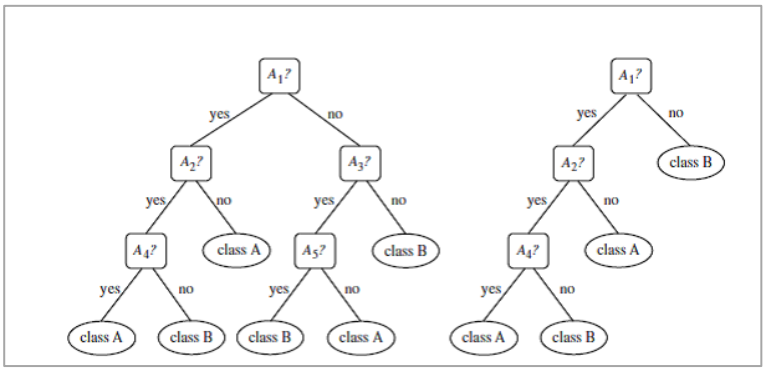
Existen básicamente dos métodos de poda:

Pré poda (Pre-pruning), se utiliza cuando algunos ejemplos son ignorados ya en la generación de la hipótesis.

Pos poda (Pos-pruning), se utiliza cuando son eliminadas algunas partes tales como ramas u hojas del árbol al final del proceso, una vez la hipótesis ya hayan sido generadas (Dietterich, 2000).

En la Figura 12 se pueden observar los cambios en el árbol de decisión antes y después del proceso de poda.

Figura 12. Árbol de decisión antes y después del proceso de poda.



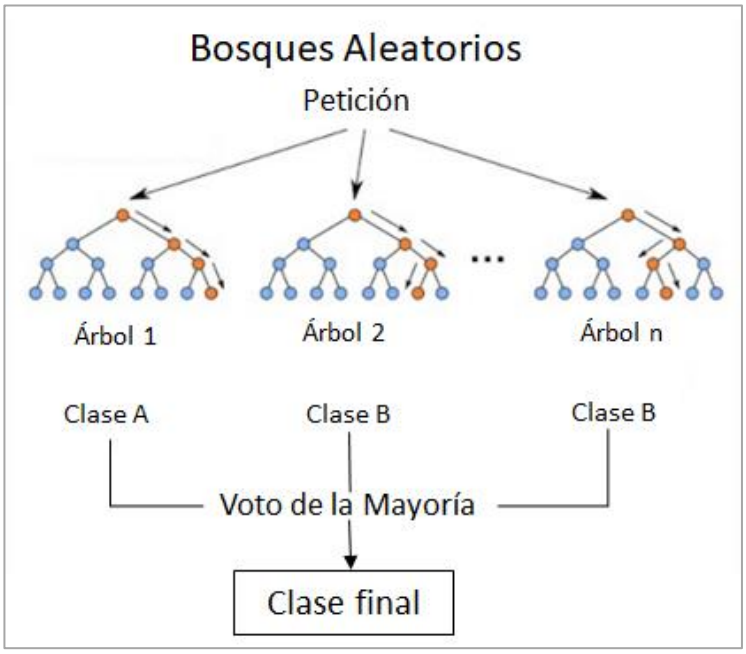
Fuente: Han et al. (2012).

2.5.4.4. BOSQUES ALEATORIOS (*RANDOM FOREST*)

Este método de aprendizaje automático es derivado de los Árboles de Decisión comentados en el epígrafe anterior. Se basa en una multitud de árboles de decisión donde la palabra aleatorio (*random)* significa que cada árbol tiene igual probabilidad de ser muestreado.

El clasificador utiliza el criterio de mayoría de votos aplicado a varios árboles y devuelve la clase con mayor número de votos (Liaw & Wiener, 2002; Segal, 2003). Los diversos árboles resultado de esta técnica pueden ser generados de forma eficiente y la combinación de grandes conjuntos llevan a modelos de gran precisión (Zhao & Zhang, 2008) (Figura 13).

Figura 13. Proceso de clasificación de los Bosques Aleatorios.



Fuente: Romain (2019).

Formalmente este método es definido por una colección de árboles $\{h_k(x)\}$, $k = 1, 2, \ldots, L$, donde h_k son muestras aleatorias independientes e idénticamente distribuidas, de manera que cada árbol vota en la clase más popular para la entrada x (Breiman, 2001; Han et al., 2012).

2.5.4.5. APRENDIZAJE PROFUNDO (*DEEP LEARNING*) Y REDES NEURONALES (*RNA*)

El Aprendizaje Profundo es una subárea del Aprendizaje Automático que investiga técnicas para simular el comportamiento del cerebro humano en tareas direccionados más específicamente a los análisis y reconocimiento de señales (como audios y habla) y contenidos visuales (como imágenes y

vídeos), aunque también se puede utilizar para textos (Zhang, Wang, & Liu, 2018).

Es necesario conocer algunos prerrequisitos para entender cómo funciona el Aprendizaje Profundo como, por ejemplo, el aprendizaje automático, el procesamiento de imágenes, clasificación, redes neuronales multicapas, aprendizaje no-supervisado, filtrado y convolución (David & Shwartz, 2014).

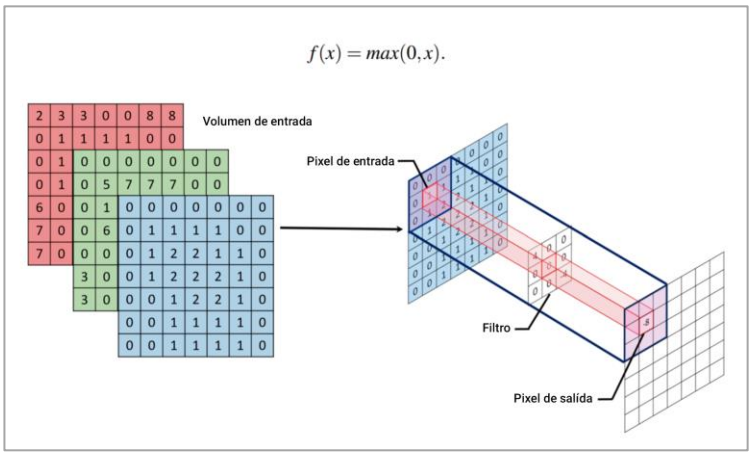
Concretamente se le puede definir como un conjunto de técnicas de Aprendizaje Automático que se utiliza a múltiples capas de procesamiento de información bajo arquitecturas jerárquicas de alto nivel para análisis de patrones y extracciones de características de los datos por medio de una Red Neural Artificial *(RNA)*(Nisbet, Miner, & Yale, 2018). En otras palabras, los métodos que utilizan el Aprendizaje Profundo buscan descubrir un modelo, por medio de un conjunto de ejemplos, y un método (con base en capas) que conduzca el aprendizaje del modelo a partir de esos ejemplos.

Al final del proceso de aprendizaje, se obtiene una ecuación capaz de recibir datos brutos como entrada y ofrecer como salida una representación adecuada para el problema en cuestión (Goodfellow, Bengio, & Courville, 2016).

Relacionado con las capas utilizadas en el Aprendizaje Profundo, mencionadas anteriormente, éstas de definen principalmente por:

1. **Capa Convolucional**: se compone por un conjunto de filtros *(kernels)* que son aplicados al objeto de estudio. Estos filtros son convoluídos[4] con los datos de entradas para obtener un mapa de características (Figura 14).

Figura 14. Proceso ejecutado por la capa convolucional.



Fuente: Romain (2019).

Son tres los parámetros que manejan el tamaño del volumen resultado de la camada convolucional: profundidad (*depth*), paso (*stride*) y zero-padding (Ujjwal, 2016). Se observa que la profundidad del volumen resultante es igual al número de filtros usados. Cada filtro es responsable por extraer características distintas en el volumen de entrada. De modo que cuanto mayor el número de filtros, mayor será el número de características extraída, sin embargo, la

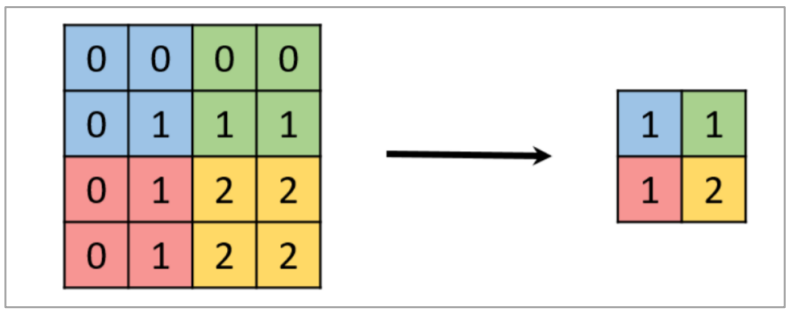
[4] Convolución es un operador matemático que transforma dos funciones f y g en una tercera función que en cierto sentido representa la magnitud en la que se superponen f y una versión trasladada e invertida de g.

complexidad computacional y el tempo de análisis también será mayor (Nisbet et al., 2018).

2. **Capa de *Pooling***: Subsecuentemente a una capa convolucional, por norma generar existe uma capa de *pooling*. Ésta se aplica para reducir el tamaño espacial de las matrices resultantes de la convolución y capturar pequeñas variantes con informaciones más representativas. Esto reduce el número de parámetros a ser aprendidos y contribuyen para el control de sobre ajuste *(overfitting*) (Karpathy, 2017).

   En esta operación, los valores de una determinada región del mapa de atributos generados por las capas convolucionales son sustituidos por alguna métrica de esta misma zona. Esta operación es conocida como *max pooling*, muy eficiente em eliminar valores inútiles de modo que reduce la dimensión de la representación y acelera el proceso computacional (Figura 15) (Goodfellow et al., 2016).

Figura 15. Proceso de max pooling aplicado a una imagen 4x4 utilizando un filtro 2x2.

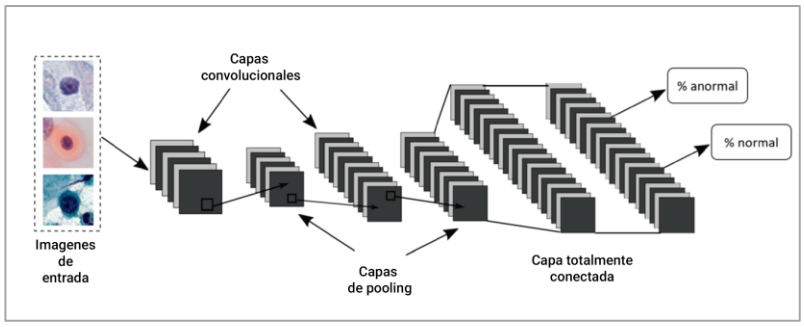


Fuente: Goodfellow, Bengio y Courville (2016).

Capa totalmente conectada: los resultados procedentes de las capas convolucionales y de *pooling* representan características extraídas de la imagen de entrada.

La capa totalmente conectada, tiene como objetivo principal utilizar la información proveniente de la aplicación de las capas anteriores para clasificar la imagen en una clase predeterminada (Figura 16).

Figura 16. Proceso extracción de características de una imagen y su posterior clasificación.



Fuente: Goodfellow, Bengio y Courville (2016).

Las capas totalmente conectadas, actúan de forma similar a uma red neural artificial convencional (*Multi Layer Perceptron ou MLP*) (Goodfellow et al., 2016).

Estas capas son formadas por unidades de procesamiento también conocidas como neuronas y el término “totalmente conectadas” indica que todas las neuronas de la capa anterior están conectadas a la capa que le sigue.

2.5.5. MÉTODOS DE PRE-PROCESAMIENTO DE DATOS

A veces es necesario preparar y limpiar los datos para que los métodos comentados anteriormente funcionen adecuadamente. Además de los procedimientos habituales de detección de anomalía, valores perdidos, eliminación de duplicados, reducción de la dimensionalidad, etc., típicos de cualquier procedimiento de MD, en el caso de esta tesis donde trabajamos con minería de textos, es necesario además recurrir a procedimientos típicos relacionados con el Procesamiento del Lenguaje Natural (PLN) como:

POS: (*Part of speech tagging* o *POS-tagging*), es cuando se asigna un valor a gramatical cada término de la sentencia. Esta técnica permite la identificación de nombres, adjetivos, adverbios entre otros elementos gramaticales que son utilizados como indicadores en la clasificación de sentimiento (Turney, 2002).

Frecuencia y términos: Se tiene en cuenta la presencia y frecuencia de uni-gramas o n-gramas[5] en los datos. En la literatura, Dave et al. (2003) y Pang et al. (2002) prueban con esta técnica y afirman que los bi-gramas y tri-gramas son mejores al clasificar polaridad en revisiones

[5] Se llama n-grama a una subsecuencia de n elementos consecutivos en una secuencia. Si n=2 se denominan bigramas; n=3, trigramas; para n>=4 entonces se llaman genéricamente n-gramas o Modelos de Markov de orden (n-1).

de productos. Para revisiones de películas los uni-gramas y bi-gramas alcanzan mejores resultados.

Opiniones de palabras o frases: Esta técnica es bastante usada para extraer sentimientos y se compone básicamente de dos enfoques: *lexicon-based or statistical-based*. Hu and Lui et al. (2004) determinaron la polaridad de grupos de datos utilizando el léxico de sentimiento SentiWordNet citado anteriormente.

2.6. HERRAMIENTAS DE MONITOREO DE MEDIOS SOCIALES WEB

Ha quedado evidente que hoy, gran parte de la interacción entre las personas en internet ocurre en los medios sociales que, su vez, son utilizados por las organizaciones para identificar el comportamiento y las opiniones de sus clientes. De modo que estos medios representan una poderosa fuente de datos para la investigación de mercados, el marketing y la toma de decisiones empresariales.

Son muchos los medios sociales útiles para esta finalidad, como son Facebook, Google +, Instagram, Twitter, foros web y sitios web de comercio electrónico. Analizar los datos de estos medios para comprender mejor por qué los clientes compran un producto o servicio, por ejemplo, tiene un roll importante en el mantenimiento de la ventaja competitiva en las organizaciones (Lee, 2018).

Según Accenture (2018) las empresas reconocen el valor del análisis de los medios sociales en la innovación, en el desarrollo de productos, en los servicios ofrecidos al cliente y en sus operaciones estratégicas. Para monitorear y analizar esta información existen diversas herramientas gratuitas y de pago que ofrecen distintos recursos. Algunas son más completas que otras y ofrecen una gran variedad de recursos para monitoreo de medios y analítica web. Otras son más especializadas en alguna tarea específica, como, por ejemplo, seguimiento de la reputación en línea, escucha externa, métricas, extracción de datos, etc. Con estas herramientas, las empresas innovan y encuentran nuevas formas de recopilar, combinar y analizar automáticamente esa información para comprender mejor a sus clientes, el mercado y su competencia, administrando relaciones web, diseñando nuevos productos, servicios, etc.

En este sentido, a continuación se describen brevemente once opciones de herramientas de monitoreo y analítica web disponibles actualmente.

Audiense: Esta herramienta se centra exclusivamente en Twitter, y como su propio nombre sugiere, su objetivo principal es identificar y comprender los datos relevantes de tu audiencia para tu negocio ampliando su alcance social. Entre sus funcionalidades más potentes se encuentran la capacidad de crear reglas automatizadas para, por ejemplo, generar respuestas o crear listas en función del comportamiento de los usuarios en Twitter, el monitoreo de miles de tuits a la vez, identificar cuáles son los intereses de tus

seguidores, su ubicación geográfica, idioma que hablan entre otros. La descarga de Audiense se da desde https://es.audiense.com/.

Brandwatch: Se le conoce por permitir a sus usuarios identificar la opinión de los consumidores acerca de cualquier tema en los medios sociales. Se trata de una herramienta muy completa tanto para grandes y complejos proyectos como para estudios más sencillos y a un nivel local. Uno de sus diferenciales, por ejemplo, es que te da la posibilidad de monitorear a Sina Weibo, la red social china similar al Facebook. También permite, por ejemplo, recuperar mensajes de Twitter desde el año de 2011. Finalmente permite integrar datos del CRM, email marketing, entre otras herramientas, para tener una visión general y amplia de todas las acciones de marketing que quieras. La descarga de Brandwatch se da desde https://www.brandwatch.com/.

Google Analytics: Es una de las herramientas de análisis y monitoreo de sitios web (ej. foros, blogs y comercios en línea) más utilizada en el mundo. Esta herramienta muestra principalmente información relacionada con volumen de visitas a un sitio web, cuánto tiempo un usuario ha estado en una página, así como las conversiones y todo su recorrido de navegación. También puede decir, por ejemplo, de qué ciudad son las personas más interesadas en un sitio web, las horas con mayor volumen de acceso, tasas de abandono, entre otros. También se puede utilizar para monitorear además de sitios web, a medios sociales como Twitter, sin embargo, no es

habitual. Para ello, Twitter y Facebook, por ejemplo, tienen sus propios recursos de analítica como son el Twitter Analytics (https://analytics.twitter.com/about) y el Facebook Analytics (https://analytics.facebook.com/). El acceso al Google Analytics se da desde https://analytics.google.com/.

Hootsuite: Es una herramienta considerada bastante completa que abarca a múltiples redes, como Twitter, Instagram, Facebook, LinkedIn, WordPress, Foursquare y Google+. Es muy conocida por sus funciones como informes semanales y su facilidad de manejo. Ofrece monitorear términos de búsqueda específicos en tiempo real lo que facilita rastrear las menciones de una marca, productos o palabras clave relevantes que te interesan. Según sus creadores sus características fundamentales son, programado de publicaciones, selección de contenido web, informes de performance, análisis de medios sociales, monitoreo, gestión de equipos además de contar con aplicaciones integradas de terceros a su ecosistema. La descarga de Hootsuite se da desde https://www.hootsuite.com/.

Meltwater: Es una plataforma integrada muy completa que ofrece soluciones como escucha y supervisión de medios en múltiples canales, analíticas sociales, identificación de tendencias y temas centrales, recuperación y acceso a información antigua, filtrado por medio social, es multilingüe, identifica los usuarios por geolocalización y realiza análisis de sentimientos. Además, permite programar y publicar mensajes a partir de un calendario de

planificación. La descarga de Meltwater se da desde https://www.meltwater.com/es/monitoreo-de-medios/.

Metricool: Es una herramienta se caracteriza por su capacidad gestión de una gran cantidad de perfiles (hasta 50 en su plan más completo). Por medio de un cuadro de control ofrece un extenso abanico de información y analíticas sobre las publicaciones. Para cada perfil que se incluya en Metricool, es posible conectar distintas fuentes webs, medios sociales, blogs y etc. Por ello, esta opción termina por ser muy versátil y recomendable para la escucha activa y monitoreo de redes a gran escala de fácil manejo y bastante intuitiva. La descarga de Metricool se da desde https://metricool.com/.

RapidMiner: Se trata de una herramienta de ciencia de datos con foco en las técnicas de minería de datos y recursos de aprendizaje automático que, además, agrega, mediante extensiones de terceros, algunos recursos de monitoreo de medios sociales y analítica web. A través de un ambiente visual de programación, RapidMiner posibilita al usuario construir de manera rápida distintos modelos de análisis de datos. Aunque no es una herramienta exclusivamente dedicada al monitoreo de medios, ofrece un extenso abanico de plugins que se le incorporan para este fin como, por ejemplo, extracción y procesamiento de datos, análisis de sentimientos, análisis de geolocalización, informes gráficos, entre otros. La descarga de RapidMiner se da desde https://rapidminer.com/.

SenticNet: Fue creado en el *MIT Media Laboratory* en el año del 2009 dentro de un proyecto de investigación industrial de Cooperative Awards in Science and Engineering (CASE). El objetivo principal de SenticNet es hacer que la información conceptual y afectiva transmitida por el lenguaje natural sea mejor entendida por las máquinas. Su enforque principal se da en las tareas de Análisis de Sentimientos utilizando el modelo de bolsa de conceptos, en lugar de simplemente contar las frecuencias de coincidencia de palabras y en la indexación semántica aprovechando los patrones lingüísticos para permitir que los sentimientos fluyan de un concepto a otro en función de la relación de dependencia entre ellos. La descarga de SenticNet se da desde https://sentic.net/downloads/.

Social Studio: Con esta herramienta es posible escuchar, analizar y publicar de forma integrada en los medios sociales. Permite crear y publicar contenidos para distintas audiencias y en diversos medios sociales, con opciones de edición, anexo de imágenes, vídeos, incluso te da la posibilidad de incluir contenido de otras aplicaciones.

Permite también realizar tareas de analítica web y escucha activa por medio de un ambiente web dinámico que, interactúa con el usuario en tiempo real. La descarga de Social Studio se da desde https://socialstudio.radian6.com/login/.

Socilbakers: Permite medir el rendimiento de cuentas en los medios sociales, comparar resultados con otros actores de la red y generar informes sobre todas sus actividades en los medios web.

Socialbakers utiliza la Inteligencia Artificial (IA) para encontrar su público objetivo en los medios sociales, el contenido que más les interesa y las personas influyentes de la red.

Es conocida por ser bastante completa y profesional. Algunos de sus principales clientes son Toyota, Orange, Desigual, Lexus, entre otros. Sus principales recursos de inteligencia competitiva son, gestión de audiencia e influenciadores, identificación y generación de contenido inteligente, analítica y monitoreo de medios sociales y herramientas de *community management.* La descarga de Socialbakers se da desde https://www.socialbakers.com/.

Weka: Se trata básicamente de una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Contiene herramientas para la preparación de datos, clasificación, regresión, agrupamiento, minería de reglas de asociación y visualización.

Aunque su enfoque no sea especialmente el monitoreo de medios sociales, Weka es muy utilizada en las tareas de analítica web por ser una herramienta gratuita, versátil, muy potente y en constante actualización por la comunidad por ser clasificada como software libre. La descarga de Weka se da desde https://www.cs.waikato.ac.nz/ml/weka/.

Tabla 1. Herramientas de monitoreo de medios sociales web (MD=Minería de Datos - AS=Análisis de Sentimientos Automático - CMS=Conexión a Medios Sociales - Geo=Geolocalización de la información - Gratuito= (S)sí, (N)no, (F)freemium, (P)prueba.

Herramienta	MD	AS	CMS	Inf. Gráficos	Multilingüe	Interfaz Web	Necesidad de Login	Gratuito	Geo	Código Libre
Audiense	si	si	si	si	si	si	si	P	si	no
Brandwatch	si	si	si	si	si	si	si	P	si	no
Google Analytcs	si	no	si	si	si	si	si	F	si	no
Hootsuite	si	si	si	si	si	si	si	F	si	no
Meltwater	si	si	si	si	si	si	si	T	si	no
Metricool	si	si	si	si	si	si	si	F	si	no
Rapidminer	si	si	si	si	si	no	si	no	si	no
SenticNet	si	no	no	si	no	si	no	F	no	no
Social Studio	si	si	si	si	si	si	si	P	si	no
Socilbakers	si	si	si	si	si	si	si	P	si	no
Weka	si	si	si	si	si	no	no	S	no	si

Fuente: Elaboración propia.

La Tabla 1 lista las herramientas descritas arriba teniendo en cuenta los diez aspectos clave en este ámbito: a) si la herramienta realiza tareas de Minería de Datos (MD) y de Análisis de Sentimiento (AS); b) si se conecta automáticamente a medios sociales (CMS); c) si ofrece informes gráficos; d) si trata con distintos idiomas; e) si ofrece una interfaz web; f) si exige login por parte del usuario; g) si es gratuita; h) si ofrece geolocalización de la información, y finalmente i) si es de código libre.

Como se observa en la tabla, se corrobora la principal motivación de esta tesis, ya mencionada en el Capítulo 1, que busca aportar una solución frente a la **escasez de herramientas de monitoreo de medios sociales que sean 100% gratuitas, que hagan tareas de Minería de datos y Análisis de Sentimientos (ej. extracción, procesamiento, visualización y clasificación de la información) en múltiples idiomas a través de una interfaz sencilla e**

intuitiva, que sean de código libre y que además estén pensadas con un enfoque técnico y de diseño desde la perspectiva del marketing.

2.7. APLICACIONES Y TRABAJOS RELACIONADOS

Son muchas las aplicaciones y trabajos relacionados con el uso de los medios sociales por parte de las empresas. Esto ocurre principalmente porque que estos medios actúan muchas veces como sistemas de recomendación, resultado de la alta interacción entre sus usuarios. En este sentido, los contenidos generados por los internautas aportan opiniones y sentimientos sobre lo que ocurre en el mundo real de modo que la información proveniente de medios sociales se constituye como uno de los valiosos pilares sobre los que se sustentan las empresas contemporáneas (Epstein, 2018; Lv, Yu, & Wu, 2018; Wamba & Carter, 2016).

Los usos y utilidades de estos tipos de estudios son múltiples y facilitan, por ejemplo, conocer la influencia de un anuncio, marca o producto en los usuarios, identificar las tendencias de mercado, conocer valoraciones globales de la propia empresa, de la competencia, entre otros recursos que brindan a las organizaciones la posibilidad de poder realizar pronósticos futuros y tomar decisiones más acertadas. Además, los medios sociales también desempeñan un rol fundamental en la calidad de las relaciones entre las marcas y sus consumidores, influyendo directamente en la gestión de su reputación (Toldos & Castro, 2013).

En este sentido son diversos los trabajos que corroboran las ventajas de las prácticas de monitoreo de los medios sociales en los más variados contextos, sobre todo de Twitter debido al dinamismo y alcance que este medio ofrece.

En el área de la **educación**, Rinaldo, Tapp y Laverie (2011) en su estudio llamado *"Learning by Tuiting: Using Twitter as a Pedagogical Tool"* intentan responder a la pregunta, ¿Pueden los profesores usar Twitter para involucrar a los estudiantes? Los autores argumentan que Twitter tiene muchos beneficios para los educadores que están interesados en involucrar a los estudiantes en el aprendizaje experiencial. Para estos autores, por medio del ambiente dinámico que ofrece Twitter, los profesores pueden comunicarse directamente con los estudiantes y generar discusiones de interés en los temas y ejemplos que demanda el curso. Del mismo modo que los especialistas en marketing usan Twitter para generar interés y debates sobre la marca, los educadores pueden usar Twitter con el mismo propósito en sus cursos. Para estos autores, Twitter es un método rápido y fácil para emitir comunicados, resolver problemas de los estudiantes y realizar tareas administrativas. Además, sus estudios sugieren que cuando los estudiantes se involucran en el uso de Twitter con el profesor, se sienten mejor preparados para ingresar en las universidades.

En la misma línea, Rath (2011) defendió el uso de Twitter en un ambiente de aprendizaje virtual, con el objetivo de determinar los diferentes niveles de participación así como la sensación de comunidad percibida por los involucrados. Suliman (2010), Duarte, Brito y Medeiros (2009) y Borau,

Ullrich, Feng y Shen (2009), en sus estudios también relacionados con la educación y los medios sociales, traen a la luz el interés en añadir Twitter a las herramientas disponibles para el profesorado, tanto en la enseñanza presencial como en la enseñanza a distancia. Así mismo, estudian la aplicación de métricas de análisis de medios sociales como ambientes colaborativos de aprendizaje. Orto estudio relevante en esta área, titulado "*A comparison of students Twitter use in a postsecondary course delivered on campus and online*" (Peters, Crane, & Costello, 2019) propuso identificar diferencias entre las percepciones de los estudiantes en el uso de Twitter como un dispositivo para el aprendizaje en el aula. Los autores se basaron en resultados de encuestas de 37 estudiantes inscritos en un curso de Sociología para determinar el uso y percepciones sobre la utilidad de Twitter como herramienta de evaluación de cursos y de conexión entre los estudiantes.

Relacionado con las métricas utilizadas en medios sociales para medir el aprendizaje, Magnani, Montesi y Rossi (2011) introducen un nuevo paradigma en base a las investigaciones en redes de micro-mensajes. Utilizan técnicas de Recuperación de Información (IR), del inglés *"Information Retrieval",* y métricas de Análisis de Medios sociales (SNA), del inglés *"Social Network Analysis*". Así pues, Bakharia y Dawson (2011) utilizan métricas de SNA como apoyo para la visualización de las relaciones entre los participantes de foros de discusión de Ambientes Virtuales de Aprendizaje (*AVAs*), permitiendo estrategias de intervención y la potenciación del aprendizaje. Prieto (2016), analiza una campaña en Twitter realizada por estudiantes de primero de Derecho, con el objetivo de

acreditar la actualidad del concepto de interactividad reflexiva y crítica en el ámbito de la educomunicación, término que utiliza el autor para fusionar las palabras educación y comunicación. El autor busca desarrollar la dimensión social, ética y política del proceso de aprendizaje a partir de un fundamento práctico y cooperativo.

En el ámbito de la **moda** los autores Abdelfattah, Galal, Hassan, Elzanfaly, y Tallent (2016), se centran en analizar las reacciones generadas por las 50 mejores marcas de moda en Instagram dadas sus 20 mejores imágenes con el mayor número de me gustas. El enfoque adoptado en este estudio es calificar la estética visual de las imágenes de moda y establecer por qué algunas marcas triunfan en los medios sociales más que otras. Los autores intentan identificar cuáles estéticas visuales más atraen a los usuarios en base a una medida a la que nombran 'Valor Social'.

En la literatura encontramos también trabajos aplicados al área de la **salud**. Kandadai et al. (2016), presentan un método para que las partes interesadas puedan evaluar y optimizar el uso de Twitter para difundir información sobre salud. Bhattacharya, Srinivasan y Polgreen (2014) investigan sobre la participación de las agencias federales de salud de los EE. UU. en Twitter, y se centran en distintos componentes como: a) el número de retuits; b) el tiempo entre el tuit de la agencia y el primer retuit y c) el tiempo entre el tuit de la agencia y el último retuit. Estos hallazgos contribuyen al desarrollo de futuros experimentos controlados con el fin de aumentar el compromiso de salud pública a través de Twitter.

Por otro lado, Pemmaraju, Thompson y Qazilbash (2017), reconocen en su trabajo que Twitter se utiliza cada vez más para la recopilación de información y la diseminación de ideas tanto en la práctica médica como en la investigación científica. En este sentido, resaltan que la creación y adopción de hashtags específicos de enfermedades por parte de las partes interesadas, ha llevado a una mayor uniformidad de las discusiones médicas y que pueden ser recuperadas y referenciadas en puntos temporales posteriores.

Según los autores, a medida que se crean nuevos hashtags específicos para las enfermedades hematológicas y oncológicas, también aumenta la red de usuarios conectados interesados en el tema. En esta misma línea, Pemmaraju, Utengen, et al. (2017) afirman que recientemente se han realizado importantes esfuerzos por parte de los usuarios para ayudar a simplificar la cantidad de información que se puede encontrar en Twitter. Estos esfuerzos han llevado a la creación de comunidades médicas específicas y han mejorado enormemente la capacidad de comprender y categorizar mejor los datos disponibles en esta red social. Específicamente, para aquellos involucrados en temas relacionados con cánceres raros, la creación de hashtags direccionados y usados de manera frecuente ha llevado a la diseminación rápida y confiable de la información. Este fenómeno, fomenta la capacidad de discutir y debatir temas de interés de manera eficiente. Un ejemplo de ello está en el campo de los neoplasmas mieloproliferativos (MPN); la creación de la etiqueta #MPNSM (neoplasias mieloproliferativas en los medios sociales) en 2015 ha facilitado las

interacciones entre las partes interesadas de la asistencia sanitaria de todo el mundo en el campo MPN.

Kotsenas et al. (2018), relatan el caso aplicado a la Clínica Mayo que hizo una inversión estratégica en medios sociales. Concretamente esta clínica implementó el uso intensivo de los medios sociales, potenciando los medios de comunicación y marketing en toda la organización para mejorar la atención a los pacientes y sus familias, avanzar en la investigación médica y ampliar el conocimiento de la marca, entre otros.

En el ámbito de la **seguridad** los autores Vomfell, Härdle y Lessmann (2018), utilizaron los datos de Twitter y del Foursquare[6] para mejorar la precisión predictiva de los delitos en la ciudad de Nueva York en comparación con el uso exclusivo de datos demográficos.

Relacionado con el área de la **sociología** Hemsley, Palmer, Dann y Balandin (2018), utilizaron los datos de los medios sociales producidos por personas que usan la Comunicación Aumentativa y Alternativa del inglés AAC *(Augmentative and Alternative Communication -AAC)* que sirven como complemento en el lenguaje oral cuando, por sí sólo, este no es suficiente para entablar una comunicación efectiva con el entorno para: (1) entender

[6] https://es.foursquare.com/ (acceso el 07/08/2019).

cómo las personas utilizan el AAC en los medios sociales, e (2) identificar mayores oportunidades para mejorar la participación e inclusión en línea.

En los estudios **sociodemográficos** también se ha utilizado Twitter como recurso de investigación. En su estudio actual titulado "*Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level"*, los autores Jiang, Li y Ye (2019) en su trabajo demuestran que no solo se puede identificar cómo los factores demográficos y socioeconómicos se relacionan con el número de usuarios de Twitter, sino que también es posible medir y hacer un mapa de cómo varía la influencia de estos factores entre distintas regiones.

En los temas relacionados a **religión**, también se han hecho estudios con el objetivo de responder si las diferentes religiones están asociadas con diferentes tendencias sociales, cognitivas y emocionales, utilizando los mensajes en los medios sociales de cristianos y budistas que viven en los Estados Unidos (Chen & Huang, 2019).

En el ámbito relacionado con las **nuevas tecnologías** también se encuentra en la literatura estudios interesantes como el *"Anticipating acceptance of emerging technologies using Twitter: the case of self-driving cars"*, llevado a cabo por los autores Kohl, Knigge, Baader, Böhm y Krcmar (2018). En este trabajo, los autores presentaron un análisis, en base a los comentarios de Twitter, respecto a los riesgos y beneficios relacionados con la aceptación de tecnologías emergentes como la de los automóviles auto conducidos (*self-driving cars*). Otro estudio actual (Araujo & Kollat, 2018), se centró en investigar los factores que conducen a la eficacia de la

Responsabilidad Social Corporativa del inglés CSR (*Corporate Social Responsibility*) en Twitter a través de estrategias narrativas de cuenta cuentos (s*torytelling)* enmarcando distintas y variadas recomendaciones prácticas de aplicaciones del CSR para las empresas.

También son muy frecuentes en la literatura la utilización de la información de los medios sociales para **desarrollo de sistemas computacionales** que interpreten y clasifiquen esta información. Los autores Noureen, Qamar, Khan y Muhammad (2018), utilizan las fotos tipo *"selfie"* de Instagram para proponer un sistema llamado InstaSent de análisis de sentimientos que también incorpora técnicas de minería de datos para clasificar estas imágenes. Para ello, utilizan información como subtítulos, hashtags y comentarios de los "*selfies*". En la misma línea, Zhan, Tu, y Yu (2018), emplearon algoritmos supervisados de aprendizaje automático para crear un clasificador que identifica a tres polaridades de opinión y seis emociones en base a los subtítulos de las imágenes de Instagram. Ducange, Fazzolari, Petrocchi y Vecchio (2019), presentan un Sistema de Apoyo de Decisiones (DSS), que puede ayudar de manera eficiente a las empresas en la gestión de campañas promocionales y de marketing en múltiples canales de medios sociales como TripAdvisor, Facebook e Instagram. El DSS propuesto monitorea continuamente esos medios y recopila los comentarios de los usuarios sobre promociones, productos y servicios. Luego, a través del análisis de estos datos el DSS estima la reputación de las marcas relacionadas y proporciona informes sobre campañas de marketing además de realizar tareas de análisis de sentimientos para identificar la polaridad de los mensajes. Finalmente, los

autores Okada, Yanagimoto, & Hashimoto (2019), también proponen un clasificador de sentimientos utilizando las revisiones de productos extraídas de la página de comercio electrónico Amazon.com y TripAdvisor.com.

Las revisiones de productos de Amazon.com frecuentemente son objeto de estudios que tienen como finalidad conocer las preferencias y opiniones de los clientes. Recientemente, los autores Aryo Prakoso, Winantesa Yananta, Fitra Setyawan y Muljono (2018), usaron procesamiento de lenguaje natural y técnicas de diccionarios léxicos, con el propósito de ayudar a Amazon.com a mejorar la calidad de servicio en base a los comentarios de estos en su sitio web. En la misma línea, Haque, Saber y Shah (2018), desarrollaron un modelo computacional que polariza las revisiones de Amazon.com y aprende constantemente con base en los mensajes polarizados anteriormente.

Relacionado con el área de la **comunicación y periodismo**, Jukes (2019) ha utilizado Twitter para identificar cómo los 10 principales periodistas de Inglaterra utilizan este medio para promover su "marca personal". Orellana-Rodriguez y Keane (2018), buscaron identificar el comportamiento y las técnicas que los periodistas y los medios de comunicación utilizan para difundir sus noticias en Twitter, y sobre los factores que afectan a la atención y el compromiso de los lectores en este medio. Los autores Crisci et al. (2018), identificaron y presentaron un conjunto de métricas basadas en los datos de Twitter para predecir la audiencia de los programas de televisión programados, donde la audiencia está muy involucrada, como ocurre con los programas de *reality X Factor* y *Pechino Express* en Italia. Las

métricas se basan en el volumen de tuits, la distribución de elementos lingüísticos, el volumen de distintos usuarios involucrados y el análisis de sentimientos de los mensajes.

En el ámbito de la **política** los estudios relacionados con el uso de los medios sociales son diversos y variados. Los autores Ramos-Serrano, Fernández Gómez y Pineda (2018), llevaron a cabo un estudio con el objetivo de investigar si los partidos políticos españoles utilizaron Twitter para desarrollar una comunicación interactiva con su audiencia, o simplemente se preocuparon por transmitir mensajes durante la campaña europea de 2014. Haro-de-Rosario, Sáez-Martín y del Carmen Caba-Pérez (2018), se basaron en Twitter y Facebook para analizar a los ciudadanos españoles y sus relaciones con el gobierno local. El objetivo principal de este trabajo fue determinar cuál de estos dos medios sociales logra mayor grado de compromiso desde diversos factores como la transparencia en línea, el estado de ánimo, el nivel de actividad en los medios sociales y la interactividad ofrecida por el sitio web del gobierno local.

Todavía en el ámbito de la política pero en una línea un poco distinta, los autores Posegga y Jungherr (2019) examinaron la agenda temática de Twitter basada en hashtags populares utilizados en mensajes que se refieren a la política, comparando la agenda de Twitter con la agenda pública y las agendas de los periódicos y programas de noticias de televisión. Finalmente, Friedland, Joseph, Swire-Thompson, Grinberg y Lazer (2019), utilizaron los mensajes generados en Twitter en el período de las elecciones presidenciales del 2016 en Estados Unidos para lidiar con los

temas relacionados con la difusión de noticias falsas (*fakenews)* y sus impactos en el proceso electoral.

Relacionado al **medio ambiente**, también hay estudios interesantes como el *"Discourse over a contested technology on Twitter: A case study of hydraulic fracturing"*, donde los mensajes de Twitter fueron utilizados para medir la aprobación o desaprobación de una técnica de perforación hidráulica denominada "fracturación hidráulica". Los autores Hopke y Simis (2017), en su trabajo analizan el discurso sobre la fracturación hidráulica en Twitter durante el período de mayor controversia pública con respecto a la aplicación de la tecnología. También analizaron el papel de los activistas y su poder de influir en Twitter.

En el ámbito de la **hotelería y turismo,** trabajos recientes utilizaron principalmente las revisiones y puntuaciones de Booking.com y Tripadvisor.com para evaluar distintos enfoques de análisis de sentimientos (Alaei, Becken, & Stantic, 2019a), así como arrojar luz sobre la potencialidad de identificar los factores que están asociados con el comportamiento de revisión del cliente, atraer a más clientes para reutilizar sus servicios y predecir el comportamiento futuro de la visita del cliente a un hotel (Park, Kang, Choi, & Han, 2018). Finalmente los autores Valdivia, Luzón y Herrera (2017), utilizaron las calificaciones de Tripadvisor.es sobre tres monumentos conocidos en España: Alhambra, Mezquita Córdoba y Sagrada Familia para extraer la polaridad de casa opinión.

Relacionado al ámbito de las **empresas**, los autores Rybalko y Seltzer (2010) estudiaron la construcción y las ventajas de las relaciones en línea

examinando cómo las empresas del Fortune 500 utilizan Twitter para facilitar la comunicación con las partes interesadas. Culnan, McHugh y Zubillaga (2010), también estudiaron la actuación de las empresas del Fortune 500 en Twitter, Facebook, blogs y foros con el fin de obtener el máximo valor comercial de los medios sociales. Para ellos, las empresas obtienen valor de los entornos virtuales (VCEs), del inglés, *"virtual customer environments"*, cuando los clientes interactúan con las organizaciones de forma regular, co-creando contenido y compartiendo el poder. Si estas relaciones tienen éxito, es esperado que los clientes se sientan como "expertos" de la compañía. En este sentido es más probable que estos clientes sean leales a los productos y servicios ofrecidos por la empresa, estén más dispuestos a probar sus nuevas ofertas y se vuelvan más resistentes a alguna información negativa relacionada con la empresa. Para ello, las empresas necesitan desarrollar estrategias de implementación basadas en tres elementos: adopción consciente de decisiones, construcción de comunidades y capacidad de absorción. La adopción consciente de decisiones exige que la empresa preste una atención especial a su contexto local, el valor que se espera conseguir y los riesgos existentes antes de decidir proseguir. En otras palabras, adopción consciente de decisiones significa adoptar la innovación "correcta" en el momento "correcto" en el camino "correcto" en todas las fases de implementación. Por otro lado, la construcción de comunidades consiste en lograr que un grupo crítico de personas que se identifica con la comunidad se mantenga involucrado. Cuanto mayor sea la participación de una persona en la red, más probabilidades habrá de que contribuya co-creando con la empresa.

Además, con el tiempo, las personas desarrollan un sentido de responsabilidad hacia la comunidad en base a sus intercambios y relaciones diversas con otros miembros.

Finalmente, la capacidad de absorción se hace indispensable para reconocer, adquirir y explotar nuevos conocimientos provenientes de la comunicación con los clientes en la red. No basta con tener una comunidad próspera con participantes activos para obtener valor de las VCEs. Es necesario que las organizaciones tengan la capacidad de procesar esta información transformándola en conocimiento competitivo para la marca (Culnan et al., 2010).

Los autores Swani, Brown y Milne (2014) basándose en las teorías de comunicación y de boca en boca, investigan cómo los mercadólogos usan Twitter en contextos diversos y predicen los factores clave que influencian las estrategias usadas en cada caso. Estos realizan un análisis de contenido longitudinal y de regresión logística para evaluar una muestra de más de 7000 tuits relacionados con las compañías de la Fortune 500, para desvelar las diferencias significativas en sus estrategias de marca y venta en las relaciones B2B, del inglés *"business to business",* y B2C del inglés *"business to consumer"*.

En el trabajo titulado *"The adoption of new technology by listed companies: the case of Twitter"* (Xiong, Nelson, & Bodle, 2017), los autores investigan la adopción de Twitter por parte de compañías australianas, para difundir información de marketing y para interactuar directamente con los consumidores. El estudio se basa en una muestra de 200 empresas que

cotizan en la Bolsa de Valores Australiana (ASX). En general, los resultados indican que el nivel de acceso de una empresa a los recursos on-line puede influir en su adopción de una nueva tecnología y la manera en que se utiliza**.**

Einwiller & Steilen (2015) reconocen que los medios sociales ofrecen numerosas posibilidades para que los consumidores y las partes interesadas puedan expresar sus quejas sobre las organizaciones, pudiendo dañar la reputación de una organización. En este sentido, su trabajo analiza cómo las grandes empresas manejan las quejas en sus páginas de Twitter y Facebook. Los resultados revelan que las empresas no están adoptando por completo las oportunidades que ofrecen los medios sociales para demostrar su voluntad de interactuar y ayudar a las partes interesadas. De modo que la capacidad de respuesta organizacional es moderada y no satisfactoria. Finalmente, también afirman que las empresas comprenden que es primordial ofrecer una acción resolutiva, conectando al demandante con alguien que pueda proporcionar una solución al problema.

En este ámbito de la gestión de críticas o quejas, Grégoire, Salle y Tripp (2015) abordan el proceso de quejas de clientes en medios sociales, y diferencian entre distintas situaciones: **a) Las buenas que representan oportunidades:** (1) cuando los clientes se quejan a la compañía en línea inmediatamente después de una falta en el primer servicio y (2) cuando los consumidores publican mensajes extraordinarios; **b) Las malas que implican riesgos:** (3) cuando los clientes debaten sobre una falta sin quejarse ante la empresa y (4) cuando los consumidores hacer llegar sus quejas a intercesores o terceros; **c) Las verdaderamente feas que**

representan el pico de amenazas en línea y las crisis públicas: (5) cuando los clientes difunden publicidad negativa a través de contenido generado por el usuario y (6) cuando los competidores responden a este contenido para robar clientes.

En la misma línea, Istanbulluoglu (2017) afirma que con la creciente popularidad de los medios sociales, comprender las conductas on-line del consumidor es cada vez más importante para los investigadores, para los profesionales y para las empresas.

Centran su estudio en uno de los aspectos críticos en el manejo de reclamos en línea: el tiempo de respuesta tanto en Twitter como en Facebook. El autor utiliza datos recopilados de consumidores que se quejaron en estos medios sociales, analizando cómo los tiempos de respuesta de las compañías en los medios sociales influyen en la satisfacción del consumidor. Los participantes del estudio declararon que esperan que las empresas respondan a sus quejas dentro de 1 y 3 horas en Twitter y dentro de 3 y 6 horas en Facebook. El análisis revela que una primera respuesta rápida y concluyente conduce a una mayor satisfacción en el manejo de las quejas.

Las acciones de las empresas en Twitter también han sido objeto de estudio. Chao y Florenthal (2016), en su trabajo titulado *"A comparison of global companies performance on Twitter and Weibo"*, comparan el desempeño de algunas compañías globales en relación a la correcta creación de contenido para clientes con diferentes antecedentes culturales y también si las empresas varían en cuanto a la utilización de la

interactividad en los sitios de micro-mensajes. Se utilizaron dos medidas validadas para analizar 548 tuits y 589 weibos publicados por American Dell, Nike, Chinese Lenovo y Li-Ning durante un período de un mes en Twitter en los EE. UU. y en Weibo en China. Los resultados finales indicaron que estas empresas se desempeñaron mejor en términos de interactividad en Weibo que en Twitter.

Relacionado con el **mercado minorista**, los autores Bhattacharjya, Ellison y Tripathi (2016) afirman que el comercio minorista electrónico está intrínsecamente relacionado con la eficacia de sus procesos logísticos que inevitablemente involucran a proveedores de servicios externos.

Es lógico que los clientes esperen que el minorista resuelva las consultas relacionadas con la entrega, incluso en las plataformas de medios sociales. De este modo, el trabajo se centra en la efectividad de las interacciones de servicio al cliente relacionadas con la logística de los minoristas electrónicos en Twitter. El objetivo principal es identificar estrategias eficaces e ineficaces para el servicio al cliente en los medios sociales.

Finalmente, en el ámbito de las **organizaciones sin ánimo de lucro**, Young (2017) examinó cómo y por qué estas organizaciones están utilizando los medios sociales. Por medio de un diseño de encuesta transversal, éstas respondieron preguntas que ilustran la adopción y utilización de los medios sociales en cinco dimensiones: las razones para usar los medios sociales, las prácticas en los medios sociales, la frecuencia, la satisfacción general y los planes futuros.

Los resultados indican que, en general, están satisfechas con los medios sociales, las usan principalmente para promover su organización y servicios, y a pesar de los recursos limitados, las organizaciones planean continuar utilizando los medios sociales en el futuro.

CAPÍTULO 3: OBJETIVOS Y ASPECTOS METODOLÓGICOS

3.1. OBJETIVOS

3.2. ASPECTOS METODOLÓGICOS

3.3. INSTRUMENTOS Y TÉCNICAS

3.1. OBJETIVOS

Como se ha comentado previamente, cada vez son más los ámbitos que reconocen la importancia hoy día del seguimiento de los medios sociales, y en particular las empresas, las cuales han visto en este recurso una oportunidad para mejorar las relaciones con sus clientes a través del conocimiento que se puede extraer de los comentarios y hábitos de uso que éstos realizan sobre sus marcas y/o productos/servicios. Por lo tanto, el seguimiento de los medios sociales se ha convertido en una potente herramienta que apoya a las organizaciones en su toma de decisiones de marketing (Parenteau et al., 2016).

La competitividad de los mercados actuales demanda a las organizaciones la necesidad de crear relaciones más estables con sus clientes, y la fidelización de éstos se ha convertido en un aspecto crucial para su supervivencia. En este sentido los medios sociales son el punto de contacto más directo y menos costoso para que se establezcan estas relaciones a través de la escucha activa y el monitoreo de los medios sociales (Ciribeli & Paiva, 2011; Lahuerta-Otero & Cordero-Gutiérrez, 2016; Lim et al., 2013). Según lo expuesto, existen herramientas y servicios webs que permiten realizar este seguimiento, obteniendo información y conocimiento a partir de los comentarios y hábitos de los usuarios en estos medios.

Sin embargo, las herramientas disponibles hoy día o bien son principalmente de pago (las más potentes), o bien de uso gratuito para un determinado número de ejecuciones (usos o llamadas). La mayoría de ellas

están focalizadas exclusivamente para el análisis de textos en el idioma inglés. Algunas son herramientas para un único propósito (como API´s para análisis de sentimientos en inglés, software exclusivamente de minería de datos, o herramientas exclusivamente de monitoreo y seguimiento de medios sociales). Hasta nuestro conocimiento, no existe una herramienta como la que pretendemos desarrollar en el seno de esta tesis doctoral. Una herramienta que permita tanto el seguimiento de medios sociales, el análisis de datos desde el punto de vista del descubrimiento de conocimiento, la generación de informes, así como la integración del análisis de sentimientos en múltiples idiomas. Una herramienta que además sea gratuita, de libre distribución y de código abierto.

Con todo esto en mente, el **objetivo general** de la presente tesis doctoral consiste en el desarrollo de una herramienta web de Inteligencia Empresarial multilingüe que permita extraer, filtrar y clasificar información procedente de medios sociales, que sea fácil de usar, de código abierto, de distribución libre, y que además ofrezca y genere conocimiento para la toma de decisiones de marketing, y que con todo ello facilite a las PyMEs este tipo de recursos, y muy especialmente a las PyMEs de Brasil, más limitadas en el acceso a este tipo de recursos. En este sentido esta tesis ha recibido apoyo económico por medio de la beca brasileña *"Douorado Pleno no Exterior"* a través de la *"Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior*" (CAPES), bajo la autoridad del Ministerio de Educación, que pretende además de fomentar el intercambio de conocimiento entre distintos países, dar soporte técnico y científico a las empresas en Brasil.

Este objetivo general se desglosa a su vez en los siguientes **objetivos específicos**:

A. Conocer las herramientas de Inteligencia Empresarial que hay en el mercado tanto de manera gratuita como comercial.
B. Estudiar y revisar las principales técnicas de Minería de Datos, Minería de Opiniones y Análisis de Sentimientos disponibles en la literatura, así como los servicios disponibles en línea.
C. Diseñar y construir la herramienta de Inteligencia Empresarial.
D. Validar la herramienta con datos reales.

3.2. ASPECTOS METODOLÓGICOS

Para alcanzar estos objetivos, se ha seguido una metodología que se divide en cuatro fases principales. La primera fase (Capítulos 1, 2, y 4) tiene carácter documental y de revisión de la literatura. Se revisa exhaustivamente el estado del arte bajo el paraguas de la Minería de Datos, la Minería de Opiniones y el Análisis de Sentimientos a través de algoritmos automáticos. Fue realizada la revisión de libros, publicaciones académicas, informes de organismos públicos, empresas del sector y expertos.

La segunda fase (Capítulo 5), presenta un carácter más práctico y se centra en los diferentes métodos, enfoques y algoritmos más frecuentes existentes en la literatura en las tareas relacionadas al Análisis de Sentimientos automático. Utilizando recursos procedentes de investigaciones académicas de referencia identificadas en la fase anterior,

junto a herramientas de Minería de Datos, comparamos el rendimiento de estos algoritmos en base a la precisión y el tiempo de análisis frente a diferentes grupos de datos en varios idiomas. Se realizan test estadísticos no paramétricos para la confirmación de hipótesis.

La tercera fase (Capítulo 6) se fundamenta en los conocimientos y resultados de las fases precedentes para crear un sistema web de inteligencia empresarial y de apoyo a decisiones que contemple aspectos como análisis de requisitos, diseño, implementación, validación y prueba. Será programado utilizando el soporte de los lenguajes de programación `R` y `Python` para el *back-end* (o etapa de análisis y trabajo de datos), y se usará `Shiny` para el *front-end* (o capa de presentación e interacción) que tendrá la forma de sistema web. Para esta fase se seguirá el típico ciclo de vida de un desarrollo software, consistente fundamentalmente en la iteración cíclica de análisis de requerimientos, diseño de prototipo, implementación, test y validación; hasta alcanzar el objetivo deseado.

La cuarta y última fase (Capítulo 7), deriva de las aplicaciones prácticas y conclusiones realizadas utilizando la herramienta creada (`LOGOS`). Serán presentados y descritos los resultados arrojados por la herramienta, en base a los análisis de tuits provenientes de dos empresas multinacionales con presencia significativa en España y Brasil. Estos tuits fueron recogidos por un período de doce meses y comprenden los idiomas español y portugués. Por último, en esta fase también presentaremos las conclusiones y discusiones de la investigación. Junto a ello, discutiremos las

implicaciones de estos hallazgos y de la utilización de la herramienta de apoyo a decisiones tanto para el marketing como para las empresas.

3.2.1. PLANIFICACIÓN TEMPORAL – DIAGRAMA DE GANTT

A continuación en la Tabla 2, se lista el conjunto de actividades llevadas a cabo en esta tesis doctoral y su planificación temporal.

Para ello, se ha utilizado el diagrama de Gantt que permite identificar el tiempo dedicado a las diferentes tareas a lo largo de los años lectivos del 2014/2015 al 2018/2019.

Tabla 2. Planificación temporal de la tesis – Diagrama de Gantt.

Actividades	2014/2015		2015/2016		2016/2017		2017/2018		2018/2019	
	1º	2º	1º	2º	1º	2º	1º	2º	1º	2º
Perfeccionar y definir las bases conceptuales y teóricas del proyecto										
Tutoría con los directores – presencial y virtual										
Realización de estancia de investigación										
Ejecución de la primera fase										
Ejecución de la segunda fase										
Ejecución de la tercera fase										
Ejecución de la cuarta fase										
Redacción de los capítulos de la tesis										
Participación en congresos y jornadas científicas, presentación de posters y publicaciones de artículos										
Depósito de la tesis										
Defensa de la tesis										

Fuente: Elaboración propria con base en el diagrama de Gantt.

3.3. INSTRUMENTOS Y TÉCNICAS

En este apartado listamos los principales instrumentos y técnicas utilizadas para llevar a cabo este trabajo:

A. Técnicas de tratamiento de datos procedentes del Procesamiento del Lenguaje Natural (PLN) ampliamente utilizadas en la literatura para tratamiento de los datos de tipo textual;
B. Herramienta de Minería de Datos `Rapidminer` Studio (varias versiones desde la 7.1), para realizar el Análisis de Sentimientos, entrenamiento y test de los algoritmos analizados;
C. Lenguajes de programación `R` y `Python` para el trabajo de manejo y análisis de datos.
D. Bases de datos como `SQLite`, `MongoDB` y `Neo4j` para el almacenamiento persistente de datos y apoyo para ciertas tareas de análisis de datos.
E. Herramientas de programación web como `Shiny` para construir la interfaz de usuario.
F. Diccionarios semánticos, léxicos y *datasets* de sentimientos divididos entre grupos de entrenamiento y prueba para realización de pruebas que determinen el rendimiento de los algoritmos de clasificación.
G. Programación y utilización de un servidor (hipatia.ugr.es) para las tareas de descarga y almacenaje de datos, y distribución de la herramienta;

H. Herramientas de construcción de páginas web para la producción del repositorio del conocimiento en línea de la tesis (http://hipatia.ugr.es/steiner).

Concretamente, por tratarse de técnicas e instrumentos diversos y concernidos a distintas áreas del conocimiento, estos recursos serán detallados de manera exhaustiva en los apartados metodológicos de los estudios que integran a esta tesis doctoral.

CAPÍTULO 4: APORTACIÓN 1 – COMPILACIÓN DE *DATASETS* DE SENTIMIENTOS MULTILINGÜES

4.1. INTRODUCCIÓN

La utilización de los medios sociales (Twitter, Facebook, Instagram, entre otros) como canal directo entre organizaciones e individuos que comparten recomendaciones, revisiones de productos y servicios, aumenta exponencialmente día a día. Su modelo de relaciones proporciona a los usuarios, además, el poder de crear y compartir ideas al instante y sin obstáculos. De modo que las empresas, encuentran en estos medios una mina de informaciones valiosas e indispensables para la mejora constante de lo que ofrecen. De esta manera, seguir y monitorear a esos medios permite saber el qué y cómo hablan sus clientes no solo de la marca o empresa, sino también de la competencia (Azorín-Richarte, Orduna-Malea, & Ontalba-Ruipérez, 2016; López, García, & Fernández, 2018; Morinaga, Yamanishi, Tateishi, & Fukushima, 2002; Bo Pang & Lee, 2008).

Ante tal escenario, y tal volumen de datos, ganan protagonismo las investigaciones y el desarrollo de las técnicas automáticas de Análisis de Sentimientos (AS) centradas en clasificar los mensajes extraídos de Twitter y de otros medios (Gaspar, Pedro, Panagiotopoulos, & Seibt, 2016; Go, Huang, & Bhayani, 2009; S. Mukherjee & Bhattacharyya, 2012; Spencer & Uchyigit, 2012). Para ello los grupos de datos llamados corpus o *datasets* son recursos fundamentales y determinantes de la calidad del AS. Estos se pueden usar para diferentes tareas: Clasificación de Sentimientos (Ding, Liu, & Yu, 2008; Sarvabhotla, Pingali, & Varma, 2011a; Wilson et al., 2009), Análisis de Subjetividad (Pang & Lee, 2004; Wilson et al., 2009), Extracción de Opiniones (Sarvabhotla et al., 2011a),

identificación de puntos de vista (Greene & Resnik, 2009), identificación de patrones (Park, Lee, & Song, 2011), entre otras.

Los corpus se pueden dividir básicamente en dos grupos: entrenamiento y prueba. El grupo de entrenamiento alberga conjuntos de mensajes evaluados de forma subjetiva y anotados manualmente indicando el sentimiento o emoción contenido en cada mensaje. Este grupo se suele utilizar como base para entrenar el clasificador de sentimientos automáticos (Jurafsky & Martin, 2009; Nakov et al., 2013; Shamma, Kennedy, & Churchill, 2009). De este modo, cuanto mayor sea la calidad del dataset de entrenamiento, mejor será la precisión en la clasificación. Una vez entrenado el clasificador automático, se utiliza el grupo de prueba para testificar la eficiencia de este. Usualmente el grupo de prueba se compone de mensajes sin clasificar y versa sobre los mismos temas tratados en el grupo de entrenamiento. Cabe resaltar que para que el Aprendizaje Automático funcione correctamente, es necesario que los algoritmos subyacentes se entrenen y aprendan a partir de los mejores recursos posibles. Es muy necesario disponer de conjuntos de entrenamiento adecuados y totalmente pertinentes al problema de aprendizaje en cuestión.

Esta tesis se centra en el AS de textos con características muy específicas (textos cortos) de dudosa sintaxis y gramática, con símbolos extraños (emoticonos fundamentalmente), en varios idiomas y de léxico ambiguo. Para que los resultados de nuestros estudios sean buenos, necesitamos disponer de conjuntos de entrenamientos con similares características a las mencionadas. Contra más conjuntos de datos de este

tipo podamos usar y más diversos sean, mejor será para nuestros fines, mejores resultados se podrán obtener y más robustos serán nuestros algoritmos ya entrenados. En la literatura existe de manera pública, algunos de estos conjuntos de datos. Colecciones cuyos textos ya han sido previamente etiquetados en base a su polaridad; se han seguido diferentes metodologías para este etiquetado y los textos proceden de diversas fuentes y de diferentes idiomas. Estos corpus se encuentran diseminados a lo largo de la literatura (en publicaciones científicas, en blogs y repositorios especializados, etc.)

Los *datasets* de entrenamiento más frecuentes, son clasificados mediante polaridad positiva, negativa y neutra o también representadas por rangos, utilizando por ejemplo el número +5 para indicar muy positivo y -5 indicando muy negativo. Sin embargo, también existen otros corpus menos frecuentes los cuales llevan etiquetas adicionales relacionada a emociones como alegría, rabia, disgusto, irrelevancia y otros (Go, Bhayani, & Huang, 2009; Román, Morera, Cámara, & Zafra, 2015; Saif, He, & Alani, 2012; Y. Yu & Wang, 2015).

Otro punto que resaltar relacionado con los *datasets* de entrenamiento encontrados en la literatura es precisamente la escasez de éstos en otros idiomas distintos al inglés. Para atender esta demanda, este estudio pretende identificar, reunir y describir la información relacionada a 25 corpus de mensajes cortos en distintos idiomas, anotados manualmente disponibles hoy en día. Entre los idiomas se dispone de 10 en inglés, 4 en español (siendo uno de ellos en español de México), 4 en el idioma portugués (siendo 3 en portugués de Brasil), 2 en

alemán, 1 en árabe, 3 en italiano y 1 en francés. Al final este compendio de *datasets* servirá de recurso científico facilitador para futuras investigaciones relacionadas con el Análisis de Sentimientos automáticos.

4.2. OBJETIVOS

Los objetivos principales que persigue este capítulo son fundamentalmente dos, 1) recopilar y usar todos los *datasets* posibles para nuestro beneficio (conseguir el mejor entrenamiento posible de nuestros algoritmos) y 2) facilitar a otros esta misma tarea. Así mismo, este estudio busca elaborar un compendio de recursos de base y de alta calidad para investigaciones relacionadas con el análisis de sentimientos automáticos en múltiples idiomas. En este sentido, objetiva de manera principal identificar, recompilar y describir la información relacionada a 25 *datasets* de mensajes cortos en 7 idiomas distintos, anotados manualmente, validados científicamente y disponibles hoy en día.

Para poder atender al objetivo principal, fueron designados los siguientes objetivos específicos.

A. Revisar la literatura relacionada con los *datasets* de mensajes cortos disponibles hoy en día.
B. Identificar, catalogar y describir los *datasets* de mensajes cortos, anotados manualmente y validados científicamente en los idiomas inglés, español, portugués, alemán, árabe, italiano y francés.

C. Elaborar un repositorio web con los avances y recursos encontrados en el estudio a fin de darle difusión conjunta a todos los *datasets* identificados y catalogados.
D. Difusión en forma de artículo científico de esta misma revisión.

En los epígrafes que siguen se detallan los múltiples recursos compilados en este trabajo, un compendio de estos en forma de tabla, así como su ubicación en la Web.

4.3. *DATASETS* DE SENTIMIENTOS

Como ha quedado patente en el apartado introductorio de este capítulo, (punto 4.1.), los corpus o *datasets* – recurso fundamental en el AS - consisten en un conjunto de entradas anotadas de manera subjetiva indicando la clasificación correspondiente a cada una de ellas.

Esta anotación tiene la función de enseñar al clasificador de sentimientos cuáles palabras o secuencias de palabras están asociadas a una determinada clase. (Jurafsky & Martin, 2009; Nakov et al., 2013; Shamma et al., 2009). De este modo, cuanto mayor la calidad del dataset, mejor será la precisión en la clasificación de los mensajes.

A continuación, se describe a cada uno de los 25 *datasets* que componen el objeto principal de este estudio.

4.3.1. *DATASETS* EN INGLÉS

En esta sección se describen las especificaciones de cada dataset de acuerdo con el idioma inglés, los autores creadores de estos recursos y algunos de los artículos científicos que han hecho referencia y uso de ellos.

4.3.1.1. 2000ENTITIES

El corpus **2000Entities** fue publicado por Mukherjee et al. (2012) y ampliamente utilizado en (Mukherjee & Bhattacharyya, 2012; Subhabrata Mukherjee et al., 2012; Sabou, Aroyo, Bontcheva, Bozzon, & Qarout, 2018). El dataset, se compone de 8507 tuits referentes a aproximadamente 2000 personalidades famosas pertenecientes a más 20 diferentes ramas de actuación como, películas, restaurantes, televisión, política, deportes, educación, filosofía, viajes, libros, tecnología, bancas, música, medio ambiente, informática, sector automovilístico, etc. Los mensajes que le componen fueron clasificados manualmente por 4 evaluadores en las categorías *positive, negative, objective-not-spam and objective-spam.* Para los análisis realizados en esta tesis, utilizamos los tuits clasificados como positivos y negativos resultando en un corpus final con 3750 tuits.

4.3.1.2. HEALTHCARE REFORM (HCR)

El corpus **HCR** , fue construido en marzo de 2010 y está compuesto por 2516 tuits de entrenamiento que contienen el hashtag “#hcr” que hace referencia al programa de salud pública *healthcare reform*, introducido

en 2010 por Barack Obama en los Estados Unidos de América (Speriosu, Sudan, Upadhyay, & Baldridge, 2011). Los mensajes fueron clasificados manualmente por 5 científicos en 5 categorías (positiva, negativa, neutra, irrelevante u otro). Luego, el corpus fue dividido en tres grupos, entrenamiento (839), desarrollo (838) y prueba (839). Este recurso fue utilizado por (Coletta, Silva, Hruschka, & Hruschka, 2014) para entrenar un clasificador con SVM combinado a las técnica de clúster C3E-SL para potenciar la clasificación de datos. Por otro lado, Saif et al. (2012) presentó un nuevo enfoque que agrega la semántica como características adicionales en el conjunto de entrenamiento y mide la correlación del concepto representativo con el sentimiento negativo / positivo.

Dos años más tarde, Saif et al. (Saif, He, Fernandez, & Alani, 2014a, 2014b) también explotan la semántica como característica adicional para potenciar la clasificación. Speriosu et al. (2011) usaron una etiqueta de propagación en un clasificador de entropía máxima entrenado en *noisy labels* y conocimientos relacionados a tipos de palabras codificadas de un léxico. Tsaka-lidis et al. (2014) propuso un clasificador entrenado en un dominio general y que es capaz de adaptarse al dominio de prueba, antes de clasificar un documento y más recientemente otros trabajos también se utilizaron de este recurso para estudios relacionados con el AS (Alsaeedi, 2019; Ankit & Saleena, 2018; Kumar & Jaiswal, 2019).Para los análisis realizados en esta tesis hemos utilizado los tuits positivos y negativos del corpus, al final, el recurso se compone de 1360 mensajes.

4.3.1.3. MOVIES - UMICH SI650

Este dataset fue creado por la Universidad de Michigan entre abril y marzo de 2011 para tareas relacionadas al AS. El corpus posee 7086 tuits de entrenamiento etiquetados manualmente como positivo y negativos. También se compone de un grupo de prueba con 33052 mensajes. Ambos grupos de datos se relacionan con películas de diferentes géneros. Entre los estudios que utilizaron este conjunto de datos, se destacan por un lado el de Dickinson et al. (2015) que aplicaron recursos como Word2Vec y Sent2Vec por medio de un modelo semántico estructurado profundo (DSSM) con estructura de convolución/agrupación (CDSSM) para formar sus representaciones vectoriales y modelos de bolsas de palabras. Por otro lado, Duncan y Zhang (Duncan & Zhang, 2015) usaron este recurso con una red neuronal para refinar las tareas de Análisis de Sentimientos. Este recurso también fue utilizado en trabajo más recientes relacionados con la emoción y la inteligencia artificial (Bari & Saatcioglu, 2018; Lochter, Pires, Bossolani, Yamakami, & Almeida, 2018). Para los análisis realizados en esta tesis utilizamos una adaptación de este corpus con 6970 mensajes de entrenamiento.

4.3.1.4. OBAMA-MCCAIN DEBATE (OMD)

El dataset **OMD** posee 3238 tuits extraídos de la red Twitter durante el primero debate presidencial entre Obama y McCain en septiembre del 2008. El recurso fue clasificado de forma manual inicialmente por 2 investigadores, luego un tercer investigador fue utilizado en el caso de

discordancias resultantes de la primera clasificación. Como resultado se obtuvieron 1196 tuits negativos, 710 positivos y 245 con valoraciones mixtas o indefinidas (Mohammad, Kiritchenko, & Zhu, 2013; Shamma, Kennedy, & Churchill, 2009). El dataset fue utilizado para evaluar diferentes métodos de aprendizaje supervisado y no supervisado (Silva, Hruschka, & Hruschka, 2014; Hangya & Farkas, 2016; Hu, Tang, Tang, & Liu, 2013; Katarya & Yadav, 2018; Kenyon-Dean et al., 2018; Saif, Fernandez, He, & Alani, 2013; Saif et al., 2012, 2014a, 2014b; Speriosu et al., 2011; Tsakalidis et al., 2014; Yang et al., 2018; Y. Zhang, Song, Zhang, Li, & Wang, 2019). En los análisis realizados en este trabajo utilizamos una adaptación del corpus con 2203 mensajes.

4.3.1.5. STANFORD DATASET

Este corpus creado y ofrecido por la Universidad de Stanford contiene 16000,000 tuits de los cuales 800000 son considerados negativos puesto que contienen el emoticono :(y 800000 considerados positivos puesto que contienen el emoticono :). Una segunda clasificación fue llevada a cabo manualmente generando en un subgrupo de 177 tuits negativos y 182 tuits positivos. El corpus publicado en (Go, Bhayani, et al., 2009), fue ampliamente utilizado en la literatura en su forma completa, aunque muchos trabajos optaron por utilizar su variación reducida anotada subjetivamente. Esta versión reducida, que será utilizada en este trabajo, fue aplicada ampliamente en tareas de aprendizaje de maquina supervisadas y no supervisadas por medio de técnicas como SVM, Naïve Bayes y N-gramas (Bravo-Marquez, Mendoza, & Poblete, 2013; Hu et al.,

2013; Naiar, Bourouis, Bouguila, & Belghith, 2018; Pilehvar, Kartsaklis, Prokhorov, & Collier, n.d.; Saif et al., 2013, 2012, 2014b; Speriosu et al., 2011; Tsakalidis et al., 2014).

4.3.1.6. SEMEVAL 2015 – TASK11

Este dataset fue creado para la *Task 11* del evento internacional llamado SemEval (*Semantic Evaluation*) 2015, *Sentiment Analysis of Figurative Language in Twitter Workshop*[7]. El acto propone una ronda de evaluaciones o tareas destinadas a explorar el significado del lenguaje a través de los sistemas computacionales de análisis semántico. Estas tareas pretenden proporcionar mecanismos emergentes a fin de identificar los problemas y soluciones para cálculos significativos en esta área del conocimiento. Este evento además, busca articular las dimensiones relacionadas con el uso del procesamiento del lenguaje natural (PNL) direccionado al Análisis de Sentimientos (Rosenthal et al., 2015).

El corpus consiste en un conjunto de tuits de lenguaje creativo ricos en metáfora e ironía. Este recurso es considerado actualmente el único disponible que proporciona una alta variedad de tuits de lenguaje figurado. Concretamente se divide en 2 grupos, entrenamiento y test. El grupo de entrenamiento posee 8000 mensajes anotados con polaridad

[7] http://alt.qcri.org/semeval2015/task11/ (acceso el 07/08/2019).

entre los rangos de -5 a +5, donde -5 significa muy negativo y +5 muy positivo. El grupo de prueba más pequeño y está compuesto por 1000 tuits.

El conjunto **SemEval 2015 - *Task 11*** fue utilizado como recurso fundamental en trabajos como Baca-Gomez et al. (2016) que lo utilizaron con un enfoque hibrido de Minería de Opiniones, y Ghosh et al. (2015) para tareas de lenguaje figurativo utilizando semejanzas de coseno como una medida de error. En los análisis realizados en esta tesis utilizamos, una adaptación del corpus con 5816 mensajes. Los tuits de puntuación -5 a -1 fueron considerados como negativos y los de puntuación 1 a 5 considerados como positivos.

4.3.1.7. ANNOTATED-US2012-ELECTION

Este corpus contiene tuits extraídos de la red social Twitter entre los meses de agosto y septiembre del 2012 en base a 21 hashtags referentes a las elecciones presidenciales del Estados Unidos del mismo año. Además de los hashtags también se extrajeron los tuits que contenían las palabras Obama, Barack and Romney. Después de la limpieza de datos que eliminó mensajes en otros idiomas y retuits, se consiguió un total de 170,000 tuits originales en el idioma inglés.

La clasificación de los mensajes de este dataset creado por (Mohammad, Zhu, Kiritchenko, & Martin, 2015) se dio por medio de

crowdsourse a través del Amazon's Mechanical Turk and CrowdFlower[8]. Se utilizaron 2 cuestionarios llamados *HITS (human intelligence tasks)* para evaluar 2042 tuits – cada mensaje fue escrito por una cuenta diferente - seleccionados de forma aleatoria y evaluados por cerca de 400 hablantes nativos de lengua inglesa.

El primer cuestionario se utilizó para determinar la presencia de emociones, estilo y el propósito del tuit. El objetivo era determinar sentimientos de oposición (hipocresía, equivocaciones, desacuerdo, ridiculizaciones, criticas), a favor (concordancia, apoyo) y otros.El segundo cuestionario es un subgrupo del primero. Se escogieron los tuits que en el primer cuestionario habían sido clasificados como emocionales o que tenían contenido emocional, en total 1889 mensajes. Luego en la clasificación de estos mensajes, se buscó identificar 8 emociones básicas: *trust, fear, surprise, sadness, disgust, anger, anticipation and joy.*Estos corpus fueron ampliamente utilizados y mencionados en trabajos recientes como (Cotelo, Cruz, Enríquez, & Troyano, 2016; Fast, Chen, & Bernstein, 2016; Mohammad, Sobhani, & Kiritchenko, 2016). En los análisis realizados en esta tesis utilizamos, utilizaremos una adaptación de este dataset con 1767 mensajes donde consideramos los mensajes clasificados como *fear, sadness, disgust y anger* como negativos y *joy* como positivos.

[8] https://crowdflower.com (acceso el 07/08/2019).

4.3.1.8. DAI-LABOR ENGLISH DATASET

Este recurso contiene 7200 tuits clasificados como negativos y positivos. Los mensajes fueron extraídos de la red en base a los emoticonos :) :-) =) ;) :] :D ^-^ ^_^ indicando polaridad positiva y :(:-(:((-.- >:-(D: :/ indicando polaridad negativa, de modo que no tratan de ningún dominio o tema específico. Cada mensaje fue evaluado por 3 anotadores humanos en tres categorías (positiva, negativa y neutra), por medio de la herramienta Amazon Mechanical Turk. La concordancia entre las evaluaciones se dio utilizando el coeficiente de Fleiss' kappa (0.430) (Fleiss, 1971). El dataset creado por Dai-Labor[9] con el apoyo de la Technical University Berlin, fue construido y presentado en el *Workshop on Knowledge Discovery, Data Mining and Machine Learning* (KDML-2012) (Narr, Hu lfenhaus, & Albayrak, 2012) y utilizado para tareas de Análisis de Sentimientos supervisados utilizando el algoritmo Naïve Bayes.

4.3.1.9. RATEITALL AND EPINIONS DATASET

Este conjunto de datos aplicado en Jakob y Gurevych (2010) y Wiegand y Klakow (2012), está anotado considerando los niveles de oración y de expresión y distingue entre polaridad previa y polaridad contextual de una expresión de sentimiento. Creado por Toprak, Jakob, & Gurevych (2010), el corpus es compuesto de reseñas de clientes (extraídas de

[9] http://www.dai-labor.de/ (acceso el 07/08/2019).

Rateitall.com y Epinions.com) en dos dominios diferentes: universidades en línea y servicios en línea. Abarca a 240 revisiones de universidades (2786 oraciones) y 234 revisiones de servicio (6091 oraciones) considerando la expresión de opinión en el nivel de oración desde diferentes aspectos, como la polaridad, la fuerza, el modificador, el titular y objetivo.

4.3.1.10. REDBULL TWITTER SENTIMENT DATASET (RSD)

Este dataset es una aportación científica del propio autor de esta tesis doctoral y está publicada en el Libro Conmemorativo del X Aniversario del Máster en Marketing y Comportamiento del Consumidor (Steiner-Correa, 2017). Este recurso fue creado a partir de los tuits identificados con la etiqueta #givesyouwings relacionada con la principal campaña publicitaria de la marca de bebidas energéticas RedBull.

El corpus se compone de dos grupos, uno de entrenamiento compuesto por 100 mensajes únicos clasificados subjetivamente como positivos, neutros y negativos; y un grupo de prueba con 423 tuits. Todos los tuis fueron extraídos de la red social Twitter. Las evaluaciones de los mensajes fueron realizadas de manera online a través de la aplicación web Surveymonkey[10] por 152 estudiantes de la Universidad de Granda y conocedores de la marca, siendo un 36% hombres y un 64% mujeres de

[10] https://es.surveymonkey.com/ (acceso el 07/08/2019).

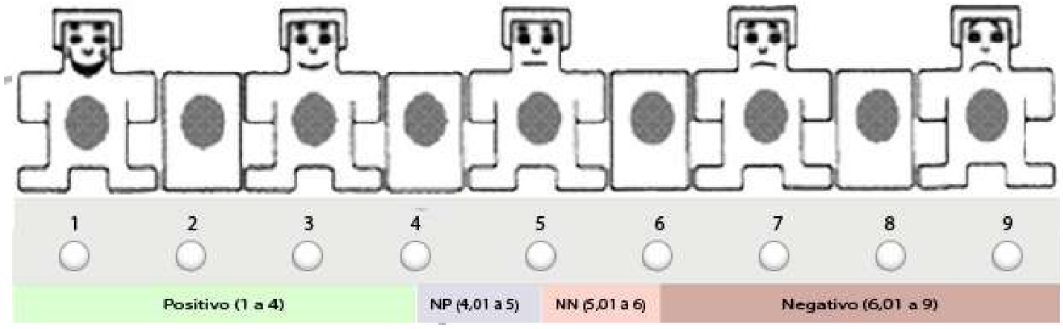
edades comprendidas entre 18 y 44 años y en su mayoría, con educación superior.

Para ello se utilizó la escala pictográfica de valencia del *Self Assessment Manikin*, SAM (Hodes, Cook, & Lang, 1985), que mide el grado de valencia a través de representaciones gráficas de figuras humanoides que van desde un rostro sonriente hasta un rostro serio.

A través de las figuras, la escala da un valor de 1 a 9, donde 1 significa valencia o polaridad positiva y 9 valencia o polaridad negativa (Figura 17). Entre estos dos extremos la escala también te permite dar puntuaciones intermedias en las figuras centrales o entre las figuras.

Finalmente, para acotar la polaridad de los tweets, fueron establecidos cuatro rangos de valoración: Positivo [1-4], Neutro-Positivo (4-5], Neutro-Negativo (5-6] y Negativo (6-9].

Figura 17. Escala de evaluación SAM.



Fuente: Adaptations a partir de (Hodes et al., 1985).

4.3.2. *DATASETS* EN PORTUGUÉS

En esta sección se describen las especificaciones de cada dataset de acuerdo con el idioma portugués, los autores creadores de estos recursos y algunos de los artículos científicos que han hecho referencia a ellos.

4.3.2.1. NOTÍCIAS-GLOBO (BRASIL)

Este dataset contiene 661 mensajes cortos sin un dominio específico. Las consideraciones fueron extraídas de la página web www.globo.com y abarcan tanto el contexto nacional de Brasil como al internacional. Los mensajes fueron evaluados por 2 anotadores con experiencia lingüística en las categorías alegría, disgusto, miedo, rabia y tristeza. El recurso creado y utilizado en (Dosciatti, Ferreira, & Paraiso, 2013) es resultado del proyecto llamado *Emoções.BR* que estudia el Análisis de Sentimientos en textos escritos en portugués de Brasil. Para los análisis realizados en esta tesis, fue necesario adaptar el corpus convirtiendo los tuits clasificados como “alegría” en positivos y los clasificados como “disgusto, miedo, rabia y tristeza” como negativos.

4.3.2.2. POLÍTICA (BRASIL)

El corpus **Política** creado y utilizado en (Nascimento et al., 2015), contiene 567 tuits de entrenamiento extraídos entre agosto y septiembre del 2011 y se relacionan con la situación política en Brasil en aquella época. Los mensajes fueron clasificados por 3 diferentes

investigadores en las categorías positivo, negativo y neutro. Este recurso no está disponible públicamente, pero puede ser conseguido mediante solicitud directa a sus creadores. Para los análisis realizados en esta tesis, consideramos a los tuits clasificados como positivos y negativos y el dataset final posee 530 mensajes.

4.3.2.3. ENTRETENIMENTO (BRASIL)

El corpus ***Entretenimento*** contiene 384 tuits de entrenamiento extraídos entre agosto y septiembre del 2011 y se relacionan a la oferta de ocio y cultura de Brasil. Creado y utilizado en (Nascimento et al., 2012) sus mensajes fueron clasificados por 3 diferentes investigadores en las categorías positiva (P), negativa (N) y neutra (NEU). Como en el corpus *Política*, los autores tuvieron especial cuidado al evaluar los casos de ironías y abreviaciones encontrados en el texto. Este recurso no está disponible públicamente, pero puede ser conseguido mediante solicitud directa a sus creadores. Para los análisis realizados en esta tesis, utilizamos una adaptación del corpus con 360 mensajes.

4.3.2.4. DAI-LABOR PORTUGUESE DATASET

Este recurso contiene 1800 tuits clasificados como negativos y positivos. Éstos no tratan de ningún dominio o tema específico puesto que fueron extraídos de la red en base a los emoticonos :) :-) =) ;) :] :D ^-^ ^_^ indicando polaridad positiva y :(:-(:((-.- >:-(D: :/ indicando polaridad negativa. Con relación al idioma, el dataset abarca el idioma

portugués general sin cernirse a un país o región específica. Cada mensaje fue evaluado por 3 anotadores humanos por medio de la herramienta Amazon Mechanical Turk en tres categorías (positiva, negativa y neutra). La concordancia entre las evaluaciones se dio utilizando el coeficiente de Fleiss' kappa (0.408) (Fleiss, 1971). El dataset creado por Dai-Labor[11] con el apoyo de la Technical University Berlin, fue construido y presentado en el *Workshop on Knowledge Discovery, Data Mining and Machine Learning* (KDML-2012) (Narr, Hu lfenhaus, & Albayrak, 2012) y utilizado para tareas de Análisis de Sentimientos supervisados utilizando el algoritmo Naïve Bayes.

4.3.3. *DATASETS* EN ESPAÑOL

Para el idioma español, presentamos un compilado de 4 *datasets*. Los corpus (General-TASS, Social-TV-TASS y STOMPOL-TASS), fueron creados y presentados en el TASS 2014 (Román et al., 2015). Este evento se configura como un taller de evaluación experimental para el análisis de opiniones centrado en el idioma español. Además, se configura como un acto satélite de la conferencia anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). El cuarto corpus llamado SpanishCorpus3100, tiene como particularidad principal la variación del

[11] http://www.dai-labor.de/ (acceso el 07/08/2019).

idioma español. Este recurso se centra específicamente en el español de México.

A continuación, describimos a cada uno de ellos.

4.3.3.1. GENERAL-TASS

Este dataset contiene más de 68.000 tuits: 10% destinado al entrenamiento y 90% para prueba. Los mensajes fueron extraídos de la red social Twitter entre noviembre de 2011 y marzo de 2012. Estos tuits se refieren a personalidades y celebridades relacionadas con el mundo de la política, economía, comunicación y cultura. Según los creadores, "aunque el contexto de la extracción tiene un sesgo centrado en España, la nacionalidad diversa de los autores incluyendo a personas de España, México, Colombia, Puerto Rico, EE. UU y muchos otros países, hace que los corpus alcancen una cobertura global en el mundo de habla española" (Román et al., 2015)Los tuits de entrenamiento fueron manualmente clasificados en 6 categorías: *strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) y no sentiment (NONE)*. Además, el dataset fue utilizado como recurso en numerosos trabajos como (Perea-Ortega & Balahur, 2014; Vilares, Doval, Alonso, & Gómez-Rodríguez, 2014) en experimentos basados en reemplazos de rasgos, técnicas de aprendizaje profundo (*deep learning*) tales como preentrenamiento no supervisado, e incorporación de palabras específicas de sentimientos. Para los análisis realizados en esta tesis, consideramos los P+ y P como positivos y los N+ y N como negativos, generando un corpus formado por 5053 mensajes.

4.3.3.2. SOCIAL-TV-TASS

El corpus **Social-TV** fue colectado durante la final de la Copa del Rey en España en 2014. Fueron extraídos los tuits desde 15 minutos antes, hasta 15 minutos después del término del partido entre el equipo Real Madrid y el F.C. Barcelona el 16 de abril del 2014. El dataset fue ampliamente utilizado en (Hurtado & Pla, 2014; Roncal & Urizar, 2014; Vilares et al., 2014) y está compuesto de 1773 tuits de entrenamiento y 1000 de prueba que fueron clasificados manualmente en 3 categorías: positive (P), neutral (NEU) y negative (N). El recurso fue ampliamente utilizado en trabajos que análisis de n-gramas, técnicas de aprendizaje supervisado profundo y SVM. Para los análisis realizados en esta tesis, utilizamos una adaptación del corpus con 1343 mensajes.

4.3.3.3. CONJUNTO STOMPOL-TASS

El **STOMPOL** (*Spanish Tweets for Opinion Mining at aspect level about POLitics*) se compone de tuits extraídos entre 23 y el 24 de abril del 2014 y se relaciona con diversos ámbitos conectados a la política como economía, educación y otros. El dataset ampliamente utilizado en (Cumbreras, Cámara, Román, & Morera, 2016; Salvatore et al., 2015; Vilares, Doval, Alonso, & Gómez-Rodríguez, 2015), se compone de 784 tuits de entrenamiento y 500 de prueba que fueron ubicados en 3 categorías *positive* (P), *neutral* (NEU) y *negative* (N). Los mensajes fueron clasificados manualmente por 2 diferentes evaluadores y un tercero en caso de que las evaluaciones precedentes fueran antagónicas. El recurso fue utilizado en diferentes tareas con clúster sociolingüístico y técnicas

de *Deep Learning*. En este trabajo utilizamos una adaptación del corpus con 598 mensajes.

4.3.3.4. SPANISHCORPUS3100 - SPANISH MEXICAN DATASET

Este dataset contiene exclusivamente mensajes en español mexicano. El proceso de etiquetado fue realizado por seis personas y dividido en 2 tandas diferentes. En la primera tanda los mensajes fueron evaluados en tres categorías: positivo, neutro y negativo. En la segunda tanda los mismos mensajes fueron evaluados en cinco categorías: muy positivo, positivo, neutro, negativo y muy negativo. La concordancia entre las evaluaciones se dio utilizando el coeficiente de Krippendorff's alpha (Krippendorff, 2011). Como resultado se obtuvieron 2 *datasets* que fueron utilizados como recurso principal por sus creadores en (Baca-Gomez et al., 2016). Finalmente, cabe mencionar que estos corpus no están disponibles públicamente, pero pueden ser conseguidos mediante solicitud directa a sus creadores. Para los análisis realizados en esta tesis, utilizamos una adaptación del corpus con 2602 mensajes.

4.3.4. *DATASETS* EN ALEMÁN

En esta sección se describen las especificaciones de cada dataset de acuerdo con el idioma alemán, los autores creadores de estos recursos y algunos de los artículos científicos que han hecho referencia a ellos.

4.3.4.1. GERMAN SENTIMENT DATASET

Compuesto por 500 mensajes en alemán, este recurso creado por (Momtazi, 2012) es considerado el primer corpus de mensajes cortos para su idioma. Sus menciones rescatadas de diferentes medios sociales como Facebook y blogs versan sobre celebridades alemanas del mundo de la música. Debido a su carácter precursor, el dataset ha sido utilizado como referencia en trabajos recientes como (Scholz, Conrad, & Hillekamps, 2012; Shalunts, Backfried, & Prinz, 2014).

El corpus fue anotado por 3 hablantes nativos de alemán utilizando las medidas de 0 a -3 para evaluaciones negativas, y de 0 a +3 para evaluaciones positivas.

Cada mensaje recibió dos diferentes clasificaciones, polaridad y fuerza. La concordancia entre las evaluaciones se dio utilizando los 2 coeficientes: *Krippendorff's alpha* (Krippendorff, 2011) y *Fleiss' kappa* (Fleiss, 1971). Para los análisis realizados en esta tesis, utilizamos una adaptación del corpus con 358 mensajes.

4.3.4.2. DAI-LABOR GERMAN DATASET

Este corpus se compone de 1800 tuits clasificados como negativos y positivos. Los mensajes sin dominio específico fueron extraídos de la red en base a los emoticonos :) :-) =) ;) :] :D ^-^ ^_^ indicando polaridad positiva y :(:-(:((-.- >:-(D: :/ indicando polaridad negativa. La clasificación se dio por 3 anotadores humanos a través de la herramienta

Amazon Mechanical Turk y se dividió en tres categorías (positiva, negativa y neutra).

La concordancia entre las evaluaciones se dio utilizando el coeficiente de Fleiss' kappa (0.419) (Fleiss, 1971). El dataset creado por Dai-Labor con el apoyo de la Technical University Berlin, fue construido y presentado en el *Workshop on Knowledge Discovery, Data Mining and Machine Learning* (KDML-2012) (Narr, Hu lfenhaus, & Albayrak, 2012) y utilizado para tareas de Análisis de Sentimientos supervisados utilizando el algoritmo Naïve Bayes.

En los análisis realizados en esta tesis, utilizamos una adaptación del corpus con 1556 mensajes.

4.3.5. DATASET EN ÁRABE

4.3.5.1. MODERN STANDARD ARABIC (MSA) TWITTER DATASET

Este dataset abarca mensajes en el idioma árabe moderno *(Modern Standard Arabic (MSA)*) y en el Dialecto Jordano. Fue creado por (Abdulla, Ahmed, Shehab, & Al-Ayyoub, 2013) y ampliamente utilizado en la literatura reciente (Al-Kabi, Al-Ayyoub, Alsmadi, & Wahsheh, 2016; Al-Twairesh, Al-Khalifa, & Al-Salman, 2015; Araujo, Pereira, Reis, & Benevenuto, 2016; Obaidat, Mohawesh, Al-Ayyoub, AL-Smadi, & Jararweh, 2015). El recurso se compone de 2000 tuits (1000 positivos y

1000 negativos) recopilados por medio del *Tweet Crawler*[12]. Cada tuit fue clasificado por 2 expertos humanos y un tercero en caso de empate. El corpus trata estrictamente de 2 tópicos: política y arte.

4.3.6. DATASET EN ITALIANO

De los 3 corpus en italiano que comentaremos en este trabajo, 2 (**TWNews y TWSpino**) son resultado de un proyecto llamado Senti-TUT[13] que busca desarrollar recursos lingüísticos centrados en la ironía. Ambos corpus que describimos a continuación se centran en tuits sobre política y de expresiva carga irónica. Creados por (Bosco, Patti, & Bolioli, 2015), fueron ampliamente utilizados en la literatura (Basile, Bolioli, Nissim, Patti, & Rosso, 2014; Chafale & Pimpalkar, 2014; Ravi & Ravi, 2015) para realizar experimentos de AS relacionados con la ironía en el idioma italiano.

El tercer corpus fue creado para la tarea **SENTIPOLC** - *SENTIment POLarity Classification* propuesta en el EVALITA 2014[14]. Esta campaña trata de evaluar el procesamiento de lenguaje natural y las herramientas de voz para el italiano. Su objetivo general es promover el desarrollo de

[12] http://www3.nd.edu/~dwang5/courses/spring15/assignments/A1/Assignment1_SocialSensing.html (acceso el 07/08/2019).

[13] http://www.di.unito.it/~tutreeb/sentiTUT.html (acceso el 07/08/2019).

[14] http://www.evalita.it/2016/tasks/sentipolc (acceso el 07/08/2019).

tecnologías de la lengua y del habla italiana, donde los diferentes sistemas y métodos puedan ser evaluados de una manera consistente.

La tarea Sentipolc se divide en 3 subtareas, *subjectivity classification, polarity classification and irony detection*.

4.3.6.1. TWNEWS

Los tuits que componen este recurso fueron recolectados de la red entre el 16 de octubre del 2011 al 3 de febrero del 2012, período siguiente al que Mario Monti fue nominado para reemplazar a Silvio Berlusconi como primer ministro del estado (Bosco et al., 2015). Por medio de filtros como "mario monti/#monti", "gobierno monti/#monti", "profesor monti/#monti", se consiguió un total de 3.228 mensajes únicos.

El proceso de clasificación se llevó a cabo por 5 anotadores humanos (2 hombres y 3 mujeres) que clasificaron los tuits en 5 categorías: POS (positivo), NEG (negativo), HUM (irónico), MIXED (positivo y negativo a la vez) y NONE (ninguno de los anteriores).

Para los análisis realizados en esta tesis, utilizamos una adaptación del corpus con 1692 mensajes.

Este recurso actualmente no está disponible en la red, pero puede ser conseguido directamente con los autores.

4.3.6.2. TWSPINO

El corpus se construyó en base a los mensajes de la sección Twitter del blog Spinoza (http://www.spinoza.it). Este blog es popular en Italia por tratar los temas políticos empelando tono sátiro en sus comentarios. Se extrajeron los mensajes publicados entre julio de 2009 y febrero de 2012. Después de eliminar los tuits con publicidad (1,5%), se logró conseguir un corpus con 1.159 mensajes únicos (Bosco et al., 2015).

El proceso de clasificación se dio de la misma manera que el TWNews Dataset. Cinco anotadores (2 hombres y 3 mujeres), que clasificaron los mensajes en 5 categorías: POS (positivo), NEG (negativo), HUM (irónico), MIXED (positivo y negativo a la vez) y NONE (ninguno de los anteriores). Para los análisis realizados en esta tesis, utilizamos una adaptación del corpus con 823 mensajes. Este recurso actualmente no está disponible en la red, pero puede ser conseguido directamente con los autores.

4.3.6.3. SENTIPOLC TASK - EVALITA 2014

El corpus con 7410 tuits incluye tanto tuits sobre política (topic = 1) como genéricos (topic = 0) elegidos al azar y se compone de los ítems: *“idTwitter” (Twitter status id), “sbj” (subjectivity), “opos” (positive overall polarity), “oneg” (negative overall polarity), “iro” (irony), “lpos” (positive literal polarity), “lneg” (negative literal polarity), “top” (topic), “text” (Twitter message).* Cada tuit ha sido clasificado por 2 anotadores expertos y por un tercero en caso de desacuerdo.

Este recurso creado y presentado en (V. Basile et al., 2014) fue ampliamente utilizado en la literatura (P. Basile, Basile, Nissim, & Novielli,

2015; Basile & Novielli, 2014; Castellucci, Croce, Cao, & Basili, 2014). Para los análisis realizados en esta tesis, utilizamos una adaptación del corpus con 3578 mensajes.

4.3.7. DATASET EN FRANCÉS

4.3.7.1. DAI-LABOR FRENCH DATASET

Este recurso contiene 1797 tuits clasificados como negativos y positivos. Éstos no tratan de ningún dominio o tema específico puesto que fueron extraídos de la red en base a los emoticonos :) :-) =) ;) :] :D ^-^ ^_^ indicando polaridad positiva y :(:-(:((-.- >:-(D: :/ indicando polaridad negativa. Con relación al idioma, el dataset abarca el idioma francés general sin cernirse a un país o región específica. Cada mensaje fue evaluado por 3 anotadores humanos por medio de la herramienta Amazon Mechanical Turk en tres categorías (positiva, negativa y neutra). La concordancia entre las evaluaciones se dio utilizando el coeficiente de Fleiss' kappa (0.244) (Fleiss, 1971). El dataset creado por Dai-Labor[15] con el apoyo de la Technical University Berlin, fue construido y presentado en el Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012) (Narr, Hu lfenhaus, & Albayrak, 2012) y utilizado para tareas de Análisis de Sentimientos supervisados utilizando el algoritmo Naïve Bayes.

[15] http://www.dai-labor.de/ (acceso el 07/08/2019).

4.3.8. TABLA DESCRIPTIVA – COMPILADO DE *DATASETS*

En la Tabla 3 presentamos una descripción resumida y las principales características los corpus mencionados anteriormente. La Tabla está dividida en 7 columnas que especifican el nombre de cada dataset, el idioma de los mensajes, el dominio o tema que abarca cada recurso, una breve descripción con el número de mensajes, número de evaluadores y filtros utilizados para extracción, las categorías en las cuales se realiza la clasificación, los grupos de clasificación (entrenamiento y test y finalmente las referencias a los creadores de cada recurso.

Los *datasets* conseguidos bajo autorización de los autores están disponibles en la página web de esta tesis en http://hipatia.ugr.es/steiner/index.php/sentiment-datasets-2/.

Tabla 3. Tabla descriptiva – Compilado de *datasets*.

Dataset	Idioma	Tema dominante	Descripción	Clase de clasificación	Tipo	Cortesía de
2000Entities	Inglés	Películas, restaurantes, televisión, política, deportes, etc.	8,507 tuits recolectados de un total de aproximadamente 2000 entidades distintas de 20 sectores diferentes	Positivo, negativo, objetivo-no-spam y objetivo-spam	Entrena miento	Mukherjee et al. 2012 Disponible directamente con sus creadores
Healthcare Reform (HCR)	Inglés	Healthcare reform	2516 tuits con el hashtag "#hcr" (healthcare reform)	Positivo, negativo, irrelevante u otros	Entrena miento y Prueba	Saif et al. 2012; Speriosu et al.2011 https://bitbucket.org/speriosu/updown/downloads [acceso en julio, 2019]
Movies - UMICH SI650	Inglés	Películas	40,138 tuits compilados por la Universidad de Michigan para tareas de Análisis de Sentimientos	Positivo y negativo	Entrena miento y Prueba	University of Michigan SI650 https://inclass.kaggle.com/c/si650winter11 [acceso en julio, 2019]
Obama-McCain Debate (OMD)	Inglés	Política	3,238 tuits sobre los debates presidenciales televisivos em Estados Unidos en septiembre del 2008	Positivo, negativo y mixto o desconocido	Entrena miento	Shammas et al. 2009 https://bitbucket.org/speriosu/updown/downloads [acceso en julio, 2019]
Stanford Twitter Dataset	Inglés	Genérico	1600000 tuits sobre emociones	Positivo y negativo	Entrena miento y Prueba	Go et al. 2009a http://help.sentiment140.com/for-students [acceso en julio, 2019]

SemEval_2015 Task 11	Inglés	Genérico	9000 tuits clasificados con puntuaciones de sentimientos entre el rango -5...+5	Positivo, neutro y negativo	Entrenamiento y Prueba	Rosenthal et al. 2015 – SemEval 2015 Task11 http://alt.qcri.org/semeval2015/task11/index.php?id=data-and-tools [acceso en julio, 2019]
Annotated-US2012-Election-Tuits	Inglés	Política	2,042 tuits clasificados por 400 hablantes nativos en inglés extraídos entre agosto y septiembre del 2012	Confianza, miedo, sorpresa, tristeza, disgusto, enojo, anticipación y alegría.	Entrenamiento	Mohammad et al. 2015 http://saifmohammad.com [acceso en julio, 2019]
DAI-Labor English dataset	Inglés	Genérico	7,200 tuits con base en emociones, anotado por 3 investigadores	Positivo, neutro y negativo	Entrenamiento	Narr et al. 2012 http://dainas.aot.tu-berlin.de/~andreas@dai/sentiment [acceso en julio, 2019]
Rateitall and Epinions Dataset	Inglés	Productos y servicios	240 revisiones de universidades y 234 revisiones de servicios	Positivo, neutro y negativo	Entrenamiento	Toprak et al., 2010 Disponible directamente con sus creadores
RedBull Twitter Sentiment Dataset (RSD)	Inglés	Bebida Energética	100 tuits clasificados por 152 estudiantes, extraídos en mayo del 2014 con base en el hashtag #givesyouwings de la empresa RedBull	Positivo, neutro y negativo	Entrenamiento y Prueba	Steiner et al. (2016, http://hipatia.ugr.es/steiner/index.php/redbull-sentiment-dataset-rsd/ [acceso en julio, 2019]
Noticias Globo	Portugués (Brasil)	Notíticas sobre Brasil y el mundo	661 mensajes cortos extraídos de www.globo.com	Alegría, disgusto, miedo, enojo y tristeza	Entrenamiento	Dosciatti et al. 2013 http://www.ppgia.pucpr.br/~paraiso/mineracaodeemocoes/ [acceso en julio, 2019]

Política	Portugués (Brasil)	Política	567 tuits clasificados por 3 investigadores distintos	Positivo, neutro y negativo	Entrena miento	Nascimento et al. 2015 Disponible directamente con sus creadores
Entertainment	Portugués (Brasil)	Ocio, Arte y Cultura	384 tuits clasificados por 3 investigadores distintos	Positivo, neutro y negativo	Entrena miento	Nascimento et al. 2015 Disponible directamente con sus creadores
DAI-Labor Portuguese Dataset	Portugués	Genérico	1,800 tuits con emoticonos, clasificados por 3 investigadores distintos	Positivo, neutro y negativo	Entrena miento	Narr et al. 2012 http://dainas.aot.tu-berlin.de/~andreas@dai/sentiment [acceso en julio, 2019]
General-TASS	Español	Política, economía, comunicación y cultura	68,000 tuits extraídos desde noviembre del 2011 hasta marzo del 2012	6 clases: (P+) (P) (NEU) (N) (N+) y (ninguna)	Entrena miento y Prueba	Román et al. 2015 - TASS 2014 Disponible directamente con sus creadores
Social-TV-TASS	Español	Deportes	2,773 tuits durante la final de la Copa del Rey entre el Real Madrid y el Barcelona F. C. en abril del 2014.	Positivo, neutro y negativo	Entrena miento y Prueba	Román et al. 2015 - TASS 2014 Disponible directamente con sus creadores
STOMPOL-TASS	Español	Política	1284 tuits extraídos entre el 23 y el 24 de abril del 2014	Positivo, neutro y negativo	Entrena miento y Prueba	Román et al. 2015 - TASS 2014 Disponible directamente con sus creadores
Spanish Corpus3100	Español Mexicano	Genérico	3100 tuits clasificados por 6 investigadores distintos	5 clases (P+) (P) (NEU) (N) (N+)	Entrena miento	Baca-Gomez et al. 2016 Disponible directamente con sus creadores

Modern Standard Arabic (MSA)	Arábico dialecto jordano	Política	2.000 tuits clasificados por 3 investigadores distintos	2 clases: positivo y negativo	Entrenamiento	Assiri et al. 2015 https://archive.ics.uci.edu/ml/datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis [acceso en julio, 2019]
German Sentiment dataset	Alemán	Cantantes y músicos alemanes	500 mensajes cortos clasificados por 3 alemanes nativo hablantes	Puntuaciones entre -3...+3	Entrenamiento	Momtazi 2012 Disponible directamente con sus creadores
DAI-Labor German dataset	Alemán	Genérico	1,800 tuits con emoticonos, clasificados por 3 investigadores distintos	Positivo, neutro y negativo	Entrenamiento	Narr et al. 2012 http://dainas.aot.tu-berlin.de/~andreas@dai/sentiment [acceso en julio, 2019]
TWNews	Italiano	Política	3,228 tuits con tono irónico, extraídos entre el 16 de octubre del 2011 y el 3 de febrero del 2012, clasificados por 3 investigadores distintos	5 clases: POS, NEG, HUM, MIXED y NONE	Entrenamiento	Bosco et al. 2015 Disponible directamente con sus creadores
TWSpino	Italiano	Política	1,159 tuits con tono irónico, extraídos entre julio del 2009 y febrero del 2012, clasificado por 3 investigadores distintos	5 clases: POS, NEG, HUM, MIXED y NONE	Entrenamiento	Bosco et al. 2015 Disponible directamente con sus creadores
SENTIPOLC task - Evalita 2014	Italiano	Política y temas genéricos	7410 tuits con tono irónico clasificados por 3 investigadores distintos	Subjetividad, polaridad global positiva, polaridad global negativa, ironía, polaridad literal positiva, polaridad literal negativa	Entrenamiento	Basile et al. 2014 Disponible directamente con sus creadores

DAI-Labor French Dataset	Francés	Genérico	1,797 tuits con emoticonos clasificados por 3 investigadores distintos	Positivo, neutro y negativo	Entrenamiento	Narr et al. 2012 http://dainas.aot.tu-berlin.de/~andreas@dai/sentiment [acceso en julio, 2019]

Fuente: Steiner-Correa, Viedma-del-Jesus, y Lopez-Herrera (2018).

4.4. CONCLUSIÓN

En este trabajo hemos proporcionado una descripción general sobre 25 *datasets* de mensajes cortos clasificados de manera subjetiva y actualmente disponibles a la comunidad científica. Estos recursos fueron extraídos en su mayoría de la red social Twitter y anotados manualmente según su polaridad o emoción.

Algunos de los corpus además de un grupo de entrenamiento, cuentan con un grupo de mensajes para prueba. Los *datasets* reunidos se dividen en 7 idiomas: 10 en inglés, 4 en español (uno en variante lingüística del español de México), 4 en el idioma portugués (3 en la variante lingüística del portugués de Brasil), 2 en alemán, 1 en árabe, 3 en italiano y 1 en francés. Además, tratan de dominios variados como política, futbol, cine, salud, revisiones de productos, servicios y otros. Los mensajes de entrenamiento y prueba compilados en este trabajo suman un total de 1.778,081 tuits de los cuales 1.681,618 mensajes están en inglés, 75.157 están en español, 3.412 están en portugués, 2.000 en árabe, 2.300 en alemán, 11.797 en italiano y 1.797 en el idioma francés. Adicionalmente de la variabilidad de idiomas y dominios que tratan los *datasets*, hay que resaltar variabilidades técnicas como las relacionadas con el tipo de clasificación. Algunos son clasificados como positivos, neutros y negativos, otros utilizan rangos numéricos [1…5] y por fin algunos clasifican los mensajes con base en emociones como alegría, rabia, disgusto, tristeza, entre otros.

Este estudio fue publicado en forma de artículo en medio científico contrastado: revista internacional Soft Computing - *A fusion of Foundations,*

Methodologies and Applications con factor de impacto JCR (2.367)[16] en (Steiner-Correa, F., Viedma-del-Jesus, M. I., & Lopez-Herrera, A. G. (2018). A survey of multilingual human-tagged short message datasets for sentiment analysis tasks. Soft Computing, 22(24), 8227–8242. https://doi.org/10.1007/s00500-017-2766-5). Además, los datasets de publicación autorizada por sus creadores, están reunidos y accesibles en el repositorio de esta tesis doctoral en (http://hipatia.ugr.es/steiner/index.php/sentiment-datasets-2/).

Por último, el extenso volumen de mensajes y la variedad de *datasets* en múltiples idiomas ofrecidos por este compendio, así como la publicación del artículo en medio científico contrastado, hacen cumplir los objetivos de este trabajo centrados en recompilar recursos y ofrecer información vital para las posteriores investigaciones científicas relacionadas con el Análisis de Sentimientos por medio del aprendizaje automático.

[16] https://www.springer.com/engineering/computational+intelligence+and+complexity/journal/500 (acceso el 07/08/2019).

CAPÍTULO 5: APORTACIÓN 2 - COMPARATIVA DE ALGORITMOS AUTOMÁTICOS DE APRENDIZAJE

5.1. INTRODUCCIÓN

5.2. OBJETIVOS

5.3. METODO DE INVESTIGACIÓN

5.4. RESULTADOS

5.5. DISCUSIÓN

5.6. CONCLUSIONES

5.1. INTRODUCCIÓN

En este capítulo de la tesis, nos centramos en las técnicas de Análisis de Sentimientos (AS) relacionados al aprendizaje automático. Se comparan en un contexto multilingüe los cuatro algoritmos más frecuentemente utilizados en la literatura *(Naïve Bayes, Random Forest, Support Vector Machine y Decision Tree)* para la tarea de clasificación de sentimientos. Luego se contrastan sus resultados a nivel de precisión y tiempo para determinar cuál de ellos presenta mejores resultados de rendimiento. Finalmente se obtienen conclusiones sobre su aplicación al contexto multilingüe, con el objetivo de conocer qué algoritmo es que presenta mejor rendimiento según el idioma del texto a analizar. Este conocimiento se usará en la construcción del sistema LOGOS que se establece en la siguiente fase de esta tesis doctoral (Capítulo 6).

Para realizar esta tarea se probó cada algoritmo con los 21 *datasets*[17] de mensajes cortos en distintos idiomas (descritos en el Capítulo 4 de esta tesis): 8 en inglés, 4 en español (uno de ellos en español de México), 4 en el idioma portugués (tres de ellos en portugués de Brasil), 2 en alemán y 3 en italiano. Cada *dataset* varía según componentes como el tamaño, entidad o tema, entorno, extensión de los mensajes, tipo de clasificación y método de extracción entre otros. Concretamente, estos análisis determinarán cuál algoritmo ofrece mejor rendimiento para cada uno de estos cinco idiomas.

[17] Se descartaron 4 de los datasets recopilados, por no complicar las condiciones necesarias para la experimentación diseñada.

Este estudio previo tiene característica determinante, puesto que permitirá elegir de manera correcta cuál algoritmo se debe utilizar al analizar los mensajes/textos en distintos idiomas a fin de obtener los mejores resultados en las tareas de AS que serán implementados en la creación del sistema web de apoyo a decisiones propuesto en los objetivos de esta tesis doctoral.

5.2. OBJETIVOS

5.2.1 OBJETIVO GENERAL

El objetivo general de este estudio se centra en revelar y comparar los resultados relacionados al rendimiento de cuatro algoritmos frecuentemente utilizados en la literatura, en las tareas de clasificación de sentimientos. Para ello, serán contrastados sus resultados de precisión de clasificación y tiempo de respuesta frente a 5 idiomas distintos para, determinar cuál recurso presenta los mejores niveles de bondad en las tareas de AS para cada idioma y también de manera general en multi-idioma.

5.2.2 OBJETIVOS ESPECÍFICOS

Para poder atender al objetivo principal, fueron designados los siguientes objetivos específicos.

A. Estudiar y revisar la literatura de los procedimientos, métodos y aplicaciones del Análisis de Sentimientos a través de algoritmos automáticos.

B. Revisar la literatura para identificar los cuatro algoritmos frecuentes en la literatura designados a las tareas de AS.

C. Diseñar un sistema que posibilite determinar el rendimiento con base en la precisión de clasificación y el tiempo de respuesta de los algoritmos utilizados en las tareas de AS utilizando recursos de Minería de Datos (MD) y Procesamiento del Lenguaje Natural (PLN).

D. Realizar pruebas estadísticas para determinar un ranking de rendimiento de los algoritmos según idioma (inglés, español, portugués, alemán e italiano).

E. Realizar pruebas estadísticas para determinar un ranking de rendimiento de los algoritmos en general (considerando los cinco idiomas).

5.3. METODO DE INVESTIGACIÓN

Los puntos que describiremos a continuación abarcan las herramientas y métodos utilizados para medir el rendimiento de los algoritmos de aprendizaje automático en las tareas de AS. Para establecer el rendimiento de los cuatro algoritmos, serán medidos los niveles de bondad en relación con la clasificación y el tiempo de análisis y cada uno de ellos frente a distintos *datasets* en múltiples idiomas. Finalmente, se aplicarán las

debidas pruebas estadísticas para establecer rankings en base a la clasificación y al tiempo de análisis de estos algoritmos.

A continuación, describimos la fase experimental llevada a cabo.

5.3.1. MEDIDAS DE RENDIMIENTO

La calidad de clasificación de los algoritmos en la clasificación de sentimientos puede ser medida principalmente por 3 indicadores: exactitud *(accuracy)*, precisión *(precision)* y exhaustividad *(recall)*, calculados por medio de las siguientes ecuaciones:

Accuracy: $$\frac{TP + TN}{TP + TN + FP + FN}$$

Precision: $$\frac{TP}{TP+FN}$$

Recall: $$\frac{TP}{TP + FN}$$

Donde TP, FN, FP y TN significan respectivamente el número de instancias verdadero-positivas *(true positive - TP)*, el número de instancias falso-negativas *(false negative - FN)*, el número de instancias falso-positivas *(false positive - FP)* y el número de instancias verdadero-negativas *(true negative TN)*.

La Exactitud (*accuracy* - Acc) indica el porcentaje de concordancia entre la clasificación manual y la clasificación real generada por el método. La

precisión (Prec) indica el valor predictivo positivo, relacionados con los casos recuperados. La exhaustividad ("*recall*" - Rec) indica la fracción de los casos pertinentes que son recuperaron (Zoonen & Meer, 2016). Sin embargo, hay una cuarta medida llamada *F-measure* (F1) que representa la medida harmónica de los indicadores de precisión (precisión) y exhaustividad (recall). Este indicador es representado por la siguiente ecuación:

$$F1 = \frac{2(precision * recall)}{(precision + recall)}$$

En las subsecciones siguientes se muestran los resultados obtenidos de manera agrupada por idioma (subsección 5.4.1.1) y de manera global (subsección 5.4.1.2).

Además de las medidas de rendimiento con relación a la bondad de la clasificación, también es útil evaluar cuán rápidos son los algoritmos en las fases de entrenamiento como de test. Es por lo que se evaluará también tiempo.

5.3.2. PROCEDIMIENTO EXPERIMENTAL

Para estructurar y ejecutar toda la tarea experimental y medir el rendimiento de los algoritmos en los procesos de aprendizaje supervisado,, se utilizó el software `RapidMiner` Studio© en su versión 7.1 (https://rapidminer.com/products/studio). Se trata de un software de pago, aunque tiene una licencia para ámbito académico. Esta licencia académica tiene algunas limitaciones como son la cantidad de registros a

usar en cada ejecución (hasta 20.000) o la imposibilidad de ejecución en paralelo (solo un proceso a la vez). Limitaciones que sin embargo no han condicionado la experimentación llevada a cabo en esta tesis. `RapidMiner` dispone de una interfaz visual para ejecutar los procesos de tratamiento de datos, minería web, aprendizaje automático y otros recursos, como los procedentes del PLN. Su estructura se articula a través cajas con funciones específicas llamadas operadores y cada operador realiza una determinada función en el tratamiento de la información. Además, `RapidMiner` posibilita agregar una amplitud de extensiones de otras aplicaciones (y lenguajes de programación) como por ejemplo `R`, `Python` y `Weka` incorporando sus funciones a su entorno. RapidMiner permite incluso la incorporación mediante interconexión vía API (*Application Programming Interfaz*) con servicios desarrollados por terceros, como por ejemplo la extensión de pago "AYLIEN Análisis de Texto" que permite extraer y analizar informaciones provenientes de textos como artículos de noticias, comentarios en medios sociales y reseñas en sitios web, así como realizar análisis de sentimientos sobre ellos.

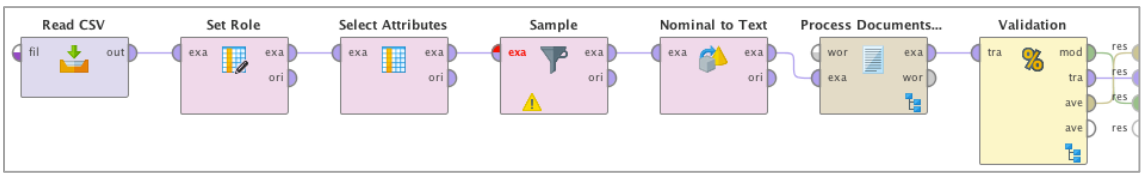
Los operadores en Rapidminer se dividen en 7 grandes grupos: a) importación y exportación de dados (*Data Access*), que permiten lectura y grabación en diversos formatos, así como la conexión para almacenamiento en la nube como Dropbox y Amazon; b) transformación de dados (*Blending*), que permiten funcionalidades como trabajo sobre filas (registros) y columnas (atributos), la definición de tipos de datos, el particionado y la unión de datos, entre otros, c) limpieza de datos (*Cleansing*), con utilidades como normalización, filtrado, eliminación de duplicados, trabajo (imputación) sobre valores perdidos, detección de

valores anómalos, etc., d) modelado de datos (*Modeling*) con utilidades para la aplicación de técnicas predictivas, de segmentación-clasificación, asociativas, optimización, etc., e) validación (*Validation*) con módulos para la medición del rendimiento, la ejecución de la experimentación, la validación cruzada, etc., f) cuantificación (Scoring), que permite la incorporación de bloques (módulos) para medición de rendimiento de la ejecución de los diferentes modelos disponibles en el grupo Modeling, y g) donde se incluirían otras utilidades y para la incorporación de plugins y extensiones.

Todos estos aspectos que van desde su interfaz intuitiva a la extensa opción de análisis y recursos que dispone, hacen de `RapidMiner` una herramienta que destaque frente a demás opciones en este ámbito, lo que justifica su utilización para las tareas propuestas en este trabajo.

Para la experimentación llevada a cabo en este trabajo adaptamos el modelo creado y utilizado por Gonçalves & Fernandes Brito (2013), y que está compuesto por 6 operadores de primer nivel (Figura 18) que se describen en los epígrafes a continuación.

Figura 18. Modelo de operadores de Rapidminer.



Fuente: Elaboración propia a partir de Gonçalves y Fernandes Brito (2013).

5.3.1.1. SISTEMA DE OPERADORES

Los tres primeros operadores que vemos en la Figura 18 se encargan de leer dos datos y determinan el atributo que será utilizado en la evaluación:

- ***Read CSV***: este operador lee el *dataset* de entrenamiento con las entradas previamente clasificadas. Los *datasets* de entrenamiento serán un subconjunto de los *datasets* presentados y comentados en el Capítulo 4.
- ***Set Role***: este operador se utiliza para cambiar la función de uno o más atributos en un grupo de datos. Se puede cambiar los atributos por tipo, por ejemplo, de nominal a binominal, o numérico para polinómica entre otros.
- ***Select attributes*****:** este operador selecciona qué atributos del grupo de datos deben mantenerse y qué atributos no deben ser considerados. Se utiliza en los casos en que no se requiere todos los atributos de un grupo de datos y ayuda a seleccionar solamente los atributos requeridos. Este operador permite reducir la dimensionalidad del *dataset*, y en general gracias a ello se agilizan las tareas de entrenamiento y test.

El cuarto operador llamado "***Sample***", ayuda en el equilibrado de datos, igualando el número de entradas clasificadas como positivas al número de entradas clasificadas como negativas. Suele haber un desbalanceo en los diferentes *datasets* hace un grupo (positivos o negativos).

El quinto operador se llama "***Nominal to text*****.** Este operador cambia el tipo de atributo de nominal para texto. Este procedimiento es

importante para que el programa sepa cuales son las columnas que contienen los textos a ser analizados.

El sexto operador llamado "***Process Document from Data***", genera vectores de palabras a partir de las cadenas de atributos. Además, se encarga de los procesos de preparación y tratamiento de la información (PLN) por medio de operadores a un segundo nivel.

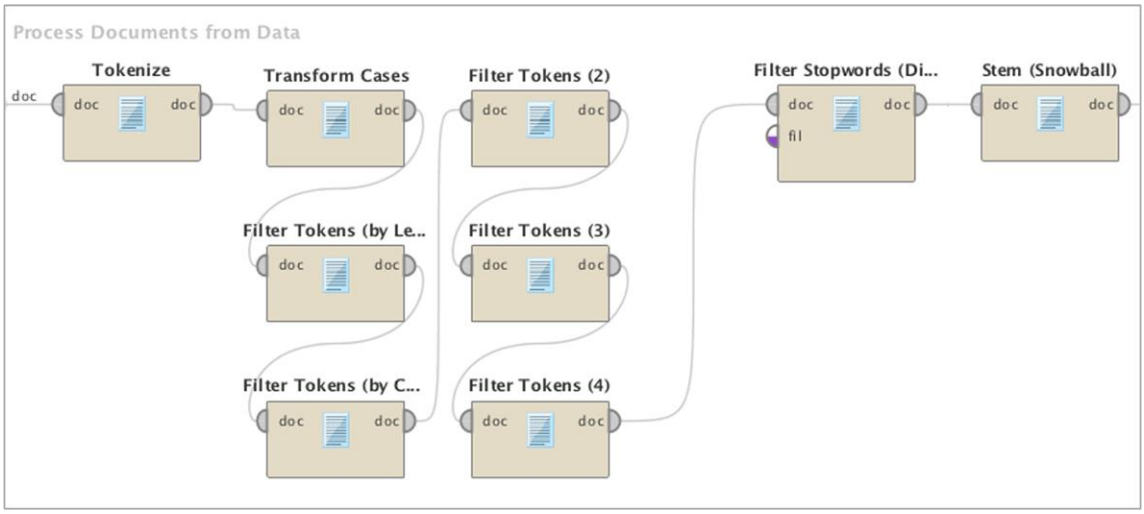
Esta etapa del proceso es fundamental para mejorar el rendimiento a la hora de trabajar con los datos ya que reduce el ruido y simplifica la información.

Para esta tarea fueron utilizamos los siguientes operadores (Figura 19):

- ***Tokenize*** que divide el texto de un documento en una secuencia de tokens. Esto hace posible la aplicación de filtros que simplifiquen el texto como limpieza de dobles espacios en blanco, expresiones no relevantes como http://, entre otros.
- ***Transform cases*** que transforma todas las palabras en mayúsculas encontradas en el texto en minúsculas (también es posible la inversa).
- ***Filter Tokens (by Length)*** que permite determinar la longitud en caracteres de las palabras que se debe considerar en el análisis. Para este trabajo, consideramos las palabras con longitud mínima de 3 caracteres y máxima de 25.
- **Filter Tokens (*by Content)*** que desconsidera palabras o símbolos como o "@", por ejemplo.

- ***Stopwords*** que elimina las palabras vacías como artículos, preposiciones entre otras palabras que no son relevantes para el estudio.
- ***Stem*** que reduce las palabras a su formato raíz antes de añadir afijos flexivos.

Figura 19. Modelo de operadores de segundo nivel *Process Documents from Data* en Rapidminer.

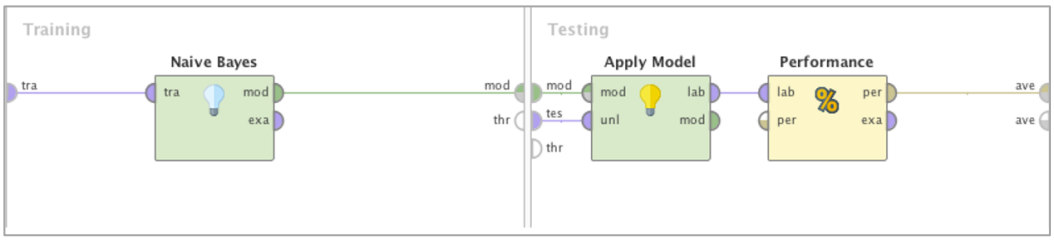


Fuente: Elaboración propia a partir de Gonçalves y Fernandes Brito (2013).

A continuación se aplica el esquema de ponderación de términos TF-IDF (*term frequency – inverse document frequency*) (Salton, 1975; Salton & Yang, 1973; Salton, Yang, & Yu, 1975) que genera una medida numérica de relevancia de cada palabra del documento. Por último, los resultados son transformados en una matriz de documentos de palabras que sirve como entrada óptima para el algoritmo a ser aplicado (van Zoonen & van der Meer, 2016). Como resultado de este proceso se crea un vector de representación que será utilizado por el séptimo operador de nombre Validación ("*Validation"*). El operador "***Validation"***, así como el operador "***Process Document from Data"***, también permite la utilización de

operadores a un segundo nivel y se compone de dos subprocesos llamados entrenamiento (*training*) y test (*testing*) (Figura 20).

Figura 20. Modelo de operadores de segundo nivel *Validation* en Rapidminer.



Fuente: Elaboración propia a partir de Gonçalves y Fernandes Brito (2013).

El primer subproceso (*training)*, recibe los vectores creados en el paso anterior, aplica el algoritmo elegido y crea el modelo (las configuraciones en `RapidMiner` utilizadas para cada uno de los cuatro algoritmos utilizados en este trabajo se observan en la Tabla 4). El segundo subproceso (*testing*), divide la muestra en 15 submuestras de entrenamiento con el mismo número de ejemplos rotulados.

A continuación 14 submuestras son utilizadas para entrenar el clasificador y la submuestra restante se utiliza como grupo de prueba, para verificar la tasa de aciertos del modelo. Este proceso se llama "*Cross-validation"* (*15-fold cross-validation*) (Rezende, 2005) que genera la clasificación y eficiencia del modelo por medio de los indicadores de precisión (*precision*), exactitud (*accuracy*), y exhaustividad (*recall*).

Al final de cada análisis, además de los indicadores citados anteriormente, la herramienta también reporta la cantidad de procesos realizados, la fecha, hora y tiempo demandado para su ejecución (Figura 21).

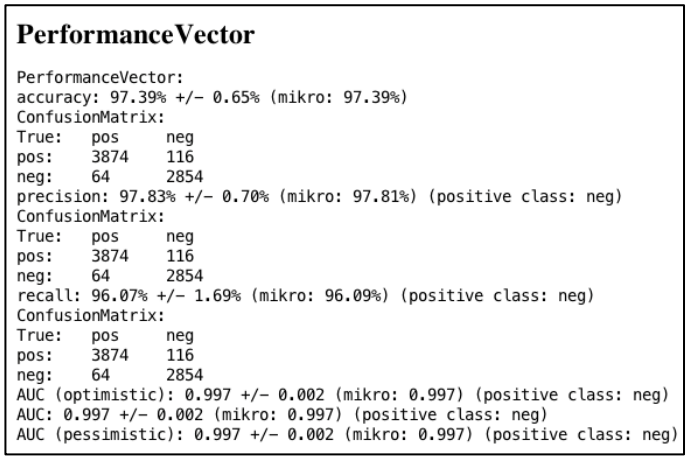
Tabla 4. Configuración de los algoritmos en Rapidminer. Las siglas representadas en las Tablas hacen referencia a los algoritmos de aprendizaje y significan respectivamente, NB=Naïve Bayes, RF=Random Forest, SVM=Support Vector Machine y DT=Decision Tree.

Especificaciones	NB	RF	SVM	DT
Laplace correction	True	-	-	-
Number of trees	-	10	-	-
Criterion	-	grain_ratio	-	grain_ratio
Maximal depth	-	20	-	20
Apply pruning	-	true	-	true
Confidence	-	0.25	-	0.25
Apply prepruning	-	true	-	true
Minimal gain	-	0.1	-	0.1
Minimal leaf size	-	2	-	2
Minimal size for split	-	4	-	4
Number of prepruning alternative	-	3	-	3
Guess subset ratio	-	true	-	-
Voting strategy	-	confid. vote	-	-
Kernel cache	-	-	200	-
Convergence epsilon	-	-	0.001	-
Max interations	-	-	100000	-
Scale	-	-	true	-
L pos	-	-	1.0	-
L neg	-	-	1.0	-

Fuente: Elaboración propia a través del Rapidminer.

En general se recomienda aplicar el proceso de "*cross-validation*" a 10 sub muestras (Chisholm & Hofmann, 2016; Han et al., 2012). Sin embargo, preferimos trabajar con 15 submuestras para obtener una mayor fiabilidad de los resultados.

Figura 21. Resultados de *accuracy*, precisión, *recall*, y rendimiento del vector.



Fuente: Elaboración propia a través del Rapidminer.

5.3.3. TEST ESTADÍSTICOS

La forma clásica utilizada en la comparación sobre múltiples conjuntos de datos es a través del método estadístico ANOVA. Sin embargo, este realiza muchas suposiciones sobre los datos que usualmente no se cumplen en el aprendizaje automático.

Las pruebas no paramétricas pueden utilizarse para comparar los resultados de diferentes algoritmos de aprendizaje automático (García & Herrera, 2008; García et al., en prensa).

Dado que las pruebas no paramétricas no requieren condiciones explícitas para su realización, se recomienda que la muestra de resultados se obtenga siguiendo el mismo criterio, es decir, calcular la misma agregación (promedio, modo, etc.) sobre el mismo número de ejecuciones para cada algoritmo y problema.

En particular, aplicaremos la prueba no paramétrica de Friedman (Demsar, 2006; García & Herrera, 2008; García et al., en prensa; Sheskin, 2003). Esta prueba se puede utilizar para ver si existen diferencias estadísticas significativas entre los algoritmos de aprendizaje automático utilizando la medida de rendimiento en cuestión (exactitud, precisión, exhaustividad y tiempo). El estadístico de Friedman se describe en la Figura 22.

Figura 22. Test no paramétrico de Friedman.

Supongamos que tenemos k clasificadores y n conjuntos de datos. Para calcular el test de Friedman realizamos un ranking de los diferentes algoritmos para cada conjunto de datos por separado. En caso de empate promediamos los rankings. r_i^j corresponderá al ranking del algoritmo j sobre el conjunto de datos i. Bajo la hipótesis nula, el estadístico se calcula como:

$$\chi_F^2 = \frac{12n}{k(k+1)} \left(\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right)$$

donde $R_j = \frac{1}{n} \sum_i r_i^j$ corresponde al promedio de los algoritmos sobre el conjunto de datos i.

Fuente: Benítez, Escudero y Kanaan (2013).

Si se detectan diferencias entonces es necesario aplicar un test posterior para determinar dónde exactamente se encuentran esas diferencias. En nuestro caso usaremos el post test de Holm.

Hay que decir que asumiremos en ambos casos (Friedman y Holm) un valor de satisfacción ($\alpha = 0.05$).

5.4. RESULTADOS

Los resultados procedentes de los análisis fueron ordenados según el idioma y revelan los indicadores relacionados con el rendimiento de los algoritmos en relación con la clasificación (Sección 5.4.1) y el tiempo necesario para realizar cada análisis (Sección 5.4.2). A continuación, detallamos a cada uno de ellos.

5.4.1. RENDIMIENTO DE LOS ALGORITMOS (CLASIFICACIÓN)

En las Tablas 5-9 se presentan los resultados de las cuatro medidas de clasificación usadas (Acc, Prec, Rec y F1) referentes a los análisis de los 4 métodos aplicados respectivamente a los 5 idiomas, así como la media y la desviación típica respectivas. La Figuras 23 y 24 muestra de manera gráfica los indicares de Acc y F1 de los algoritmos en relación con cada *dataset* analizado.

Tabla 5. Indicadores de clasificación en tanto por ciento para *datasets* en inglés.

		Naïve-Bayes				Random-Forest				SVM				Decision-Tree			
	Nº tuits	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
2000Entities	3750	50.08	60.00	0.50	0.99	49.92	49.9	40.25	44.55	49.96	49.90	20.50	29.06	49.75	49.53	26.67	34.67
Hcr	1360	99.01	98.07	100.0	99.02	53.9	77.09	51.47	61.72	100.0	100.0	100.0	100.0	71.56	99.95	43.14	60.26
Movies	6970	92.27	99.07	85.62	91.85	50.00	50.00	66.67	57.14	97.39	97.26	97.54	97.39	49.97	49.75	6.67	11.76
Omd	2203	57.36	60.83	40.73	48.79	50.72	50.78	47,00	48.81	64.39	68.89	52.69	59.71	55.59	53.14	95.90	68.38
Stanford	359	75.41	73.48	81.05	77.07	53.94	53.8	55.51	54.64	79.40	77.05	84.82	80.74	67.21	61.51	93.17	74.09
Semeval2015	5816	54.11	55.99	38.76	45.80	50.08	50.19	20.16	28.76	60.47	67.41	40.78	50.81	61.47	74.13	46.20	56.92
US2012	1767	77.32	74.08	85.15	79.23	52.21	52.06	56.00	53.95	81.60	79.24	86.32	82.62	64.51	89.59	33.13	48.37
DAI_EN	7200	77.57	87.41	64.43	74.18	50.09	50.12	40.00	44.49	83.23	86.37	78.92	82.47	56.46	53.48	99.48	69.56
Media		72.89	76.12	62.03	64.62	51.36	54.24	47.13	49.26	**77.06**	78.27	70.20	**72.85**	59.57	66.39	55.55	53.00
Desviación típica		17.81	17.10	33.13	31.71	1.75	9.33	14.02	10.24	17.62	16.53	28.85	24.39	7.97	19.46	35.74	20.98

Fuente: Elaboración propia a través del Rapidminer.

Tabla 6. Indicadores de clasificación en tanto por ciento para *datasets* en español.

		Naïve-Bayes				Random-Forest				SVM				Decision-Tree			
	Nº tuits	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
General-TASS	5053	70.48	72.00	67.14	69.48	50.11	50.11	53.42	51.71	74.54	71.37	82.22	76.41	57.81	90.77	17.51	29.35
Stompol-TASS	598	60.16	59.33	67.49	63.14	51.83	51.72	54.55	53.09	61.99	59.74	75.55	66.72	54.01	68.44	22.64	34.02
SocialTV-TASS	1343	62.56	59.47	78.92	67.82	50.44	50.55	40.18	44.77	74.43	73.50	76.43	74.93	59.67	85.06	24.09	37.54
SpanishCorpus3100	2602	80.30	76.56	87.37	81.60	50.88	51.08	41.98	46.08	86.35	85.43	87.74	86.56	60.25	95.46	21.62	35.25
Media		68.38	66.84	75.23	70.51	50.82	50.87	47.53	48.91	**74.33**	72.51	80.49	**76.16**	57.94	84.93	21.47	34.04
Desviación típica		9.09	8.79	9.77	7.87	0.75	0.69	7.50	4.10	9.95	10.52	5.67	8.14	2.82	11.79	2.82	3.45

Fuente: Elaboración propia a través del Rapidminer.

Tabla 7. Indicadores de clasificación en tanto por ciento de los *dataset*s en portugués.

		Naïve-Bayes				Random-Forest				SVM				Decision-Tree			
	Nº tuits	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Noticias-Globo	661	58.55	59.01	57.25	58.11	51.00	51.44	36.33	42.58	61.60	60.44	67.67	63.85	58.74	70.40	30.91	42.95
Politica	530	66.82	64.61	74.80	69.33	53.56	53.88	49.28	51.47	69.99	64.41	90.13	75.12	65.52	60.43	91.68	72.84
Entretenimiento	360	60.38	65.07	48.97	55.88	50.52	50.51	53.29	51.86	66.75	66.52	68.16	67.33	62.24	57.77	92.01	70.97
DAI_PT	1800	78.02	73.34	88.47	80.19	50.97	51.21	41.32	45.73	82.89	83.87	81.68	82.76	58.07	54.45	99.22	70.31
Media		65.94	65.51	67.37	65.88	51.51	51.76	45.06	47.91	**70.31**	68.81	76.91	**72.27**	61.14	60.76	78.46	64.27
Desviación típica		8.80	5.90	17.71	11.21	1.38	1.47	7.65	4.53	9.07	10.35	10.95	8.44	3.44	6.87	31.89	14.25

Fuente: Elaboración propia a través del Rapidminer.

Tabla 8. Indicadores de clasificación en tanto por ciento de los *dataset*s en alemán.

		Naïve-Bayes				Random-Forest				SVM				Decision-Tree			
	Nº tuits	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
DAI_GE	358	83.74	80.57	89.08	84.61	50.71	50.66	54.23	52.38	85.99	88.03	83.43	85.66	62.34	57.12	99.23	72.5
GSData	358	56.57	58.88	43.15	49.80	50.35	50.27	60.67	54.98	64.21	69.39	50.36	58.36	52.63	51.98	82.12	63.66
Media		70,16	69,73	66,12	67,21	50,53	50,47	57,45	53,68	**75,10**	78,71	66,90	**72,01**	57,49	54,55	90,68	68,08
Desviación típica		19.21	15.34	32.48	24.61	0.25	0.28	4.55	1.84	15.40	13.18	23.38	19.30	6.87	3.63	12.10	6.25

Fuente: Elaboración propia a través del Rapidminer.

Tabla 9. Indicadores de clasificación en tanto por ciento de los *dataset*s en italiano.

		Naïve-Bayes				Random-Forest				SVM				Decision-Tree			
	Nº tuits	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Sentipolc	3578	68.14	71.85	59.64	65.17	49.97	49.98	60.06	54.55	70.04	80.65	52.66	63.71	56.79	53.72	98.27	69.46
TWNews	1692	68.86	71.33	63.15	66.99	50.00	50.00	54.02	51.93	72.86	72.71	73.52	73.11	60.17	57.00	90.54	69.95
TWSpino	823	67.96	72.02	58.83	64.76	50.70	50.66	53.81	52.18	65.01	82.33	38.47	52.43	58.87	55.07	96.72	70.18
Media		68.32	71.73	60.54	**65.64**	50.22	50.21	55.96	52.89	**69.30**	78.56	54.88	63.08	58.61	55.26	95.18	69.86
Desviación típica		0.48	0.36	2.30	1.19	0.41	0.39	3.55	1.45	3.98	5.14	17.63	10.35	1.70	1.65	4.09	0.37

Fuente: Elaboración propia a través del Rapidminer.

Los resultados descritos en la Tabla 5 relacionados con el idioma inglés muestran que, con relación a *accuracy*, el método SVM presenta mejor clasificación en el 75% (6 de 8) de los *dataset*s analizados, seguido por el método NB con 12,5% (1 de 8) y DT con 12,5% (1 de 8). En relación con el *F-measure*, SVM también presenta los mejores resultados en 62% (5 de 8) de los casos, seguido por el método DT con un 25% (2 de 8), y RF con un 13% (1 de 8).

Los resultados descritos en la Tabla 6 relacionados al idioma español, demuestran que, en relación con la *accuracy*, el algoritmo SVM presenta mejor clasificación en 100% (4 de 4) de los *dataset*s. Lo mismo ocurre con relación al *F-measure* puesto que también presenta los mejores resultados en 100% (4 de 4) de los casos.

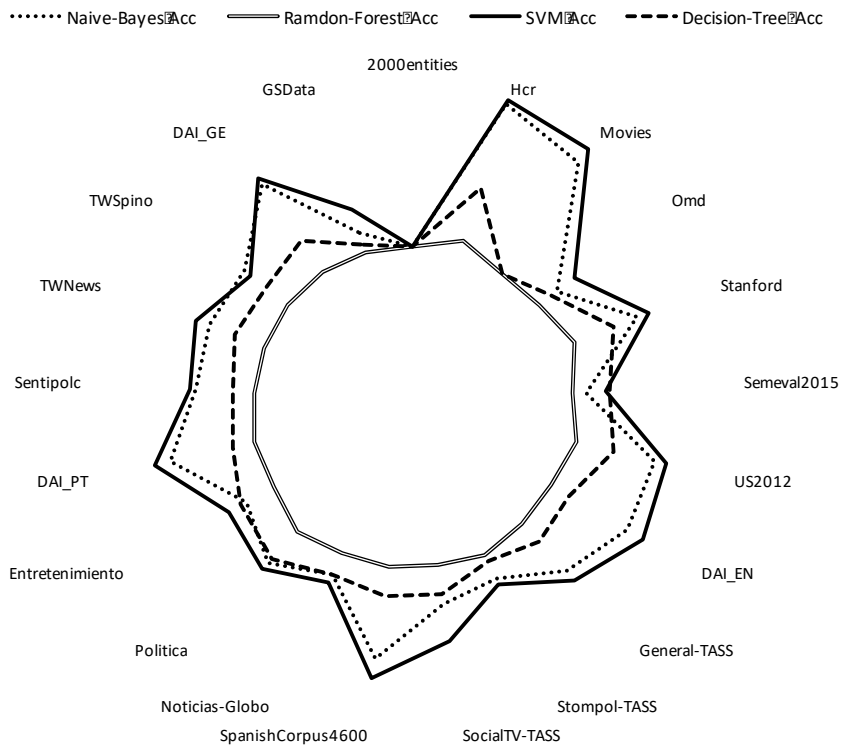
Los resultados descritos en la Tabla 7 relacionados al idioma portugués, demuestran que, con relación a la *accuracy*, el algoritmo SVM presenta mejor clasificación en 100% (4 de 4) de los análisis. Con relación al *F-measure* el SVM presenta los mejores resultados en 75% (3 de 4) de los casos seguido por el DT con 25 % (1 de 4).

Los resultados descritos en la Tabla 8 relacionados al idioma alemán, demuestran que, en relación con la *accuracy*, el algoritmo SVM presenta mejor clasificación en los 2 *dataset*s estudiados. Con relación al *F-measure* los métodos SVM y DT empatan puesto que presentan los mejores resultados en 50% (1 de 2) de los casos.

Los resultados descritos en la Tabla 9 relacionados al idioma italiano demuestran que, en relación con la *accuracy*, el algoritmo SVM presenta mejor clasificación en un 66% (2 de 3) de los análisis, seguido por el NB con 33% (1 de 3). Con relación al *F-measure* el método DT presenta los mejores resultados en un 66% (2 de 3) de los casos seguido por el método SVM con 33% (1 de 3).

En las Tablas 5-9 también se indican las medias y la desviación típica que corroboran la superioridad del SVM en todos los idiomas con base en el indicador de *accuracy.* Con base en el *F-measure* el SVM también es superior en todos los idiomas con excepción del idioma italiano.

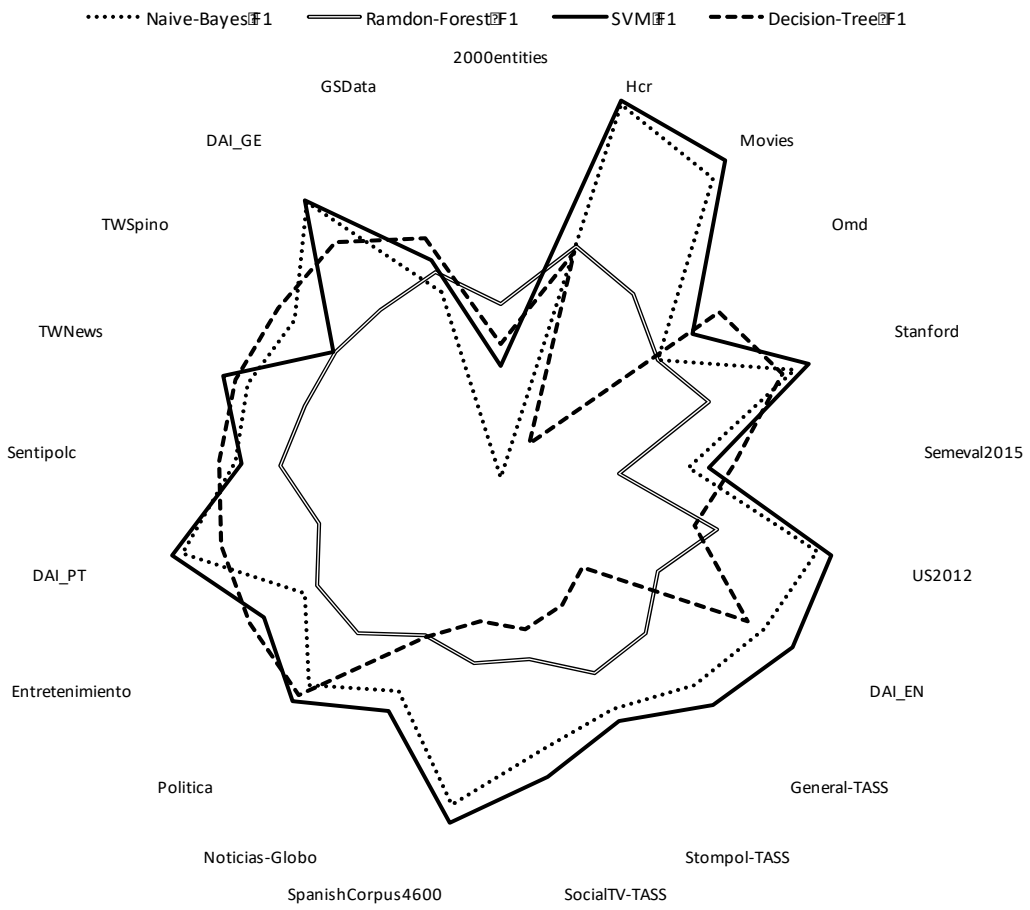
Figura 23. Rendimiento de clasificación normalizado de los algoritmos por *dataset* indicador *accuracy***.**



Fuente: Elaboración propia.

En una representación radial (Figuras 23 y 24), se observa como el método SVM, de manera general, envuelve en rendimiento a los demás algoritmos, indicando su superioridad en este contexto. Este fenómeno es más preeminente en la medida *accuracy* que en el *F-measure*.

Figura 24. Rendimiento de clasificación normalizado de los algoritmos por *dataset* indicador *F-measure*.



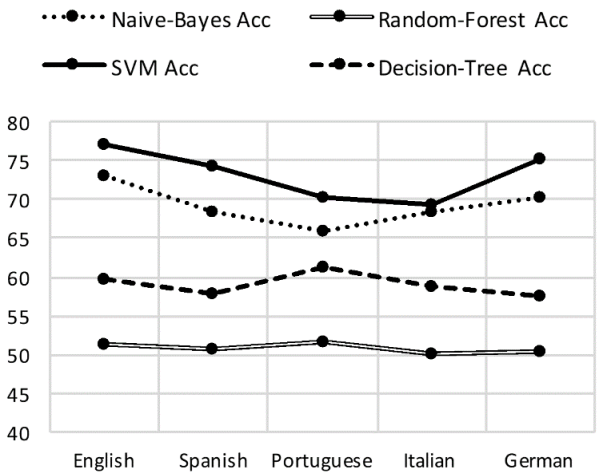
Fuente: Elaboración propia.

Se observa también en el caso específico del dataset TWSpino que, por algún motivo, el SVM se comporta de manera distinta a los demás métodos, puesto que predice con menos eficiencia que el método NB en relación con

el indicador de *accuracy*, y con relación a *F-measure* predice con menos eficiencia que el NB y el DT.

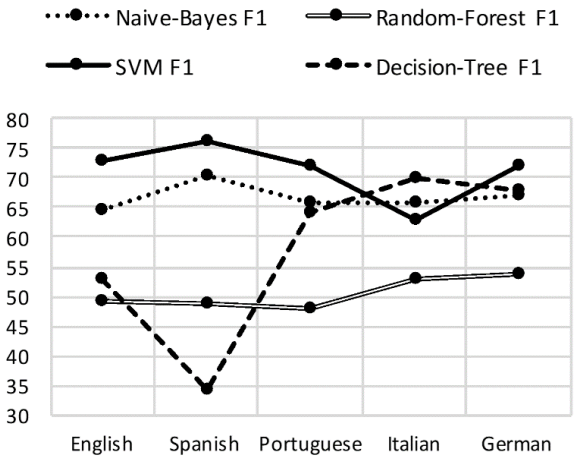
Otro fenómeno interesante ocurre con el dataset 2000entities donde todos los algoritmos presentan de forma anormal niveles muy bajos de clasificación tanto en *accuracy* como en *F-measure*. Además, con relación al *F-measure* ocurre algo inédito, puesto que el RF presenta el mejor resultado en relación con los demás algoritmos.

Figura 25. Rendimiento de clasificación normalizado de los algoritmos según idioma. Indicador *accuracy.*



Fuente: Elaboración propia.

Figura 26. Rendimiento de clasificación normalizado de los algoritmos según idioma. Indicador *F-measure*.



Fuente: Elaboración propia.

Finalmente, se observa en la Figura 25 que con relación a la medida *accuracy* el algoritmo SVM presenta visiblemente los mejores resultados en todos los idiomas. Sin embargo, en el idioma italiano la diferencia en relación con el NB es mínima. Con relación al indicador *F-measure* el método SVM presenta los mejores resultados en todos los idiomas con excepción al idioma italiano donde los métodos DT y NB obtienen mejores resultados (ver Figura 26).

5.4.2. RENDIMIENTO DE LOS ALGORITMOS (TIEMPO)

Las Tablas 10-14 presentan los resultados (medidos en segundos) referentes al tiempo de análisis requerido por cada algoritmo en la ejecución de las tareas de entrenamiento y test, frente a los *dataset*s en inglés, español, portugués, alemán e italiano respectivamente, así como la media y la desviación típica respectivas.

Tabla 10. Tiempo de análisis para los *dataset*s en inglés (segundos).

Data-set	NB	RF	SVM	DT
2000Entities	01	04	01	05
Hcr	13	207	84	397
Movies	01	02	24	12
Omd	04	73	26	20
Stanford	01	05	01	03
Semeval2015	07	85	02	16
US2012	03	37	17	15
DAI_EN	49	376	195	285
Media	**9,88**	98,63	43,75	94,13
Desviación típica	16,33	131,14	66,89	155,39

Fuente: Elaboración propia.

Tabla 11. Tiempo de análisis para los *dataset*s en español (segundos).

Data-set	NB	RF	SVM	DT
General-TASS	67	963	196	241
Stompol-TASS	01	09	01	04
SocialTV-TASS	04	68	10	17
SpanishCorpus4600	18	266	98	84
Media	**22,50**	326,50	76,25	86,50
Desviación típica	30,58	438,34	91,04	108,80

Fuente: Elaboración propia.

Tabla 12. Tiempo de análisis para los *dataset*s en portugués (segundos).

Data-set	NB	RF	SVM	DT
Noticias-Globo	03	42	08	12
Politica	01	11	02	05
Entretenimento	01	04	01	02
DAI_PT	14	208	71	56
Media	**4,75**	66,25	20,50	18,75
Desviación típica	6,24	95,93	33,81	25,18

Fuente: Elaboración propia.

Tabla 13. Tiempo de análisis para los *dataset*s en alemán (segundos).

Data-set	NB	RF	SVM	DT
DAI_GE	07	117	25	29
GSData	01	07	02	03
Media	**4,00**	62,00	13,50	16,00
Desviación típica	4,24	77,78	16,26	18,38

Fuente: Elaboración propia.

Tabla 14. Tiempo de análisis para los *dataset*s en italiano (segundos).

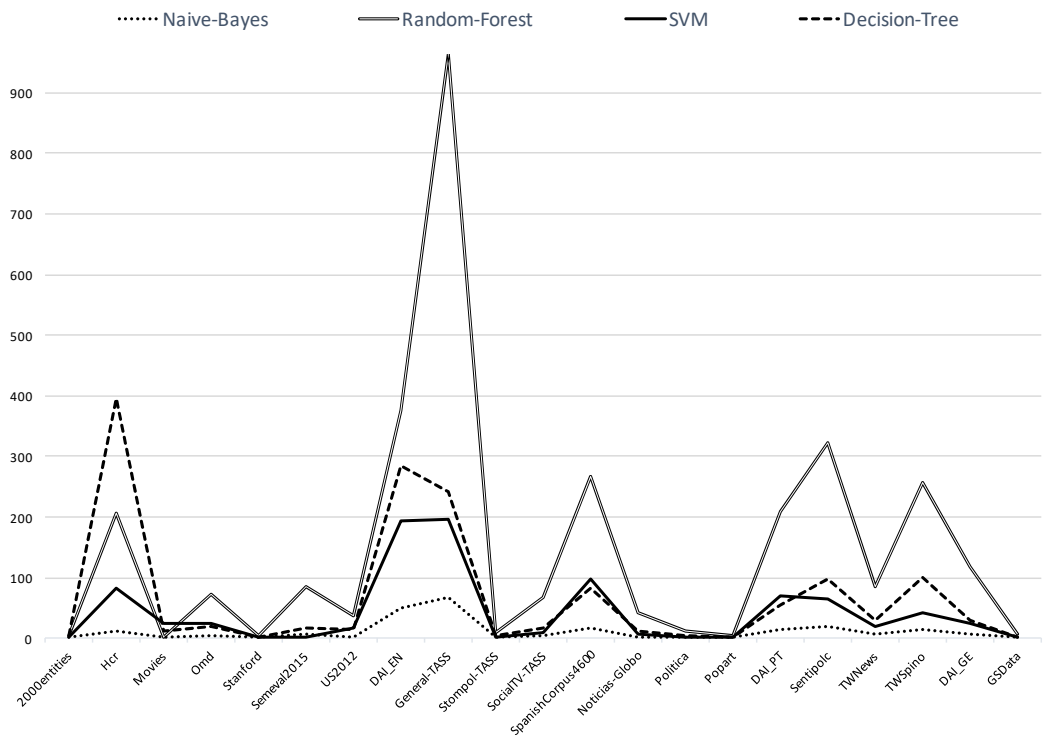
Data-set	NB	RF	SVM	DT
Sentipolc	21	323	66	97
TWNews	08	86	20	30
TWSpino	14	257	42	100
Media	**14,33**	222,00	42,67	75,67
Desviación típica	6,51	122,32	23,01	39,58

Fuente: Elaboración propia.

Los resultados referentes al tiempo de análisis destacan al algoritmo NB como el más rápido en el 95% de los casos. Sin embargo, el algoritmo SVM también presenta los mejores resultados empatando con el algoritmo NB en 4 de los 21 *dataset*s analizados (2000Entities, Stanford, Stompol-TASS y Entretenimento). Relacionado con el método DT, este ocupa el tercer puesto siendo el más rápido en apenas 5% de los casos. Finalmente, los valores expresados por la media corroboran la superioridad con diferencia del algoritmo NB para cada uno de los idiomas.

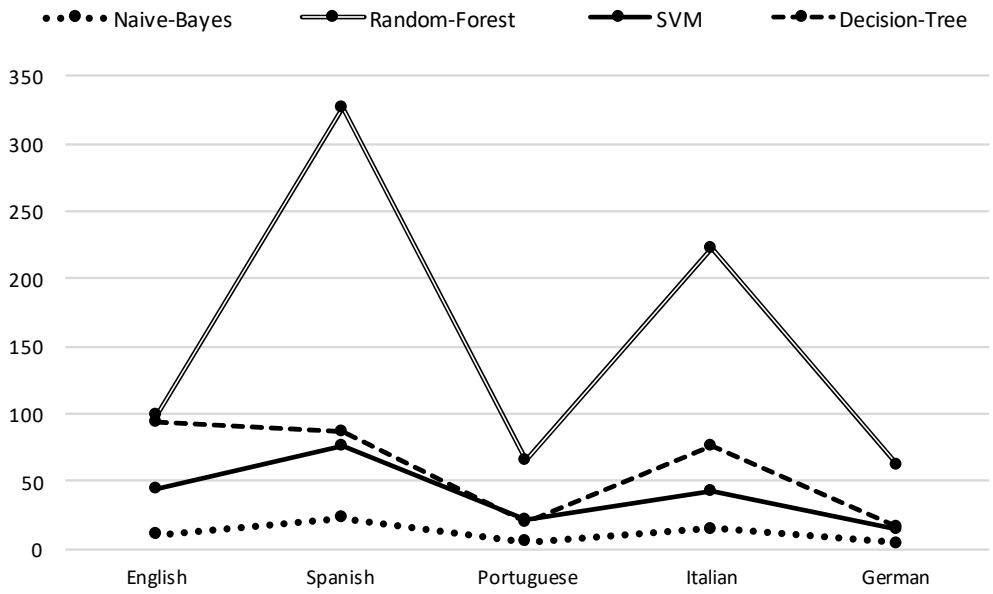
Las Figuras 27 y 28 representan gráficamente los resultados descritos con anterioridad a separados por *dataset* (Figura 27) y agrupados según el idioma (Figura 28).

Figura 27. Rendimiento de clasificación de los algoritmos según el *dataset* (tiempo).



Fuente: Elaboración propia.

Figura 28. Rendimiento de clasificación de los algoritmos según el idioma (tiempo).



Fuente: Elaboración propia.

En los resultados agrupados por idiomas (Figura 28) se corrobora que el algoritmo NB destaca como el más rápido en realizar los análisis de entrenamiento y test. Se observa que el algoritmo SVM es muy superior al DT en los idiomas inglés e italiano y muy poco superior en los idiomas portugués español y alemán. Finalmente, también se observa que el método RF es el algoritmo más lento en todos los idiomas. Sin embargo, en el idioma inglés es donde más se acerca al DT.

5.4.3. RANKING DE RENDIMIENTO (CLASIFICACIÓN)

Como se observa en las Figuras 23 y 24 el rendimiento de los cuatro algoritmos comparados es muy dispar según la medida que se analice. Así por ejemplo, con respecto a la medida *accuracy* (Figura 23) el método

Random-forest es para cualquier dataset usado el que peor y más pobre rendimiento tiene. SVM es que funciona mejor en 18 de los 21 datasets, seguido por NB. Decision-tree es quizás el tercer algoritmo con respecto a esta medida. Con respecto a la medida *F-measure* (Figura 24) no existe un algoritmo claramente vencedor, con rendimientos fluctuantes entre los diferentes datasets. En esta sección comprobaremos si los resultados obtenidos por los cuatro algoritmos comparados en torno al rendimiento en clasificación (Sección 5.4.1) son estadísticamente significativos. Para comparar estos resultados, utilizaremos, como ya se comentó (Sección 5.3.3) una prueba de comparación múltiple. En una prueba de comparación múltiple, primero es necesario comprobar (usando una prueba como la de Friedman) si todos los resultados obtenidos por los algoritmos presentan alguna desigualdad. Si la desigualdad es detectada el siguiente paso consistirá en identificar dónde se produce esa desigualdad, y para ello, como ya se comentó el método de Holm.

Se realizaron los análisis estadísticos sobre las medidas *accuracy* y *F-measure*. Hay que indicar que no es necesario analizar por separado las medidas *precisión* y *recall* ya que ambas están combinadas en la medida *F-measure*. Primero se hace un análisis global teniendo en cuenta el conjunto de los 21 datasets usados como entrenamiento y test. Los resultados arrojados por el test de Friedman muestran total significación para las medidas *accuracy* y *F-measure*, con p-valores de 8.822E-11 y 1.471E-5, respectivamente. Esto permite descartar por completo la hipótesis de igualdad de rendimiento entre los cuatros algoritmos comparados. Existen diferencias significativas de rendimiento (no todos los algoritmos son igual

de buenos). Los resultados arrojan mejores ranking medios para SVM frente al resto de algoritmos para ambas medidas (ver Tabla 15).

Tabla 15. Ranking de precisión y *F-measure* del dataset multilingüe.

Accuracy	
SVM	1.375
Naïve-Bayes	1.917
Decision-Tree	2.833
Random-Forest	3.875
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 51.949. p-valor calculado por la prueba de Friedman: **8.822E-11**.	

F-measure	
SVM	1.625
Naïve-Bayes	2.292
Decision-Tree	2.625
Random-Forest	3.458
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 25.100. p-valor calculado por la prueba de Friedman: **1.471E-5**.	

Fuente: Elaboración propia.

A la vista de los resultados de la Tabla 15, aplicamos la prueba de Holm para comparar el algoritmo con mejor ranking (SVM) con respecto a los otros tres.

Para mostrar los resultados de esta prueba, presentaremos las tablas asociadas al procedimiento de Holm (Tabla 16 para *Accuracy* y Tabla 17 para *F-measure*), en la que se muestran todos los cálculos.

En estas tablas se ordenan los algoritmos con respecto al valor de z obtenido. Así, utilizando la distribución normal, podemos obtener el valor p correspondiente asociado a ella y éste puede compararse con el valor p asociado α/i en la misma fila de la tabla para mostrar si la hipótesis asociada de igual comportamiento es rechazada a favor del algoritmo con mejor ranking o no.

Tabla 16. Test de Holm para *accuracy* para el dataset multilingüe.

i	Algoritmo	z	p	α/i	Hipótesis frente a SVM
3	Ramdon-Forest	6.708204	1.970344E-11	0.016666	Rechazada
2	Decision-Tree	3.913119	9.111162E-5	0.025000	Rechazada
1	Naive-Bayes	1.453444	**0.146101**	0.050000	**Aceptada**

Fuente: Elaboración propia.

Tabla 17. Test de Holm para *F-measure* para el dataset multilingüe.

i	Algoritmo	z	p	α/i	Hipótesis frente a SVM
3	Ramdon-Forest	4.919350	8.683228E-7	0.016666	Rechazada
2	Decision-Tree	2.683282	0.007290	0.025000	Rechazada
1	Naive-Bayes	1.788854	**0.073638**	0.050000	**Aceptada**

Fuente: Elaboración propia.

Como se puede observar, la aplicación de la prueba de Holm determina que el rendimiento de SVM es claramente superior al rendimiento mostrado por los algoritmos Randon-Forest y Decision-Tree para ambas medidas. El test rechaza la hipótesis de igualdad frente a estos algoritmos. Sin embargo, el test no ha podido determinar diferencias significativas, con α = 0.5, con respecto a Naive-Bayes. Los resultados no son lo suficiente potentes como para que el test detecte la diferencia. En lo que sigue se realizan test de Friedman y Hold individualizados para determinar dónde, en qué idioma, se producen diferencia y dónde no (Tablas 18-32).

En la Tabla 18 se observa como el test de Friedman descarta la igualdad de rendimiento (p = 5.153E-4) para *accuracy* y no para *F-measure* con respecto al idioma inglés. El procedimiento de Holm (Tabla 19) nos indica que las diferencias se encuentran con respecto a los algoritmos Random-Forest y Decision Tree. Sin embargo, Holm no es capaz de detectar diferencias estadísticas entre SVM (el algoritmo con el mejor ranking) y NB (el algoritmo con el segundo mejor ranking).

Tabla 18. Ranking de precisión y *F-measure* de datasets en inglés.

Accuracy	
SVM	1.444
Naïve-Bayes	1.888
Decision-Tree	2.888
Random-Forest	3.777
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 17.666. p-valor calculado por la prueba de Friedman: **5.153E-4**.	

F-measure	
SVM	1.666
Naïve-Bayes	2.444
Decision-Tree	2.666
Random-Forest	3.222
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 6.733. p-valor calculado por la prueba de Friedman: **0.081**.	

Fuente: Elaboración propia.

Tabla 19. Test de Holm para *accuracy* para el dataset en inglés.

i	Algoritmo	Z	P	α/i	Hipótesis frente a SVM
3	Ramdon-Forest	3.834058	1.260465E-4	0.016666	Rechazada
2	Decision-Tree	2.373464	0.017622	0.025000	Rechazada
1	Naive-Bayes	0.730297	**0.465209**	0.050000	**Aceptada**

Fuente: Elaboración propia.

Con respecto al idioma español, Friedman observa diferencias de rendimiento entre los cuatro algoritmos (Tabla 20). Holm (Tablas 21 y 22) nos indica que estadísticamente no es posible diferenciar el rendimiento de SVM (el algoritmo con mejor ranking) con respecto a los métodos DT y NB. Sí que se observa estadísticamente mejor rendimiento de SVM con respecto a RF. Mismo razonamiento se puede hacer con respecto al idioma portugués (Tablas 23, 24 y 25).

Tabla 20. Ranking de precisión y *F-measure* de datasets en español.

Accuracy	
SVM	1.200
Naïve-Bayes	1.799
Decision-Tree	3.000
Random-Forest	4.000
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 14.04. p-valor calculado por la prueba de Friedman: **0.003**.	

F-measure	
SVM	1.200
Naïve-Bayes	1.799
Decision-Tree	3.000
Random-Forest	4.000
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 14.04. p-valor calculado por la prueba de Friedman: **0.003**.	

Fuente: Elaboración propia.

Tabla 21. Test de Holm para *accuracy* para el dataset en español.

i	Algoritmo	z	p	α/i	Hipótesis frente a SVM
3	Ramdon-Forest	3.429286	6.051723E-4	0.016666	Rechazada
2	Decision-Tree	2.204541	**0.027486**	0.025000	**Aceptada**
1	Naive-Bayes	0.734847	**0.462433**	0.050000	**Aceptada**

Fuente: Elaboración propia.

Tabla 22. Test de Holm para *F-measure* para el dataset en español.

i	Algoritmo	z	p	α/i	Hipótesis frente a SVM
3	Ramdon-Forest	3.429286	6.051723E-4	0.016666	Rechazada
2	Decision-Tree	2.204541	**0.027486**	0.025000	**Aceptada**
1	Naive-Bayes	0.734847	**0.462433**	0.050000	**Aceptada**

Fuente: Elaboración propia.

Tabla 23. Ranking de precisión y *F-measure* de datasets en portugués.

Accuracy	
SVM	1.599
Naïve-Bayes	2.200
Decision-Tree	2.400
Random-Forest	3.800
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 7.800. p-valor calculado por la prueba de Friedman: **0.050**.	

F-measure	
SVM	1.599
Naïve-Bayes	2.199
Decision-Tree	2.200
Random-Forest	4.000
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 9.719. p-valor calculado por la prueba de Friedman: **0.021**.	

Fuente: Elaboración propia.

Tabla 24. Test de Holm para *accuracy* para el dataset en portugués.

i	Algoritmo	z	p	α/i	Hipótesis frente a SVM
3	Ramdon-Forest	2.694439	0.007051	0.016666	Rechazada
2	Decision-Tree	0.979796	**0.327187**	0.025000	**Aceptada**
1	Naive-Bayes	0.734847	**0.462433**	0.050000	**Aceptada**

Fuente: Elaboración propia.

Tabla 25. Test de Holm para *F-measure* para el dataset en portugués.

i	Algoritmo	z	p	α/i	Hipótesis frente a SVM
3	Ramdon-Forest	2.939388	0.003289	0.016666	Rechazada
2	Decision-Tree	0.734847	**0.462433**	0.025000	**Aceptada**
1	Naive-Bayes	0.734847	**0.462433**	0.050000	**Aceptada**

Fuente: Elaboración propia.

Con relación al idioma alemán, los resultados arrojados por los cuatros algoritmos no son los suficientemente potentes en ninguna de las dos medidas (*accuracy* y *F-measure*) como para que las diferencias puedan ser detectadas por Friedman (Tabla 26).

Tabla 26. Ranking de precisión y *F-measure* de datasets en alemán.

Accuracy	
SVM	1.000
Naïve-Bayes	2.000
Decision-Tree	3.000
Random-Forest	4.000
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 6.000. p-valor calculado por la prueba de Friedman: **0.112**.	

F-measure	
SVM	1.500
Decision-Tree	2.000
Naïve-Bayes	3.000
Random-Forest	3.500
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 3.000. p-valor calculado por la prueba de Friedman: **0.392**.	

Fuente: Elaboración propia.

Con respecto al italiano Friedman detecta diferencias de rendimiento con respecto a *accuracy* y no con respecto a *F-measure* (Tabla 27). Un análisis más detallado mediante Holm nos muestra que las diferencias se encuentran con respecto al algoritmo Random-Forest. Holm no detecta diferencias estadísticamente significativas entre el algoritmo con mejor ranking (SVM) y los algoritmos NB y DT (Tabla 28).

Tabla 27. Ranking de precisión y *F-measure* de datasets en italiano.

Accuracy	
SVM	1.333
Naïve-Bayes	1.666
Decision-Tree	3.000
Random-Forest	4.000
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 8.200. p-valor calculado por la prueba de Friedman: **0.042**.	

F-measure	
Decision-Tree	1.333
SVM	2.333
Naïve-Bayes	2.333
Random-Forest	4.000
Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: 6.599. p-valor calculado por la prueba de Friedman: **0.086**.	

Fuente: Elaboración propia.

Tabla 28. Test de Holm para *accuracy* para el dataset en italiano.

i	Algoritmo	Z	P	α/i	Hipótesis frente a SVM
3	Ramdon-Forest	2.529822	0.007051	0.016666	Rechazada
2	Decision-Tree	0.979796	**0.327187**	0.025000	**Aceptada**
1	Naive-Bayes	0.734847	**0.462433**	0.050000	**Aceptada**

Fuente: Elaboración propia.

5.4.4. TEST ESTADÍSTICO SOBRE EL TIEMPO

En la Figura 28 se observa como NB es siempre el algoritmo en media más rápido con relación al tiempo (entrenamiento + test). El algoritmo que más tiempo consume de todos los algoritmos es Random-Forest. En la Tabla 29 se muestra el resultado del test estadístico de Friedman sobre los diferentes idiomas y sobre el conjunto completo de dataset (columna multilingüe).

Tabla 29. Ranking de tiempo de análisis. Estadístico de Friedman (distribuido según chi-cuadrado con 3 grados de libertad: inglés: 11.633 y P-valor: **0.008**; español: 13.380 y P-valor: **0.003**; portugués: 13.380 y P-valor: **0.004**; alemán: 6.000 y P-valor: **0.112**; italiano: 9.000 y P-valor: **0.029**; Multilingüe: 49.737 y P-valor: **1.169E-10**.

Inglés		Español		Portugués	
Naïve-Bayes	3.611	Naïve-Bayes	3.899	Naïve-Bayes	3.900
SVM	2.333	SVM	2.900	SVM	2.900
Decision-Tree	2.500	Decision-Tree	2.200	Decision-Tree	2.200
Random-Forest	1.556	Random-Forest	1.000	Random-Forest	1.000

Alemán		Italiano		Multilingüe	
Naïve-Bayes	4.000	Naïve-Bayes	4.000	Naïve-Bayes	3.812
SVM	3.000	SVM	3.000	SVM	2.666
Decision-Tree	2.000	Decision-Tree	2.000	Decision-Tree	2.313
Random-Forest	1.000	Random-Forest	1.000	Random-Forest	1.208

Fuente: Elaboración propia.

Relacionado al tiempo de análisis el algoritmo NB revela tener los mejores resultados tanto por idioma cuanto de manera general. Este algoritmo viene seguido por el SVM, DT y RF que ocupan el segundo, tercer y cuarto lugar sucesivamente.

5.5. DISCUSIÓN

En los resultados obtenidos se observa que el método SVM presenta mejor clasificación en todos los idiomas frente a los demás métodos analizados. El algoritmo NB ocupa el segundo puesto en clasificación seguido por los algoritmos DT y RF. Con relación al tiempo de análisis, los primeros lugares se invierten, puesto que el método NB presenta los menores tiempos de análisis, seguido por SVM, DT y RF.

En términos generales, con relación a la medida *accuracy*, el método SVM es en media 5,7% mejor que el NB. Con relación al *F-measure*, SVM es en media 4,3% mejor que el NB. Sin embargo, con relación al tiempo de análisis NB es en media 113,6% más rápido que el método SVM.

En la Tabla 30 obsérvese la diferencia porcentual entre los algoritmos SVM y NB en las categorías *accuracy*, *F-measure* y tiempo, agrupados por idioma.

Considerando el indicador *accuracy*, el método SVM presenta resultados superiores en todos los idiomas. La menor diferencia entre SVM y el NB está en el idioma italiano (1,4%) y la mayor diferencia en el idioma español (8,3%).

Tabla 30. Diferencias porcentuales entre SVM y NB (*accuracy, F-measure y tiempo*).

Idioma	Accuracy	F-measure	Time
Inglés	SVM (5,5%) > NB	SVM (8,1%) > NB	NB (126,6%) más rápido que SVM
Español	SVM (8,3%) > NB	SVM (5,1%) > NB	NB (108,8%) más rápido que SVM
Portugués	SVM (6,4%) > NB	SVM (6,2%) > NB	NB (124,7%) más rápido que SVM
Italiano	SVM (1,4%) > NB	NB (2,6%) > SVM	NB (99,4%) más rápido que SVM
Alemán	SVM (6,8%) > NB	SVM (4,6%) > NB	NB (108,5%) más rápido que SVM
Medias	**SVM (5,7%) > NB**	**SVM (4,3%) > NB**	**NB (113,6%)** más rápido que SVM

Fuente: Elaboración propia.

Considerando el indicador *F-measure*, el método SVM presenta resultados superiores en todos los idiomas con excepción al italiano. En este idioma el NB predice un 2,6% mejor que el SVM. Para los demás idiomas apréciese la menor diferencia entre los dos algoritmos en el idioma alemán (4,6%) y la mayor diferencia en el idioma inglés (8,1%).

Considerando el indicador tiempo, el algoritmo NB es el más rápido en todos en todos los idiomas. La menor diferencia entre ellos está en el idioma italiano donde NB es 2,97 veces más rápido que SVM (99,4%) y la mayor diferencia está en el idioma inglés donde es 4,43 veces más rápido que SVM (126,6%).

5.6. CONCLUSIONES

El Análisis de Sentimientos utiliza dos principales recursos para clasificar sentimientos, a) abordaje léxica en base a diccionarios léxicos de sentimientos y b) aprendizaje automático, que sucede a través de datos previamente rotulados utilizados para entrenamiento de un modelo predictivo a través de algoritmos automáticos (Tsytsarau & Palpanas, 2012).

En este capítulo, por un lado, se hizo una revisión del estado de arte sobre ambos enfoques y sus respectivas técnicas. Por otro lado, a través de la aplicación de técnicas de Aprendizaje automático se estudió cuál de los 4 algoritmos frecuentes en la literatura, presenta mejor performance con relación a la precisión de clasificación y el tiempo de análisis. Para llevar a cabo esa tarea, se utilizaron un total de 21 *datasets* en 5 idiomas distintos, manualmente etiquetados por investigadores del área y recuperados a partir de la literatura disponible (Steiner-Correa et al., 2018). Además, hemos creado un repositorio online donde están disponibles, para uso de la comunidad científica y aficionados, todos los *datasets* utilizados en este trabajo.

Los resultados consecuentes de los más de 500 análisis realizados en este trabajo ponen de manifiesto que, respecto al tiempo requerido por los algoritmos en realizar los análisis, el método NB presenta con diferencia los mejores resultados y viene seguido de los algoritmos SVM, DT y RF.

Por un lado, respecto a la precisión de clasificación, el algoritmo SVM ocupa el primer puesto en el ranking seguido por NB, DT y RF. Al considerar las diferencias de los resultados de clasificación y tiempo entre los algoritmos NB y SVM, se verifica que el SVM entrega una clasificación mejor que el NB *(accuracy 5,7% y F-measure 4,3%)*. Sin embargo, a cambio exige un gran sacrificio en tiempo ya que es 113,6% más lento.

Por otro lado, agrupando los resultados según idioma, se constata que el algoritmo SVM presenta mayor dificultad en predecir correctamente los datos en italiano frente a los demás idiomas. Cabe mencionar que la calidad del *dataset* de entrenamiento es clave para conseguir buenos resultados,

de modo que se abren precedentes para investigaciones futuras con grupos de entrenamientos aún más amplios en el idioma italiano a fin de corroborar este fenómeno y validarlo científicamente.

Finalmente, como los resultados conseguidos con este estudio se ha revelado el algoritmo que clasifica mejor y más rápido relacionado con cada uno de los 5 idiomas analizados. Este descubrimiento permite implementar en el sistema web LOGOS, recursos de inteligencia artificial que decidan de manera automática cuál algoritmo utilizar, según el idioma que se analiza, para ofrecer los mejores resultados. Debido a los resultados obtenidos, justo con los test estadísticos realizados, se posibilitará al usuario – analista la elección del algoritmo que mejor desempeño presente según sus necesidades. Así podrá elegirse entre a) obtener resultados más precisos y tiempos de análisis más lentos (SVM), b) o bien decantarse por NB, algo menos preciso pero mucho más rápido. LOGOS incorporará ambos algoritmos como posibilidades de análisis, dado que ambos métodos son equivalentes a nivel de clasificación desde el punto de vista estadístico..

CAPÍTULO 6: APORTACIÓN 3 - DESARROLLO DE LOGOS: UNA HERRAMIENTA BASADA EN EL ANÁLISIS DE SENTIMIENTOS MULTILINGÜE COMO APOYO A LA TOMA DE DECISIONES DE MARKETING.

6.1. INTRODUCCIÓN

Este capítulo atiende al objetivo específico (C) de diseñar e implementar un software que integre recursos de minería de datos, análisis de sentimientos y monitoreo de medios sociales en contextos multilingües.

En los ítems que siguen, describimos las fases de conceptualización y diseño, de implementación, las pruebas realizadas, así como la metodología seguida para todo ello.

6.2 ANÁLISIS DE REQUERIMIENTOS

Se busca diseñar y desarrollar un sistema web que actúe como cuadro de mando y que sea utilizado libremente por los analistas de marketing, para que de manera intuitiva, puedan sacar informes que auxilien en la tomada de decisiones estratégicas para/en las PyMEs.

Los requerimientos básicos que se desean del sistema son:

Requerimientos funcionales:

RF.1 Diseñar y desarrollar un cuadro de mando intuitivo, que permita recolectar los mensajes (de Twitter y otros medios sociales) relacionados a una determinada cuenta de usuario, hashtag o cualquier palabra en general.

RF.2 Proporcionar la visualización y la descarga de los mensajes y reenvíos en ficheros que puedan ser manejados por aplicaciones de hojas de cálculo.

RF.3 Posibilitar la visualización y la descarga de listados, gráfica y rankings de los hashtags más frecuentes en los mensajes obtenidos.

RF.4 Facilitar la visualización y descarga de listados, gráficas y ranking de los usuarios más mencionados en los mensajes, de manera individual o por parejas.

RF.5 Posibilitar la visualización geográfica de procedencia de los mensajes, cuando esta esté disponible, así como el porcentaje de mensajes geolocalizables.

RF.6 Proporcionar la aplicación de análisis de sentimientos, mediante el enfoque del aprendizaje automático por medio del método más adecuado según el idioma con el que se vaya a trabajar en cada caso.

RF.7 El sistema detectará el idioma automáticamente y aplicará el idioma más adecuado.

RF.8 Creación de nubes de palabras con distintas configuraciones, según su frecuencia encontrada en los mensajes.

RF.9 Suministrar la visualización y descarga de listados, gráficas y ranking de los dispositivos utilizados en la publicación de los mensajes.

RF.10 Posibilitar la visualización y descarga de listados y gráficas, del perfil de los usuarios que realizaron las publicaciones, así como el

número de mensajes publicados por cada uno de ellos, el número de seguidores y su idioma.

RF.11 Realizar el análisis de los mensajes en varios idiomas, entre los que se debe incluir el inglés, y el español y portugués (por cuestiones relacionadas con la beca, como ya se comentó en secciones anteriores).

Requerimientos tecnológicos:

RT.1 El sistema tiene que ser un sistema basado en web.

RT.2 El sistema tiene que estar desarrollado con tecnologías libres, tanto en los lenguajes de programación como en el almacenamiento persistente de datos.

RT.3 El sistema tiene que ser liberado en forma de código libre.

RT.4 Los informes tienes que ser generados en varios formatos libres como PDF, LaTeX, HTML, SVG y CSV.

6.3. MÉTODOLOGÍA

Una vez conocidos los objetivos y requerimientos funcionales y tecnológicos que debe tener el sistema, y para poder determinar qué métodos, visualizaciones, tratamiento de datos ayudarían más al usuario final (analista) para la toma de decisiones en marketing, se investigó sobre literatura previa de análisis de sentimientos en medios sociales, minería de textos, monitoreo de medios sociales y procesamiento del lenguaje natural.

Se investigó sobre aplicaciones web que se encuentran en el mercado actual, qué es lo que ofrecen estas herramientas, sus costos, tipos de gráficas que muestran, con qué medios sociales están conectadas, etc. (ver Tabla 1 e información asociada en el capítulo 2).

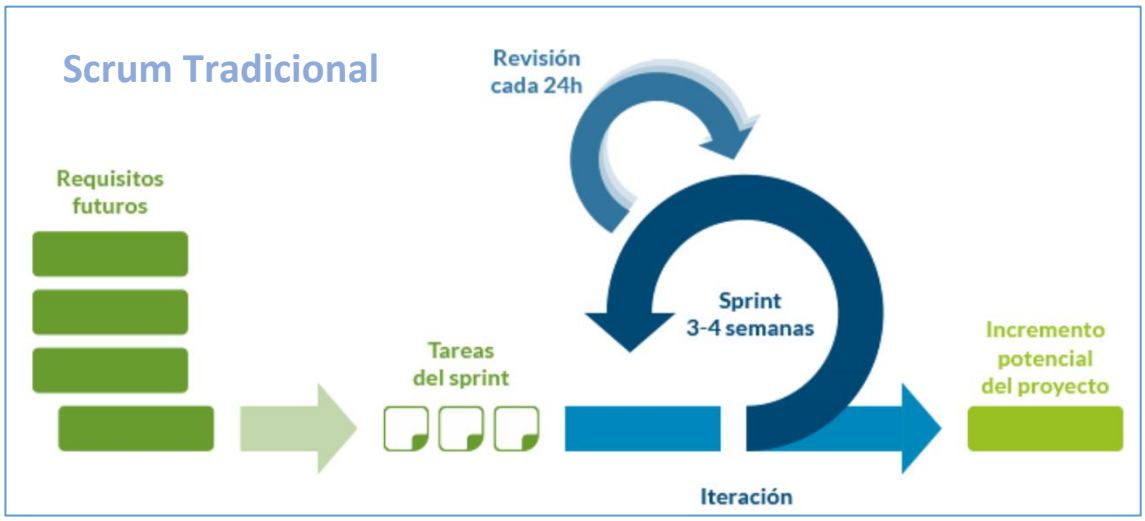
Se investigó también qué tecnologías serían las más adecuadas para el desarrollo e implementación a fin de satisfacer los requerimientos definidos en los objetivos. Esto contemplaba definición del lenguaje de programación (se analizaron `PHP`, `Python` y `R`), base de datos para almacenamiento persistente (se analizaron `MongoBD`, `MySQL`, Texto Plano), frameworks (`YII`, `Shiny`), y otros recursos requeridos por parte del servidor (como servidor de ficheros, servidor web, etc.). Se analizó la funcionalidad de los mencionados lenguajes de programación y sistemas de almacenamiento, así como su curva de aprendizaje y facilidad de mantenimiento a largo plazo. Más adelante se comentan las decisiones tomadas.

El diseño y desarrollo de la aplicación web se dividió en varias fases (releases) utilizando la metodología Agile[18], y el framework de gestión de proyectos Scrum (Robert & Brown, 2004). A continuación se hace una breve explicación del ciclo de vida del proyecto a aplicar (ver Figura 29).

[18] Fuente: http://www.dmk-innovations.de/en/agile-development/ (acceso el 07/08/2019).

En cada iteración *(o sprint*) se analiza si lo planificado resultó factible o no; internamente cada sprint cuenta con su análisis, diseño, desarrollo, pruebas según el *sprint planning* previamente definido, de esta manera el proyecto se convierte en un organismo vivo capaz de cambiar por sprint según se requiera o el usuario final lo necesite, con el objetivo de tener un mínimo producto viable.

Figura 29. Proceso de Scrum Framework para gestión de proyectos



Fuente: https://lorbada.com/blog/2017/02/10/diferentes-metodologias-agiles/.

La primera fase o iteración se centró en la conceptualización del producto (Sección 6.4), se realizó un diseño preliminar de la interfaz de usuario (Sección 6.5), se diseñó su imagen visual (Sección 6.6) y un test de funcionalidad mínima (Sección 6.7).

A continuación se realizó una validación más exhaustiva, con un caso práctico que permitiera analizar y contrastar los resultados de la aplicación.

Teniendo en cuenta lo comentado anteriormente con respecto a la beca (CAPES) que sustenta la presente tesis doctoral, se decidió realizar el caso práctico de validación con empresas con relevancia en Twitter y posicionadas en los marcados de España y Brasil. Los principales resultados y las conclusiones más destacadas obtenidas por esta versión preliminar (prototipo) de LOGOS se muestran en detalle en el Capítulo 7.

Fruto de lo aprendido con el uso de esta primera versión de LOGOS, y con los resultados obtenidos en su validación, se ha llegado a una serie de conclusiones, algunas limitaciones y varias acciones de mejora a tener en cuenta en la siguiente iteración del desarrollo. Todas ellas se relatan en la Sección 6.8 del presente capítulo.

Hay que decir que la versión que se presenta en esta memoria de tesis es la versión prototipo, y que se describe a continuación. LOGOS continuará en evolución y mejora, pero ya fuera del ámbito de este documento de tesis.

6.4. FASE DE CONCEPTUALIZACIÓN

En esta fase se llevó a cabo la primera de las iteraciones.

Inicialmente se plantearon las tareas de investigación correspondientes de búsqueda de artículos, libros u otras aplicaciones relacionadas con el proyecto. Este paso previo fue fundamental para determinar cuáles características, recursos y arquitectura, debería ofrecer LOGOS para que

fuese una solución funcional flexible que facilite, al público en general y a las PyMEs en particular, la puesta en marcha y la ejecución de análisis de datos de medios sociales de una manera fácil y adaptable a las necesidades particulares de cada empresa, a bajo o nulo costo (ver Tabla 1 e información asociada en el capítulo 2). Para ello, tras una exhaustiva revisión de la literatura, se definió que el sistema debería ser capaz de realizar distintas tareas relacionadas con la Minería de Datos y también con el Análisis de Sentimientos. Además, debería conectarse a medios sociales de manera automática para proceder con la descarga de datos. Que estuviera adecuado para procesar la información y generar distintos informes, que fuese multilingüe, es decir, con capacidad de procesamiento de datos en distintos idiomas utilizando automáticamente el algoritmo que mejor resultados ofrece según el idioma que se analice, que se pudiera manejar de manera intuitiva por medio de una interfaz web de control y finalmente, que fuese gratuita para uso de cualquier persona o empresa.

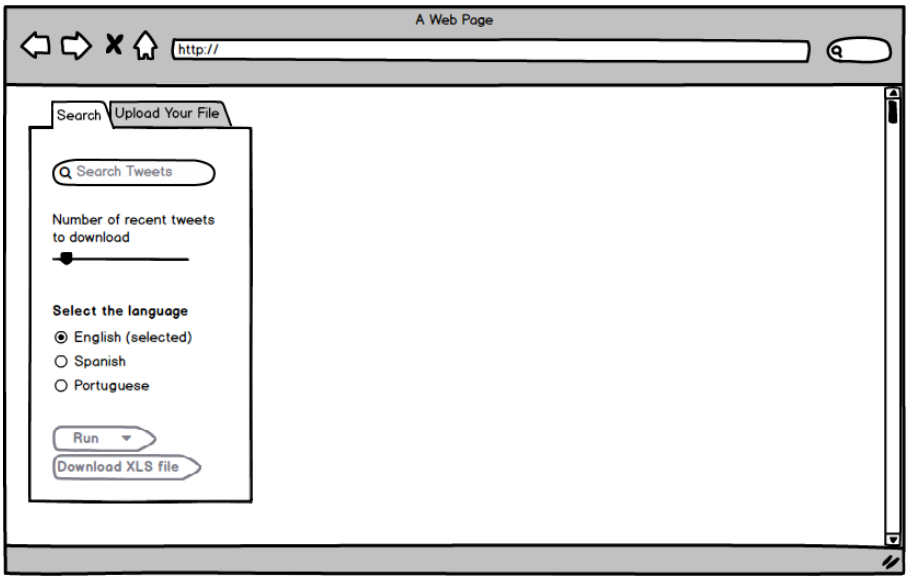
Con base en las investigaciones previas realizadas, se decidió construir `LOGOS` utilizando el Framework web `Shiny` (Chang, Cheng, Allaire, Xie, & McPherson, 2015). Este recurso tiene `R` como lenguaje de programación nativo. Ambos recursos se comentarán en detalle en la implementación del prototipo (sección 6.6.).

El siguiente paso fue definir el idioma y el ambiente gráfico visual de la aplicación. En cuanto al idioma de la interfaz, se optó inicialmente por inglés debido a su amplitud de uso a nivel mundial. Sin embargo, rápidamente se

implementaron las traducciones de la interfaz a otros idiomas empezando por el español y portugués.

En cuanto al ambiente gráfico visual, la interfaz de usuario debería atender a los conceptos minimalistas a nivel de diseño e intuitivos a nivel de usabilidad. En este sentido, se decidió crear un cuadro de mando general a la derecha de la pantalla, y a la izquierda ubicar las diferentes pestañas de informes. Este cuadro de mando general tendría, en esta primera fase (ver Figura 30), dos pestañas llamadas "*Search"* y "*Upload Your File".* La primera (*Search)*, posibilitaría la inserción de la búsqueda por palabra, cuenta o hashtag, el número máximo de tuits que se desea descargar y el idioma de los mensajes. También tendría insertados dos botones, "*Run"* que dispara el proceso y el botón "*Donwload .XLS File"* que genera el fichero .xls con los mensajes descargados del Twitter.

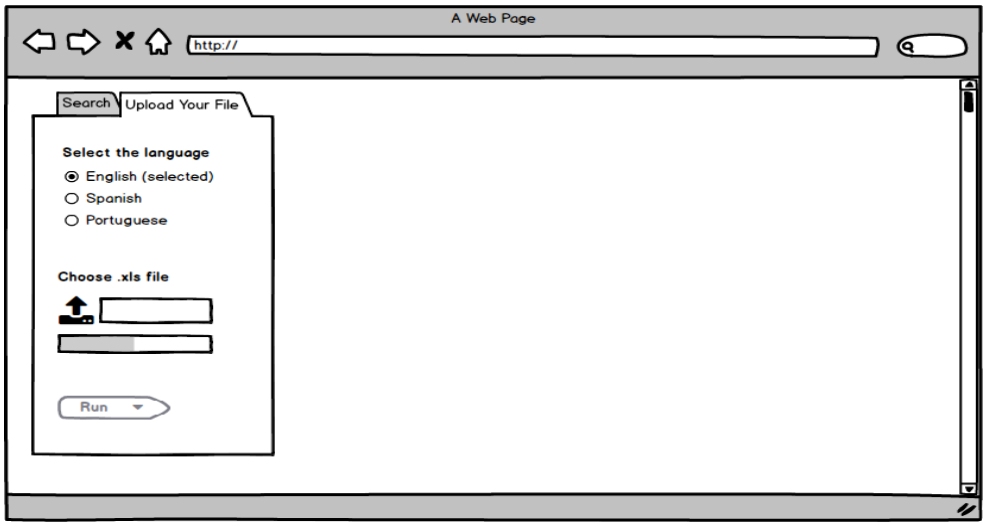
Figura 30. Mockup de interfaz del usuario del cuadro de mando, pestaña "*Serach"*.



Fuente: Elaboración propia.

La segunda pestaña es la que posibilita la subida de ficheros descargados anteriormente por la herramienta (Figura 31). El cuadro de mando de la pestaña "*Upload Your File"*, permitiría elegir el fichero de datos a subir, así como indicar su idioma correspondiente. Es necesario indicar que en esta temprana fase de diseño la herramienta no realiza detección automática de idioma en los textos a analizar, y es necesario definir el idioma contenido en el fichero para que la herramienta reconozca las acentuaciones y otras características específicas de cada lenguaje. De este modo se consigue amplificar la calidad de los informes y sobre todo del Análisis de Sentimientos que será aplicado.

Figura 31. Mockup de interfaz del usuario del cuadro de mando, pestaña *"Upload Your File".*

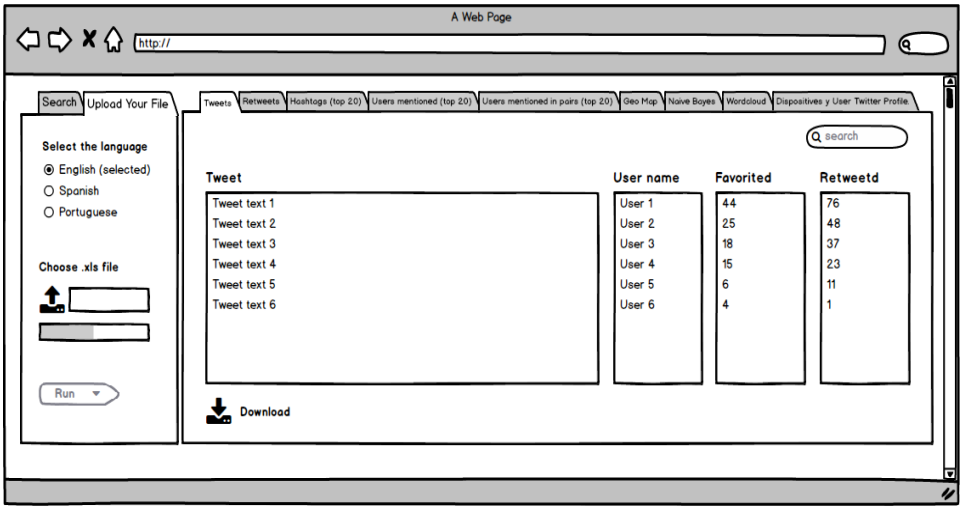


Fuente: Elaboración propia.

Como se mencionó anteriormente, situadas a la derecha de la pantalla están ubicadas las pestañas de informes. En total eran nueve las pestañas, en esta fase de conceptualización, que proporcionarían a los analistas información sobre distintas características encontradas en los mensajes:

Tweets: Esta pestaña mostrará el mensaje de texto contenido en los tuits descargados, el número de veces que cada mensaje fue marcado como favorito, el número de veces que cada mensaje fue replicado en la red y el nombre de la cuenta del usuario que generó el mensaje original. Además, posibilita ordenar los tuits según cada una de estas características (Figura 32). Finalmente, también ofrece opción de buscar palabras en los mensajes y de descargar los datos exclusivos de esta pestaña.

Figura 32. Mockup de interfaz del usuario de la pestaña *"Tweets"*.

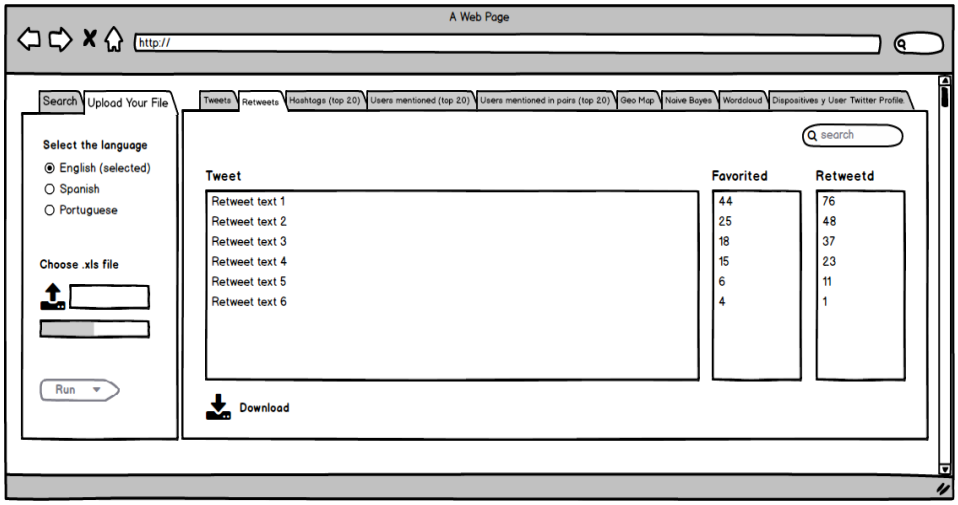


Fuente: Elaboración propia.

Retweets: Esta pestaña mostrará el mensaje de texto contenido en los retuits descargados, el número de veces que cada mensaje fue marcado como favorito, el número de veces que cada mensaje fue replicado y el nombre de la cuenta del usuario que generó este mensaje. Además, posibilita ordenar los tuits según cada una de estas características (Figura

33). Finalmente, también ofrece opción de buscar palabras en los mensajes y de descarga de los datos exclusivos de esta pestaña.

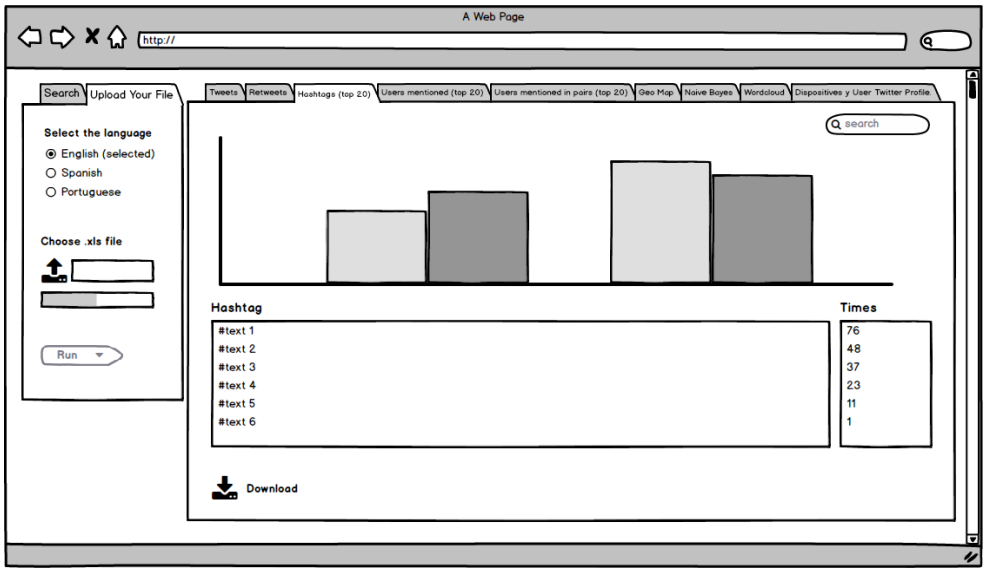
Figura 33. Mockup de interfaz del usuario de la pestaña *"Retweets"*.



Fuente: Elaboración propia.

***Hashtags (top 20)*:** Esta pestaña, por un lado, mostrará los 20 hashtags más frecuentes en los mensajes y los ordenará en formato de ranking (de mayor a menor en función de su frecuencia). Por otro lado, construirá una gráfica en base al número de veces que cada etiqueta aparece en los mensajes (Figura 34). Además, es posible interactuar con la gráfica para saber cuál etiqueta (*hashtag*) pertenece a cada barra de valor. Finalmente, también ofrecerá opción de buscar palabras en los mensajes y descargar los datos exclusivos de esta pestaña.

Figura 34. Mockup de interfaz del usuario de la pestaña *"Hashtags (top 20)".*

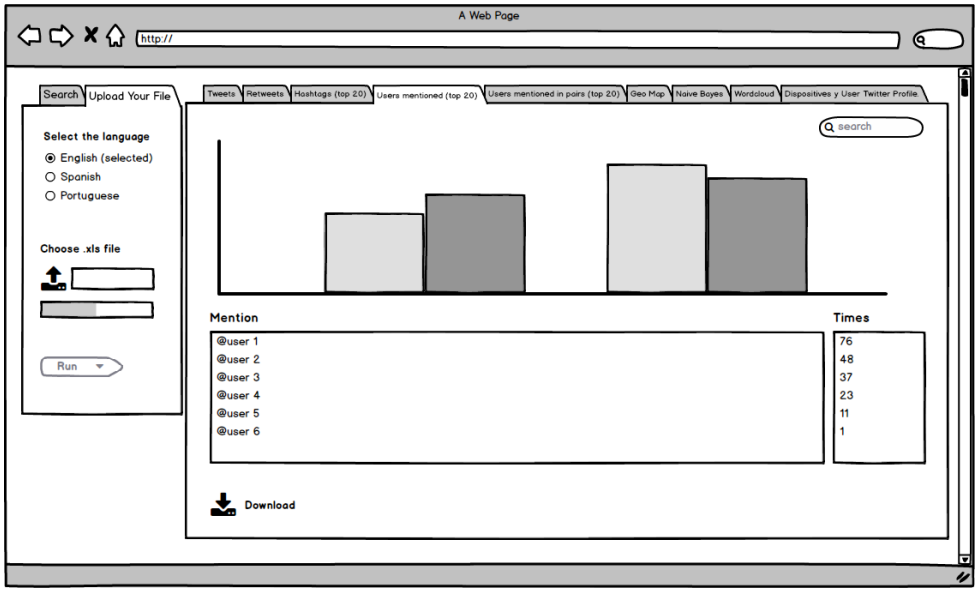


Fuente: Elaboración propia.

***Users mentioned (top 20)*:** Esta pestaña, por un lado, muestrará los usuarios más mencionados en los mensajes y los ordena en formato de ranking. Por otro lado, construirá una gráfica en base al número de veces que cada usuario aparece en los mensajes (Figura 35).

Además, el analista podrá interactuar con la gráfica para saber cuál nombre de usuario corresponde a cada barra de valor. Finalmente, también ofrecerá opción de buscar palabras en los mensajes y de descarga de los datos exclusivos de esta pestaña.

Figura 35. Mockup de interfaz del usuario de la pestaña "*User mentioned (top 20)*".

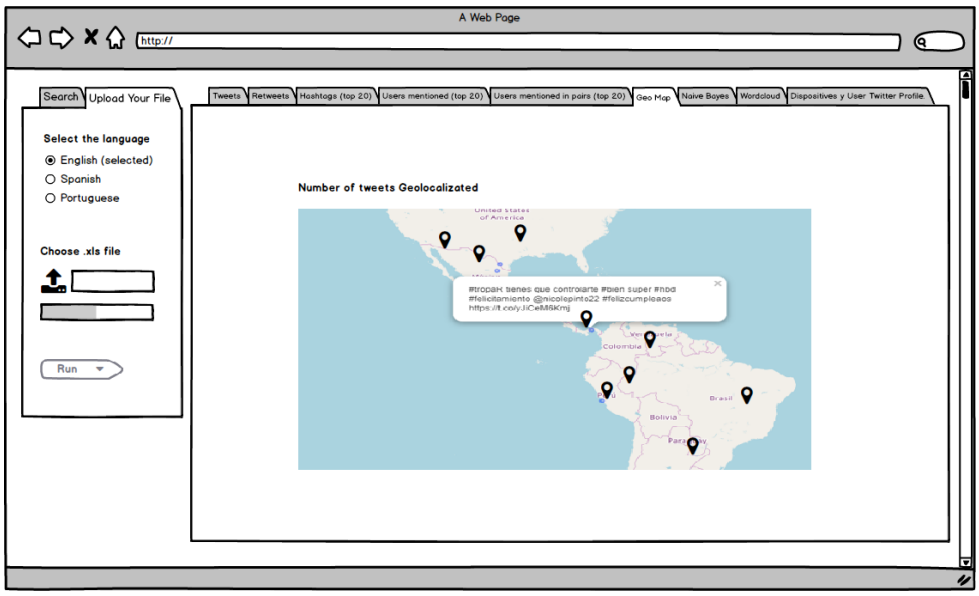


Fuente: Elaboración propia.

Geo Map: Esta pestaña mostrará la ubicación geográfica de los tuits en base a la latitud y longitud encontrada en los mensajes y los pintará en el mapa del mundo. También indicará el número de mensajes geolocalizables así como el porcentaje que estos representan frente a los mensajes no geolocalizables (Figura 36).

Es necesario indicar que la información de longitud y latitud no está siempre disponible, ya que es necesario que el usuario haya dado su consentimiento para que su dispositivo y/o aplicación compartan esta información. También es imprescindible que el dispositivo en cuestión tenga posibilidad de geolocalizarse (por ejemplo, que disponga de GPS y éste esté activo).

Figura 36. Mockup de interfaz del usuario de la pestaña *"Geo Map"*.

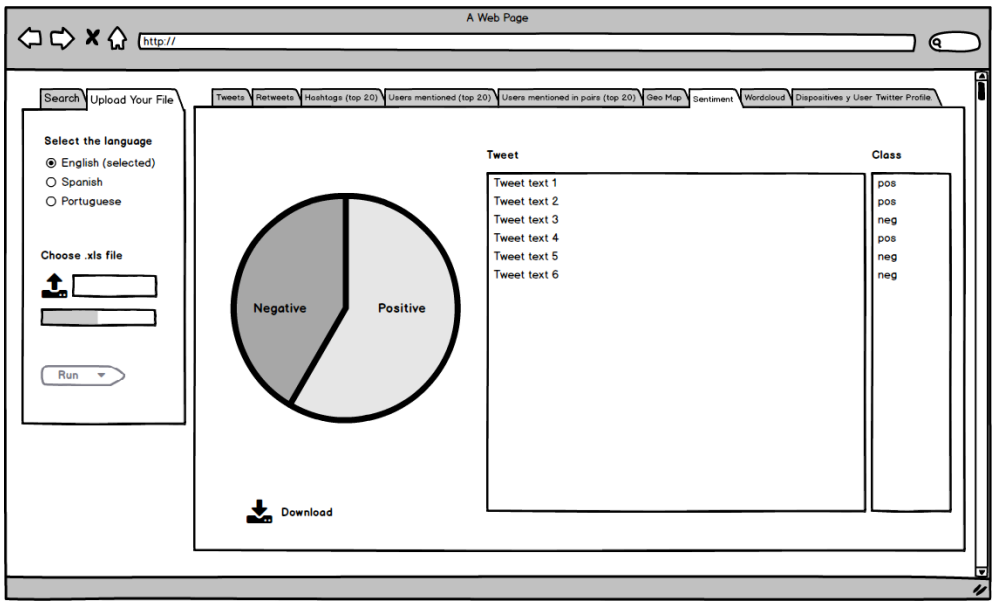


Fuente: Elaboración propia.

Sentiment: Esta pestaña se encargará de realizar los procesos relacionados con el Análisis de Sentimientos automáticos. Para esta tarea, fue elegido, para realizar el análisis de todos los idiomas, el algoritmo de aprendizaje Naïve Bayes en base a sus buenos resultados de clasificación y tiempo de análisis presentados en el Capítulo 5. De este modo, una vez realizado el análisis, se muestran el listado de los mensajes según su clasificación de sentimientos (positivo o negativo) y la gráfica correspondiente.

Como en las demás pestañas, ésta también presentará la opción de buscar palabras y de descarga de los datos exclusivos de la misma (Figura 37).

Figura 37. Mockup de interfaz del usuario de la pestaña *"Sentiment"*.

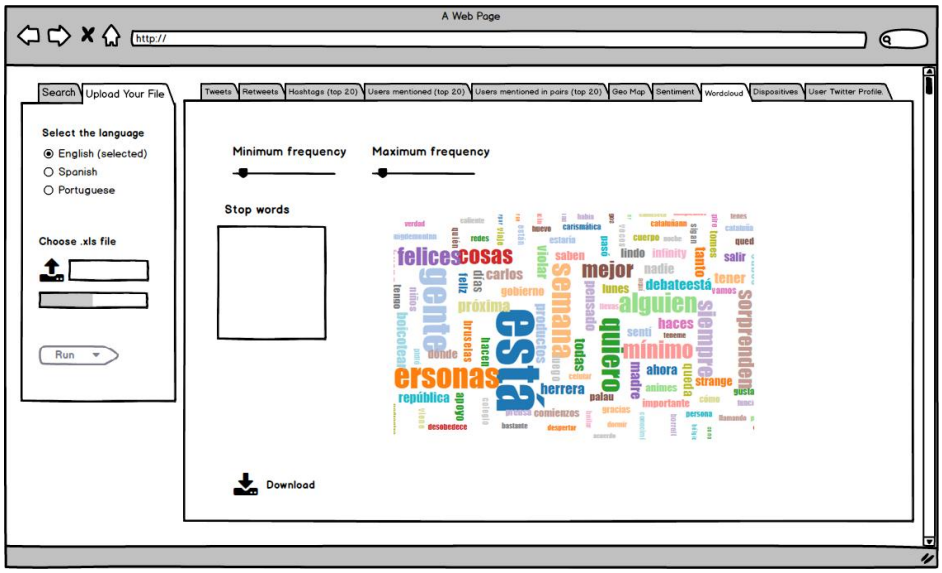


Fuente: Elaboración propia.

WordCloud: En esta pestaña el usuario podrá visualizar la nube de palabras generadas por la herramienta en base a la frecuencia de términos encontrada en los mensajes. En la nube, las palabras cobrarán tamaño según su frecuencia. Cuanto mayor sea la frecuencia de una palabra, mayor será su tamaño. Cuanto menor su frecuencia, menor será su tamaño. En este sentido, con el objetivo de generar una mejor visualización y adaptación de la nube a los usuarios analistas, se decidió insertar tres opciones de configuración. La primera, determina la frecuencia mínima que una palabra debe de tener para pertenecer a la nube. La segunda opción, da al usuario la posibilidad de indicar el número máximo de palabras que compondrán la nube. Finalmente, la tercera opción ofrecerá al analista un cuadro para la inserción específica de palabras vacías (*stop words)*, que no quiera que figuren en la nube o que no sean relevantes al objeto de estudio.

Por último, esta parte de la interfaz también ofrecerá las opciones de buscar palabras y de descargar los datos exclusivos de la misma (Figura 38).

Figura 38. Mockup de interfaz del usuario de la pestaña "*Wordcloud*".

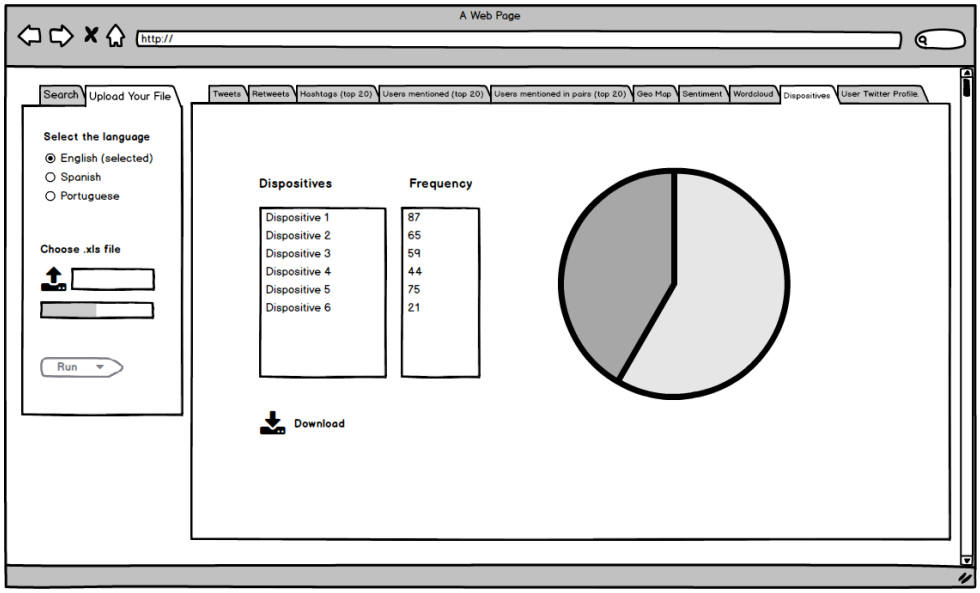


Fuente: Elaboración propia.

Dispositives: En la pestaña llamada "*dispositives*" (dispositivos), la herramienta analizará los datos y generar una gráfica que posibilite al analista, visualizar el porcentaje de los sistemas operativos, dispositivos apps más utilizados en la publicación de los mensajes.

Esta pestaña también ofrecerá la posibilidad de ordenar los dispositivos según su frecuencia y, por último, también estarán disponibles las opciones de buscar palabras y de descargar los datos exclusivos de la misma (Figura 39).

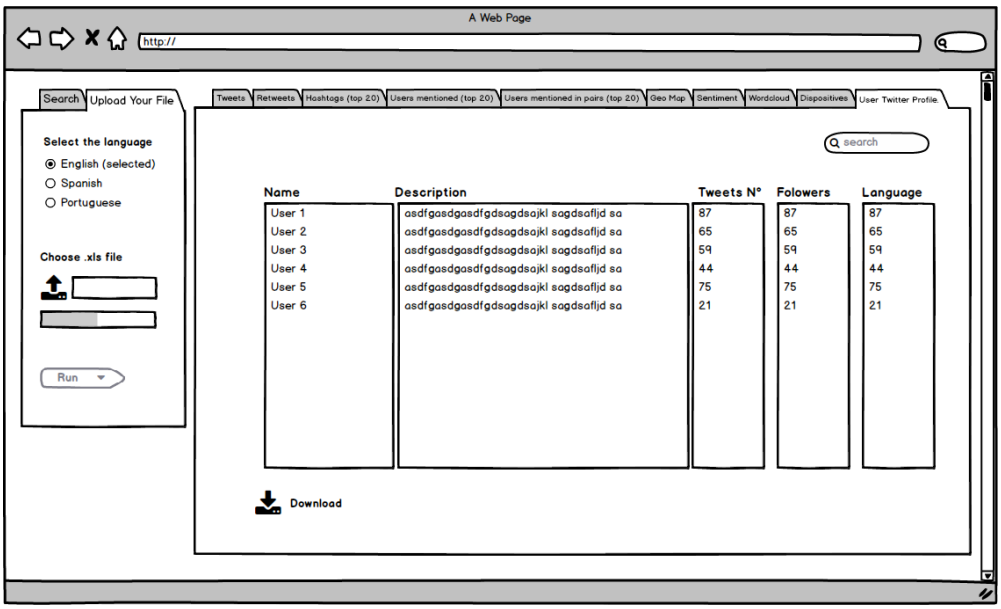
Figura 39. Mockup de interfaz del usuario de la pestaña *"Dispositives".*



Fuente: Elaboración propia.

User Twitter Profile: Esta pestaña tendrá como objetivo principal, revelar la información disponible en los perfiles de los usuarios, que publicaron los menajes objeto del estudio. De este modo, se muestran 5 columnas: "*Screen Name"* (nombre del usuario), "*Description"* (descripción que el usuario tiene en su perfil), "*Tweets Count"* (número de tuits que ha publicado en la red), "*Followers" (*número de seguidores que posee) y, por último, "*Language"* (que indica el idioma que habla el usuario autor del mensaje, o que suele utilizar en sus publicaciones). Además, la herramienta ofrece, la opción de ordenar a los usuarios según cada uno de esos parámetros, con el fin de facilitar una mejor visualización y comprensión de los datos. Finalmente, ésta pestaña, también ofrece la opción de buscar palabras y de descargar los datos exclusivos de la misma (Figura 40).

Figura 40. Mockup de interfaz del usuario de la pestaña *"User Twitter Profile"*.



Fuente: Elaboración propia.

6.5. IDENTIDAD VISUAL

El proceso de identidad visual en este caso se ocupa en definir el nombre y el logotipo del sistema. Con relación al nombre, se buscó una palabra que integrara los conceptos de gestión y manejo del lenguaje y de la información con el objetivo de aportar conocimiento. En este sentido, después de realizadas investigaciones previas, se eligió el nombre de origen griego, LOGOS (*λόγος)*. Son innumerables las definiciones para esta palabra. Sin embargo, todas convergen en un mismo sentido. Su significado según el Diccionario filosófico marxista (1946:179-180) significa, pensamiento, discurso y razón. Según el Diccionario filosófico abreviado (1959:301) significa pensamiento, concepto, palabra y razón. El Diccionario filosófico (1965:282) la define como, palabra, pensamiento e intelección.

Según el diccionario de filosofía (1984:260) significa, palabra, pensamiento, razón y ley. Finalmente, para el autor Soler (2017), LOGOS significa la palabra en cuanto meditada, reflexionada o razonada, es decir: "razonamiento", "argumentación", "habla" o "discurso" que también puede ser entendida como: "inteligencia", "pensamiento" y "sentido"). Para él, su raíz estaría probablemente en el indoeuropeo *legō*, que tiene el sentido de "recoger junto", imponiendo a ese recoger un "criterio", por lo tanto, derivaría, tanto en el griego como en el latín, en el sentido de recoger, seleccionar, elegir.

Una vez definido el nombre, se procedió a la confección del logotipo del sistema. Esta etapa requirió del desarrollo y aplicación de habilidades creativas y de producción gráfica adquiridas durante el proceso.

La parte gráfica del logotipo se compone del nombre, función de la herramienta y tres formas geométricas en tres diferentes colores interconectadas entre sí (Figura 41). Las formas, representan a los datos, los colores indican su variedad y la conexión que se establece entre ellos, representa a los informes de apoyo a decisiones que genera la herramienta por medio de la minería de datos, PLN e inteligencia artificial.

Finalmente, se creó un video breve de presentación de la conceptualización de LOGOS, donde se muestra su interfaz y algunas de sus principales características que se puede ver desde el enlace https://www.youtube.com/watch?v=LVYDuE8m8R4.

Figura 41. Logotipo del sistema LOGOS.

Fuente: Elaboración propia.

6.6. IMPLEMENTACIÓN DEL PROTOTIPO

Con el advenimiento y la consolidación de las plataformas de software libre (*open source)* que funcionan bajo las licencias GPL (*GNU General Public License)*, el número de alternativas de softwares estadísticos viene aumentando exponencialmente. En este sentido el software R[19] ha ganado extensa popularidad tanto en la comunidad científica, cuanto en el sector público/privado (Muenchen, 2014). El mismo autor, demuestra los avances

[19] https://www.r-project.org/about.html (acceso el 07/08/2019).

del software y su versatilidad por medio de los numerosos artículos científicos que emplean el lenguaje de programación en R. El uso de las plataformas del tipo abiertas (*open source),* libres y seguras, en general son atractivas por dos motivos principales. En primer lugar, permiten conocer al detalle los procedimientos realizados por el software. El hecho de que sean de código abierto permite a los usuarios con alguna experiencia en su lenguaje de programación, desarrollaren sus propias aplicaciones. En R, por ejemplo, para conocer el método utilizado por determinada función, basta con digitar el nombre de ésta para obtener su código fuente. Otro punto muy favorable a las plataformas abiertas se relaciona con la reducción de la burocracia establecida para la adquisición del software por los usuarios sean públicos o privados. De modo general, para adquirir sistemas que requieran licencias, las PyMEs deben iniciar un proceso de solicitudes que involucran diversos departamentos, entre ellos el de compras. Esto puede tornar el proceso moroso y retardar el acceso al software en cuestión. En el caso de los softwares libres, el usuario puede descargar el sistema sin ninguna pega en cualquier terminal

6.6.1. LA PLATAFORMA R

La plataforma R es un software del tipo libre y disponible a todos los sistemas operativos. Su descarga puede ser realizada a través del enlace oficinal del proyecto R, www.r-project.org. Se trata de una plataforma multi objetiva que permite el análisis y la visualización de datos a niveles estadísticos de diferentes complexidades.

Este software, también pode ser utilizado como plataforma de desarrollo de aplicaciones específicas e inéditas o híbridas mezclando códigos inéditos con aplicaciones ya existentes en ostro softwares. R cuenta con un módulo básico de instalación que promueve un ambiente de desarrollo base, una consola para la realización de análisis que componen un interfaz gráfico para visualización de los resultados. Su módulo básico puede complementado por paquetes (*packages)* adicionales para realización de tareas específicas en áreas más generales como las matemáticas o más específicas como la ecología, el financiero, el procesamiento geoespacial entre otros. Este software credo en 1995, hoy en día es responsabilidad gerencial de un grupo de individuos denominado "R-core" y es mantenido por la fundación-R a cuál está soportada por diversas empresas de distintos segmentos. Según Muenchen (2014) R es el software más utilizado en competiciones de análisis de datos a nivel mundial, como por ejemplo el Kaggle[20].

6.6.2. LA PROGRAMACIÓN EN R

En R es posible utilizar diversas técnicas de programación. Esta característica se da gracias a que su código fuente se aprovecha de rutinas establecidas en Fortran, C y en su propio lenguaje R. El Fortran es un lenguaje de programación informático utilizado fundamentalmente para

[20] www.kaggle.com (acceso el 07/08/2019).

las matemáticas y tareas de cálculos científicos. Desarrollada en 1954 por John Backus (1924 -2007) y su equipo, tuvo mucho éxito por ser el lenguaje de programación pionero para dichas tareas (Rodríguez Sala, Santamaría Sala, Rabasa Dolado, & Martínez Bonastre, 2003). Ya el C, se trata de un lenguaje que gracias a su construcción se permite crear conjuntos de instrucciones para que el ordenador as realice.

Desarrollada en 1972 por Dennis Ritchie de los Bell Lab para uso en el sistema operacional Unix. En su momento fue ampliamente aceptada por ofrecer a los programadores el máximo en control y eficiencia. Su objetivo principal era facilitar la creación de programas extensos con errores reducidos por medio de los paradigmas de la programación algorítmica y procedimental. Sin embargo, por tratarse de un lenguaje poderoso, robusto y flexible, conocer todos sus detalles y "trucos", requiere un estudio criterioso y profundo, aunque también por trabajar con sintaxis simplificadas, también puede ser manejada por programadores iniciantes. Por estos motivos, C hoy en día es uno de los lenguajes de programación preferido en el desarrollo de sistemas y softwares de base, pero también puede ser usado para construir programas más complexos.

En el ámbito académico, este lenguaje es ampliamente utilizado en investigaciones científicas y como instrumento de aprendizaje para la construcción de algoritmos (Gardener, 2012; Lafaye de Micheaux, Drouilhet, & Liquet, 2013; Schildt, 1997; Teetor, 2011).

6.6.3. EL PAQUETE SHINY PARA R

Según comentado anteriormente, el software, en esta primera iteración, fue desarrollado utilizado el paquete de funciones o también llamado "*framework"* de aplicaciones Shiny (*R package version 0.12.2*), que permite acceder a todo el contenido del software R (version 3.2.2) vía interfaz web.

El Shiny combina los recursos computacionales del software R con la interactividad de la Web moderna permitiendo al usuario poder interactuar de forma rápida y sencilla con los análisis de los datos. Por este motivo, este recurso es una excelente opción para el uso de la computación a nivel científico.

Para la creación de la interfaz del programa, fueron utilizadas algunas funciones principales como por ejemplo la "navbarPage", usada para crear barras superiores y laterales que dividen el aplicativo en diferentes secciones. También fue utilizada la función "mainPanel", para la creación del cuadro de mando general con el objetivo de ofrecer una interfaz más organizada, agradable e intuitiva para el usuario.

Finalmente, para la visualización de los informes, en especial los que se basan en las frecuencias, aplicamos los recursos de graficas de barra, *wordcloud* (nube de palabras), tablas y listas.

6.6.4. ARQUITECTURA DE LA HERRAMIENTA

Para establecer la arquitectura de la herramienta, en primer lugar, se estableció un flujo de trabajo específico que posibilitó ordenar la ejecución de los procedimientos informáticos necesarios para la extracción, preprocesamiento y visualización de los datos. A seguir detallaremos cada uno de ellos.

Extracción: Para el caso específico de Twitter, el proceso de extracción de las informaciones se hace por medio de una API con diversas opciones. En el contexto de desarrollo Web, una API se trata de un conjunto de tipos de requisiciones HTTP juntamente con sus respectivas definiciones de respuesta. En aplicaciones de redes sociales online es común encontrarnos con API´s que listan los amigos de un usuario, sus objetos, sus comunidades y etc. Otra ventaja de las API´s está en que ofrecen los dados en formatos estructurados como XML y JSON.

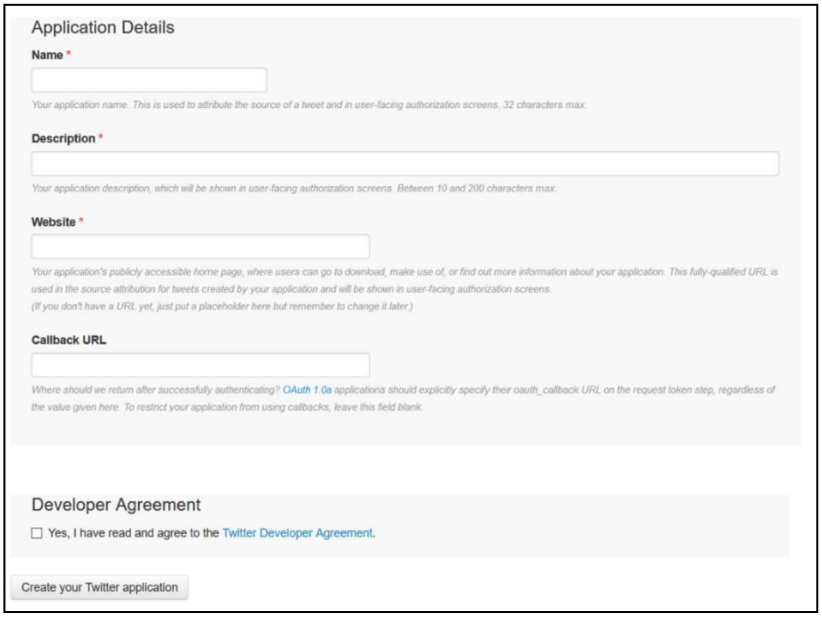
Twitter por su parte, ofrece por medio de su API la posibilidad de colectar 5000 tuits a través de una única requisición. Esta misma recolecta de información a través de otros sitios Web convencionales, necesita de centenares de requisitos. Son varias las redes que adoptan la utilización de API´s, incluyendo Twitter, Flickr, YouTube, Google Mapas, Yahoo Mapas y etc. Con tantas API´s existentes, se hace común ver aplicaciones que se utilizan de dos o más API´s para crear un nuevo servicio, a esto se le llama Mashup Pipes. En el caso de Twitter, este trabaja con 2 tipos de API´s, "*Rest y Streaming".* La API REST permite al programador de lectura y escrita de dados del Twitter, la extracción de todos (o los que se permita) los datos

almacenados en la red. Ya la API *Streaming* dispone a los dados de Twitter en tiempo real. En el ámbito de este trabajo, la API *REST* fue la más apropiada una vez que necesitamos el acceso también a tuits publicados en el pasado.

Configuración de la API de Twitter: Esta es la primera parte que se necesita configurar correctamente para poder utilizar los servicios que ofrece Twitter a través de su API. Para ello es necesario obtener los llamados Tokens, estos Tokens son códigos que identifican a un usuario de Twitter y sus aplicaciones conectadas con su API.

Primeramente, debemos acudir a https://apps.twitter.com/ y crear una nueva aplicación. En la Figura 42 se observan algunos de los campos que se solicitan para poder crearla.

Figura 42. Detalles para la creación de la aplicación de Twitter.



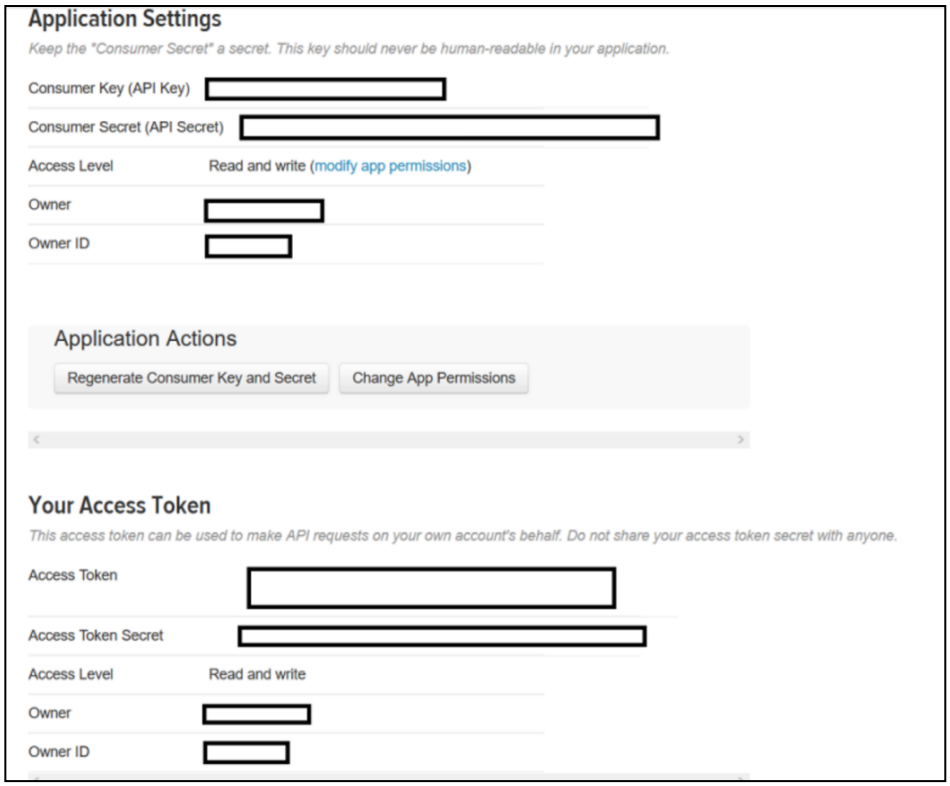
Fuente: https://apps.twitter.com/.

Con la aplicación ya creada tendremos que ir a la pestaña "*Keys and Acces Tokens*" en la que tendremos una serie de datos como se pueden ver en la Figura 43.

De estos los que nos interesan son: "*Consumer Key (API Key), Consumer Secret (API Secret), Access Token y Access Token Secret*"

Una vez tengamos estos códigos, se nos permite acceder a la API de Twitter para enviar y recibir datos.

Figura 43. Configuración de la aplicación de Twitter.



Fuente: https://apps.twitter.com/.

En R nos encontramos un paquete ya creado llamado twitteR[21] que nos permite realizar diferentes operaciones mediante funciones prediseñadas, para poder realizar las operaciones es necesario utilizar los códigos que en la sección anterior se han configurado y obtener nuestra autorización. Dicha autorización se obtiene utilizando la siguiente función, junto con los correspondientes tokens:

- *setup_twitter_oauth (consumer_key=consumer_key*
- *consumer_secret=consumer_secret*
- *access_token=access_token*
- *access_secret=access_secret*

Una vez que tenemos nuestra autorización, podemos comenzar a utilizar el resto de las funciones que podemos encontrarlas en la documentación de la librería.

Para nuestro problema en cuestión utilizamos la función "*searchTwitter"*, esta función permite recuperar los últimos Tuits publicados. Para emplearla es necesario pasarle una serie de parámetros:

- El término que se quiere buscar, por ejemplo *#nike.*
- La cantidad de Tuits que queremos recuperar como máximo (entre 1 y 3200).
- El idioma que queremos que tengan los Tuits recuperados.

[21] Jeff Gentry (2015). twitteR: R Based Twitter Client. R package version 1.1.9. https://CRAN.Rproject.org/package=twitteRproject.org/package=twitteR (acceso el 07/08/2019).

Con esta función se recuperan los Tuits bajo los parámetros deseados. La Tabla 31 indica la información que cada tuit colectado contiene. Una vez descargados los datos se procede a la siguiente sección que se encarga del tratamiento de estos datos.

Tabla 31. Información colectada de cada Tuit.

"text"	Indica el texto del mensaje escrito por el usuario
"favorited":	Indica si el tuit ha sido marcado como favorito
"favoriteCount"	Indica el número de veces que el tuit ha sido marcado como favorito
"replyToSN"	Indica el nombre de usuario del usuario que está en el campo "reply to"
"created"	Indica la fecha y hora de la creación del mensaje
"truncated"	Indica si este estatus fue truncado
"id"	Indica el número único de identificación del mensaje
"replyToUID"	Indica el ID de usuario del usuario que está en el campo "reply to"
"statusSource"	Indica el sistema operativo o aplicación utilizada para publicar el mensaje
"screenName"	Indica el nombre de la cuenta en Twitter utilizada para publicar el mensaje
"retuitCount"	Indica el número de veces que el mensaje fue replicado por otros usuarios
"isRetuit"	Indica si el mensaje se trata de un mensaje replicado por otros usuarios
"retuited"	Indica si el mensaje fue replicado por otros usuarios
"longitude"	Indica los datos geográficos de longitud cuando disponibles
"latitude"	Indica los datos geográficos de latitud cuando disponibles

Fuente: Elaboración propia.

Tratamiento de datos: Para atender el planteamiento de nuestro sistema, fue necesario realizar diferentes procesamientos para obtener la información deseada en cada sección.

El primer paso ha sido discriminar los Tuits y Retuits con el objetivo de mostrarlos en secciones separadas, gracias a que Twitter identifica los Retuits como *'RT'* se discriminaron aquellos con esa etiqueta para obtener los Tuits originales y Retuits.

Seguidamente para la sección de Hashtags, fue necesario obtener todos los hashtags que aparecen en todos los mensajes y contarlos, de modo que estipulamos buscar cualquier texto que comience con *'#'*, además de contar las veces que aparece en todo nuestro *dataset*.

En cuanto a los usuarios mencionados se procedió de manera similar a los de los hashtags, pero esta vez discriminando por '@' que es como Twitter identifica a los usuarios dentro de su red.

Para realizar los usuarios que aparecen de manera conjunta fue necesario no solo discriminar y contar las veces que aparece un usuario en nuestro *dataset*, sino que, se debía realizar este conteo en base a grupos indeterminados de usuarios en los que mínimamente aparecieran dos usuarios.

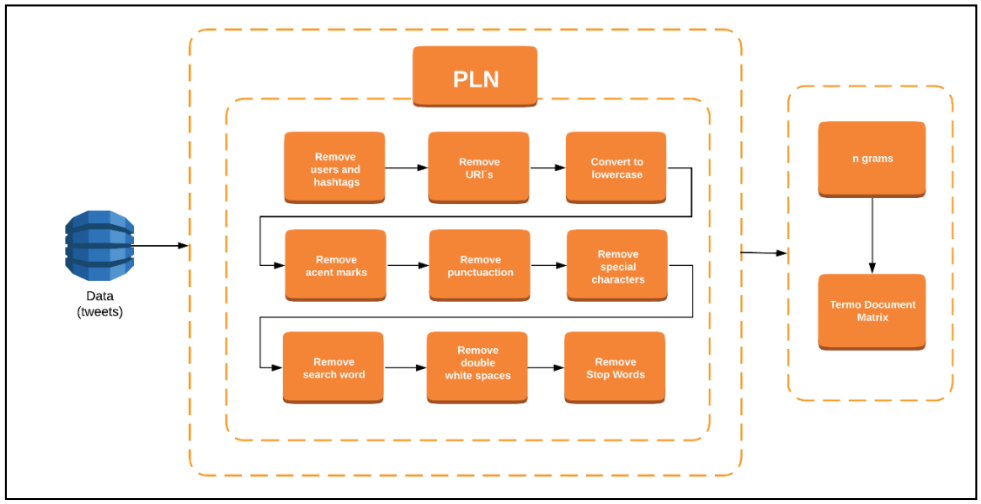
Para realizar el mapa de geolocalización, se utilizaron las informaciones de latitud y longitud disponible por la API de Twitter. En adición se determinó que se mostrara el texto del mensaje correspondiente.

Para la pestaña que analiza el Análisis de Sentimientos, el proceso es algo más complejo ya que, es necesario realizar tratamientos relativos al procesamiento de lenguaje natural (PLN) de manera más exhaustiva con el fin de simplificar o "limpiar" el texto de los Tuits. Los procesos de PLN son de extremada importancia, ya que eliminan posibles "ruidos" que afectan directamente en la comprensión del texto por parte del algoritmo, además de retardar el tiempo de análisis. En este sentido, se aplica un limpiado de texto para eliminar datos como: url´s, aquellos prefijos como 'RT' o 'via',

eliminar menciones, caracteres ilegibles y etc. (Figura 44). Seguidamente se entrenan diferentes modelos para cada idioma (español, portugués e inglés) que posteriormente se encargaran de predecir los nuevos tuits.

La pestaña *"Wordcloud" (*nubes de palabras) nos permite conocer sobre qué se está hablando en nuestro *dataset* realizando un conteo de palabras según una determinada frecuencia, de modo que de un vistazo rápido podemos saber sobre qué se está hablando.

Figura 44. Flujo de preprocesamiento de datos de LOGOS.



Fuente: Elaboración propia.

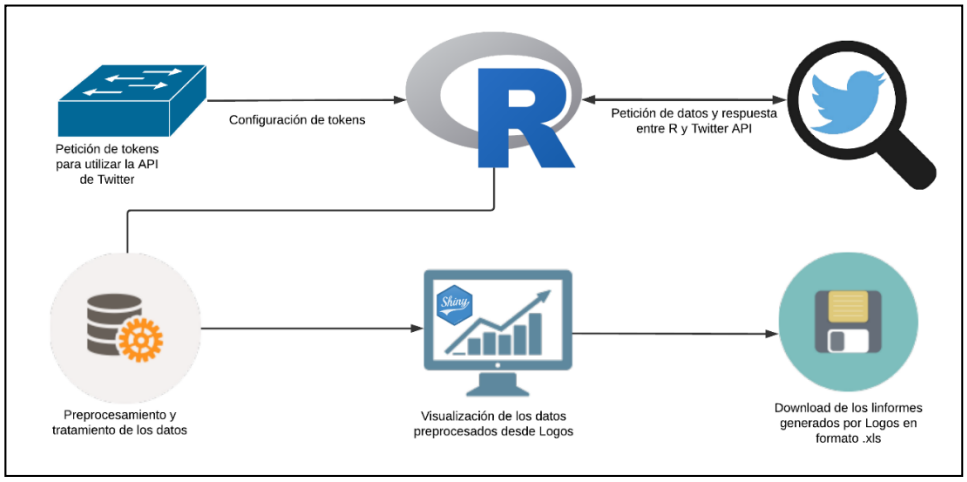
En cuanto a la pestaña de dispositivos, ésta se realiza en base la columna *"statusSource"* para obtener la fuente desde la que se generó cada Tuit.

Por último, los perfiles de los usuarios de Twitter nos muestran información sobre los usuarios que se encuentran en nuestro *dataset*, nos permite saber su dirección, idioma, gustos u otros datos de interés que el usuario haya puesto disponible en su perfil.

Salida de datos: Cada una de las secciones tratadas anteriormente tienen su espacio en Shiny, accesible en cada pestaña. Estos datos son aquellos por los que el usuario ha realizado su búsqueda o bien ha subido un archivo en formato Excel previamente descargado.

Finalmente, en la Figura 45 se puede observar la arquitectura de la herramienta según el flujo de trabajo (*workflow)* que hemos detallado en este epígrafe.

Figura 45. Arquitectura de la herramienta LOGOS.



Fuente: Elaboración propia.

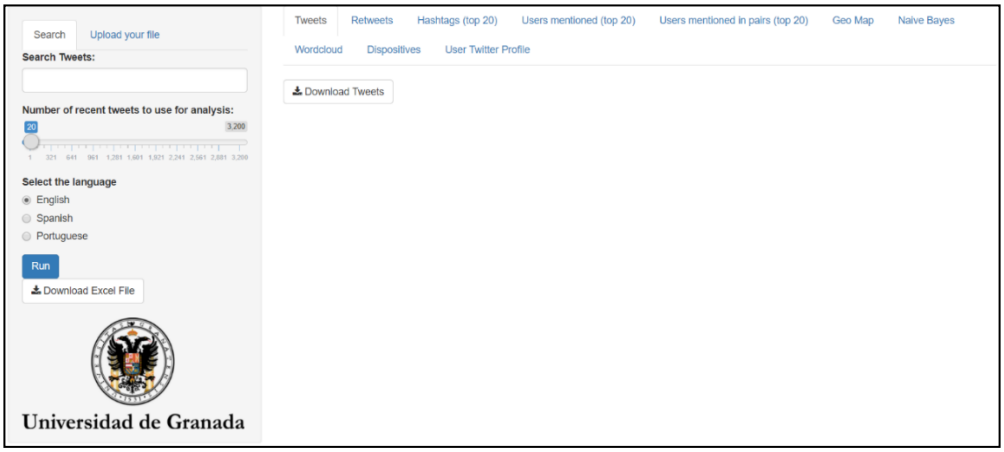
6.7. PRUEBAS DEL PROTOTIPO

Finalizada la etapa de programación, el sistema debe pasar por la etapa de prueba y posterior validación. Esta etapa se nutre de las anteriores y realiza la integración y certificación final del sistema. Para ello se llevó a cabo la puesta en marcha de la herramienta utilizando como ejemplo para la consecución de los informes, la etiqueta (#bien) en el idioma español.

A continuación, se muestra la forma de la interfaz del prototipo del sistema, así como los distintos informes generados por LOGOS.

En la (Figura 46) se puede observar la vista general del software con el cuadro de mando situado en la parte izquierda del navegador y las pestañas de informes situadas en la parte derecha.

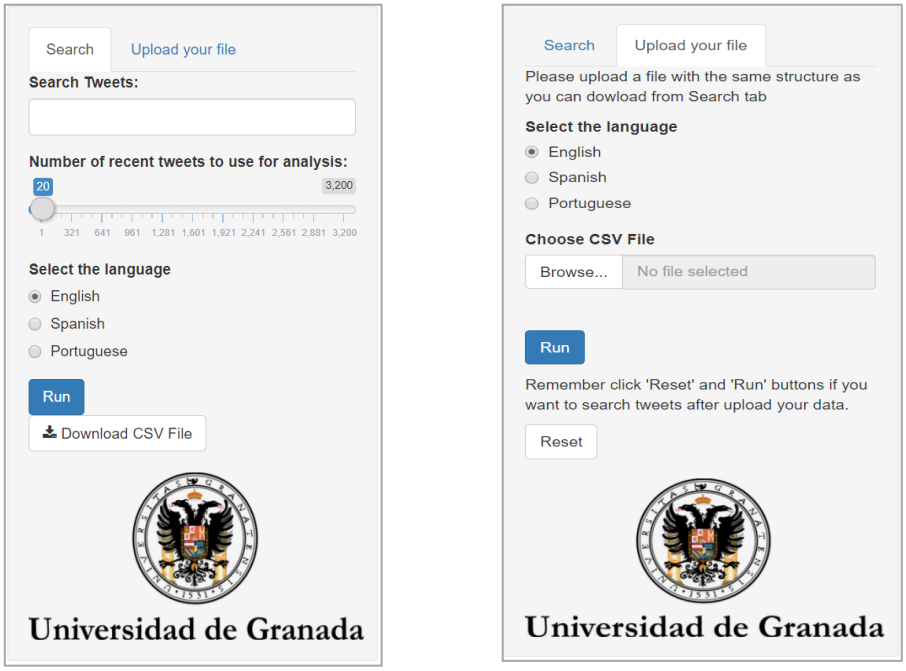
Figura 46. Interfaz del usuario de LOGOS.



Fuente: http://hipatia.ugr.es:3838/logos/.

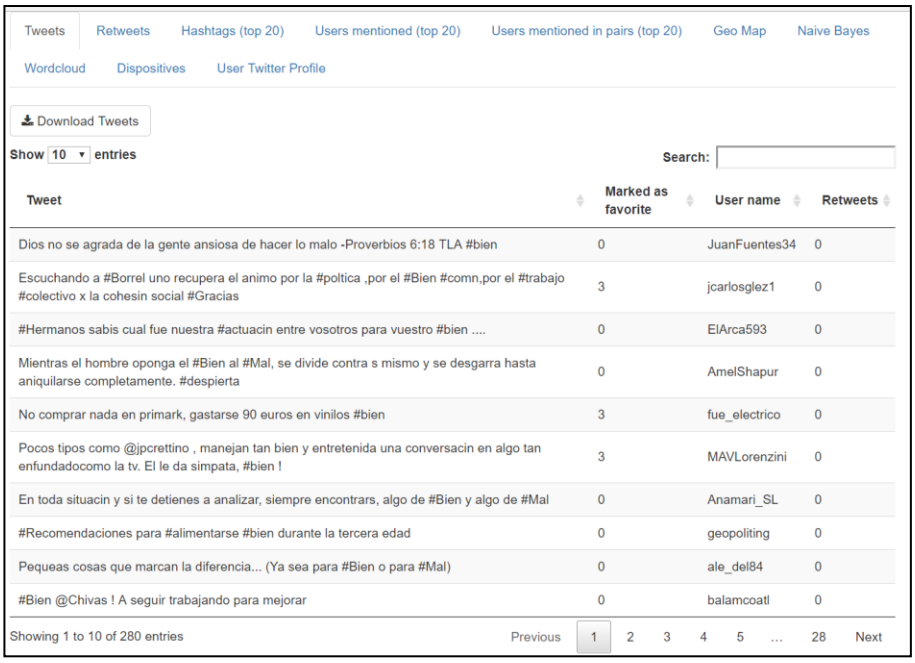
En la Figura 47 se muestra el cuadro de mando de la pestaña "*Search"* y el cuadro de mando de la pestaña "*Upload Your File".* En las Figuras 48-57 a continuación, se aprecia la forma final de las pestañas de informes, "*Tweets, Retweets, Hashtags (top20), Users mentioned (top 20) Users mentioned in pairs (top 20), Geo Map, Sentiment, Wordcloud, Dispositives y User Twitter Profile*".

Figura 47. Interfaz del usuario: cuadro de mando *"Serach" y "Upload Your File".*



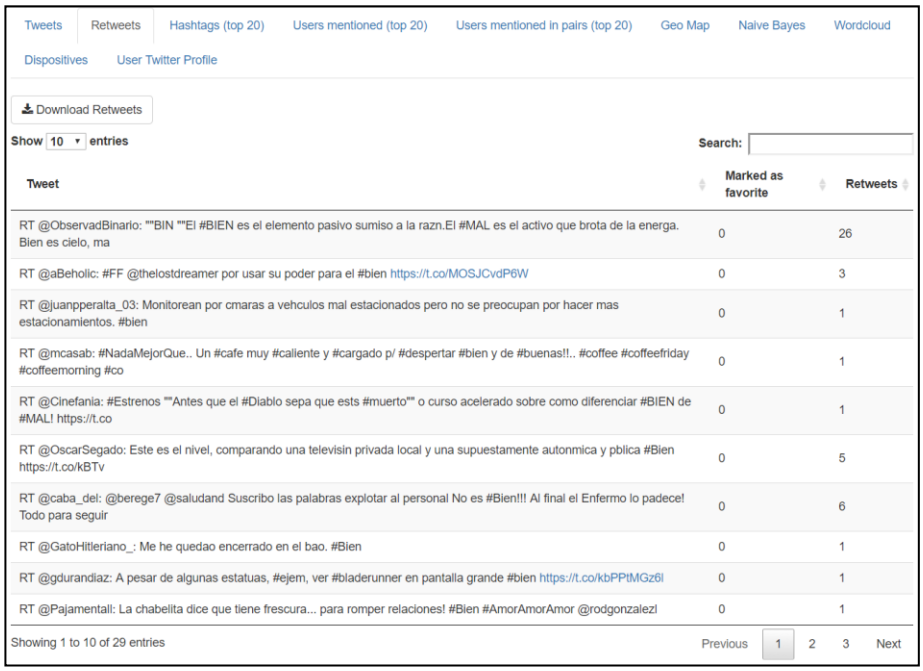
Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 48. Interfaz del usuario de la pestaña *"Tweets".*



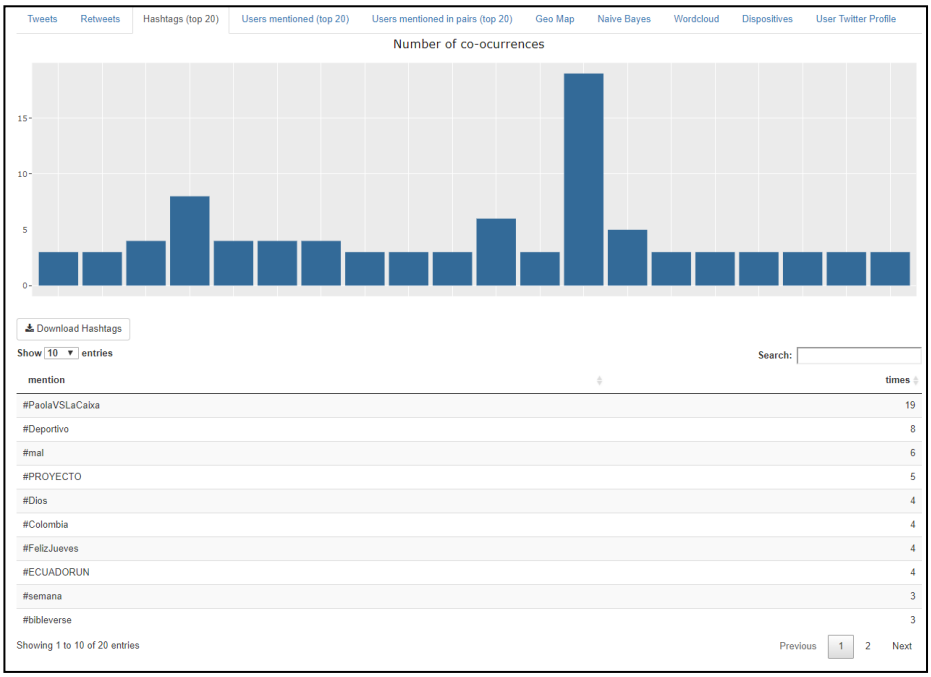
Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 49. Interfaz del usuario de la pestaña "*Retweets*".



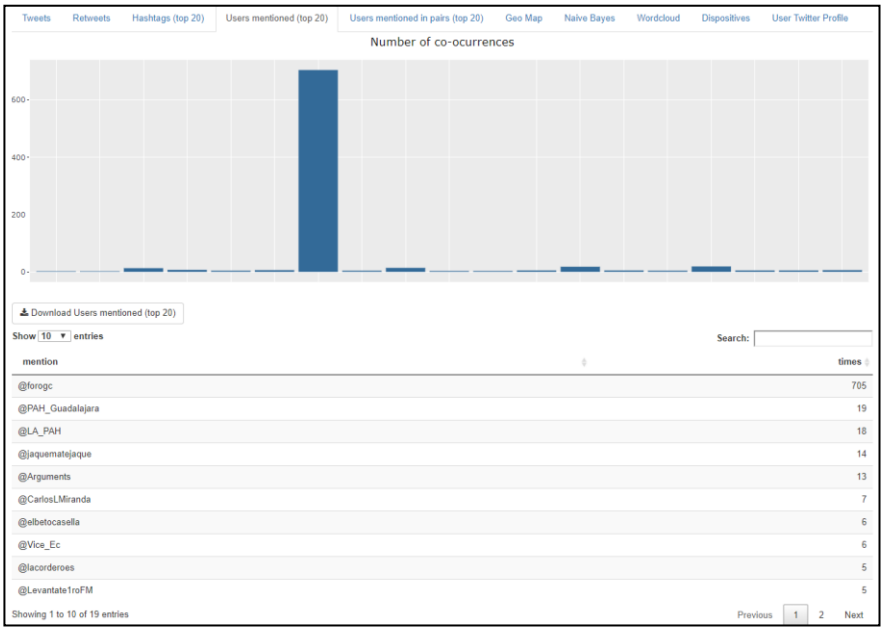
Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 50. Interfaz del usuario de la pestaña "*Hashtags (top 20)*".



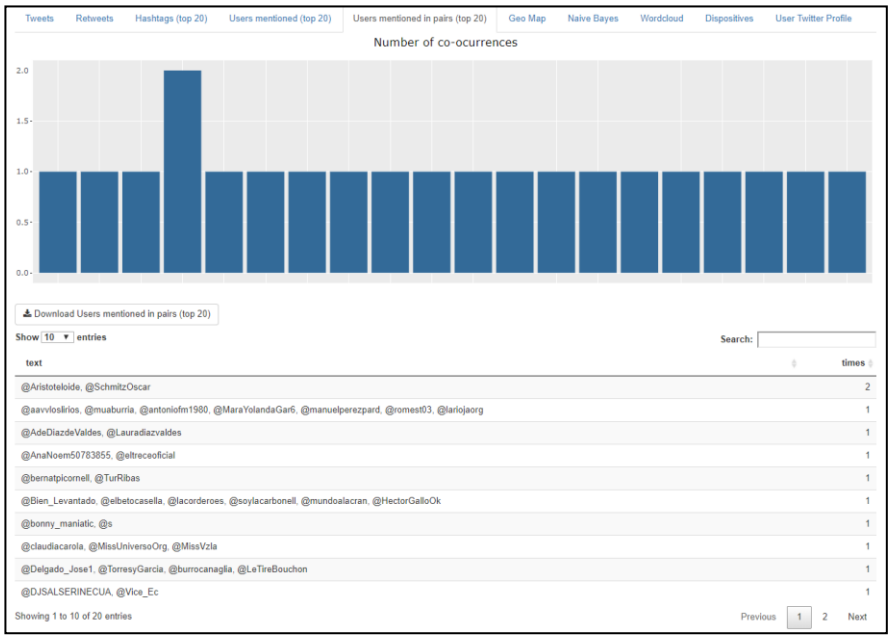
Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 51. Interfaz del usuario de la pestaña "*Users mentioned (top 20)*".



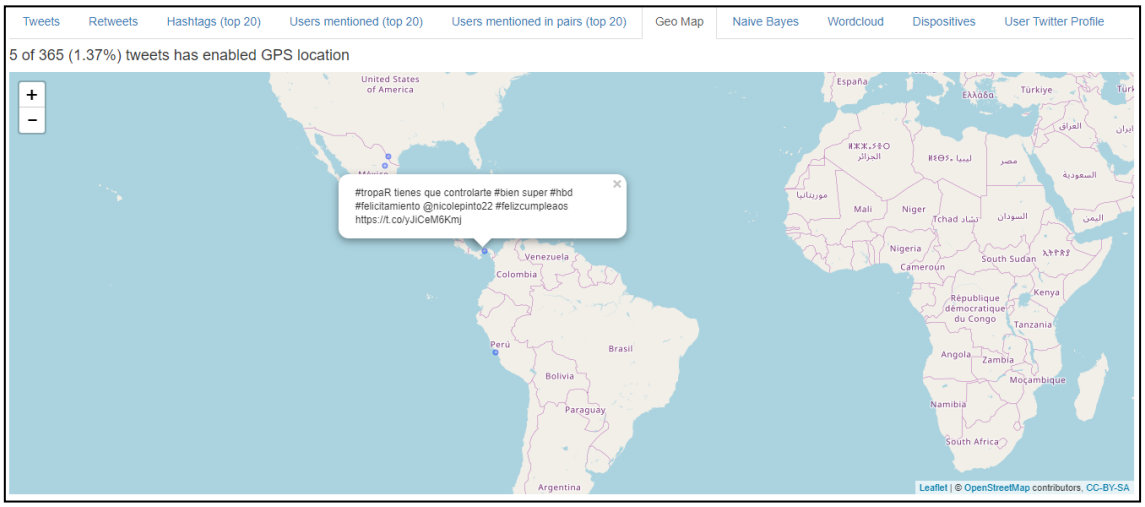
Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 52. Interfaz del usuario de la pestaña "*Users mentioned in pairs (top 20)*".



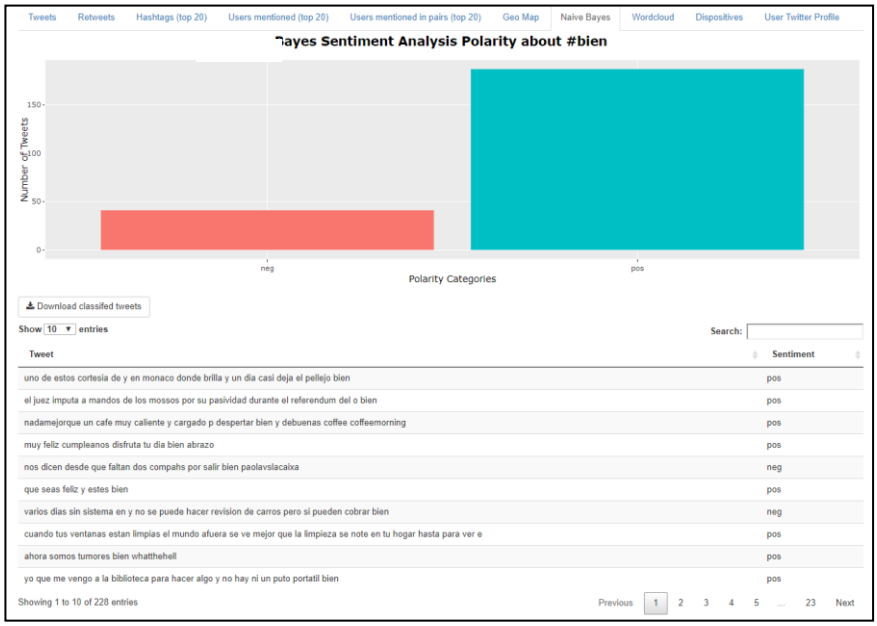
Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 53. Interfaz del usuario de la pestaña "*Geo Map*".



Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 54. Interfaz del usuario de la pestaña "*Sentiment*".

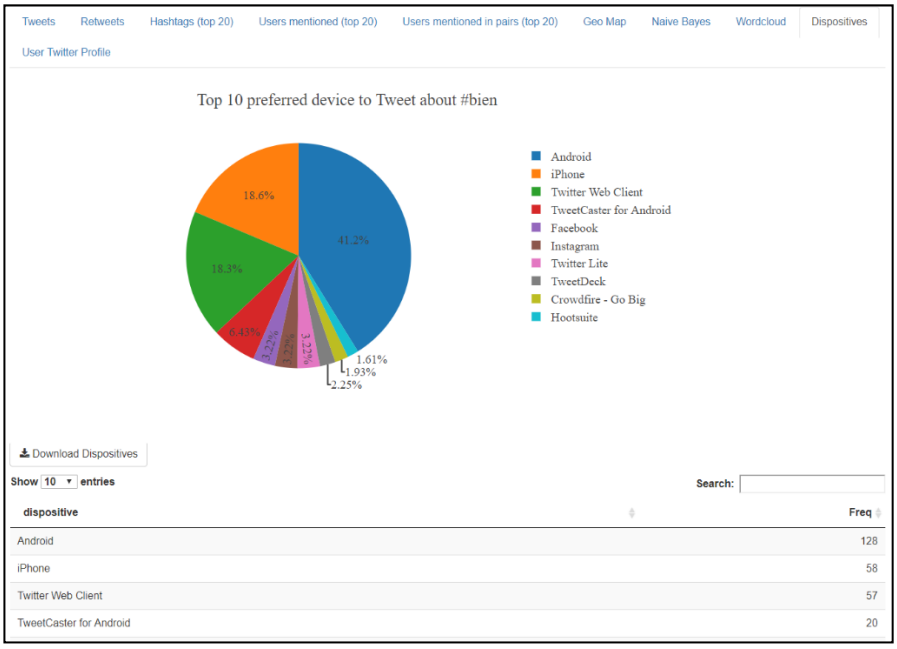


Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 55. Interfaz del usuario de la pestaña "*Wordcloud*".

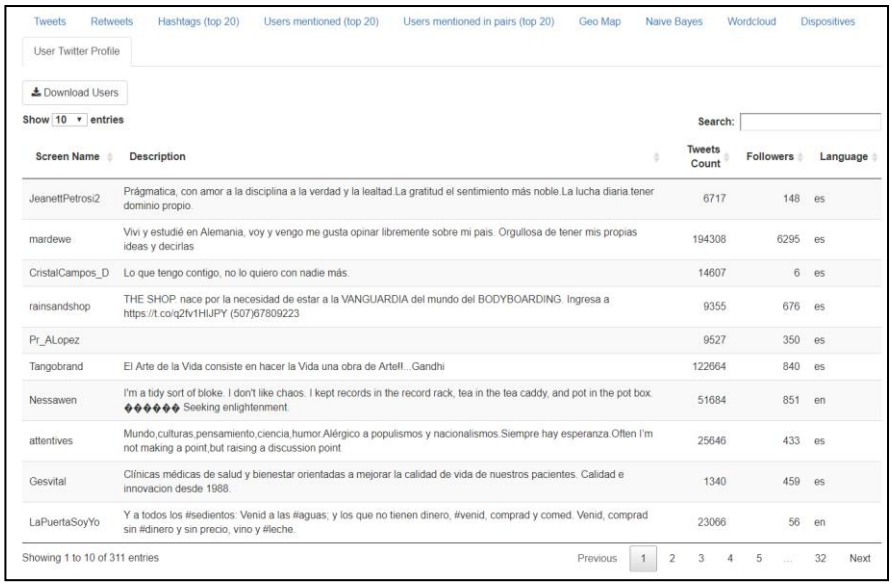
Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 56. Interfaz del usuario de la pestaña *"Dispositives"*.



Fuente: http://hipatia.ugr.es:3838/logos/.

Figura 57. Interfaz del usuario de la pestaña *"User Twitter Profile"*.



Fuente: http://hipatia.ugr.es:3838/logos/.

6.8. CONCLUSIONES, LIMITACIONES Y EXTENSIONES

Como hemos podido apreciar a lo largo de este capítulo, la versión prototipo de LOGOS ofrece a los usuarios una interfaz minimalista e intuitiva y se compone por distintas pestañas de informes que tienen como objetivo proporcionar una lectura de los datos que auxilian en la toma de decisiones de marketing en las empresas en general y en las PyMEs en particular.

Concretamente, la herramienta está compuesta inicialmente por un cuadro de mando fijo, que permite tanto la búsqueda de mensajes en base a una cuenta de usuario, #hashtag o palabra cualquiera, como también la subida (*upload)* de ficheros para que sean analizados por la aplicación.

En el primer caso, el sistema se conecta a Twitter por medio de la API "*Rest"* y extrae los mensajes disponibles. Luego, aplica diferentes clases de

preprocesamiento de datos, para que se adecuen a los diferentes informes ofrecidos por el sistema.

Además del cuadro de mando, dispone nueve diferentes pestañas. Cada pestaña ofrece un informe específico sobre los datos. Éstos son presentados en formatos de listas y gráficas que pueden ser ordenados según su importancia o frecuencia. La versión prototipo de LOGOS también permite la descarga de cada informe en el formato de ficheros que puedan ser tratados por aplicaciones de hojas de cálculo.

Las nueve pestañas de informes que componen esta versión inicial de LOGOS son: *"Tweets, Retweets, Hashtags (top20), Users mentioned (top 20), Geo Map, Sentiment, Wordcloud, Dispositives y User Twitter Profile".* Estos recursos en conjunto proporcionan a) la visualización y la descarga de los tuits y retuits en ficheros que puedan ser manejados por aplicaciones de hojas de cálculo; b) la descarga de listados, gráfica y ranking de los hashtags más frecuentes en los mensajes y de los usuarios más mencionados en los mensajes, tanto solos como por pares; c) la visualización geográfica de los mensajes, y el porcentaje de mensajes geolocalizables; d) la aplicación del Análisis de Sentimientos que clasifica los mensajes en positivos o negativos; e) la creación de nubes de palabras con distintas configuraciones; f) la creación del ranking de dispositivos utilizados en la publicación de los mensajes; g) la identificación de los perfiles de los usuarios que realizaron las publicaciones, el número de tuits publicados por cada uno de ellos, el número de seguidores y su idioma y finalmente h) permite la ejecución de todos estos análisis en 3 idiomas distintos, inglés, español y portugués.

De este modo, LOGOS ofrece a los analistas datos valiosos respecto a los usuarios, productos y servicios entre otros. En el Capítulo 7 se muestran los principales resultados de un análisis detallado que fue realizado a modo de validación de la herramienta, y que muestra la potencia que tiene esta versión inicial de la herramienta.

Como resultado de ese caso de aplicación y uso se obtuvieron una serie de conclusiones, algunas limitaciones y algunas acciones de mejora que pasamos a enumerar:

- La herramienta debería ser sensible al idioma en que están escritos los textos y debería tener mecanismos internos tanto para detectar el idioma de cada texto como de aplicar el algoritmo de análisis de sentimientos que mejor se adapte al idioma en cuestión.
- Ampliar el número de idiomas de análisis y de idiomas de la interfaz.
- Aunque la herramienta cuenta con algoritmos de análisis de sentimientos ya pre-entrenados en varios idiomas, debería no obstante dar la posibilidad al analista de usar otros mecanismos y servicios de terceros de análisis de sentimientos como: *Google Analytics* (https://analytics.google.com/analytics/web/), *Meaninngcloud* (https://www.meaningcloud.com/es), *Indico* (https://indico.io), etc.
- Seguir ahondando en el estudio de otros mecanismos de aprendizaje automático, como por ejemplo *Deep Learnig*, para

aumentar la precisión y la rapidez, tanto en la fase entrenamiento como en la de ejecución en tiempo real.

- En esta primera versión, LOGOS solo descarga informes parciales (un informe por cada una de las nueve pestañas programadas), y eso no facilita el análisis y/o seguimiento de los datos y resultados. Pensamos resultaría más fácil e intuitivo disponer de informes más interactivos, como los hipertextuales y autoreferenciados. Sería también ideal disponer de gráficos automáticamente exportados por la herramienta, y de alta calidad.
- Sería conveniente disponer de análisis más profundos, que involucren más cantidad de datos, más tipos de datos y más conexiones entre ellos. Sería conveniente migrar el sistema de almacenamiento actual (CSV) a otro más versátil, y en este sentido se propone el uso de MongoDB para el almacenamiento persistente de datos, y Neo4J para la explotación de las interconexiones entre los datos.
- Esta primera versión de LOGOS se ha centrado en la red social Twitter, pero todos los desarrollos son, a priori, extrapolables al uso de datos procedentes de otros medios sociales. Se propone por tanto incorporar el acceso a otros medios como Instagram, Facebook y/o Google+.
- Se ve conveniente que la herramienta pudiera dar soporte simultáneo a varios análisis y/o analistas de manera simultánea, y por lo tanto se añade como necesidad la incorporación de espacios

de trabajo internos diferenciados (al menos un espacio de trabajo por analista registrado en el sistema).

- Con respecto al registro, se consideran necesarios mecanismos de registros seguros mediante la incorporación de mecanismos de autenticación cifrados.
- Sería conveniente dotar al sistema de mecanismos para el acopio automático de datos. En esta versión prototipo es el analista el que deber estar pendiente del proceso de descarga de datos, y eso no es viable en análisis a gran escala.
- Se considera necesario dotar al sistema en sí de mecanismos de autogestión, como la implementación de mecanismos de copia de seguridad; respaldo de datos, configuraciones de análisis y los propios informes.

CAPÍTULO 7: APORTACIÓN 4 – APLICACIÓN PRÁCTICA DEL SISTEMA WEB, LOGOS

7.1. INTRODUCCIÓN

En la actualidad, es visible el crecimiento exponencial de usuarios y clientes que se lanzan en la Web divulgando sus impresiones sobre los más variados temas y contextos. El incremento de información, resultado de este fenómeno, despertó la atención de las organizaciones respecto a los contenidos generados por sus usuarios en las redes. Estas empresas entienden que en los medios sociales se puede descubrir, a un coste mínimo, información relevante sobre sus clientes, sobre la competencia y también sobre su propia organización.

El usuario contemporáneo desea hablar y ser escuchado. Además, es consciente de que los medios sociales amplifican el alcance de su voz. Por este motivo los productos y servicios de buena calidad terminan por ser bien valorados por los usuarios, los que con sus comentarios contribuyen a la generación de su buena reputación. Por otro lado, posibles problemas con los consumidores pueden hacerse más visibles en las redes, en la medida que éstas asumen el rol de una especie de "centro de reclamaciones" (Nascimento et al., 2015). Por eso, Telles (2010), pone de manifiesto el lugar que los medios sociales ocupan en la creación y en el mantenimiento de la reputación de marca, puesto que atienden tareas como la atención al consumidor, solicitud e información de servicios, gestión de reclamaciones, sugerencias respecto a productos y servicios ofrecidos por la empresa, entre otros. En este sentido, Chen y Zimbra (2010) comentan que el movimiento comunicativo proporcionado por los medios sociales facilita conocer las preferencias, evaluaciones,

sentimientos y opiniones de un gran número de usuarios sobre contenidos, productos, servicios, entidades, personas, entre otros. Por ello, el monitoreo de los medios sociales es esencial, ya que permite múltiples tareas como obtener datos para identificar, medir, cualificar y cuantificar perfiles de usuarios, descubrir posibilidades de acción en los ambientes virtuales, además de adelantarse a posibles situaciones de crisis y prevenir daños a la marca. Este contexto demanda a las organizaciones ser cada vez más conscientes de la importancia del "escuchar" de manera activa a los medios sociales para conocer qué se habla sobre las organizaciones, en qué sentido se habla (positivo o negativo) y de qué manera esto puede influir en la reputación de la marca (M. A. Russell, 2013; Safko & Brake, 2010; Serrano-Cobos, 2014; Torres, 2009; Tsytsarau & Palpanas, 2012).

Así en este nuevo contexto, se requiere que las empresas actuales manejen una gran cantidad de datos que aporten en última instancia conocimiento útil a las organizaciones (Rud, 2009). En este sentido, las soluciones derivadas de la Inteligencia Empresarial (*Business Intelligence)* (BI) a través de técnicas de Minería de Datos y de Análisis de Sentimientos, ofrecen la posibilidad de recogida, almacenamiento y análisis de datos relacionados con las empresas y sus clientes provenientes de los medios sociales on line. Estas soluciones son capaces de transformar datos aparentemente sin relevancia, en información útil que apoye la toma de decisiones de las organizaciones (Sabanovic & Søilen, 2012), proporcionando información que permita una mejor comprensión sobre la actividad de la empresa y los factores externos influyentes (Cordero-Guzmán & Rodríguez-López, 2017).

Como se apuntaba en el capítulo 2 de esta tesis, el BI permite transformar grandes cantidades de datos en información de calidad que apoye la toma de decisiones empresarial, ofreciendo una visión sistémica del negocio (Reginato & Nascimento, 2007). En la misma línea, Shmueli, Patel, y Bruce (2011) definen el BI como un conjunto de conceptos y métodos que apoyan a la toma de decisiones en los negocios, transformando el dato en información y la información en conocimiento. Según Loshin (2013), el BI se define como un conjunto de metodologías, procesos, arquitecturas y tecnologías que dan significado a los datos primarios, transformándolos en información útil. Chaudhuri, Dayal y Narasayya (2011) presentan el BI como la fusión entre la tecnología y sus distintos procesos interacción. En definitiva, estas definiciones apuntan a que en la práctica el *Business Intelligence* tiene funcionalidades propias que implican distintas herramientas de a) extracción de datos; b) sistema de gestión de bases de datos; c) desarrollo de informes *(Reporting Services - RS)*; d) mecanismos de mineración de datos *(Data Mining - DM)* y e) mecanismos de análisis y de procesamiento en línea *(OLAP)* (Aruldoss et al., 2014; Chen & Wang, 2010; Eckerson, 2008; Lee et al., 2016; Lee & Widener, 2016; Li et al., 2008; Zeng et al., 2012).

Para Moss y Atre (2003), el BI facilita información activa que puede ser visualizada en tiempo real, en la ubicación correcta y que además asiste fácilmente al proceso de decisión, una tarea muchas veces compleja que no se debe realizar de manera simplista. En este sentido, es vital el control del análisis de los datos, sobre todo porque debido al crecimiento exponencial de la información disponible en la Web resulta difícil identificar qué

información es realmente relevante. Además, los datos no siempre son obtenidos de fuentes lo suficientemente estructuradas, de manera repetida o con el mismo formato y diseño, lo que añade aún más complejidad al análisis (Brand, 2013; Peters, Wieder, Sutton, & Wakefield, 2016; Watson, 2010). En algunos casos, cuando los datos no son estructurados, el proceso de estructuración de la información puede ser muy costoso y demandar demasiado tiempo a las organizaciones, haciendo de este paso algo inviable al ser llevado a la práctica (Inmon & Linstedt, 2014).

En este sentido, la evolución del BI tradicional al BI 2.0 se da con el objetivo de permitir que la información contenida en la Web 2.0 sea introducida de manera estructurada en los repositorios de datos de las organizaciones. Este cambio dotó de agilidad a los recursos de análisis de la información, facilitando de manera considerable los procesos de decisión.

Como apuntan Lusch, Liu y Chen (2010), el BI 2.0 busca mejorar el desempeño de los procesos de toma de decisiones, reduciendo el tiempo entre la ocurrencia de un evento en el ambiente transaccional y la toma de decisiones resolutivas.

Finalmente en la misma línea, Nelson (2010) defiende que el BI 2.0 supone alejarse del almacén de datos ortodoxo para dar paso a un nuevo sistema que responde a la necesidad de relacionar información proveniente de distintas fuentes de forma rápida.

7.1.1. El BUSINESS INTELIGENCE EN LAS ORGANIZACIONES

Desde una perspectiva histórica, las primeras soluciones de BI emergen en la década de los setenta cuando los primeros softwares de análisis de datos surgieron en el mercado. En la siguiente década aparecieron las hojas de cálculo como por ejemplo el Lotus 123 y su evolución, el Microsoft Excel, que a principios de los años 90 se dieron a conocer como sistemas ejecutivos de información. Estos sistemas permitían un acceso rápido a la información interna y externa además de responder a las necesidades de apoyo en la toma de decisiones de su época (Rasmussen, Goldy, & Solli, 2002).

En virtud de su poder de análisis y de apoyo estratégico, estos softwares despertaron en muy poco tiempo la atención de los diferentes tipos de sectores empresariales. No obstante, debido a su complejidad en aquel momento, el manejo de tales recursos demandaba de competencias específicas por parte de los profesionales. En los inicios de los 2000, estas aplicaciones ganan en facilidad de uso, en potencia y en la optimización de los procesos, ofreciendo una respuesta cada vez más inmediata a sus usuarios. En la actualidad, las herramientas de BI ya están bastante consolidadas y consideradas por los gestores como herramientas de apoyo operacional imprescindible para las decisiones estratégicas del negocio (Solomun Negash, 2004).

En el BI, el protagonista es la tecnología que permite recuperar, grabar, manipular y analizar la información. Las trasformaciones de los sistemas de almacenaje y tratamiento de datos son prueba de esto, ya que pasamos del

Data Warehouse (DW) al *Data Mining* (DM) y finalmente al *Web Mining* (WM). Para Petrini y Pozzebon (2009), el BI parte de dos conceptos centrales: a) recabar, analizar y distribuir información y b) sustentar las decisiones estratégicas y operacionales de las organizaciones a medio y largo plazo. Entre las principales ventajas de su utilización están: a) la reducción de costes; b) el aumento de ingresos; c) las mejoras en las relaciones, lo que incide en la satisfacción del cliente; y d) la simplificación en la comunicación dentro de la organización y el soporte a las decisiones (Arnott, Lizama, & Song, 2017; Arnott & Pervan, 2014; Elbashir, Collier, & Davern, 2008; Wieder & Ossimitz, 2015).

7.1.2. SISTEMA DE APOYO A DECISIONES

Relacionado con el proceso de apoyo a decisiones por medio de sistemas de información, Delibašić et al. (2015) afirman que los diferentes niveles de una organización requieren de distintas necesidades de información para apoyar sus decisiones. Para estos autores, las decisiones pueden clasificarse en estructuradas, no estructuradas y semiestructuradas. Las decisiones estructuradas son frecuentes y rutinarias y cuentan con procedimientos preestablecidos y no novedosos. En contrapartida, las decisiones no estructuradas exigen el uso del sentido común por parte del responsable, por lo que son mucho más subjetivas ya que dependen de su capacidad de evaluación, entendimiento y experiencia. Son decisiones no rutinarias donde no hay respuestas predeterminadas para el problema.

En general, las decisiones no estructuradas son más frecuentes en los niveles organizacionales más altos, mientras que los problemas estructurados son más comunes en los niveles más bajos de la empresa. Finalmente, las decisiones semiestructuradas tienen características de los dos tipos de decisiones comentadas anteriormente, de modo que, una parte del problema tiene una respuesta clara, predeterminada y precisa, mientras otra parte del problema no la tiene.

Los sistemas de BI cuando son utilizados como Sistemas de Apoyo a Decisiones (SAD), se basan en modelos informáticos que ofrecen soluciones a problemas semiestructurados y no estructurados. En términos prácticos, el propio usuario puede generar consultas de acuerdo con sus necesidades y relacionar la información ofreciendo una nueva perspectiva que contribuya al descubrimiento de lo que busca.

Los SAD responden de manera rápida y son capaces de dar apoyo directamente a todos los tipos específicos de decisiones con estilos y necesidades diferentes (Arnott & Pervan, 2016). Para los autores Delibašić et al. (2015), los SAD también proporcionan un proceso de toma de decisiones mejorado ya que ofrecen un mejor entendimiento del negocio, muestran un mayor número de alternativas para una decisión, presentan la capacidad de implementar análisis ad hoc o aleatorios, ofrecen respuestas más rápidas a las situaciones previstas, proporcionan una comunicación mejorada, vuelven el trabajo en equipo más eficaz, además de ofrecer un mejor control y ahorro de tiempo y de costes.

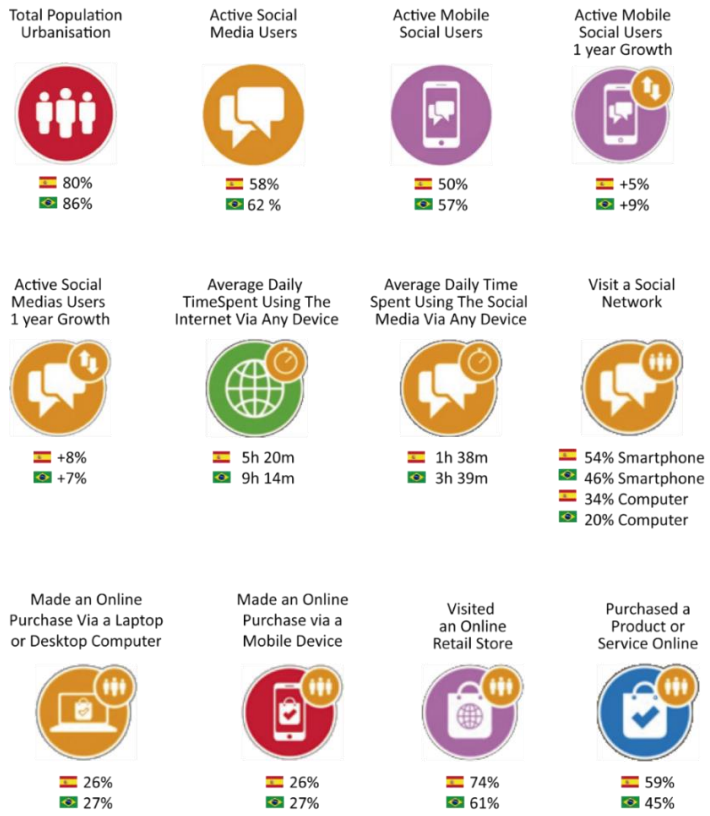
Ante lo anteriormente expuesto, este estudio propone la validación del sistema web de inteligencia empresarial y de apoyo a decisiones LOGOS, desarrollado en el marco de esta tesis doctoral, con el fin de corroborar su desempeño y utilidad para las empresas, para la comunidad científica y para la sociedad. Para ello, se llevarán a cabo cuatro aplicaciones prácticas (dos en el idioma español y dos en el idioma portugués de Brasil), con el propósito de reunir y transformar datos aparentemente irrelevantes de la red de micro-mensajes Twitter, en conocimiento competitivo que apoye y potencie la toma de decisiones de las organizaciones.

La elección de estos dos idiomas, relacionados con los mercados de España y Brasil, se dio principalmente por tres motivos. El primer motivo se justifica debido a la escasez y, en consecuencia, a la necesidad latente de herramientas gratuitas pensadas para actuar como soporte a decisiones de marketing, basadas en técnicas de Minería de Datos y Análisis de Sentimientos que estén adaptadas a diversos idiomas.

El segundo motivo está relacionado al alto número de características compartidas entre estos dos mercados. En general, los mercados de España y Brasil tienen considerables semejanzas que favorecen, por ejemplo, los análisis relacionados con el benchmarking, entre otros. Según el reporte 2019 Digital Yearbook, las semejanzas entre España y Brasil en lo referente al consumo de internet y el uso de medios sociales es alta. Como se observa en la Figura 58, España y Brasil guardan consonancia en muchos aspectos tales como el porcentaje de la población en áreas urbanas, el porcentaje de usuarios de medios sociales, el porcentaje de usuarios que acceden a los

medios sociales por medio de smartphones, el porcentaje de incremento de usuarios de medios sociales con relación al año anterior y el porcentaje de usuarios que realizaron visitas y compras en tiendas online. En los ítems referentes a la realización de una compra online utilizando ordenadores de sobremesa (*desktops*), portátiles *(laptops*) y teléfonos inteligentes *(smartphones)*, los resultados son prácticamente iguales siendo la diferencia entre ambos países de apenas un 1%. Por otro lado, las mayores diferencias están en los ítems relacionados con el tiempo de utilización de los medios sociales. Según este informe, el ciudadano brasileño dedica casi el doble del tiempo a los medios sociales en comparación con el ciudadano español.

Figura 58. Indicadores semejantes entre España y Brasil del 2019 Digital Yearbook.



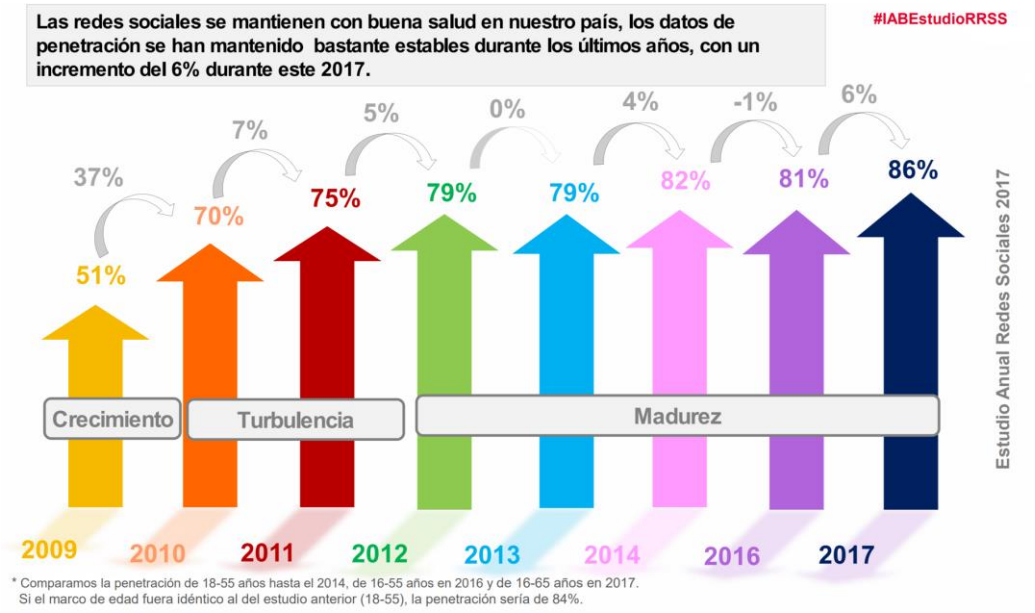
Fuente: Hootsuite (2019).

Otro estudio relevante realizado por Interactive Advertising BureauI (IAB, 2017), identifica a España como un mercado de medios sociales con carácter comercial y maduro aunque todavía en expansión.

Según este estudio, los medios sociales se mantienen con buena salud en el país, ya que los datos de penetración se han mantenido bastante estables durante los años. El estudio también afirma que: a) entre un 11-24% de los usuarios hablan de compras realizadas, crean eventos y contactan con el servicio al cliente de una marca utilizando los medios sociales; b) un 83% de los internautas declara ser fan o seguir marcas a través de los medios sociales de los cuales, un 39% informa hacerlo con intensidad; c) para un 25% de los usuarios la presencia en los medios sociales aumenta la confianza en la marca; d) un 52% afirma haber sido influido por los medios sociales en sus compras; e) un 53% suele buscar información en alguna red social antes de realizar compras por internet; f) un 39% realiza comentarios, expone sus problemas o dudas sobre sus compras por internet, y g) un 77% les parece muy o bastante interesante recibir atención al cliente y soporte técnico por medio de los medios sociales (Figura 59).

Por otra parte, el escenario de los medios sociales brasileño, además de semejante al español, es prometedor. Según el reporte *Social Media Trends* 2018, llevado a cabo por la empresa Rockcontent, un 94,4% de los usuarios utiliza los medios sociales diariamente, un 57% dedica más de 3 horas al día en los medios sociales, un 81% busca información sobre su profesión o área de actuación y un 57,2% busca información sobre productos y servicios en las redes (Rockcontent, 2018).

Figura 59. Utilización de los medios sociales en España 2017.



Fuente: IAB (2017).

Desde el prisma de la administración pública, el gobierno brasileño manifiesta un creciente apoyo para fomentar las iniciativas relacionadas con la utilización de los medios sociales, tanto por las empresas como por los usuarios. No obstante, todavía son pocas las empresas capaces de hacer un uso eficaz de estas redes que les permita alcanzar sus objetivos empresariales (Xavier & Nunes, 2014). Según los autores Ananda, Hernández-García y Lamberti (2016), en la mayoría de las ocasiones, las estrategias en los medios sociales on line todavía son pautadas utilizando la intuición o por medio de ensayo y error. Esto ocurre quizás por la falta de sistemas gratuitos y de fácil uso para las empresas, principalmente para aquellas de mediano y pequeño tamaño, que muchas veces no pueden prescindir de un analista de marketing para minerar los medios sociales, incrementando con ello considerablemente sus costes.

Finalmente, el tercer motivo sobre la elección del mercado español como elemento de análisis está relacionado con el hecho de que la tesis se desarrolla en España y está sostenida por los recursos ofrecidos por el sistema de enseñanza español. Y, por otro lado, se ha optado por complementar el análisis con el mercado brasileño debido a que esta tesis ha sido apoyada por una beca brasileña a través de la CAPES *(Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior).*

Se trata de un organismo brasileño bajo la autoridad del Ministerio de Educación que busca por medio de becas universitarias, además de fomentar el intercambio de conocimiento entre distintos países, dar soporte a las empresas en Brasil.

Por último, una vez establecidos los mercados, se ha optado por analizar datos de Twitter asociados a las cuentas de las marcas Nike y Samsung, debido a su gran presencia en esta red y al extenso abanico de productos y servicios comercializados activamente por ambas empresas en los mercados de España y Brasil.

7.2. OBJETIVOS

Para la realización de este estudio fueron determinados **dos** objetivos generales y **diez** objetivos específicos que serán detallados a continuación.

7.2.1. OBJETIVOS GENERALES

El presente estudio considera los siguientes **objetivos generales**:

a) Realizar una aplicación práctica del sistema LOGOS, creado en esta tesis doctoral, a través de la recogida, filtrado y análisis de la información generada por los usuarios de Twitter relacionada con las cuentas @Nike_Spain, @Nikebrasil, @SamsungEspana y @SamsungBrasil.
b) Demonstrar y discutir las ventajas de la utilización de LOGOS en la búsqueda de informaciones (*insights)* que apoyen a la toma de decisiones en las organizaciones.

7.2.2. OBJETIVOS ESPECÍFICOS

Los objetivos generales comentados arriba se desglosan en los siguientes **objetivos específicos**:

a) Identificar los temas de mayor interés tratados por los usuarios detectando las palabras más usadas y obteniendo su nube de palabras;
b) Detectar las etiquetas hashtags más usadas por los usuarios y generar nube de palabras para identificar los temas y las campañas de mayor interés tratados por los usuarios;
c) Identificar los mensajes más importantes de la red según los usuarios.

d) Identificar y establecer un ranking de mensajes seleccionados como favoritos por los usuarios;
e) Identificar las cuentas de usuarios más mencionadas en los mensajes.
f) Identificar geográficamente el origen de los mensajes a través de la creación de un mapa de geolocalización;
g) Identificar y clasificar a los usuarios más activos y con mayor influencia en base al número de seguidores;
h) Determinar la polaridad de los mensajes por medio de técnicas de Análisis de Sentimientos a través de algoritmos automáticos;
i) Crear nubes de palabras para los tuits generales, los tuits clasificados como negativos y los tuits clasificados como positivos;
j) Identificar los dispositivos y aplicaciones de conexión más utilizados por los usuarios al momento de publicar un tuit.

7.3. MÉTODO DE RECOGIDA DE DATOS

Este apartado recoge los aspectos metodológicos considerados para la obtención de las muestras de datos utilizadas en este estudio. Para ello, se definieron las cuentas oficiales en Twitter pertenecientes a las empresas Nike y Samsung tanto en España como en Brasil.

Respecto a la extracción de los tuits y retuits, ésta se realizó por un período de doce meses, concretamente de mayo de 2016 a abril de 2017, por medio de la función de búsqueda disponible en LOGOS. De este modo, se obtuvieron cuatro muestras distintas, siendo dos de la empresa Nike,

una en el idioma español a través de la cuenta (@Nike_Spain), y otra en portugués de Brasil a través de la cuenta (@Nikebrasil); y dos de la empresa Samsung, una en el idioma español a través de la cuenta (@SamsungEspana) y otra en portugués de Brasil a través de la cuenta (@SamsungBrasil).

En la Tabla 32 se describen respectivamente el nombre de las cuentas, el nombre de las empresas, el idioma oficial que se practica en cada cuenta, el número de tuits, el número de retuits y el número total de mensajes obtenidos (tuits + retuits) en el periodo comentado (mayo de 2016 a abril de 2017).

Tabla 32. Resumen del muestreo de las cuentas @Nike_Spain, @nikebrasil, @SamsungEspana y @SamsungBrasil.

Cuenta	Empresa	Idioma	Nº tuits	Nº retuits	Nº total
@Nike_Spain	Nike	Español	4043	6027	10070
@nikebrasil	Nike	Portugués	6936	20072	27008
@SamsungEspana	Samsung	Español	4261	25000	29261
@SamsungBrasil	Samsung	Portugués	22733	27474	50207

Fuente: Elaboración propia desde LOGOS.

Una vez reunidas las muestras, estas fueron llevadas a LOGOS para la realización de las tareas de Procesamiento del Lenguaje Natural, Minería de Datos, Análisis de Sentimientos y posterior construcción de los informes de apoyo de decisiones ofrecidos por el sistema.

En los epígrafes siguientes se presentan y se discuten los principales resultados del estudio.

7.4. RESULTADOS

En esta sección se presentan los resultados más relevantes obtenidos con relación a las marcas Nike y Samsung (España y Brasil), a partir de los análisis que ofrece el sistema LOGOS. Concretamente, se presenta un análisis integrado a partir de los siguientes aspectos: a) los mensajes de mayor impacto/alcance en la red; b) su geolocalización; c) su polaridad (positiva o negativa); d) los temas más relevantes (*trending topics)*; e) los dispositivos de publicación utilizados en las publicaciones; así como f) los usuarios más relevantes de la red según cada muestra.

El objetivo principal es desvelar información a partir de los datos que ofrezca conocimiento fundamentado sobre posiciones de mercado, acciones de marketing realizadas, *inputs* para la formulación de estrategias de marketing, peticiones de los usuarios/clientes, identificación de problemas con productos/servicios y terceros, entre otras cuestiones.

En definitiva, se pretende ofrecer una visión en conjunto de diversas ideas y lecturas sobre los datos (**destacadas en negrita**), que puedan ser útiles para las empresas y su toma de decisiones.

Los datos brutos provenientes de LOGOS a partir de los cuales se han elaborado los siguientes informes, están disponibles en la web de esta tesis doctoral y pueden ser consultados y descargados en http://hipatia.ugr.es/steiner/index.php/nikesamsung-datasets/.

7.4.1. INFORME DE MARKETING: NIKE

7.4.1.1. MENSAJES MÁS IMPORTANTES

La frecuencia de replicaciones y la indicación de "me gusta" de los mensajes de la muestra @Nike_Spain, revelaron algunas de las acciones de marketing realizadas por la marca en España entre mayo de 2016 y abril de 2017. Concretamente, un análisis a este nivel pone de manifiesto la importancia del llamado **Marketing de Influencia** (Lu & Seah, 2018), **el uso de personas consideradas influyentes para difundir información, conseguir captar la atención y el compromiso de la audiencia**. Por un lado, la influencia de una persona depende básicamente de dos factores: credibilidad y alcance. Por otro lado, la probabilidad de que el destinatario sea influenciado depende a su vez de cuatro factores: relevancia (la información correcta), timing (en el momento adecuado), alineación (en el lugar correcto) y confianza (la persona adecuada) (Smith, Kendall, Knighton, & Wright, 2018). Como espectadores, somos menos "racionales" ante alguien a quien admiramos. A través de la asociación positiva que se produce entre dicha persona y el producto promocionado por ésta, transferimos bondades al producto y se activan mecanismos emocionales que influyen en nuestro comportamiento (Romero & Castello-Martinez, 2017).

Desde el punto de vista del marketing de influencia, a continuación, se mencionan las cuatro acciones de la marca de mayor alcance/relevancia que han sido detectadas a partir de la muestra analizada. Respecto a la primera acción, la marca promovió un sorteo por medio del jugador

(*gamer)* @vegetta777 que tiene en Twitter más de 5,41 millones de seguidores y en YouTube más de 24 millones de suscriptores hoy en día. La promoción consistió en que los seguidores deberían personalizar en la página web de Nike España un modelo de las zapatillas según su gusto. Las 5 zapatillas que más gustasen a @vegetta777, serían regaladas a sus creadores como premio. Esta campaña generó en la red más de 10.000 "me gusta" y más de 1.500 retuits en Twitter. Si consideramos también la red YouTube, el video de la campaña tiene más de 58.000 indicaciones de "me gusta", más de 950.000 visualizaciones y más de 5.000 comentarios.

Figura 60. Tuit relacionado con la acción de marketing con el youtuber/gamer sTaXx.

Fuente: https://t.co/MUIxUeaRx6 (acceso el 07/08/2019).

La segunda acción, de corte más social, se lleva a cabo con el jugador de baloncesto Pau Gasol. Concretamente, el jugador publica un mensaje donde habla de la satisfacción de poder apadrinar la reforma de su primera cancha de baloncesto en colaboración con @Nike_Spain. La tercera acción se dirige especialmente al público joven de la marca. En ésta, el jugador de fútbol del *F.C. Barcelona*, Paco Alcácer, va hasta una tienda oficial de la marca en Barcelona para dar una charla y firmar productos de la marca. Finalmente, respecto a la cuarta acción identificada, el youtuber y gamer @bystaxx publica un mensaje con una foto dando las gracias a la marca por haberle enviado algunos regalos. Debido al número de seguidores que tiene en sus

medios sociales, el mensaje obtuvo más de 581 réplicas y más de 4000 indicaciones de "me gusta" (Figura 60). Hay que destacar que un mensaje replicado puede ser visualizado por todos los seguidores del usuario que lo ha replicado, lo que incrementa exponencialmente su alcance.

Estas cuatro acciones, aunque bastante diferentes, tienen como punto en común la utilización de personas consideradas influyentes. **La primera acción utiliza un youtuber/gamer para conseguir que su audiencia vaya hasta la página web de la marca con el objetivo de vivir la experiencia de la personalización de unas zapatillas Nike. Hay que mencionar que los productos personalizados son considerados premium y cuestan bastante más que los no personalizados. De modo que la marca con esta iniciativa impulsa junto a sus clientes la vivencia de experiencias positivas con la marca a través de la personalización, fomentando a la vez la utilización y el conocimiento de este recurso.** La segunda acción utiliza la figura de un jugador de baloncesto de gran reputación en España para promover acciones sociales en el país. La tercera recurre a un jugador de uno de los equipos con más seguidores en España para atraer a su público joven (actuales y futuros consumidores) a una de sus tiendas. Por último, la cuarta acción, del mismo modo que la primera, también se basa en el uso de un youtuber/gamer para que haga publicidad de algunos de los productos de la marca.

Por otro lado, en la muestra @nikebrasil se observan acciones muy parecidas a las empleadas en España. En esta ocasión, la marca utiliza el jugador de fútbol Gabriel Barbosa y promueve acciones de marketing a

través de la etiqueta "#VemJunto" (Figura 61), y el jugador Cristiano Ronaldo con la acción "#mostracomofaz".

Un aspecto que destacar de la cuenta de @nikebrasil, es la gran cantidad de mensajes identificados con tono irónico. En estos mensajes, los internautas de forma irónica expresan quejas relativas al elevado precio de los productos de la marca. Dos de los cinco mensajes más replicados de la muestra van en este sentido. Que estos mensajes tengan tantas réplicas y "me gusta" por parte de sus consumidores, puede ser un **indicador de que los usuarios brasileños, de manera general, estén en desacuerdo con la política de precios practicados por la marca en el país. Este dato es importante, puesto que puede influir directa y negativamente en el e-WoM y en la reputación de Nike en la red (Figuras 62 y 63). En este sentido, se recomienda que la marca estudie en profundidad el mercado de Brasil a fin de entender sus distintas clases sociales y adaptar su estrategia de precios a la realidad de mercado del país. Así como considerar la viabilidad de crear líneas de productos más accesibles para hacer frente a esta situación.**

Figura 61. Tuit de la cuenta @gabigol promoviendo la campaña y el #VemJunto.

Fuente: https://twitter.com/gabigol/status/756974934182756352 (acceso el 07/08/2019).

Figura 62. Mensaje de tono irónico respecto los altos precios de la marca, escrito por @LeaoNicolly.

Fuente: https://t.co/8hv1GXAvHU (acceso el 07/08/2019).

Figura 63. Mensaje de todo irónico de @caiodmg relatando dificultades económicas para adquirir el producto de @nikebrasil.

Fuente: https://twitter.com/caiodmg/status/800686231235530754 (acceso el 07/08/2019).

7.4.1.2. TEMAS MÁS COMENTADOS: *"TRENDING TOPICS"*

Los temas más comentados en Twitter, también llamados "*trending topics"*, provienen de las menciones de etiquetas tipo hashtags. En este sentido, para la muestra @Nike_Spain, se observa que los temas más comentados en orden de importancia son: *#justdoit,* que se refiere directamente a la principal campaña publicitaria y slogan de la marca; *#eslahora* y *#somosblanquinegros* que están directamente relacionados con el club de fútbol *Burgos*; #mercurial y #hipervenom, que se refieren a modelos específicos de zapatillas de fútbol de la marca; y *#nrc* que está

relacionado con la app Nike Run Club[22]. Estos *"trending topics"* pueden indicar que **el slogan de la marca es conocido y puesto en práctica por los internautas en sus mensajes; que el club de fútbol *Burgos* se hace notar en la red por lo que la marca podría monitorear y evaluar esos hashtags para considerar posibles acciones futuras de marketing direccionadas; que de todos los productos de la marca, los internautas parecen interesarse mayoritariamente por las zapatillas Mercurial y las Hipervenom; y finalmente, que la aplicación de Running de la marca también les parece un tema relevante.**

Respecto a la muestra @nikebrasil, los temas de mayor interés en orden de importancia están relacionados con la campaña publicitaria *#VemJunto,* que utiliza a personas famosas de Brasil como cantantes o deportistas entre otros, para motivar a los internautas a descargar la app de NIKE y utilizar las rutas propuestas por ellos. La etiqueta #sneakerdasemana, que es utilizado por la marca para promover algún producto específico cada semana. A continuación, encontramos la etiqueta *#mostracomofaz*, también creado por la marca, que se relaciona con las zapatillas del modelo Mercurial y se utiliza para mostrar los logros de los jugadores patrocinados por la marca. **Respecto a esta etiqueta, sucede algo en particular que merece atención. Debido a su doble sentido morfológico, este hashtag ha sido utilizado por los internautas para la realización de bromas en torno a su significado. Por**

[22] https://www.nike.com.br/corrida/app-nike-plus (acceso el 07/08/2019).

Figura 64. Mensajes estilo "juego de palabras" relacionadas al #mostracomofaz.

Fuente: https://twitter.com/Gomarez_/status/736521004630036480 (acceso el 07/08/2019).

ejemplo, uno de los usuarios utiliza la etiqueta *#mostracomofaz ("enséñame cómo se hace")* haciendo referencia a la receta secreta de la hamburguesa "cangreburguer" del personaje "Don Cangrejo" de la serie televisiva Bob Esponja. Otro usuario, utiliza la etiqueta seguida de la frase "para ganar tu corazón" (Figura 64). Situaciones como estas desvirtúan y agregan ruido en la comunicación de la acción de marketing en la red. En este sentido, para evitar sucesos como éste, se deben considerar las distinciones morfológicas de las palabras utilizadas en las acciones de marketing para cada país. Por otro lado, si la marca hubiese detectado este fenómeno a tiempo, podría haber procedido con alguna acción de marketing valorando la creatividad de los internautas, haciéndoles sentirse importantes y valorados, además de estimular el e-WoM positivo.

Otro dato que merece atención especial se relaciona con la posición en el ranking, según el grado de importancia, que ocupa la etiqueta #justdoit en la muestra @nikebrasil. Mientras que en la muestra @Nike_Spain esta etiqueta ocupa el primer puesto de la lista, en la muestra @nikebrasil ocupa el noveno lugar. Lo que puede ser un indicador de que para los brasileños el #justdoit no tiene tanta relevancia como para los españoles o ha dejado paso a otros temas/promociones para ellos más importantes, como

#vemjunto, #sneakerdasemana y otros mencionados anteriormente que aparecen encabezando la lista. Por lo que se abre un frente bastante interesante de investigación sobre este tema en cuestión.

7.4.1.3. GEOLOCALIZACIÓN

Los resultados relacionados con la geolocalización de las muestras de Nike (@Nike_Spain y @nikebrasil), indican de manera general una muy baja adhesión del recurso por parte de los usuarios. En la muestra @Nike_Spain, apenas 6 mensajes (0,14% de la muestra) llevaban la información de geolocalización y en la muestra @nikebrasil apenas 5 (0,07% de la muestra). La opción de ubicar en el mapa dónde estaba el usuario en el momento de publicar un determinado mensaje, representa una información que es de gran interés para las organizaciones puesto que trae consigo la posibilidad de identificar diversos aspectos sociodemográficos de los usuarios. Por ejemplo, este recurso puede ser indicativo del nivel de educación y de poder de compra de los usuarios, el idioma que hablan o si viven más cerca de la costa o en zonas montañosas.

Esos datos ayudan a mapear y segmentar a los clientes facilitando la aplicación del CRM, así como la creación de acciones de marketing dirigidas y con características específicas según el público objetivo (*target)* que se desea atender. De este modo, aunque LOGOS haya sido capaz de identificar y geolocalizar en el mapa a estos tuits (ej. Figuras 65 y 65), el escaso volumen de mensajes geolocalizables no es representativo.

Esta limitación podría superarse con la realización de acciones de marketing dirigidas por la marca que estimulen la activación de la geolocalización por parte de los clientes. Se sugiere la posibilidad de utilizar elementos de gamificación que estimulen el "*check-in*" de los usuarios en los establecimientos de la marca. Cuantos más *"check-in",* más importancia el usuario adquiere en la red.

Como recompensa, se podría ofrecer algún tipo de premio tal como ofertas en productos o en establecimientos específicos. Esto ampliaría el número de mensajes geolocalizables y como consecuencia se generarían más datos y análisis más consistentes.

Figura 65. Ubicación del primer mensaje geolocalizable de la muestra de tuits de @Nike_Spain.

Fuente: Elaboración propia desde LOGOS.

Figura 66. Ubicación del segundo mensaje geolocalizable de la muestra de tuits de @nikebrasil.

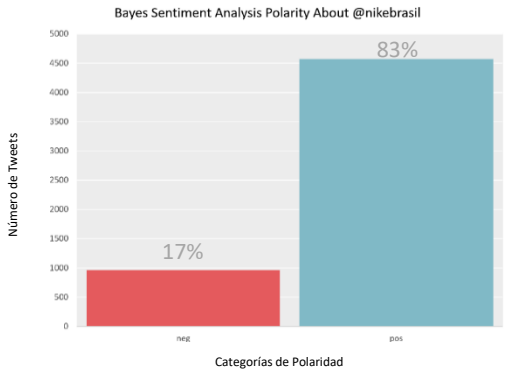
Fuente: Elaboración propia desde LOGOS.

7.4.1.4. ANÁLISIS DE SENTIMIENTOS

La aplicación de este recurso en la muestra de mensajes de la marca Nike, ha puesto de manifiesto de manera general una superioridad de los comentarios positivos frente a los negativos tanto para la muestra de @Nike_Spain como para @nikebrasil. Sin embargo, también es cierto que

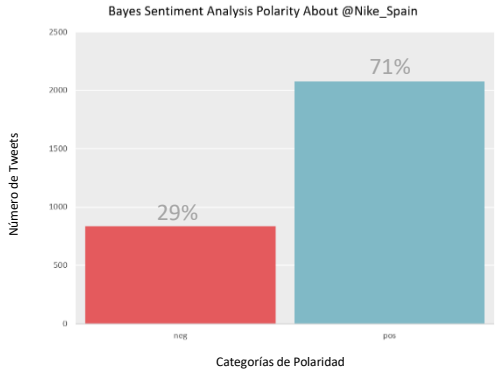
la muestra @nikebrasil, con un 83% de mensajes positivos y 17% negativos, muestra resultados más positivos que la muestra @Nike_Spain, con 71% de mensajes positivos y 29% de negativos (Figuras 67 y 68). A continuación, se detallan ambos grupos de mensajes (positivos y negativos) de cada muestra.

Figura 67. Grafica de la pestaña *Sentiment* de la muestra de tuits de @nikebrasil.



Fuente: Elaboración propia desde LOGOS.

Figura 68. Grafica de la pestaña *Sentimen*t de la muestra de tuits de @Nike_Spain.



Fuente: Elaboración propia desde LOGOS.

- **Mensajes positivos**: **@Nike_Spain**

La Figura 69 muestra la nube de frecuencia de palabras con base en los tuits clasificados como positivos en la muestra de @Nike_Spain. El tamaño de las palabras varía según su frecuencia. Cuanto más frecuente es una palabra en los textos analizados, mayor será su tamaño en la nube indicando de manera visual su relevancia en los análisis. En este sentido, las palabras más importantes son: "*gracias" (143 menciones),* que puede indicar una actitud positiva de los usuarios hacia la marca, además de la correcta clasificación de estos mensajes por parte del algoritmo que se

refuerza por, obras palabras de tono positivo que salen en la nube como "*bien" (58), "mejor" (47), "preciosas" (23)* o *"bonita" (12),* entre otras*. En segundo puesto, está la palabra "nuevas" (131),* que se refiere a los nuevos productos y modelos de la marca. Así mismo, aparecen palabras relacionadas con productos específicos comercializados por Nike como, "*camiseta" (112) (palabra que ocupa el tercer puesto del ranking de frecuencia), "zapatillas" (49), "Pegasus" (38) y "running" (76)*. **Hay que resaltar que, según la nube de los comentarios positivos, la palabra "*camisetas*" (112) genera muchos más comentarios en la red que la palabra *"zapatillas"* (49), por lo que los productos referidos a camisetas producen más tráfico de comentarios en la red que los productos relativos a zapatillas vendidos por la marca.**

Figura 69. Nube de frecuencia de palabras en base a los tuits clasificados como positivos de la muestra @Nike_Spain.

Fuente: Elaboración propia desde LOGOS.

A modo de ejemplo, en la Tabla 33, se muestran tres mensajes relacionados con las palabras *"gracias"*, *"nuevas"* y *"camiseta"* que ocupan las primeras tres posiciones en el ranking según su frecuencia de aparición. A continuación, analizamos cada uno de estos mensajes con el fin de entender su contexto e identificar posibles *insights* que ayuden en la toma de decisiones de marketing.

Tabla 33. Listado de tuits relacionados con las tres palabras más frecuentes en los tuits clasificados como positivos de la muestra de @Nike_Spain.

Palabra		Tuit
gracias	1a	El #DiaSinCoches en Bcn lo hemos aprovechado muy bien con @Nike_Spain probando las #nikestructure20 #nikezoom… https://t.co/ru6GIwFj6E
	1b	La que liastéis ayer en el #Runwithdabiz fue muy grande. Sois unos cracks, mil gracias @Nike_Spain
	1c	Ligeras!! Súper cómodas!! Y preciosas!! Casi casi corren solas!! Regalazo, Muchas gracias @Nike_Spain… https://t.co/GeD8
nuevas	2a	Descansando en la habitación, probando mis nuevas AirMax1 Flyknit, la más ligera #thelightest1 #airmax @Nike_Spain https://t.co/zHBcSsUvTF
	2b	La verdad es que @Nike_Spain lo tiene fácil para que la nueva camiseta del @Atleti me guste más que la de la temporada pasada..
	2c	Muy bonita la nueva equipación que le ha hecho @Nike_Spain a la @SDP_1922: https://t.co/hFHVOBqR5X
camiseta	3a	@ANAALThani @MalagaCF ya que ningún jugador nuevo ha sacado la nueva camiseta, yo me ofrezco voluntario para modelo, de guay @Nike_Spain
	3b	Las nuevas camisetas del Sporting, con las banderas de Asturias y de Gijón son muy guapas. ¡Felicidades @Nike_Spain!
	3c	Boceto rápido de lo que esperemos que sea la nueva camiseta del @Atleti@Nike_Spain

Fuente: Elaboración propia desde LOGOS.

Según la Tabla 33, en relación con la palabra *"gracias",* el tuit 1a fue publicado en la celebración del día sin coches[23] en Barcelona, concretamente el 22 de septiembre de 2016. En el video aparece el usuario corriendo por la calle "Vía Layetana", que cruza el centro de la ciudad de Barcelona, y en su tuit promueve el estreno de sus nuevas zapatillas nikestructure20. En el tuit 1b, el usuario indica que lo pasó muy bien en la carrera nocturna promovida por Nike, llamada #Runwithdabiz, donde

[23] http://mobilitat.ajuntament.barcelona.cat/es/dia-sin-coches (acceso el 07/08/2019).

participa el chef de cocina español David Muñoz. Finalmente, en el tuit 1c, la usuaria manifiesta un sentimiento de alegría por unas zapatillas que le han regalado. Lo hace por medio de adjetivos y expresiones muy positivas con relación al producto.

En estos tres casos se observa claramente la actitud positiva hacia la marca con relación a productos, servicios y promociones. **Los clientes se muestran comprometidos e involucrados con la marca puesto que actúan, en alguna medida, como "embajadores" de la misma. De modo que más acciones de marketing en la misma línea, que busquen estimular en los usuarios este tipo de sentimientos positivos, tienden a ser acciones de éxito.**

Respecto a los tuits relacionados con la palabra *"nuevas"*, se puede apreciar en el mensaje 2a que el jugador Diego Llorente, del club de fútbol *Real Sociedad*, luce sus nuevas AirMax1 Flyknit (Figura 70). Este mensaje fue replicado 33 veces y obtuvo 98 "me gusta". El tuit 2b, aunque se haya clasificado como positivo, lleva consigo un tono sarcástico, ya que manifiesta un sentimiento de rechazo del usuario con relación al modelo de las camisetas del club *Atleti* de la temporada del año anterior. El tuit 2c, se refiere a lo bella que es la nueva camiseta de SD Ponferradina. Este mensaje obtuvo

Figura 70. Mensaje y foto relacionadas al tuit 2a de la Tabla 25.

Fuente: https://t.co/zHBcSsUvTF (acceso el 07/08/2019).

129 retuits y 154 "me gusta". **Este grupo de tuits es interesante porque pueden ser un indicador del éxito del e-WoM relacionado con dos productos específicos de la marca (zapatillas AirMax1 Flyknit y la nueva camiseta del SD Ponferradina). Así mismo, tuits como el 2c también ponen de manifiesto la dificultad del Análisis de Sentimientos en identificar y clasificar correctamente mensajes con tono de sarcasmo/ironía.**

Finalmente, en relación con la palabra "*camiseta*", el tuit 3a es muy interesante, ya que el usuario se ofrece para posar como modelo de las nuevas equipaciones que sacaría la marca en aquel año. **Ideas *(insights)* como esta pueden ser utilizadas por la marca en acciones de marketing que elijan bajo algunos aspectos a clientes de la marca que puedan posar como modelos de las equipaciones de sus equipos. Este tipo de acciones tienden a generar un volumen de datos significativo en la red, estimulando el e-WoM y evidenciando a la marca en la red, además de crear conexión y compromiso con su público a un bajo coste.**

El tuit 3b trata de felicitar a Nike por las banderas de Asturias y de Gijón que llevan las nuevas camisetas del Sporting. **Aquí, se destaca el hecho de que el cliente además de felicitar a la marca, indica la característica específica que le gusta del producto (las banderas de Asturias y de Gijón). Esta es una idea que quizás podría ser replicada con éxito en otros productos.**

En el mensaje 3c, el usuario propone un modelo de colores para la equipación del equipo de Atleti FC. **Este caso es muy interesante porque se observa a una persona que parece ser un entusiasta del Atleti FC, por lo**

que además se podría deducir que muy probablemente conoce a su equipo y se preocupa por él. Es muy probable también que invierta dinero yendo a los partidos y comprando los productos de la marca como por ejemplo sus equipaciones. Este usuario dice directamente a la marca los colores que a él le gustaría ver en la equipación de su equipo del alma. Mensajes como éste ponen de relieve la importancia de los medios sociales en eliminar intermediarios en la comunicación y en las relaciones entre cliente y empresa, además de destacar la importancia de la escucha activa de los medios sociales. Este caso puede servir como idea *(insight)* para acciones de marketing que, por ejemplo, animen a los clientes a enviar dibujos con sugerencias de colores y modelos que les gustarían ver en las equipaciones de sus equipos.

- **Mensajes positivos**: **@nikebrasil**

Respecto a las palabras más frecuentes en la muestra de mensajes positivos de @nikebrasil (Figura 71), se observan similitudes con la muestra @Nike_Spain, y aparecen por ejemplo las palabras *"camisa" (225)*, *"tênis"* (212) y *"obrigado" (140)* entre las palabras más frecuentes. Además, al igual que ocurre en @Nike_Spain, la palabra "*camisa*" genera bastante más

Figura 71. Nube de frecuencia de palabras en base a los tuits clasificados como positivos de la muestra @nikebrasil

Fuente: Elaboración propia desde LOGOS.

comentarios en la red que la palabra "*tênis*". **Otro aspecto que merece especial atención es la presencia de la palabra "*patrocina*" (150) que hace referencia a la palabra "patrocinar". Esta palabra no aparece en la muestra de @Nike_Spain, lo que puede indicar que estamos ante un fenómeno particular del mercado brasileño, en el cual podría resultar interesante profundizar los estudios sobre este fenómeno.**

La Tabla 34, muestra ejemplos concretos relacionados con las tres palabras más importantes de la muestra (*camisa, tênis y patrocina*).

Tabla 34. Listado de tuits relacionados con las tres palabras más frecuentes en los tuits clasificados como positivos de la muestra de @nikebrasil.

Palabra		Tuit
Camisa	1a	Parabéns @nikefutebol @nikebrasil e @alimeira e direção do @SCInternacional pela camisa mais bonita da história do Inter!
	1b	Rapaz pq nao fizeram a camisa em degradê da seleção @CBF_Futebol @nikefootball @nikefutebol @nikebrasil o meiao ficou massa parabens
	1c	A @nikebrasil podia muito bem trazer as camisa de treino do @Atleti tb pras loja ne.
Tênis	2a	@nikebrasil pare de fazer tênis maravilhosos pq o meu bolso está vazio. Obrigado
	2b	olá @nikebrasil, gostaria de verificar com vocês se esse tênis foi feito com tecido extraído do vestido das fadas da floresta encantada
	2c	@nikebrasil estou sem palpites e nem falas! ESSE TÊNIS É UM SONHO! Parabéns, vou cada vez divulgando seus produtos! https://t.co/pY8pKyRBMM
Patrocina	3a	se gentileza gera gentileza, me patrocina aí, por gentileza @nikebrasil
	3b	Queria postar essa foto pra mostrar essa tecnologia Dri-Fit dps de um banho de chuva, alô @nikebrasil Patrocina nós https://t.co/sRQoDMMXYy
	3c	@nikebrasil me patrocina, me machuquei e ficou igual ao símbolo de vocês https://t.co/OLznrvwb3H

Fuente: Elaboración propia desde LOGOS.

Según la Tabla 34, en el tuit 1a, el usuario felicita a la marca por haber producido la camiseta más bonita de la historia del equipo Sport Club Internacional. En el tuit 1b, el usuario pregunta a la marca si es posible

aplicar un efecto difuminado en las camisetas de la equipación de Brasil. El tuit 1c versa sobre una petición/sugerencia relacionada con la camiseta de entrenamiento del Club Atleti de futbol. Específicamente, el usuario pide que se la traigan para venderla en las tiendas de Brasil.

Concretamente los 3 mensajes hacen referencia a ideas, peticiones, sugerencias y elogios de los usuarios hacia las camisetas de la marca, corroborando de nuevo la importancia de practicar la escucha activa en los medios sociales para poder identificar y entender mejor las necesidades de los clientes.

Con relación a los tuits relacionados con la palabra "*tênis*" (zapatilla), el tuit 2a define las zapatillas de la marca como maravillosas y se queja al mismo tiempo de no tener las condiciones económicas para adquirirlas. El mensaje 2b es bastante interesante, ya que el usuario en un primer momento pregunta si las zapatillas están hechas con tejido de vestido de hadas de algún bosque encantado. Lo que en un primer momento parece ser un cumplido a la marca. Sin embargo, cuando buscamos este mensaje en Twitter, se revela un segundo mensaje complementario donde el usuario afirma que, incluso estando hechas de este material, no se justifica el alto precio del producto. Nike, no se ha manifestado sobre el tema.

Nótese que los mensajes 2a y 2b, aunque clasificados correctamente como positivos, tras una interpretación subjetiva, se puede identificar un reclamo relacionado a los precios de los productos de Nike en Brasil. Por lo que se sugiere, que debería de ser un tema para ser considerado por

parte de Nike en Brasil ampliando las investigaciones sobre este tema en la red.

En el tuit 2c el usuario dice haberse quedado sin palabras al ver las zapatillas Jordan. Se refiere a ellas con la expresión "son de sueño" e indica que seguirá haciendo publicidad y compartiendo cosas sobre los productos de la marca en sus medios sociales. Este tipo de relato indica y refuerza lo que la marca hace bien. **Relatos como este dan credibilidad y podrían ser utilizados en campañas publicitarias y acciones de marketing que exploten esas recomendaciones y refuercen los aspectos en los cuales la marca destaca según sus clientes.**

Finalmente, respecto a los tuits que incluyen la palabra "*patrocina*", en el mensaje 3a, el usuario de una manera muy creativa utiliza la expresión "*gentileza genera gentileza*" para indicar que si la marca le patrocinara generaría gentileza. En el tuit 3b el usuario hace una foto para demostrar lo bien que funciona la tecnología Dri-fit de su camiseta después de que le lloviera encima. Al final aprovecha la oportunidad para pedir patrocinio a la marca. Por último, en el tuit 3c el usuario pide que la marca le patrocine pues lleva el logotipo de la marca en la piel, resultado de un golpe (Figura 72). **Estos tuits son muy interesantes puesto que, al pedir patrocinio de**

Figura 72. Mensaje y foto relacionadas al tuit 3c de la Tabla 26.

Fuente: https://twitter.com/MaariFletcher_/status/781237921017987072/photo/1 (acceso el 07/08/2019).

Nike utilizando la red social, los usuarios se hacen valer de toda su creatividad revelando ideas *(insights)* que podrían ser utilizadas para la elaboración de campañas promocionales diversas. En el tuit 3a, por ejemplo, el usuario utiliza un creativo juego de palabras recurriendo al apelo emocional y social que trae consigo la palabra "*gentileza*". En el 3b, intenta intercambiar publicidad por patrocinio, y finalmente en el tuit 3c el internauta reclama patrocinio de la marca porque según él, lleva su logotipo en el cuerpo.

- **Mensajes negativos: @Nike_Spain**

Figura 73. Nube de frecuencia de palabras en base a los tuits clasificados como negativos de la muestra @Nike_Spain.

Fuente: Elaboración propia desde LOGOS.

En la muestra de mensajes negativos de @Nike_Spain, las 3 palabras de mayor frecuencia/relevancia (Figura 73) son "*camiseta*" (63), "*tienda*" *(28) y "zapatillas"* (27). Nótese que estas palabras están directamente relacionadas con dos productos y con puntos de ventas de la marca. **Otro fenómeno interesante es que las palabras "camiseta" y "zapatillas" que salen en este listado, también encabezan el listado de palabras positivas. Esto puede ser un indicativo de que los temas relacionados con las camisetas y zapatillas se encuentran entre los más**

relevantes cuando se habla de los productos de Nike en Twitter, sea para elogiar, quejarse, opinar o simplemente aportar alguna sugerencia, siendo según los datos la palabra *"camiseta"* más relevante que *"zapatillas"*. Como puede observarse en la figura anterior, también aparecen algunas palabras de connotación negativa tales como "*mal" (20), "fea" (12) y "malas" (13),* que corroboran la correcta aplicación de las técnicas y procesos de Análisis de Sentimientos por parte de LOGOS.

A continuación, se muestran a modo de ejemplo mensajes relacionados con las 3 palabras más frecuentes (Tabla 35).

Tabla 35. Listado de tuits relacionados con las tres palabras más frecuentes en los tuits clasificados como negativos de la muestra de @Nike_Spain.

Palabra		Tuit
Camiseta	1a	@25p @FCBarcelona @Nike_Spain Ni idea. No parece lógico que no puedas ponerte el nombre de tu ídolo en la camiseta de tu equipo
	1b	No hace falta que hagáis cosas raras para la camiseta del Sporting @Nike_Spain ,simplemente respetar la historia
	1c	@Aylenmadrid9 @Nike_Spain Si el traje de entrenamiento es asiiii. Me temo lo peor con la camiseta del GLORIOSO ATLETICO DE MADRID.
Tienda	2a	Estoy en la tienda de @Nike_Spain en la Gran Vía de Madrid y no tienen nada del @Atleti. Nada. Ni una puñetera camiseta. No existe.
	2b	Es curioso que, siendo de la misma marca, -@Nike_Spain- los precios de la tienda del @Atleti sean mucho mayores que los del @FCBarcelona_es.
	2c	@RodolfoHache @LaLiga @Nike_Spain @Nike Parece el típico balón de plástico que compras en la tienda de los chinos de la esquina.
Zapatillas	3a	@Uaretheoneblog @ASOS_ES @Nike_Spain y tan en serio... Ahora tengo unas zapatillas de 90€ para tirar a la basura... ??
	3b	@Nike_Spain Segun Nike unas zapatillas que apenas se pusieron es normal que le salga un agujero. Ni corrí con ellas https://t.co/slkXVG9bgw
	3c	Pedir unas zapatillas y que te lleguen con la caja con la etiqueta fotocopiada u medio deformada, bastante desconten… https://t.co/rmYDgJrfCu

Fuente: Elaboración propia desde LOGOS.

En relación con la palabra *"camiseta"* se puede observar en el tuit 1a un reclamo por parte del usuario que manifiesta su descontento al no poder personalizar la camisa de su equipo con el nombre de su ídolo. **Este mensaje informa a la marca sobre la falta de información objetiva en su página web con relación a las camisas personalizadas, un producto premium que merece de gran atención. Para solucionarlo, una de las opciones sería que la marca indicase en su web de personalización de camisetas los motivos por los que no se pueden utilizar determinados nombres en estos productos**. En el tuit 1b, el usuario pide que la marca respete la historia del Sporting FC y que no haga "cosas raras" en la camiseta del equipo. El mensaje 1c, también está relacionado con la camiseta de un equipo. El usuario manifiesta su descontento con la equipación de entrenamiento del F.C. Atlético de Madrid y proyecta su frustración respecto a las futuras camisetas de la equipación oficial del club.

Respecto a la palabra "*tienda*", el mensaje 2a informa a Nike sobre la falta de camisetas del FC Atleti en la tienda de Nike ubicada en la Gran Vía de Madrid. En el tuit 2b, el usuario hace una reflexión sobre la diferencia de precio que existe entre las tiendas de Nike para los productos del Atleti y del F.C. Barcelona. En el tuit 2c el usuario critica el aspecto del balón de Nike y lo compara con un balón de plástico de las tiendas de productos de China. **Nótese que en estos mensajes el usuario actúa como un informante de posibles problemas en las tiendas de la marca tales como la falta de productos, precios incongruentes entre tiendas y productos que le parecen de muy mala calidad. Todo eso de manera gratuita.**

Figura 74. Mensaje y foto relacionadas al tuit 3b de la Tabla 27.

Fuente: https://twitter.com/mazingerastur/status/749963631207604225 (acceso el 07/08/2019).

Finalmente, relacionado con la palabra "*zapatillas*", en el mensaje 3a el cliente manifiesta su descontento al tener que tirar a la basura unas zapatillas que le costaron 90€. En el tuit 3b (Figura 74), otro cliente se queja de la calidad de las zapatillas de la marca que, con poco uso, ya presentan un agujero en uno de los laterales. Por último, en el mensaje 3c (Figura 75) el cliente comenta sobre la mala calidad de la etiqueta de la caja de sus zapatillas Nike.

Estos tuits relacionados con las zapatillas vendidas por la marca, de manera general informan de posibles problemas de fabricación de modelos específicos, así como en la manipulación del embalaje de las zapatillas compradas en la web.

Este tipo de información, ofrecida gratuitamente por los clientes, dan pistas y permiten a las marcas que practican la escucha activa en la red, identificar posibles problemas e intervenir en estas situaciones lo más pronto posible.

Figura 75. Mensaje y foto relacionadas al tuit 3c de la Tabla 27.

Fuente: https://twitter.com/DanCarter7_/status/781493212238184452 (acceso el 07/08/2019).

- **Mensajes negativos: @nikebrasil**

Del mismo modo que en @Nike_Spain, en la muestra de mensajes negativos de @nikebrasil (Figura 76), las 2 palabras más relevantes según los resultados son *"camisa"* (142) y *"tênis"* (141). **Cabe resaltar que, en este caso, a diferencia de @Nike_Spain donde la frecuencia de la palabra "camiseta" (63) es muy superior a la de la palabra "zapatillas" (27), en la muestra @nikebrasil, su frecuencia es bastante similar, por lo que en Brasil ambas palabras tienen una connotación negativa bastante similar.**

Figura 76. Nube de frecuencia de palabras en base a los tuits clasificados como negativos de la muestra @nikebrasil.

Fuente: Elaboración propia desde LOGOS.

En el tercer lugar de la lista según frecuencia está la palabra "*comprar*" (123), en el cuarto lugar la palabra "*air*" (118), que está relacionada con una línea de productos de la marca, y en quinto la palabra "*site*" (101) que se refiere a la página web de la marca. **El listado de palabras negativas fundamentalmente trae información relacionada con posibles "cosas que la marca no hace bien". En este sentido, es muy interesante atender a estos resultados ya que indican posibles aspectos a mejorar por parte de la marca.**

A continuación, se realiza a modo de ejemplo un breve análisis de las tres primeras palabras identificadas como negativas con mayor frecuencia. No obstante, desde el punto de vista de la marca podría ser interesante y necesario realizar análisis más amplios que revelen el por qué otras palabras como *"quero" (74), "producto" (59), "loja" (59), o "meses" (54)* resultan negativas.

Tabla 36. Listado de tuits relacionados con las tres palabras más frecuentes en los tuits clasificados como negativos de la muestra de @nikebrasil.

Palabra		Tuit
Camisa	1a	@nikebrasil comprei a camisa dos Patriots, paguei super caro e a qualidade deixou a desejar. Na primeira lavagem já ficou zuada!
	1b	Por que todas minhas camisas do @Flamengo da @adidasbrasil soltaram números e patrocínios?As da @nikebrasil e da @OlypikusBR isso ñ ocorria.
	1c	Só acho q já passou da hora da @CBF_Futebol tomar uma atitude e a @nikebrasil fabricar camisas da seleção feminina!
Tênis	2a	Eu queria saber se a garantia de só 3 meses é apenas no site da @nikebrasil ou se o tênis em uma loja física também tem esse tempo ridículo
	2b	Comprei o AIR PEGASUS+ 30 na loja online da @nikebrasil - com menos de 6 meses, o tênis direito abriu um buraco e o tênis esquerdo também
	2c	eu até tento comprar tênis da @nikebrasil, mas daí eu vejo a linha @adidasbrasil...e o preço mais acessível...e compro da adidas
Comprar	3a	Olha a @nikebrasil me enviando email sobre tênis de corrida no valor de R$ 699,90... como se eu pudesse comprar, né
	3b	Tristeza essa vida de não conseguir comprar tênis feminino pois não oferecem meu número. #MulheresTambemCalçam40 @nikebrasil @adidasbrasil
	3c	O maior problema é quando vc tem um cupom de desconto.e hora o site aceita.hora não..valew @nikebrasil pode deixar que deixarei de comprar !

Fuente: Elaboración propia desde LOGOS.

Por tanto, respecto a los mensajes relacionados con las 3 palabras negativas más frecuentes (Tabla 36), el tuit 1a trata de una queja relacionada con la calidad del material de una camiseta de la marca. Concretamente, el usuario afirma que pagó muy caro por una camiseta del equipo "Patriots" que ha quedado dañada con el primer lavado. **Este**

mensaje podría indicar una señal de alerta a la marca sobre posibles problemas en un producto específico. Disponer de informaciones como estas resulta muy útil a la marca que puede encontrar soluciones de manera ágil, evitando problemas en el futuro de mayor proporción.

En el tuit 1b (Figura 77), el usuario se queja de la calidad de una camisa fabricada por la marca de la competencia Adidas y afirma que las camisetas de Nike tienen más calidad. **Este caso, LOGOS identifica correctamente en este tuit mayor carga negativa que positiva, por lo que lo ha clasificado como negativo. Sin embargo, al analizar el texto del mensaje subjetivamente se observa que es un mensaje positivo para la marca, ya que el usuario realiza el comentario negativo sobre su principal competidor. Además, expone cuál es el aspecto específico en el que las camisas de Nike, según su experiencia, son mejores que las de Adidas (no se despegan los números). Otro aspecto que cabe mencionar es que, en los comentarios provenientes de este mensaje en Twitter, otros clientes también se quejaron del mismo problema, lo que puede ser un indicativo de que no sea un problema aislado. En este caso en especial, Adidas no se ha manifestado en la red. Por su parte Nike, de haber identificado este**

Figura 77. Mensaje y foto relacionadas al tuit 1b de la Tabla 28.

Fuente: https://twitter.com/fernandojff3105/status/757857500339118080 (acceso el 07/08/2019).

tuit, podría haber respondido al cliente agradeciendo por su aportación y reforzando el cuidado que tiene Nike en la fabricación de sus camisetas y principalmente con los materiales utilizados.

Finalmente, el mensaje 1c representa un reclamo a la marca para que fabrique y ponga a la venta las camisetas de la selección femenina de futbol de Brasil. **Obsérvese que este tuit trae a la luz un importante reclamo con relación a la falta de una línea de productos específica (camisas de la selección femenina de fútbol), que debería ser considerada por la marca.**

Por su parte, el mensaje 2a representa una crítica al poco tiempo de garantía (3 meses) que ofrece la marca en sus zapatillas. Además, el usuario también pregunta si comprar en la web de la marca o en la tienda física tiene efecto directo sobre el tiempo de la garantía, **indicando quizás que la marca debería profundizar la investigación sobre reclamos como éste en la red. En el caso de ser un reclamo recurrente, ésta podría barajar la posibilidad de indicar más claramente la información sobre la garantía de sus productos**.

En el tuit 2b, el usuario afirma que, con apenas 6 meses de uso, sus zapatillas modelo Air Pegasus+ 30 se han estropeado muchísimo hasta el punto de tener agujeros. **Comentarios de este tipo, caso sean recurrentes, sugieren a la marca la necesidad de reflexionar sobre el porqué del deterioro tan acelerado de sus zapatillas, y la consideración de una posible ampliación de la garantía de al menos algunos de sus productos**. En el tuit 2c el usuario dice que le gustaría comprar zapatillas de Nike, pero al ver las de Adidas a mejor precio acaba por comprar a la competencia.

Este tuit trae a la luz puntos a considerar respecto al precio de sus zapatillas. Quizás una línea más económica de zapatillas pudiera solucionar el problema.

Finalmente, el tuit 3a hace referencia a un correo publicitario de Nike que recibió el cliente. En éste se muestran unas zapatillas a un precio que él no se puede permitir pagar. En el mensaje el cliente deja claro que no debería recibir este tipo de correos ya que lo que le ofrecen está fuera de sus posibilidades económicas. **Este mensaje sugiere dificultades relativas a la segmentación de los clientes de la marca. Caso sean recurrentes, estos tipos de errores, además de no promover la venta al cliente, le genera un sentimiento de frustración. Con una mejor segmentación y una línea más económica de productos (idea proveniente del tuit 2c), la marca podría generar publicidad electrónica más segmentada y con promociones más acordes al poder de compra de cada cliente.**

En el mensaje 3b la clienta se queja de no encontrar zapatillas de su talla número 40 y hace un reclamo a la marca sobre la necesidad de más unidades de zapatillas femeninas de su talla. **Esto podría estar relacionado con posibles problemas en la tienda, problemas de logística o también de demanda de fabricación, que como sean recurrentes deberían ser investigados y considerados por la marca**. Por último, en el tuit 3c, el usuario afirma que dejará de comprar productos de Nike puesto que su cupón de descuento a veces se acepta en la web y a veces no (Figura 78). **Cabe resaltar que en este caso la marca actuó correctamente escuchando al cliente y contactando con él para intentar solucionar el tema.**

Figura 78. Mensaje y foto relacionadas al tuit 3c de la Tabla 28.

Fuente: https://twitter.com/ECAlmeida/status/801245864781983744 (acceso el 07/08/2019).

En resumen, como ha quedado evidente en los epígrafes anteriores, las nubes y listados de frecuencia de palabras han permitido identificar específicamente a qué productos, servicios, campañas publicitarias, entre otros, se refieren los usuarios cuando hablan "positiva o negativamente" de la marca. **Estos análisis permitieron establecer un indicador relacionado con los puntos fuertes y débiles de la empresa desde la perspectiva y vivencia experimental de sus propios clientes. Los resultados demuestran que, por un lado, de los comentarios positivos es posible extraer posibles ideas *(insights)* para acciones de marketing que resalten y promuevan lo que la marca ya hace bien. Por otro lado, los comentarios negativos dan pistas de dónde, según sus clientes, puede haber debilidades relacionadas con la marca y en qué aspectos podrían mejorar. De este modo, las organizaciones podrían, entre otras cosas, actuar para corregir sus errores, anticiparse a problemas y promocionar acciones de marketing para gestionar o cambiar una mala percepción que pueda haber sido generada en los usuarios a través de sus experiencias con los productos y servicios ofrecidos por la organización**.

7.4.1.5. DISPOSITIVOS DE PUBLICACIÓN

Los tres dispositivos de publicación más utilizados por los usuarios de Twitter cuando publican sobre Nike son similares tanto en España, como en Brasil (Figuras 79 y 80). Estos internautas publican mayoritariamente en dispositivos Android: 47,9% (@Nike_Spain) y 47,4% (@nikebrasil). En el segundo lugar de la lista están los dispositivos iPhone de la marca Apple: 31,2% (@Nike_Spain) vs 31,1% (@nikebrasil). Estos resultados son acordes con el *share* de mercado para los sistemas Android y iPhone[24], aunque de manera no proporcional ya que, hoy en día, el sistema Android tiene un *share* de mercado entre un 75% y un 80% y el iOS que se utiliza en los móviles iPhone, entre un 15% y un 20%. **Eso puede ser un indicador delimitador relacionado con aspectos económicos de los usuarios de Nike en Twitter**. Finalmente, el tercer puesto de la lista es para la aplicación de gestión de cuentas en Twitter llamada Twitter Web Client: 13,6% (@Nike_Spain) vs 16,2% (@nikebrasil). **Estas aplicaciones suelen ser utilizadas por profesionales del marketing digital o gestores de medios sociales (*community mannagers)* para gestionar cuentas de carácter más comercial que privado. Lo que pude ser un indicador de la cantidad de mensajes provenientes de cuentas conectadas a empresas que buscan promover, vender e informar sobre sus productos y servicios en la red.**

[24] http://gs.statcounter.com/os-market-share/mobile/worldwide (acceso el 07/08/2019).

Figura 79. Grafica de los diez dispositivos más utilizados en base a los tuits de la muestra @Nike_Spain.

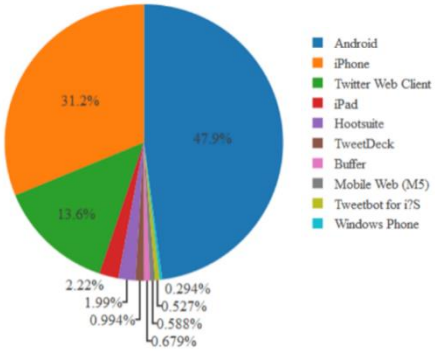
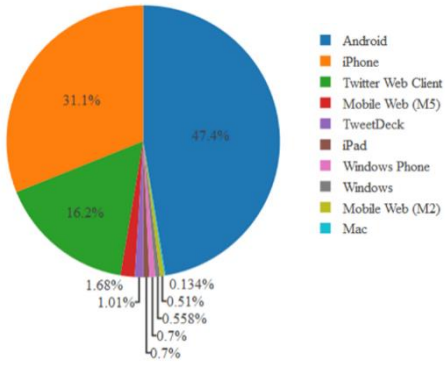


Figura 80. Grafica de los diez dispositivos más utilizados en base a los tuits de la muestra @nikebrasil.



Fuente: Elaboración propia desde LOGOS.

7.4.1.5. USUARIOS MÁS RELEVANTES

Las cuentas más relevantes fueron establecidas considerando dos variables: en primer lugar, el número de seguidores, y en segundo lugar el número de publicaciones. **Este tipo de información que ofrece LOGOS permite definir las cuentas con características de "*influencers*" a las cuales la empresa debería de cuidar y acompañar además de buscar promover acciones de marketing con ellas.**

En las Tablas 37 y 38 se muestran las cinco cuentas con más seguidores según el número de publicaciones para las dos muestras de Nike (@Nike_Spain y @nikebrasil).

Tabla 37. Listado de las cinco cuentas con más seguidores y publicaciones de la muestra @Nike_Spain**.**

Nombre	Tuits Pub.	Seguidores	Seguido por la marca
vegetta777	518	5412105	Sí
FCBarcelona_es	444	13005825	Sí
paugasol	196	7384863	Sí
Atleti	59	3842171	Sí
Cosmopolitan_es	2	1925823	No

Fuente: Elaboración propia desde LOGOS.

Tabla 38. Listado de las cinco cuentas con más seguidores y publicaciones de la muestra @nikebrasil.

Nombre	Tuits Pub.	Seguidores	Seguido por la marca
camilaloures	399	787514	No
cleytu	249	2055096	No
NesterTuits	142	961957	No
IgurRibeiro	39	534832	No
bebeto7	1	588161	No

Fuente: Elaboración propia desde LOGOS.

Cabe resaltar que, en las dos muestras, los primeros puestos están ocupados por youtubers; @vegetta777 en @Nike_Spain y @camilaloures en @nikebrasil; de modo que su capital digital favorece la realización de acciones de marketing con ellos, así como con las demás cuentas que componen el ranking. Por otro lado, es importante también monitorear a las cuentas con estas características (seguirlas en Twitter) por si publican algún mensaje negativo que involucre a la marca. A este respecto, queda patente al observar las Tablas 29 y 30 que @Nike_Spain presenta mejores resultados que @nikebrasil, ya que sigue a cuatro de las cinco cuentas con más seguidores de la muestra.

En @nikebrasil, la situación es bastante distinta y preocupante, puesto que la marca no sigue a ninguna de las cinco cuentas con más seguidores de la muestra, las cuales juntas suman casi cinco millones de usuarios (4927560). **Es decir, si alguna de estas cinco cuentas publica algo negativo sobre Nike, y no mencionan @nikebrasil en el texto del mensaje, casi**

cinco millones de usuarios podrían verlo y la marca no se enteraría de lo sucedido. De modo que Nike debería seguir a estas cuentas en Twitter. Al mismo tiempo sería interesante también llevar a cabo acciones de marketing que las involucre de alguna manera, teniendo en cuenta el alcance proveniente de su capital digital en la red.

7.4.2. INFORME DE MARKETING: SAMSUNG

7.4.2.1. MENSAJES MÁS IMPORTANTES

La frecuencia del indicador de "me gusta" y el número de retuits en los mensajes de la muestra @SamsungEspana, ponen de manifiesto la influencia en la red de youtubers como "@jonanperrea" y "@Alvaro845". Por lo que se sugiere la necesidad por parte de la marca de monitorear a sus cuentas en Twitter, además de la posibilidad de realizar acciones de marketing con esos youtubers. Por otro lado, el número de replicaciones de algunos mensajes revela otra estrategia de marketing que parece funcionar bastante bien en Twitter.

Consiste en la realización de sorteos y promociones que requieren que el usuario replique dicha promoción en su muro de mensajes para poder participar de ellas. De este modo, la marca utiliza los propios usuarios en la red para ampliar el alcance de la promoción/sorteo, a la vez que se hace conocer como empresa en la red. Además de replicar el mensaje, otro requerimiento es que el usuario siga a la cuenta de la empresa en Twitter,

de modo que la marca obtiene más seguidores y amplía su capital digital y alcance en la red.

En la muestra de @SamsungBrasil la situación es bastante similar a la de @SamsungEspana comentada anteriormente. La frecuencia de "me gusta" y el número de retuits indican principalmente la eficacia de acciones de marketing que se basan en el uso de influenciadores en la red. En este caso destacan los siguientes aspectos: a) un "twittero" llamado Luan Lovato, muy conocido por publicar en Twitter viñetas humorísticas (conocidas popularmente como memes).

Concretamente, en su mensaje más popular, él pide un teléfono móvil a Samsung Brasil para poder seguir publicando sus memes sobre las olimpiadas de Rio de Janeiro 2016, evento que se realizaba en aquellos días, ya que su teléfono se había estropeado. **A pesar de que este mensaje tuvo el mayor número de "me gusta" de la muestra, estar entre los cuatro más replicados y ser apoyado por muchos otros usuarios que también solicitaban a Samsung que atendiese sus peticiones, Samsung no se manifestó al respecto. El que la marca no se manifieste ante un Tuit con tanta relevancia en la red puede influir negativamente en la reputación de esta, ya que puede mostrar indiferencia respecto a los requerimientos de su audiencia en línea. Por otro lado, Samsung podría haber aprovechado la situación para promocionarse a través de este twittero, como por ejemplo, darle el teléfono móvil que pidió y que a cambio todos sus memes sobre las olimpiadas llevasen la inscripción "patrocinado por Samsung" y el logotipo de la marca o "producido por el nuevo GalaxyS7",**

modelo de lanzamiento de la marca en estos momentos; b) la cantante de música Pop Anitta, que produjo un video clip con el patrocinio de @SamsungBrasil y; c) la actriz, cantante y exmodelo brasileña Melanie Nunes Fronckowiak, que fue invitada por @SamsungBrasil al evento de lanzamiento del nuevo GalaxyS7 en la ciudad de Nueva York.

7.4.2.2. TEMAS MÁS COMENTADOS: *"TRENDING TOPICS"*

Las etiquetas tipo hashtag revelan los temas más comentados por los internautas en Twitter también llamados *"trending topics"*. En este sentido, en la muestra @SamsungEspana en el periodo de obtención de los datos, se observa que esos temas versan principalmente sobre el teléfono móvil de lanzamiento de la marca "*#galaxys7"*.

Concretamente sobre dos características específicas de este teléfono puestas en evidencia a través de los hashtags "#superusunagua" y "#superusuncámara", y que se refieren a la protección al agua y a la cámara de fotos y videos del GalaxyS7. **Por un lado, la campaña de marketing "superusun" promueve cuatro características innovadoras del GalaxyS7: batería, realidad virtual, protección al agua y cámara de fotos**[25]**. Sin embargo, las que más comentarios generaron en la red fueron las**

[25] https://www.youtube.com/watch?v=Ody_VKog0E4&index=2&list=PLYYUmABqKYgOXfKla0-73EYfjgVFem44z (acceso el 07/08/2019).

relacionadas con la protección al agua y la nueva cámara de fotos. Este dato puede ser un indicador de la preferencia/interés de los usuarios por estas dos características frente a las demás. Finalmente, cabe mencionar que los hashtags "#blackfriday" (790 menciones) y "#beepfriday" (788 menciones) también salen como temas en alza. El "beepfriday" se refiere a la promoción que la tienda Beep Informática promueve en el período de rebajas del Black Friday. Nótese que la pequeña diferencia de menciones entre esos dos hashtags ("#beepfriday" y "#blackfriday") indica el éxito de la tienda informática respecto a la estrategia de divulgación de su promoción, ya que la palabra Black Friday se refiere a un evento a nivel mundial y la etiqueta #blackfriday suele ser utilizada por todas las empresas que comercializan en estas fechas.

Para participar del sorteo, los usuarios debían seguir a la cuenta de la tienda en Twitter y replicar el mensaje de la promoción en su muro. De este modo, utiliza a los propios usuarios para ampliar el alcance de la promoción, aumentar su número de seguidores en Twitter y potenciar su capital digital en la red.

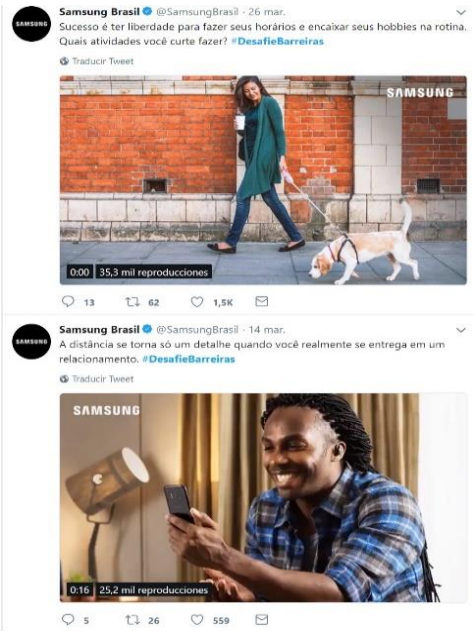
En la muestra @SamsungBrasil, aunque la etiqueta #GalaxyS7 esté presente y ocupe el segundo lugar del listado, en primer lugar, aparece la etiqueta *#desafiebarreiras,* relacionado a una campaña de comunicación desarrollada especialmente para Brasil con el objetivo de estrechar la relación con los consumidores.

Según la directora de marketing de la División de dispositivos móviles de Samsung Brasil, el objetivo de esta campaña era animar a los consumidores

Figura 81. Promoción #desafiebarreiras de @SamsungBrasil ampliada a los juegos olímpicos.

Fuente: https://twitter.com/SamsungBrasil/status/767420749699248128 (acceso el 07/08/2019).

Figura 82. Promoción #desafiebarreiras de @SamsungBrasil.



Fuente: https://twitter.com/SamsungBrasil/status/978371539665072128 (acceso el 07/08/2019).

a romper barreras, actuando lado a lado con ellos y reforzando el compromiso y las relaciones entre el cliente y la marca. Para ella, "el propósito de dicha campaña es contribuir a que todos los clientes alcancen sus metas, independientemente de los desafíos a los que se sometan diariamente[26]" (Figura 81). **Un punto interesante por resaltar es que la**

[26] http://www.clubedecriacao.com.br/ultimas/desafie-barreiras/ (acceso el 07/08/2019).

etiqueta *#desafiebarreiras* terminó por ser utilizado en conjunto con otros hashtags relacionados con los juegos olímpicos realizados en Brasil en 2016 como, por ejemplo, *"#bra, #ouro y #voleibol"* (Figura 82), que ocupan el tercer, cuarto y quinto lugar entre los temas más comentados de la muestra. Cuatro de los cinco hashtags más mencionados por los usuarios, se refieren exclusivamente a campañas publicitarias llevadas a cabo por la marca, lo que indica el éxito de estas campañas en la red.

7.4.2.3. GEOLOCALIZACIÓN

Los resultados de geolocalización obtenidos a partir de los mensajes que componen las dos muestras de Samsung (@SamsungEspana y @SamsungBrasil), indican una muy baja adhesión a este recurso por parte de los usuarios. En la muestra @SamsungEspana, se han podido geolocalizar tan solo 8 mensajes (0,18% de la muestra) y en la muestra @SamsungBrasil apenas 88 (0,38% de la muestra). De este modo, aunque LOGOS identifica y geolocaliza en el mapa a estos mensajes (ej. Figuras 83 y 84), el bajo volumen de tuits posibles de localizar no es representativo.

En este sentido, así como en el informe de Nike que presenta resultados similares a los de Samsung, esta situación puede ser resuelta con acciones de marketing dirigidas por la marca que fomenten la activación de la geolocalización por parte de los clientes, utilizando por ejemplo elementos de gamificación que estimulen el "*check-in*" de los usuarios en los establecimientos de la marca. Cuantos más *"check-in",* más

importancia el usuario adquiere en el juego. Como recompensa, se podría ofrecer algún tipo de premio como ofertas en productos o en establecimientos específicos. Esto ampliaría el número de mensajes geolocalizables y como consecuencia se generarían más datos y análisis más consistentes.

Figura 83. Ubicación del primer mensaje geolocalizable de la muestra de tuits de @Nike_Spain.

Fuente: Elaboración propia desde LOGOS.

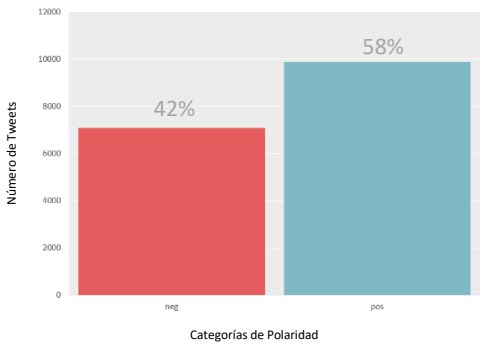
Figura 84. Ubicación del segundo mensaje geolocalizable de la muestra de tuits de @nikebrasil.

Fuente: Elaboración propia desde LOGOS.

7.4.2.4. ANÁLISIS DE SENTIMIENTOS

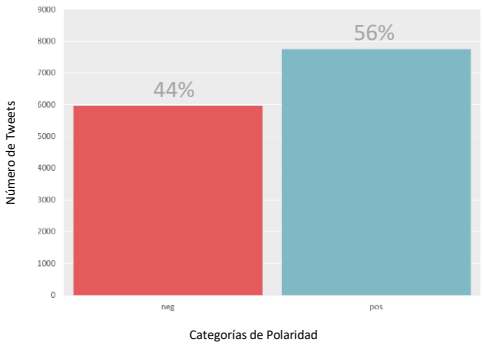
La clasificación de polaridad de los mensajes relativos a la marca Samsung - @SamsungEspana (58% positivos vs 42% negativos) y @SamsungBrasil (56% positivos vs 44% negativos) -, indican en cada caso una superioridad de los comentarios positivos frente a los negativos (Figuras 85 y 86). Estos datos ofrecen una perspectiva sobre la dirección general de los comentarios relacionados a Samsung en Twitter tanto para España como para Brasil.

Figura 85. Grafica de la pestaña Sentiment de la muestra de tuits de @SamsungEspana.



Fuente: Elaboración propia desde LOGOS.

Figura 86. Grafica de la pestaña Sentiment de la muestra de tuits de @SamsungBrasil.



Fuente: Elaboración propia desde LOGOS.

- **Mensajes positivos: @SamsungEspana**

En los mensajes positivos de la muestra @SamsungEspana (Figura 87), la palabra de mayor frecuencia/relevancia es *"gracias"* (1026), que puede indicar, además de una actitud positiva de los usuarios hacia la marca, la correcta clasificación de estos mensajes por parte del algoritmo. En el segundo lugar del ranking, con una frecuencia aproximadamente 4 veces menor que la palabra "gracias", está la palabra "*galaxys7*" (960). A

Figura 87. Nube de frecuencia de palabras en base a los tuits clasificados como positivos de la muestra @SamsungEspana.

Fuente: Elaboración propia desde LOGOS .

continuación, aparece la palabra "*móvil*" (643) relacionada directamente a una línea de productos especifica de la marca. Finalmente aparecen las palabras "*blackfriday*" (562) y "*gears*" (421) que hacen referencia respectivamente a la promoción anual Blackfriday y a las gafas de realidad virtual de la marca.

En la Tabla 39, se listan tres mensajes relacionados con cada una de las palabras que ocupan las tres primeras posiciones del ranking. A continuación, analizamos cada uno de estos mensajes con el fin de identificar insights que ayuden en la toma de decisiones.

Tabla 39. Listado de tuits relacionados con las tres palabras más frecuentes en los tuits clasificados como positivos de la muestra de @SamsungEspana.

Palabra		Tuit
gracias	1a	@SamsungEspana buenos dias! Queria saber cuando se actualizaran las samsung galaxy tab s2 a android M. Gracias
	1b	Las fotos de la #LimpiezaDeCostas fueron hechas con una gran cámara. Gracias @SamsungEspana. #VoluntariosEcomar
	1c	Gracias @SamsungEspana por tener cargadores en los aeropuertos ❤
Galaxys7	2a	@aitor_gl10: Estoy enamorado de mi @SamsungEspana #GalaxyS7edge
	2b	@SamsungEspana que ganas tengo de probar el #SamsungGalaxyS7Edge y el #galaxynote7
	2c	Espectacular foto tomada a las 23h en #railaybeach en #Krabi Tailandia con @SamsungEspana #galaxys7 en modo PRO https://t.co/g8KobWkp8d
móvil	3a	Genial Arturo Fernández...los beneficios del servicio de pago móvil @SamsungEspana Pay: https://t.co/xp19oykSDw https://t.co/RZ1yb6eNjK
	3b	@SamsungEspana Dios necesito este teléfono. Cada móvil nuevo que sacáis me dejáis flipando. ??
	3c	@SamsungEspana gracias por crear el móvil perfecto ????

Fuente: Elaboración propia desde LOGOS.

En relación con la palabra *"gracias"*, en el tuit 1a el usuario solicita información sobre la fecha de una actualización de software para su Tablet

modelo S2 de la marca. En el tuit 1b, la Fundación Ecomar da la gracias a Samsung por haber cedido las cámaras de fotos con las cuales se fotografió la acción voluntaria #LimpiezaDeCostas realizada con los niños de la fundación.

Acciones como esta, además de tener un coste pequeño para la marca, ayudan y son muy bien vistas por la comunidad. En este caso en concreto, quizás la marca podría haber solicitado poner en las fotos una pequeña indicación de que estas fueron sacadas desde una cámara Samsung y así promocionar que contribuyeron con el proyecto.

En el tuit 1c, el usuario da las gracias a la marca por poner a disposición cargadores de móviles en los aeropuertos de manera gratuita. **Este mensaje demuestra el éxito de esta acción de Samsung en los aeropuertos, por lo que la marca podría investigar la recurrencia de mensajes relacionados con esta acción con el objetivo de considerar su expansión, así como invertir en publicidad relacionada a esta acción.**

En el tuit 2a, el usuario manifiesta de manera muy positiva su "enamoramiento" por los GalaxyS7 Edge fabricados y comercializados por la empresa. En el mensaje 2b el usuario deja claro las ganas que tiene de probar el modelo GalaxyS7 Edge y el Galaxynote7.

Finalmente, en el mensaje 2c, el usuario publica una foto realizada con un GalaxyS7 en la playa de Railaybeach en Krabi, Tailandia, utilizando el modo PRO de la cámara del móvil (Figura 88). **Esos mensajes demuestran la actitud positiva de los usuarios respecto al modelo GalaxyS7 Edge y**

también a la calidad de su cámara de fotos. A este respecto, una idea *(insight)* que podría ser considerado, se relaciona con la realización de una acción animando a los usuarios a que publiquen en la red, bajo una etiqueta específica (ej. #myS7photo), sus mejores fotos tomadas con la cámara del GalaxyS7 utilizando el modo automático y el modo profesional. Acciones como éstas, podrían además de estrechar relaciones con los clientes, ayudan a generar contenido en la red. Además, con eso, la marca podría disponer de un repositorio de fotos inéditas y únicas que servirían para futuras campañas publicitarias sobre la calidad de la cámara del GalaxyS7.

Figura 88. Mensaje y foto relacionadas al tuit 2c de la Tabla 31**.**

Fuente: https://twitter.com/gemaesantiago/status/772472184879001600 (acceso el 07/08/2019).

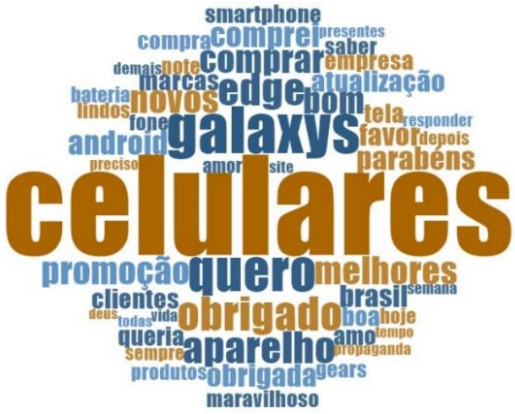
Por último, respecto a los tuits que contienen la palabra “móvil”, en el mensaje 3a, la marca lanza un divertido vídeo tutorial de Samsung Pay protagonizado por Arturo Fernández. El conocido actor español de amplia trayectoria teatral y cinematográfica muestra lo seguro y fácil que es utilizar el nuevo servicio de pago móvil de la marca. El tuit 3b, versa sobre lo “flipado” que se queda el usuario con cada nuevo modelo de móvil que saca la marca. Finalmente, en el tuit 3c el usuario agradece a la marca la creación de móviles tan perfectos**. Estos dos últimos mensajes indican la**

satisfacción de los usuarios con los dispositivos móviles que la marca saca cada año al mercado. Estos datos pueden ser un indicativo para la compañía de que van por buen camino y una confirmación de que, según sus usuarios, lo están haciendo bien.

- **Mensajes positivos: @SamsungBrasil**

Respecto a las palabras más frecuentes en la muestra de mensajes positivos de @SamsungBrasil (Figura 89), se observan algunas similitudes con la muestra @SamsungEspana. Así, las palabras *"celulares" (*705) *y "galaxys7"* (328) encabezan la lista. **Otro dato que merece especial atención es la presencia de la palabra "*quero*" (283) que hace referencia a desear/apetecer.**

Figura 89. Nube de frecuencia de palabras en base a los tuits clasificados como positivos de la muestra @SamsungBrasil.



Fuente: Elaboración propia desde LOGOS.

La Tabla 40 muestra ejemplos concretos de mensajes relacionados con las tres palabras más frecuentes de la muestra (*celulares, galaxys7* y *quero*).

Tabla 40. Listado de tuits relacionados con las tres palabras más frecuentes en los tuits clasificados como positivos de la muestra de @SamsungBrasil.

Palabra		Tuit
celular	1a	@SamsungBrasil Melhores celulares, sem dúvida
	1b	boa tarde preciso de auxílio para desbloquear o aparelho de celular s através da digital ele bloqueou e não permite desblo
	1c	obrigada por terem criado um celular tão lindo e fofo amando e descobrindo meu novo samsungs edge
galaxys7	2a	@SamsungBrasil Minha Amada samsung sua linda faça uma promoção pra gente poder ganhar o #GalaxyS7 pq pra comprar na raça tá difícil amores
	2b	O #GalaxyS7 possue a melhor câmera dos smartphones! Já quero trocar meu A5 @SamsungMobileBR @SamsungBrasil
	2c	Haja bateria quando a conversa é boa. #GalaxyS7 https://t.co/EUPoZKCt5c
quero	3a	@SamsungBrasil Eu sei disso! Estou apaixonada por ele ❤️ Quero muito'
	3b	Eu quero um SamsungS7 Edge. Preciso pagar meus cartões e ver se consigo comprar até o final do ano. @SamsungBrasil olhai por mim.
	3c	@melfronckowiak @SamsungBrasil quero essa câmera

Fuente: Elaboración propia desde LOGOS.

Según la Tabla 40, en el tuit 1a, el usuario muestra su satisfacción con los móviles de la marca cuando dice que son los mejores móviles sin lugar a duda. En el mensaje 1b, el usuario pide ayuda para realizar el desbloqueo del móvil a través de la huella dactilar ya que no le permite desbloquear su móvil. **Este mensaje, además de ser importante al representar una demanda de un cliente y, en el caso de ser recurrente, un posible problema con el sistema de huella dactilar de un modelo específico pone de manifiesto de nuevo un error en la clasificación por parte del algoritmo. Errores de este tipo pueden ocurrir debido a que los algoritmos no aciertan siempre. Concretamente como ha quedado patente en el estudio llevado a cabo en el Capítulo 5 de esta tesis, los algoritmos por general tienden a clasificar correctamente entre un 70% y 85% de las veces.**

En el tuit 1c, la usuaria da las gracias a la marca por fabricar algo tan guapo y dice que está *"amando"* descubrir su nuevo Samsung Edge.

En el tuit 2a, el usuario utiliza un tono muy amable para pedir a la marca que haga una promoción para la compra del GalaxyS7 Edge, pues no logra acumular la cantidad de dinero necesaria para comprarlo. **En este caso se muestra explícitamente un reclamo de un usuario para que la marca haga promociones para el galaxys7 Edge, de modo que ésta, una vez que identifique que se trate de un reclamo recurrente, podría considerar algunas promociones o facilidades específicas para Brasil con el objetivo de atender a las características particulares de compra del mercado brasileño.**

En el mensaje 2b, el usuario afirma que el GalaxyS7 tiene la mejor cámara de fotos y que pretende cambiar su móvil modelo A5 al GalaxyS7 lo antes posible. El tuit 2c, resalta la calidad de la batería del GalaxyS7 y S7 Edge en la medida que el usuario dice que está hecha para aguantar largas conversaciones. **Estos dos últimos tuits ponen de manifiesto según los usuarios aspectos muy positivos de la marca respecto a dos características específicas del Galaxy S7 y S7 Edge, la calidad de la cámara y la duración de la batería. Estos relatos ofrecen retroalimentación (*feedback)* a la marca, revelando características que el cliente valora y desea en un teléfono móvil y que por tanto la marca podría tener en cuenta como aspectos importantes para sus próximos modelos.**

Figura 90. Mensaje y foto relacionadas al tuit 3c de la Tabla 32.

Fuente: https://twitter.com/wiakzei/status/760264347436388352 (acceso el 07/08/2019).

Finalmente, con relación a la palabra "*quero*", en el tuit 3a la usuaria dice estar enamorada del GalaxyS7 Edge y que desea mucho tenerlo. El tuit 3b es bastante similar puesto que el usuario dice desear mucho un Galaxy S7 Edge, pero que antes necesita quitar sus deudas con las tarjetas de crédito para poder comprar el móvil. **Estos dos tuits, refuerzan la necesidad de Samsung de considerar promociones y facilidades de compra específicas para el mercado de Brasil ya que los comentarios en este sentido son recurrentes.** Finalmente, en el tuit 3c, el usuario afirma que necesita tener la cámara de fotos del GalaxyS7 Edge. **Este tuit refuerza el éxito de la cámara de fotos del Galaxys7 Edge, un cumplido recurrente entre los clientes. Otro aspecto interesante es que el comentario se da a partir de una foto proveniente de una acción de marketing de Samsung con la actriz Mer Fronckowiak que fue invitada por la marca a participar del lanzamiento del Galaxys7 en la ciudad de Nueva York (Figura 90).**

- **Mensajes negativos: @SamsungEspana**

En la muestra de mensajes negativos de @SamsungEspana, se observa que tres de las palabras de mayor frecuencia/relevancia (Figura 91) tratan

de temas conexos: "*Galaxy*" (621), "*edge*" (452) y "*móvil*" (318). **Se puede considerar como indicativo de que, en el período de análisis, esos son los temas que más incomodan a los usuarios de Samsung. Otro fenómeno interesante es que la palabra "galaxy" también es una de las que encabezan el listado de palabras positivas. Este resultado puede sugerir que los temas relacionados con el móvil modelo Galaxy son los más relevantes para los usuarios de la marca en Twitter, los cuales hablan tanto positiva como negativamente sobre este producto.**

Figura 91. Nube de frecuencia de palabras en base a los tuits clasificados como negativos de la muestra @SamsungEspana.

Fuente: Elaboración propia desde LOGOS.

Respecto a los mensajes relacionados con las tres palabras más frecuentes (Tabla 41), en el tuit 1a el usuario afirma que le gustaba más un modelo anterior llamado Galaxy S que los que la marca comercializa hoy día. **En esta afirmación el cliente da a entender que la calidad de los productos de Samsung ha empeorado con el paso del tiempo. Esto abre precedente para que la marca investigue junto al cliente cuáles son los motivos de su preferencia por los móviles antiguos a fin de identificar alguna característica que pueda ser implementada en los próximos modelos de móviles de la marca.**

Tabla 41. Listado de tuits relacionados con las tres palabras más frecuentes en los tuits clasificados como positivos de la muestra de @SamsungEspana.

Palabra		Tuit
galaxy	1a	@SamsungEspana yo quiero compra el samsung galaxy S, el primero que salió y no estas cosas que sacais ahora
	1b	#GalaxyS7edge @SamsungEspana Increible movil 800€ d repente Screen parpadea y KO. N SAT 2 semanas y debo pagar móvil de reemplazo
	1c	#SuperUsunCámara @SamsungEspana estaría bien ver que tal se ve en entornos oscuros #GalaxyS7
edge	2a	@SamsungEspana tengo mi edge + 32gb con 3 meses uso averiado en su servicio tecnico 20 dias y no me dan respuesta.Ruego una solución YA!
	2b	@SamsungEspana pues el edge es la polla no se que queréis que os diga la usun esta raya un poco y no hace gracia pero el móvil es la polla
	2c	@SamsungEspana @SamsungResponde Intenté varias veces añadir dos tarjetas y no las admite. Tengo s7 edge. Pongo todos datos correctos.
móvil	3a	Le tengo cierto cariño a @SamsungEspana, pero pagar 600€ por un móvil, y que el serv. técnico tarde 15 días en repararlo, es para pensárselo.
	3b	@LG_ES Ha llegado mi nuevo móvil, Primera vez en LG, espero mejor experiencia que con @SamsungEspana #vergonzoso https://t.co/cF1zXObNUS
	3c	VAIS DE ALTA CALIDAD EN MÓVILES Y SOIS LA PUTA MAYOR BASURA, PEOR MÓVIL NO HE COMPRADO EN MI VIDA @Samsungespana ME CAGO EN VUESTROS MUERTOS

Fuente: Elaboración propia desde LOGOS.

En el tuit 1b, el usuario se queja de tener que pagar por el móvil de reemplazo de su Galaxy S7 Edge de 800€, que ha presentado problemas de parpadeo en la pantalla. **Obsérvese que el usuario no se queja específicamente sobre el problema de parpadeo de la pantalla y si de las condiciones que el servicio técnico le impone para el teléfono de reemplazo. Esta situación pone de manifiesto una cuestión importante que está relacionada con terceros, de modo que, en el caso de ser un reclame recurrente, Samsung debería considerar revisar tales condiciones a nivel nacional junto a estos terceros.**

En el mensaje 1c, el usuario hace un reclamo a la marca para que muestre cómo se comporta la cámara del Galaxy S7 en ambientes oscuros. **Reclamos**

como este podrían ser solucionados con alguna acción en la línea como la comentada anteriormente sobre la posibilidad de desarrollar un repositorio de fotos inéditas y únicas, tanto diurnas como nocturnas, tomadas con la cámara del GalaxyS7.

En el mensaje 2a, el cliente exige una solución rápida por parte de la marca para su problema. Según el usuario, su Galaxy S7 Edge con apenas tres meses de uso lleva más de 20 días en el servicio técnico y éstos no le dan ninguna respuesta. **Esta situación indica a Samsung un posible problema con el servicio técnico. De modo que se sugiere a la marca que investigue más sobre casos como éste para intentar solucionar el problema lo antes posible.**

En el tuit 2b, el cliente en primer lugar hace un cumplido al Galaxy S7 Edge y luego manifiesta su descontento con la campaña de la marca llamada #superusun que busca promover este modelo en España. Según él, la campaña no "hace gracia" y "raya un poco". **Este mensaje ofrece información directamente a Samsung España respecto a la percepción de este usuario a la campaña #superusun.** En el tuit 2c, el usuario afirma que tiene problemas para introducir el número de sus tarjetas de compra en el Galaxy S7 Edge y pide una solución a la marca.

En el tuit 3a, el cliente afirma que empieza a cuestionar su "cariño" hacia la marca debido a que su móvil lleva más de 15 días en el servicio técnico. En el mensaje 3b, el usuario afirma que ha comparado un móvil LG y espera tener mejores experiencias que con Samsung, utilizando la etiqueta

Figura 92. Mensaje relacionado al tuit 3a de la Tabla 33.

Fuente: https://twitter.com/Belisner/status/740971529937227777 (acceso el 07/08/2019).

"#vergonzoso" (Figura 92). **Cabe mencionar que LG contesta al tuit con un mensaje de bienvenida al cliente. Esta situación puede apuntar una mejor atención al cliente y escucha activa de las redes por parte de la marca de la competencia que Samsung. Aspectos como éstos resaltan la importancia de monitorear la red y oír al cliente en la web de manera activa para que situaciones como esta no sucedan y la marca pueda actuar lo antes posible ante los reclamos de su audiencia.**

Finalmente, en el tuit 3c, el usuario utiliza expresiones bastante duras como "puta mayor basura", "el peor móvil que he comprado en mi vida" y "me cago en vuestros muertos" para quejarse de la calidad de los móviles comercializados por la marca. **Cabe resaltar que en este caso la marca entró en contacto con el cliente para intentar solucionar el problema** (Figura 93).

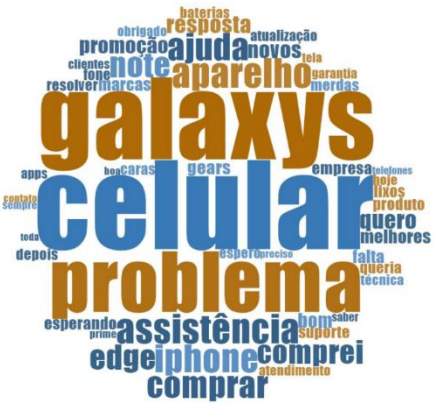
Figura 93. Mensaje relacionado al tuit 3c de la Tabla 33.

Fuente: https://twitter.com/_hambyblake/status/800074126006349828 (acceso el 07/08/2019).

- **Mensajes negativos: @SamsungBrasil**

En la muestra de mensajes negativos de @SamsungBrasil, las dos palabras más frecuentes son *"celular"* (770) y *"galaxys7"* (710), del mismo modo que en la muestra @SamsungEspana (Figura 94). Los puestos tercero y cuarto de la lista son ocupados por las palabras "*problema*" (531) (que indica a priori un correcto funcionamiento del algoritmo en la clasificación de los mensajes negativos) y la palabra "*asistencia*" (267), que hace referencia al servicio técnico ofrecido por Samsung.

Figura 94. Nube de frecuencia de palabras en base a los tuits clasificados como negativos de la muestra @SamsungBrasil.



Fuente: Elaboración propia desde LOGOS.

Como se observa en la Tabla 42, en el tuit 1a el usuario utiliza tono irónico en su discurso para quejarse de la frecuencia con la que su teléfono se queda colgado e indica que Samsung es aún peor de lo que se rumorea. En el tuit 1b, debido a un problema en el internet móvil de su teléfono modelo J5, la clienta dice que irá a demandar a la empresa y se hará rica. En el mensaje 1c, el usuario también se queja de que su móvil se queda colgado, sobre todo después de haber realizado la última actualización de software ofrecida por la marca para su Galaxy S5**. En estos tres casos los clientes indican problemas puntuales con sus teléfonos móviles. Dos mensajes**

indican un mismo problema, el teléfono se queda colgado. Sin embargo, en el caso del Galaxy S5, el cliente especifica que el error empezó tras la actualización de software que realizó. Esto puede ser un indicador de posibles problemas en el proceso de actualización para ese modelo específico, por lo cual sirve como una señal de alerta a la marca para que investigue más sobre esta situación por si llega a ser un problema recurrente.

Otra situación que debería hacer saltar las alarmas son mensajes como el 1b, en que la clienta afirma que demandará judicialmente a la marca por los problemas que viene enfrentando con su Samsung J5.

Tabla 42. Listado de tuits relacionados con las tres palabras más frecuentes en los tuits clasificados como positivos de la muestra de @SamsungBrasil.

Palabra		Tuit
celular	1a	Dizem que celular @SamsungBrasil é ruim e tals. Tudo mentira, o meu só trava umas 500 vezes no dia....
	1b	@SamsungBrasil comprei um celular chamado Galaxy J5 e ñ pega internet móvel nele. Vou botar vcs na justiça e ficar rica
	1c	A pior coisa que eu fiz foi atualizar o software do Galaxy s5 pro 6.0. @SamsungBrasil o celular não para de travar. Desligar sozinho etc...
galaxys7	2a	Maior decepção com a @SamsungBrasil ontem. Descobri que comprei um #GalaxyS7 sem ser Dual Sim. Ninguém falou nada na @ClaroBrasil
	2b	É, @SamsungBrasil Realmente bom! Pena que a principal novidade do #GalaxyS7 não cumpre o prometido: o SamsungPay ficou pro dia de são nunca!
	2c	Não acreditem no #gearv o brinde do #GalaxyS7 que nunca chega. 90 dias esperando @SamsungMobileBR @SamsungBrasil e nada.
problema	3a	Aos seguidores não recomendo @SamsungBrasil problema grave nas TV's e não fazem um recall em respeito aos seus consumidores.#ligaedesliga
	3b	Eu ia elogiar a @SamsungBrasil pela rapidez em me devolver meu celular só que arruma um problema e deixa outro no lugar constrangida
	3c	@SamsungBrasil, amanhã estou indo ao Procon abrir reclamação contra vocês e depois no juizado de pequenas causas. Problemas com dois S7

Fuente: Elaboración propia desde LOGOS.

En el mensaje 2a, el cliente dice haber tenido la mayor decepción de su vida al haber comprado un GalaxyS7 sin Dual Sim (posibilidad de operar con 2 números de teléfono). Según el tuit, el usuario dice que deberían haberle hablado sobre esta particularidad del terminal en la tienda de la Claro, operadora de telefonía móvil donde compró el teléfono (Figura 95). **Nótese que a pesar de que el error probablemente haya sido del dependiente de la tienda Claro, para el cliente, Samsung es el responsable. En este sentido, la marca podría entrar en contacto con el cliente para explicarle que el error no es suyo y ofrecerle una solución junto a su tercero. Además, Samsung podría profundizar la búsqueda por mensajes sobre este tema con el objetivo de identificar su gravedad para quizás internamente considerar reforzar esta información con sus terceros, a modo de evitar que ocurran situaciones similares en el futuro.**

Figura 95. Figura xxx: Mensaje relacionado al tuit 2a de la Tabla 34.

Fuente: https://twitter.com/marcio123rocha/status/733254409577205760 (acceso el 07/08/2019).

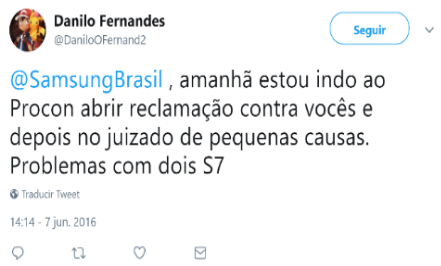
En el tuit 2b la clienta manifiesta su descontento con relación al modelo GalaxyS7 que según ella no trae el servicio Samsung Pay como le fue prometido, de modo que, para ella, la empresa miente ya que no cumple con lo que dice.

En el tuit 2c, el usuario dice que no debemos creer en lo que promete la marca, puesto que le prometieron unas gafas de realidad virtual de regalo en la compra de su GalaxyS7, y hasta el momento, 90 días después, no le habían llegado. **Cabe mencionar que la marca no se manifestó sobre esos mensajes, lo que podría traerle problemas. En dos casos específicos se habla de experiencias negativas vividas por sus clientes los cuales resaltan la incapacidad de la marca para cumplir con lo que promete. Concretamente en el caso del tuit 2b, la usuaria es una cantante con más de 133 mil seguidores hoy en día, lo que acentúa aún más el impacto negativo de su mensaje. A parte, la no manifestación de la marca respecto a esos tuits puede poner de manifiesto la indiferencia de esta ante lo que dicen sus clientes en los medios sociales. Situaciones como estas pueden ser evitadas con el monitoreo y la escucha activa de los medios sociales por parte de la marca.**

En el tuit 3a, el cliente dice a sus seguidores en la red que directamente no compren los televisores de Samsung. Según él, la marca no hace los "*recalls*" necesarios en los dispositivos y no respetan a sus consumidores. **Este tuit es interesante puesto que pone en evidencia problemas específicos relacionados con los *"recalls"* que debería efectuar la marca por si algún producto presenta algún problema de fabricación. En este caso se sugiere a la marca que identifique los productos a los que se refiere el usuario, para poder prevenir problemas futuros mayores**. **Sin embargo, en este caso la marca tampoco ha contestado al cliente en la red**.

En el mensaje 3b, el usuario dice estar molesto por el trato que ha recibido de la marca ya que ha devuelto su móvil con problemas después de una supuesta reparación. **En este mensaje el cliente atribuye a la marca un problema que en verdad es de un tercero, el servicio técnico. De modo que la marca podría pedir al cliente que le dé más información para poder intervenir junto al servicio técnico con el fin de que problemas como éste no vuelvan a pasar. Acciones de este tipo involucran al cliente como parte del proceso de mejora de la marca, el cual se puede sentir importante, útil y valorado.**

Figura 96. Mensaje relacionado al tuit 3c de la Tabla 34.



Fuente: https://twitter.com/DaniloOFernand2/status/740290883753496576 (acceso el 07/08/2019).

Finalmente, en el tuit 3c, el usuario parece bastante irritado con la marca y afirma que irá a denunciar a la empresa al órgano de defensa del consumidor *"Procon"*. Además, irá a demandar a la empresa judicialmente por problemas en dos móviles Galaxy S7. En este caso, la empresa una vez más no se ha pronunciado sobre el tema en la red (Figura 96).

Nótese en este tuit que el cliente tiene dos móviles Galaxy S7, el modelo más caro de la marca comercializado en estos momentos, por lo que estamos ante un consumidor importante. La situación puede ser delicada ya que este importante consumidor que tiene problemas con sus dos móviles no recibe ninguna atención por parte de Samsung, que deja a

entender que la marca no le escucha y/o no le hace caso. Esta es otra situación que vuelve a poner sobre la mesa la importancia vital del monitoreo de los medios sociales hoy en día. En este caso, si la marca hubiese contactado con el cliente, quizás podrían haber llegado a un acuerdo y solucionar el problema; sin embargo, la falta de respuesta puede incentivar que el cliente avance con las demandas judiciales.

7.4.2.5. DISPOSITIVOS DE PUBLICACIÓN

Los tres dispositivos de publicación más utilizados por los usuarios cuando publican sobre Samsung son algo similares tanto en España como en Brasil (Figuras 97 y 98).

Figura 97. Grafica de los diez dispositivos más utilizados en base a los tuits de la muestra @SamsungEspana.

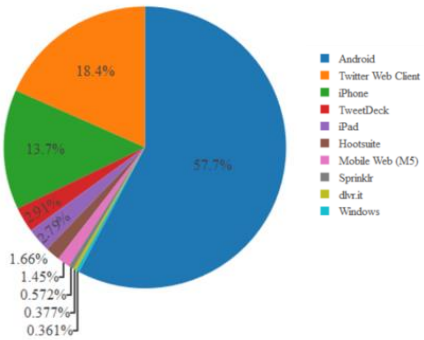
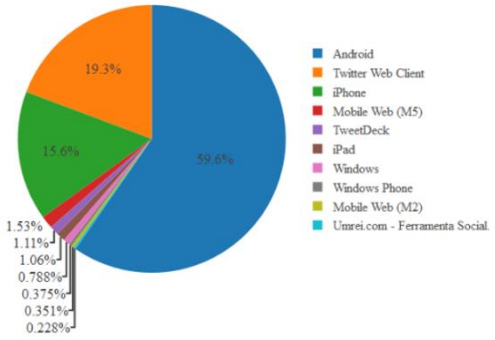


Figura 98. Grafica de los diez dispositivos más utilizados en base a los tuits de la muestra @SamsungBrasil.



Fuente: Elaboración propia desde LOGOS.

Los internautas publican principalmente desde dispositivos Android en @SamsungEspana (57,7%) y en @SamsungBrasil (59,6%). En el segundo

lugar de la lista, están los mensajes publicados desde la aplicación de gestión de cuentas en Twitter llamada Twitter Web Client con un 18,4% en @SamsungEspana y un 19,3% para @SamsungBrasil.

Estas aplicaciones, suelen ser utilizadas por profesionales del marketing digital (*community mannagers)* para gestionar distintas cuentas a la vez. Por este motivo, los tuits publicados por estas herramientas suelen estar conectados a cuentas de carácter más comercial que privado. Este aspecto puede ser un indicador de la cantidad de mensajes provenientes de cuentas comerciales que buscan promover, vender e informar sobre sus productos y servicios en la red y que estén relacionados de alguna manera a Samsung. De modo que la marca puede recurrir a este grupo de mensajes específicos para identificar a estas empresas.

Finalmente, en el tercer puesto de la lista, están los dispositivos iPhone de la marca Apple con un 13,7% en @SamsungEspana y un 15,6% en @SamsungBrasil. Estos resultados son acordes con el *share* de mercado para los sistemas Android y iPhone[27], aunque de manera no proporcional ya que, hoy en día, el sistema Android tiene una participación de mercado entre un 75% y un 80% y el iOS que se utiliza en los móviles iPhone, entre un 15% y un 20%.

[27] http://gs.statcounter.com/os-market-share/mobile/worldwide (acceso el 07/08/2019).

7.4.2.5. CUENTAS Y USUARIOS RELEVANTES

Las cuentas más relevantes fueron establecidas considerando dos variables: número de seguidores y número de publicaciones. **Este tipo de información que ofrece LOGOS permite definir las cuentas con características influenciadoras (*influencers)*, las cuales la empresa debería de cuidar y acompañar además de buscar promover acciones de marketing junto a ellas.** En las Tablas 43 y 44 se muestran las cinco cuentas más relevantes para las dos muestras (@SamsungEspana y @SamsungBrasil). **Cabe resaltar que, en ambas muestras, los primeros puestos están ocupados por cantantes; @pabloalboran en @SamsungEspana y @Anitta en @SamsungBrasil; de modo que su capital digital favorece la realización de acciones de marketing con ellos. Por otro lado, es importante también monitorear a estas cuentas (seguirlas en Twitter) por si publican algún mensaje negativo que involucre a la marca.** A este respecto la marca tanto en España como en Brasil no presenta buenos resultados. En @SamsungBrasil la marca sigue a apenas dos de las cinco cuentas más relevantes.

En @SamsungEspana, la situación es aún peor puesto que la marca sigue solamente a una de las cinco cuentas con más seguidores de la muestra. **De modo que como estas cuentas (no seguidas por la marca) publiquen algo negativo sobre Samsung y no mencionen a la cuenta de la marca en el texto del mensaje, en el caso de @SamsungEspana más de ocho millones de usuarios lo podrían ver, y en el caso de @SamsungBrasil casi once millones de usuarios y la marca no tendría conocimiento del referido mensaje. De modo que Samsung debería seguir a estas cuentas en Twitter**

y si fuera posible, llevar a cabo acciones de marketing que las involucre de alguna manera, teniendo en cuenta el alcance en la red a partir de su capital digital.

Tabla 43. Listado de las cinco cuentas con más seguidores y publicaciones de la muestra @SamsungEspana.

Nombre	Tuits Pub.	Seguidores	Seguido por la marca
mtvspain	103	2685180	Si
CristiPedroche	59	1715187	No
carlosbaute	51	2381014	No
pabloalboran	26	3003634	No
hola	4	1346373	No

Fuente: Elaboración propia desde LOGOS.

Tabla 44. Listado de las cinco cuentas con más seguidores y publicaciones de la muestra @SamsungBrasil.

Nombre	Tuits Pub.	Seguidores	Seguido por la marca
Anitta	3841	6441364	No
wsl	72	2057998	Si
LeiSecaRJ	26	1755377	No
ClaroBrasil	16	5908216	Si
rezende_evil	12	2751956	No

Fuente: Elaboración propia desde LOGOS.

7.5. CONCLUSIONES

En este capítulo hemos mostrado los principales resultados de diversas aplicaciones prácticas de la herramienta LOGOS. Se ha puesto de manifiesto su capacidad para recoger, almacenar y procesar datos de medios sociales en distintos idiomas, proporcionando a su vez de forma gratuita informes sobre los datos que permitan ayudar a las organizaciones y/o particulares en su toma de decisiones. En este caso concreto, para generar los informes que hemos mostrado en este capítulo, el sistema ha extraído los mensajes de la red social Twitter entre mayo de 2016 y abril de 2017, gestionando estos datos por medio de la aplicación de técnicas de Inteligencia Artificial, Minería de Datos, Procesamiento del Lenguaje

Natural y Análisis de Sentimientos, entre otros. Los datos obtenidos han permitido identificar los mensajes de mayor "importancia" (impacto/alcance) en la red, su geolocalización, su polaridad (positiva o negativa), además de revelar los temas más comentados (*trending topics)*, los dispositivos de publicación utilizados por los internautas, así como los usuarios más relevantes, transformando el *Big Data* en *Smart Data*. Esta información presentada en listas, tablas, gráficas, nubes de palabras y otros recursos gráficos, ofrecen al analista una perspectiva amplia de los datos, además de proporcionar una primera lectura más ágil de la información para una toma de decisiones más rápida y precisa.

Finalmente, la posibilidad de guardar los informes en varios formatos (PDF, LaTeX, HTML, RTF, CSV y Excel) favorece el manejo e intercambio de los datos por programas como hojas de cálculo (ampliamente usadas por las PyMEs) como recurso de apoyo. Además, LOGOS también facilita la subida de nuevos grupos de datos generados a partir de muestras previamente descargadas. De este modo, se pueden generar nuevos informes en base a submuestras específicas que aporten perspectivas complementarias sobre un fenómeno específico que se desee observar en los datos.

Los *insights* obtenidos a partir de los análisis realizados por LOGOS son diversos. Se ha puesto de manifiesto que la información obtenida puede ser utilizada para a) idear nuevas acciones de marketing; b) monitorear y medir el éxito de las acciones de marketing ya realizadas y las que aún están en ejecución; c) identificar los productos y servicios de éxito ofrecidos por la

marca y las características que los hacen tan buenos según los clientes; d) identificar puntos débiles en los productos y servicios ofrecidos por la marca y las características que los hacen débiles según los clientes; e) identificar problemas con terceros como la falta de productos en tiendas, prácticas de precios inadecuadas, problemas con la logística y la falta de información en los equipos de ventas; f) identificar y gestionar de manera preventiva posibles problemas tanto en el entorno físico como en el virtual; g) gestionar y prever posibles crisis; h) identificar competidores de la marca, así como sus puntos fuertes o débiles, entre otros.

En este capítulo se han puesto algunos ejemplos ilustrativos a partir de los términos más frecuentemente empleados por los usuarios en el medio social Twitter en relación con la marca y que podrían desvelar aspectos importantes para ésta. La recurrencia de este tipo de mensajes debería ser un factor para tener en cuenta por las marcas con el propósito de dar prioridad a estos asuntos. La herramienta LOGOS ayuda en este proceso identificando los aspectos más destacados dentro del conjunto de textos sujeto a análisis.

Cabe resaltar que este conocimiento ofrecido por LOGOS es obtenido a partir de la escucha activa de los medios sociales, estimulando el marketing relacional y la co-creación por la cual clientes y empresas interactúan de manera efectiva, crean experiencias juntos y consecuentemente generan valor uno para el otro.

Por último y con base a lo expuesto anteriormente, LOGOS ha demostrado su efectividad en las tareas propuestas, atendiendo a su

propósito como sistema multilingüe de inteligencia empresarial que, a través de técnicas de Inteligencia Artificial, Minería de Datos y Análisis de Sentimientos automáticos, actúa de forma gratuita como herramienta de apoyo a la toma de decisiones en las organizaciones.

CAPÍTULO 8: REPOSITORIO DE CONOCIMIENTO DE LA TESIS

8.1. INTRODUCCIÓN

8.2. EL REPOSITORIO

8.1. INTRODUCCIÓN

En este ámbito, en esta tesis doctoral se ha creado un repositorio de conocimiento web del tipo blog con el objetivo de hacer pública la información proveniente de este trabajo. Concretamente, se ha puesto a disposición pública los recursos, estudios publicados, descubrimientos y aportaciones para uso gratuito de la comunidad. Con ello, se busca contribuir con la diseminación del conocimiento práctico y científico.

Además, se pretende que este repositorio no sea un recurso estático y si en constante actualización a medida que las investigaciones provenientes de esta tesis vayan siendo ampliadas. De este modo seguirá cumpliendo con su objetivo principal de aportar conocimiento técnico y científico fundamentado a la ciencia y a la comunidad.

8.2. EL REPOSITORIO

Para la creación del repositorio se ha utilizado la aplicación en línea Wordpress[28] especializada en la creación de sitios web. Como la idea es que sea un recurso vivo y en constante expansión, se ha optado por la modalidad de blog. Esta modalidad de sitio web, permite de manera rápida

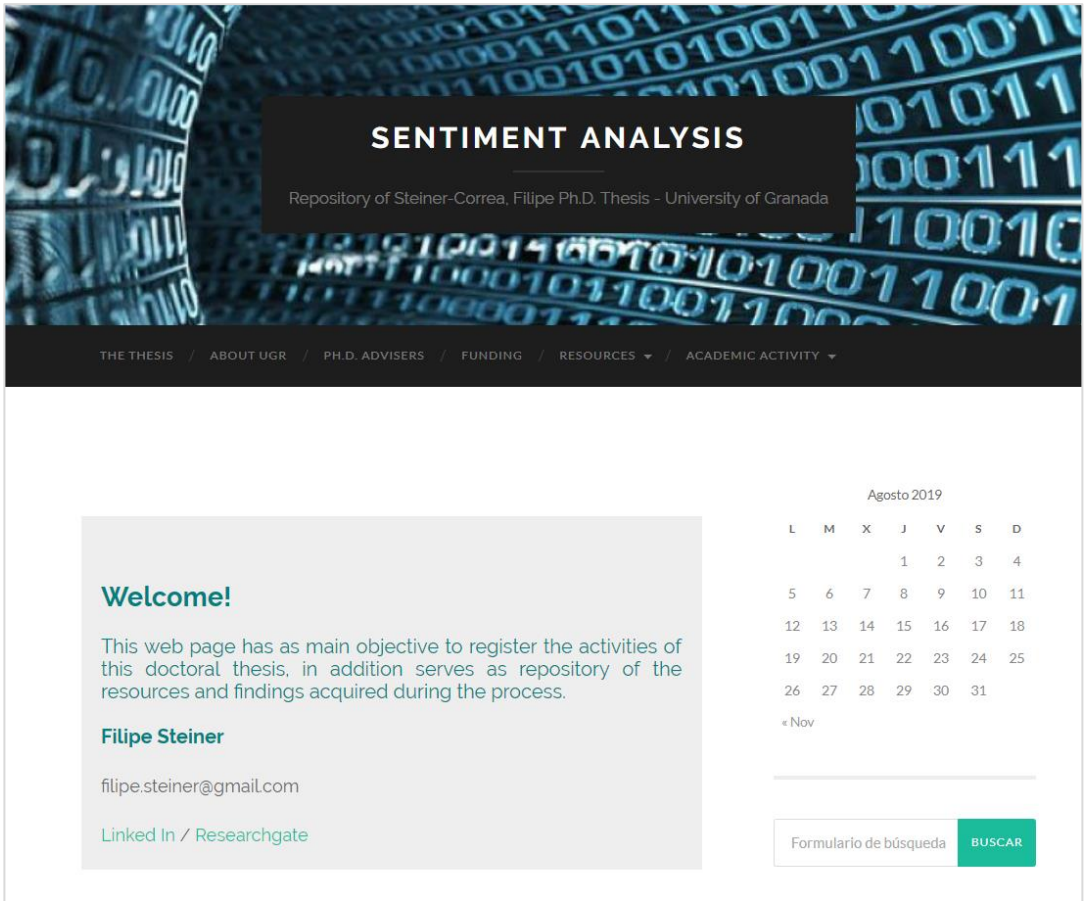
[28] www.wordpress.com/ (acceso el 07/08/2019).

y fácil la interacción y aportación de cualquier usuario de la web por medio de espacios específicos destinados a los comentarios.

El repositorio está hospedado en el servidor Hipatia, el cual está alojado en dependencias del departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada. Su ruta de acceso web es http://hipatia.ugr.es/steiner/.

En la Figura 99 se observa el aspecto visual de la página de inicio del repositorio.

Figura 99. Página de inicio del repositorio de conocimiento.



Fuente: http://hipatia.ugr.es/steiner/.

Como se observa, para que el repositorio tuviera más alcance, se estableció inicialmente el idioma inglés como idioma de publicación pudiendo ser traducido a más idiomas en el futuro.

Relacionado con los menús de opciones, estos de dividen inicialmente en 6 ítems que describiremos en los epígrafes a continuación.

8.2.1. THE THESIS (LA TESIS)

En este menú se ofrece un breve resumen de la tesis incluyendo su título, antecedentes, objetivo general y objetivos específicos además de información sobre el programa de doctorado y la línea de investigación al cual pertenece.

8.2.2. ABOUT UGR (SOBRE LA UGR)

Este menú trata de posicionar al internauta respecto a la Universidad de Granada donde se realiza esta tesis. Se expone información relevante sobre esta institución como las acreditaciones en el *European Higher Education Area (EHEA)* y su posicionamiento en el Ranking de Taiwan, Ranking de Shanghai, *y* finalmente en el Ranking de las Universidades iberoamericanas.

8.2.3. PHD. ADVISERS (SUPERVISORES DE LA TESIS)

En este apartado se ofrece información relativa a los directores de la tesis Dra. María Isabel Viedma del Jesús y Dr. Antonio Gabriel López Herrera.

8.2.4. FUNDING (SUBVENCIÓN)

Esta tesis doctoral ha sido realizada con el apoyo de la beca *"Doutorado Pleno no Exterior"* concedida por el "*Programa Observatório da Educação, de la Coordenação de Aperfeiçoamento de Pessoal de Nível Superior"* de la institución CAPES/Brasil. En este sentido, este menú contiene la información relativa a este órgano conectado al Ministerio de Educación de Brasil así como a la beca en cuestión.

8.2.5. RESOURCES (RECURSOS)

Este espacio del repositorio está reservado a los recursos creados y utilizados en las 4 aportaciones que contienen la tesis. Son ellos:

- ***Redbull Sentiment Dataset*:** se trata de un dataset de sentimientos creado por el autor de esta tesis, clasificado y validado subjetivamente por un conjunto de 152 personas. Este se compone por mensajes provenientes de Twitter en base a la etiqueta *#guivesyouwings* relacionada con la bebida energética RedBull, líder de mercado de bebidas energéticas en la actualidad.

- ***Sentiment Datasets:*** se trata de un compendio de datasets de sentimientos, validados subjetivamente. Este compendio fue utilizado en este trabajo y publicado en revista científica de índice contrastado JCR.

- ***Sentiment Lexicons*****:** son léxicos clasificados según su polaridad y utilizados en distintas técnicas de análisis de sentimientos, como por ejemplo: a) Sentistrength para el idioma inglés (Thelwall, Buckley, & Paltoglou, 2012); b) OP-Lexicon, para el idioma portugués (Souza et al., 2011) y c) iSOL, para el idioma español (Martín-Valdivia et al., 2013), entre otros.

- **`RapidMiner`:** ofrece al internauta información sobre esta herramienta utilizada en la Aportación 2 de esta tesis. En ReapidMiner se construyó un modelo de minería de datos y análisis de sentimientos para identificar el rendimiento - de precisión y tiempo - de 4 algoritmos frecuentes en la literatura.

- **LOGOS – *Open Code*:** esta opción contiene el enlace al repositorio de proyectos informáticos GitHub que alberga el código de programación de `LOGOS` así como los modelos de algoritmos entrenados.

- **Nike / Samsung *Datasets*:** en este apartado están publicados los datasets de sentimientos clasificados por LOGOS de manera automática y utilizados en la Aportación 4 de esta tesis.

8.2.6. ACADEMIC ACTIVITY (ACTIVIDAD ACADÉMICA)

En este menú están listadas las actividades académicas realizadas durante el período del doctorado. En total se dividen tres ítems:

- **Producciones:** se trata de producciones académicas llevadas a cabo durante el doctorado como artículos en revistas web, artículos científicos publicados en revistas de impacto contrastado JCR, seminarios impartidos y un capítulo de libro.

- **Conferencias, Congresos y Seminarios:** se trata de listar la asistencia a eventos académicos diversos como conferencias, congresos y seminarios durante el período de realización del doctorado.

- **Cursos técnicos y científicos:** se trata de documentar la realización de cursos técnicos y científicos realizados durante el período del doctorado.

De este modo, esta tesis doctoral por medio de este repositorio ofrece a toda la sociedad y a la comunidad científica una aportación más, compuesta por recursos y conocimiento decurrentes de muchas horas de investigación

y trabajo para que estas, lo utilicen de manera gratuita con el objetivo de seguir avanzando en los ámbitos comprendidos por este proyecto hecho realidad.

CAPÍTULO 9: PRINCIPALES CONCLUSIONES, IMPLICACIONES Y LIMITACIONES DE LA TESIS, FUTURAS LÍNEAS DE INVESTIGACIÓN, Y LECCIONES APRENDIDAS

9.1. INTRODUCCIÓN

Las diferentes aportaciones de la presente tesis doctoral son interdependientes y persiguen como objetivo principal la creación de un método y un sistema web gratuito de Inteligencia empresarial que, a través de técnicas de Minería de Datos y Análisis de Sentimientos, sirva en distintos idiomas como soporte a las decisiones estratégicas de marketing de las empresas. Con esta finalidad, **la primera aportación de esta tesis doctoral** consistió en reunir y catalogar 25 *datasets* de entrenamiento en múltiples idiomas (inglés, español, portugués, árabe, alemán, italiano y francés), considerados recursos de base en la clasificación de sentimientos por medio de algoritmos automáticos. Este compendio de recursos también fue publicado en forma de artículo científico en la revista Soft Computing - A fusion of Foundations, Methodologies and Applications, con factor de impacto JCR (2.367)[29] disponible en https://link.springer.com/article/10.1007/s00500-017-2766-5.

Como segunda aportación se utilizaron los *datasets* de la primera aportación con el propósito de identificar, tras una exhaustiva revisión de los diferentes algoritmos automáticos de Análisis de Sentimientos, cuál de los cuatro más frecuentemente considerados en la literatura ejecuta las tareas relativas al Análisis de Sentimientos de manera más precisa y rápida

[29] https://www.springer.com/engineering/computational+intelligence+and+complexity/journal/500 (acceso el 07/08/2019).

respecto a cinco idiomas distintos (inglés, español, portugués, alemán e italiano). A continuación, en **la tercera aportación** se promueve el diseño y la construcción del sistema web LOGOS, ideado para ofrecer a las organizaciones (especialmente PyMEs, brasileñas en particular) distintos informes que sirvan de apoyo a la toma de decisiones empresariales en múltiples idiomas y de forma gratuita. Finalmente, la **última aportación** tuvo como objetivo validar dicho sistema para lo que se realizaron cuatro aplicaciones prácticas con LOGOS: dos en el idioma español y dos en el idioma portugués de Brasil asociadas a las multinacionales Nike y Samsung. Este último estudio es el que corrobora la eficiencia de LOGOS como herramienta de apoyo a decisiones de marketing en múltiples idiomas, destacando sus aspectos diferenciales e innovadores respecto a las opciones disponibles en la literatura actual.

A continuación, se presentan las principales conclusiones e implicaciones de esta tesis doctoral para el ámbito académico, para el ámbito práctico de la administración de empresas y también para la sociedad. Finalmente, se discuten las principales limitaciones y las eminentes líneas de investigaciones futuras.

9.2. PRINCIPALES CONCLUSIONES DE LA TESIS

Desde la Web 2.0, también conocida como **Web Social** - caracterizada por la interdependencia de contenidos y múltiples estructuras que comparten distintos recursos con sus usuarios - emergen las **comunidades**

sociales (O'Reilly, 2007). Éstas cambiaron de manera drástica diversos aspectos de las relaciones on line entre consumidores y empresas (Berners-Lee, Fischetti, & Foreword By-Dertouzos, 2000). Para Kotler et al. (2010), esta nueva "Web Social" trajo consigo la evolución del Marketing 2.0 al Marketing 3.0, adaptado a las nuevas tecnologías y conectado a la era de la globalización, la creatividad y la participación y co-creación. Esto explica la evolución del concepto de marketing en armonía con las demandas y transformaciones de la sociedad de consumo.

9.2.1. LA EVOLUCIÓN DEL MARKETING Y SU COTEXTO ACTUAL

En la década de los cincuenta, el marketing se centraba primordialmente en el producto. A partir de la década de los setenta, empezó a centrarse también en el consumidor, de modo que gradualmente y cada vez más los profesionales del marketing buscaban entender a los clientes como seres conscientes de sus elecciones (atendiendo también a sus sentimientos y opiniones). En comparación con el **Marketing 1.0** - cuyo objetivo fundamental es vender un producto o servicio - y el **Marketing 2.0** - que tiene como principal objetivo la fidelización del cliente - el **Marketing 3.0** trabaja para conocer, reconocer e interpretar las aspiraciones y valores de los consumidores, considerando que la comunicación y las estrategias de la empresa pasan a depender del comportamiento que el consumidor asume. Se trata de una tendencia con especial atención en el consumidor, en lo que dice su **mente (Marketing 1.0)**, **su corazón (Marketing 2.0) y su espíritu** (Marketing 3.0) en la medida en que el consumidor trata no solo de elegir

los productos/servicios que mejor satisfagan sus necesidades, sino también aquellos con los que puede contribuir a lograr un mundo mejor (Kotler et al., 2010) (Tabla 45).

Tabla 45. Principales aspectos del Marketing 1.0, 2.0 y 3.0.

	Marketing 1.0	Marketing 2.0	Marketing 3.0
Objetivo	Vender productos	Satisfacer y retener a los consumidores	Hacer de este mundo uno mejor
Fuerzas propulsoras	Revolución industrial	Tecnologías de la información	Nueva ola tecnológica
Percepción del mercado por la empresa	Mercado de masas Consumidores con necesidades físicas	Consumidor más inteligente con mente y corazón	Ser humano integral, con mente, corazón y espíritu
Concepto fundamental de marketing	Desarrollo del producto	Diferenciación Posicionamiento corporativo y del producto	Valores
Directrices de marketing corporativas	Especificaciones del producto	Misión, visión y valores corporativos	Proposiciones de valor
Funcional	Funcional	Funcional y emocional	Funcional, emocional y espiritual
Interacción con los consumidores	Transacciones uno a uno	Relaciones uno a uno	Colaboración entre muchos

Fuente: Kotler et al. (2010), p. 21

El Marketing es un concepto vivo que viene evolucionando desde sus orígenes hasta la actualidad. Concretamente, estas trasformaciones son el resultado principalmente de coyunturas históricas y las demandas sociales del momento. En los últimos tiempos se han podido apreciar algunos cambios importantes como los que se mencionan a continuación (Kotler & Armstrong, 2013):

a) **El entorno económico incierto** que se establece a nivel mundial desde principios del 2008 viene cambiando las prácticas de

consumo. Ante este escenario no basta con que las empresas adecuen sus costes de producción y precios. Para destacarse de sus competidores es necesario que den un paso más en el sentido de cultivar relaciones más cercanas con los clientes por medio de la creación de "valor".

b) **La era digital** que como una avalancha tecnológica trae consigo los avances de las tecnologías de la información. La creciente capilaridad de internet y de los medios sociales están revolucionando las relaciones comerciales y el marketing. En este sentido afloran conceptos como "Marketing one2one" ,"Marketing de medios sociales", "Marketing b2b", "Marketing digital", entre otros.

c) La rápida globalización que brinda a las empresas ampliar su negocio con la apertura a nuevos mercados. Sin embargo, nuevos y diversos mercados demandan distintos públicos que requieren una compleja adaptación en la gestión estratégica. Esta nueva faceta trae consigo conceptos como el de "Marketing internacional o global" entre otros.

d) **Creciente responsabilidad ética y social** que trata sobre cuestiones sociales y principalmente medioambientales. Temas como la ética corporativa y la responsabilidad social y ambiental, asumen un importante rol para las empresas actuales. Estas

nuevas demandas traen consigo conceptos como el “Marketing medioambiental”, “Marketing ecológico”, “Marketing verde” o “Marketing sostenible”, entre otros.

Estos cambios del entorno están dando lugar a nuevas formas y prácticas de marketing. Además, se debe de considerar que estos mismos aspectos también inciden directa o indirectamente en el consumidor. Precisamente, se pueden destacar a dos principales aspectos oriundos de estos cambios (Canales & Hernández, 2013):

1. Las consecuencias de la inestabilidad económica de los últimos años han llevado a los consumidores a actuar con mayor precaución, siendo más precavidos con sus gastos. Además están más recelosos y escépticos con las prácticas de marketing realizadas por las empresas, de modo que la confianza se establece más de manera horizontal que vertical. En otras palabras, hoy en día los consumidores presentan una tendencia a confiar más en lo que dicen otros consumidores que en lo que nos cuentan las empresas. Además, internet y los medios sociales web 2.0, ofrecen a los consumidores ambientes óptimos para el intercambio de información.

2. Los avances tecnológicos y culturales dan paso a nuevos hábitos de consumo y a un nuevos tipos de consumidores que a través de aspectos como la disponibilidad de acceso a la información en cualquier lugar y momento por medio de diversos dispositivos

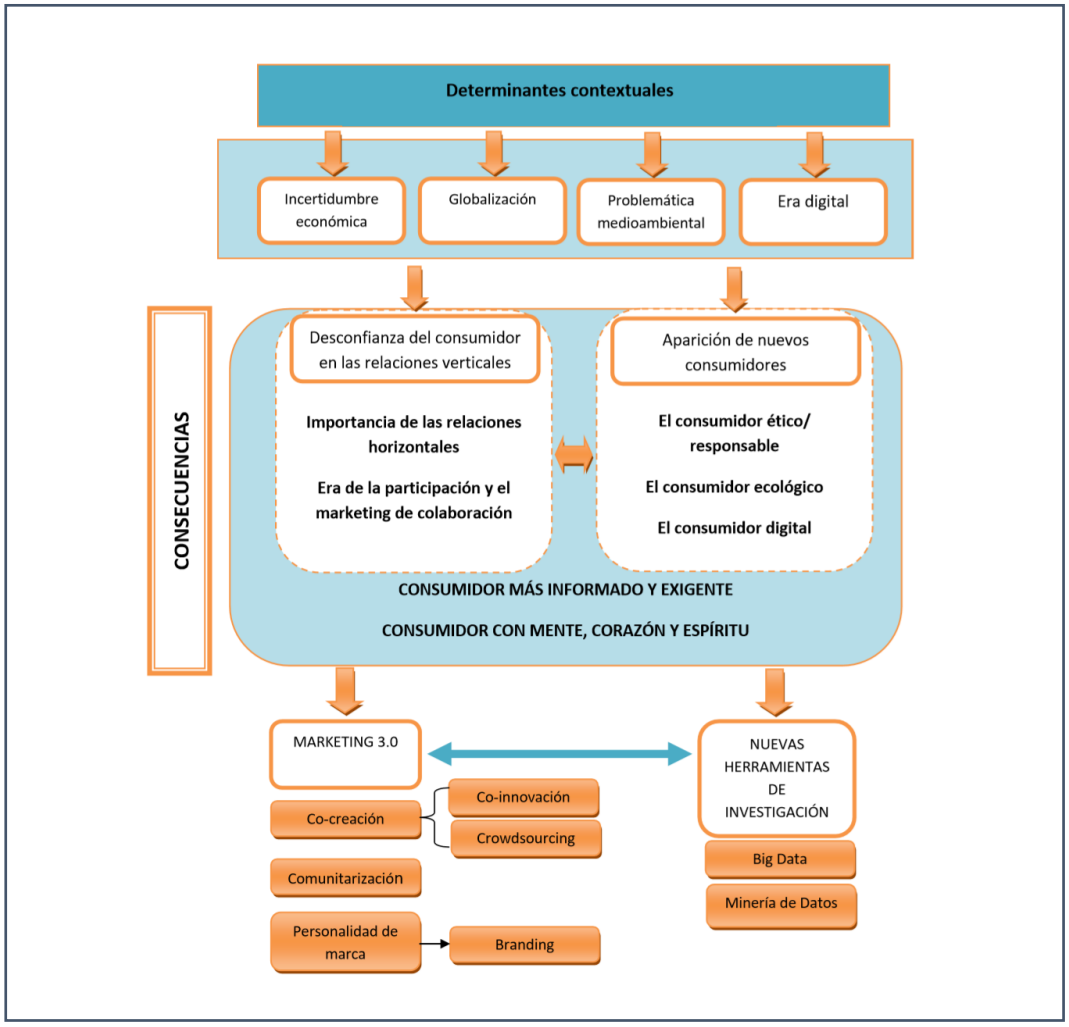
> (tabletas, teléfonos móviles, dispositivos vestibles…), se caracterizan por: a) ser más interesados, exigentes, informados y experimentados; b) críticos con las marcas, d)inmediatistas y demandantes de constates innovaciones; e) que valora prácticas más éticas y responsables, f) que quiere ser escuchado por las empresas, demandando que tengan en cuenta sus opiniones y g) que quiere que se le haga copartícipe en la creación de los productos y servicios que va a consumir.

En base a esto, se están estableciendo nuevos modelos de consumidores como el **consumidor ético/responsable** que actúa mediante pautas consideradas éticas y en consonancia con la mayoría, el **consumidor ecológico,** que prima por los aspectos del medioambiente y establece sus hábitos de consumo teniendo en mente la preservación de los ambientes y el **consumidor digital**, usuario fascinado por las tecnologías de la comunicación y sus aplicaciones de modo que recurre a éstas en sus relaciones con las marcas (Canales & Hernández, 2013).

En definitiva, estos cambios además de indicar nuevas direcciones y conceptos en marketing terminan por dar paso a nuevas formas de hacer y aplicar el marketing.

De este modo y ante todo lo mencionado, la Figura 100 ofrece un esquema visual que abarca a estos y otros conceptos destacados.

Figura 100. Nevo panorama: el futuro del marketing.



Fuente: (Viedma-del-Jesus, 2016), p. 35.

9.2.2. LOS MEDIOS SOCIALES COMO FUENTE DE INFORMACIÓN Y LA CO-CREACIÓN.

En base a lo dicho en el punto anterior, es evidente que el acceso a la información que trajo consigo Internet, ha sido un importante pilar en el cambio de los hábitos de consumo y del comportamiento del consumidor actual. Existe en internet cantidades casi incalculables de información

disponible generada por cientos de millones de personas, que a través de diversos medios sociales, pueden compartir ideas de forma rápida y sencilla. En este sentido, la Creación Colaborativa de Contenido (UGC, *User Generated Content)* (Moens, Li, & Chua, 2014) y la facilidad de interacción social proporcionada por los medios y comunidades sociales web, ampliaron y facilitaron el proceso de conexión entre los usuarios de la red, resultando en una especie de red humana de comunicación global (Kaplan & Haenlein, 2010; Lévy, 2004).

Actualmente, los usuarios de los medios sociales, caracterizados por ser creadores, innovadores y colaboradores, fueron empoderados con una vía de comunicación directa, veloz y a costes mínimos que les permite difundir contenido y hacerse oír. Este protagonismo cambió el rol en las relaciones de los consumidores con las empresas, situándolos como componente clave en la co-creación de contenido y generación de valor. Se parte de la premisa fundamental donde el consumidor/usuario es partícipe en la co-creación de valor. Es decir, el valor deja de ser algo propuesto solo por las empresas, sino que surge de la colaboración entre cliente-empresa a partir de la interacción entre las partes interesadas (*stakeholders)* (Ramaswamy & Ozcan, 2018). Este concepto trae consigo una visión más balanceada por la cual clientes y empresas participan e interactúan para co-crear experiencias y consecuentemente generar valor. Concretamente, la co-creación de valor existe cuando un servicio de calidad superior es ofrecido acorde con la percepción de valor por parte del cliente (Betz, Burkhalter, & Jung, 2019).

En definitiva, la co-creación implica co-crear valor en conjunto con el cliente, buscando satisfacer sus expectativas y demandas de manera plena. Este proceso debe de ocurrir de manera colaborativa en un ambiente de dialogo fluido como por ejemplo los medios sociales, en el que ambas partes aprendan y utilicen los recursos disponibles y sus habilidades de forma conjunta para obtener valor (Campos, Mendes, do Valle, & Scott, 2018).

Este nuevo modelo dio paso a una nueva era, donde la información procede de múltiples agentes que publican ávidamente sobre sus experiencias y/u opiniones vividas con las marcas, productos y servicios. Estas manifestaciones indican que el consumidor actual exige más, tiene preferencias específicas y desean respuestas rápidas y eficaces. Tal escenario genera lo que los autores Anderson (2006) y más recientemente Gandini (2016) llaman de "economía de reputación". Este fenómeno conjuntamente con los conceptos conectados al Marketing 3.0 (Kotler et al., 2010), hacen que cada vez sea más imprescindible para las organizaciones recoger datos, recopilar y analizar opiniones desde la web para identificar y entender lo que dicen los usuarios acerca de temas como por ejemplo un producto en particular, un candidato a la presidencia de un país, los competidores directos de su negocio, entre otros.

Otro aspecto importante sobre este fenómeno se relaciona con la posibilidad de generar fuertes vínculos afectivos y duraderos a lo largo del tiempo entre las empresas y sus consumidores. Esta relación, más cercana e intensa, facilita a las organizaciones identificar las necesidades reales de

su público, al tiempo que permite también cambios estratégicos y organizacionales sin un gran coste (Kim et al., 2016; Sung, Kim, Kwon, & Moon, 2010).

Entre los mecanismos que componen la economía de reputación está el boca oído electrónico (eWOM) (Rosario, Sotgiu, Valck, & Bijmolt, 2016), que influye de manera considerable en los clientes respecto a la elección de una marca, producto o servicio. De modo que las empresas deben considerar al eWOM como herramienta a la hora de comunicarse con sus clientes en la red, eliminando intermediarios, así como posibles "ruidos" que puedan influenciar negativamente estas relaciones.

Algunas de las principales ventajas y aplicaciones de la utilización de los medios sociales como vehículo de comunicación en las relaciones empresa/cliente son fortalecer a la marca (Ewing, Wagstaff, & Powell, 2013; Kuo & Feng, 2013; Wiertz & Ruyter, 2007; Zaglia, 2013); comunicarse directamente con grupos muy específicos y la segmentación de clientes (Chen, 2010; Flavián, Guinalíu, & Torres, 2005); detectar nuevas orientaciones así como pequeños movimientos o "cambios de humor" del mercado (Keh & Pang, 2010); mejorar las interacciones y experiencias del consumidor con la marca así como aumentar la lealtad y confianza hacia la misma (Laroche, Habibi, & Richard, 2013); conocer las percepciones de los clientes sobre los productos y servicios ofrecidos por la marca (Martínez-López, Anaya-Sánchez, Aguilar-Illescas, & Molinillo, 2015); estimular los procesos de co-creación con los clientes ampliando la propuesta de valor ofrecida (Alves, Fernandes, & Raposo, 2016); reducción de costes (Mount

& Martinez, 2014); búsqueda de nuevas formas de satisfacción de los clientes (Laroche, Habibi, Richard, & Sankaranarayanan, 2012); y medir el éxito de acciones de marketing, ya que actúan como un canal de comunicación adicional donde se desarrolla el eWOM. De modo que si logra ser positivo, puede ayudar a difundir las opiniones de los clientes de forma rápida y eficaz por la red, y si es negativo, la empresa puede verse muy dañada con la rápida propagación de estas manifestaciones (Ilhan, Kübler, & Pauwels, 2018; Seller & Laurindo, 2018).

9.2.3. MINERÍA DE DATOS, ANALISIS DE SENTIMIENTOS Y CONTRIBUICIONES RELACIONADAS

Debido a la relevancia de la información contenida en las redes, el monitoreo, la extracción y el análisis de estos datos constituyen un campo de estudio interesante, actual y en constante crecimiento. Por ello, es importante que las organizaciones dispongan de herramientas apropiadas que faciliten el proceso de descubrimiento de conocimiento (Olbrich & Holsing, 2011). Este proceso representado por la sigla KDD (*Knowledge Discovery in Databases*), que combina la extracción de patrones y el análisis de datos de manera automática, alberga soluciones esenciales para el manejo de la información como la Minería de Datos (MD) y el Análisis de Sentimientos (AS) (Deepashri & Kamath, 2017).

En líneas generales, la **Minería de Datos** (MD) es responsable de seleccionar los métodos utilizados en la localización de patrones, formación de representaciones y ajustes de parámetros de algoritmos de aprendizaje,

recursos utilizados en las tareas de Análisis de Sentimientos automáticos (Berson, Smith, & Thearling, 2000; Fayyad et al., 1996a; Johny & Scholar, 2017; Mohanty & Das, 2017). Según la literatura, sus aplicaciones más comunes son la descripción, la clasificación, la estimación o regresión, la clasificación, las agrupaciones y asociación de patrones (Chattamvelli, 2015b; Deepashri & Kamath, 2017; García-Peñalvo & Conde-González, 2017; Larose, 2014; Witten et al., 2016).

Diferentes investigaciones han demostrado la utilidad de la aplicación de las técnicas de MD en diversas áreas tales como **seguridad** (Vomfell et al., 2018), **deportes** (Bigsby, Ohlmann, & Zhao, 2019), **sociología** (Hemsley et al., 2018; Jiang et al., 2019), **religión** (Chen & Huang, 2019), **Nuevas tecnologías** (Araujo & Kollat, 2018; Kohl et al., 2018), **comunicación y periodismo** (Crisci et al., 2018; Jukes, 2019; Orellana-Rodriguez & Keane, 2018), **política** (Friedland et al., 2019; Haro-de-Rosario et al., 2018; Pond & Lewis, 2019; Posegga & Jungherr, 2019; Ramos-Serrano et al., 2018), **medio ambiente** (Hopke & Simis, 2017) **educación** (Borau et al., 2009; Prieto, 2016; Rath, 2011; Rinaldo et al., 2011; Suliman, 2010), **recuperación de Información** (IR, *"Information Retrieval")* y **métricas de Análisis de Medios sociales** (SNA, "Social Network Analysis") (Bakharia & Dawson, 2011; Magnani et al., 2011; Prieto, 2016), **salud** (Bhattacharya et al., 2014; Kandadai et al., 2016; Kotsenas et al., 2018; Pemmaraju, Thompson, et al., 2017), **empresas** (Culnan et al., 2010; Rybalko & Seltzer, 2010), ***eWOM*** (Swani et al., 2014; Xiong et al., 2017), **gestión de la reputación** (Einwiller & Steilen, 2015; Grégoire et al., 2015; Istanbulluoglu, 2017), **asociaciones**

de empresas y minoristas (Bhattacharjya et al., 2016; Chao & Florenthal, 2016), **organizaciones sin ánimo de lucro** (Young, 2017), ente otros.

Por su parte, el **Análisis de Sentimientos** (AS) es un recurso derivado de la Minería de Datos. Este campo de estudio es el que trata de analizar opiniones, sentimientos, evaluaciones, valoraciones, actitudes y emociones de las personas hacia entidades, productos, servicios, organizaciones, individuos, asuntos, eventos, temas y sus atributos (Liu, 2017; Zimbra et al., 2018).

Relacionado con su terminología, cabe mencionar que existen muchos nombres derivados que representan tareas ligeramente diferentes, por ejemplo, extracción de opiniones, minería de sentimientos, análisis de subjetividad, análisis de afecto, análisis de emociones, minería de revisión, entre otros términos. Actualmente todas estas denominaciones están bajo el paraguas del Análisis de Sentimientos o de la Minería de Opiniones (Benkhelifa & Laallam, 2018). Mientras que en la industria el término Análisis de Sentimientos se utiliza más comúnmente, en el mundo académico se consideran casi sinónimos los términos Análisis de Sentimientos y de Minería de Opiniones (Alaei et al., 2019; Anandarajan et al., 2018a; Bhadane et al., 2015; Flekova et al., 2015; Hussein, 2018; Peng et al., 2019; Ribeiro et al., 2016; Wang et al., 2019; Zhang & Liu, 2017).

El análisis de opiniones tiene un enorme campo de aplicación del que se benefician por un lado los clientes, pudiendo establecer criterios que apoyen sus decisiones de compra, y por otro lado las empresas, que obtienen el *feedback* de sus clientes que les ayuda direccionando la toma

de decisiones relacionadas al negocio. El resultado proporciona mejoras en procesos tales como: el diseño y configuración de productos o servicios; desarrollo, aplicación y control de acciones de marketing y publicidad; vigilancia tecnológica; certificación de la calidad; y seguimiento de la satisfacción y experiencia de usuario entre otras (Aggarwal, 2018; Bruseberg & Mcdonagh-Philp, 2002).

El Análisis de Sentimientos se ha considerado la base de múltiples estudios con el objetivo de clasificar la información de la web según su polaridad (positivo/negativo). Esta tarea se puede llevar a cabo a través de dos enfoques: a) el enfoque basado en el léxico, el cual a través de un sistema clasificador (M. Hu & Liu, 2004) utiliza diccionarios semánticos que contienen clases de palabras previamente rotuladas como base para la clasificación (Riloff & Wiebe, 2003; Silva et al., 2012); y b) el enfoque de aprendizaje automático, que deriva del campo de la **Inteligencia Artificial** y busca desarrollar recursos computacionales competentes en adquirir conocimientos de forma automática. Los algoritmos de aprendizaje automático pretenden desarrollar conclusiones genéricas a partir de un conjunto de ejemplos/datos llamados corpus o *datasets* y toman decisiones en base a soluciones consideradas adecuadas utilizadas en problemas anteriores (Mitchell, 1997; Weiss & Kulikowski, 1991).

En otras palabras, los algoritmos utilizan *datasets* ya clasificados para entrenarse y aprender de manera automática sobre un concepto y luego replicar lo aprendido a nuevos grupos de datos. De este modo, **cuanto mayor es la calidad del *dataset* utilizado en el entrenamiento del**

algoritmo, mejor será la precisión en la clasificación de los mensajes. Sin embargo, estos *datasets* son recursos muy escasos en la literatura, ya que para que sean aceptados científicamente deben ser clasificados subjetivamente por un mínimo de tres expertos y luego sometidos a pruebas estadísticas de validación. Con el objetivo de contribuir a esta necesidad en la literatura se **planteó la primera contribución de esta tesis doctoral.** El estudio realizado ha proporcionado los recursos base para la clasificación de sentimientos por medio de algoritmos automáticos, llamados *datasets* de sentimientos. **El objetivo principal fue recopilar distintos *datasets* de sentimientos en múltiples idiomas, clasificados de manera subjetiva y validados científicamente.** De este modo, fueron compilados 25 *datasets* de mensajes cortos, clasificados subjetivamente y validados por la comunidad científica. En total, se consideraron 7 idiomas, incluyendo 10 *datasets* en inglés (1.681,618 mensajes), 4 en español (75.157 mensajes), 4 en el idioma portugués (3.412 mensajes), 2 en alemán (2.300 mensajes), 1 en árabe (2.000 mensajes), 3 en italiano (11.797 mensajes) y 1 en francés (1.797 mensajes). Además, dichos *datasets* tratan de temas variados como política, fútbol, cine, salud y revisiones de productos/servicios, entre otros.

El gran volumen de mensajes (1.778.081) y la variedad de *datasets* en múltiples idiomas ofrecidos por este compendio, ofrece información vital y robusta que sirve de base para posteriores investigaciones científicas relacionadas con el Análisis de Sentimientos por medio del aprendizaje automático. Los resultados de este estudio han sido publicados en la revista "*Soft Computing*", ubicada hoy en día en los índices *Journal Citation Report*

(JCR) (IF 2.367), *Source Normalized Impact per Paper (SNIP)* (IF 1.110), *SCImago Journal Rank (SJR)* (IF 0.593) y en el *Google's h5 Index* (41 pts.). La publicación está disponible en https://link.springer.com/article/10.1007/s00500-017-2766-5.

Por otro lado, los algoritmos utilizados en el aprendizaje automático son variados y utilizan reglas matemáticas distintas para entrenarse. Esta variabilidad hace que tengan mejor o peor "rendimiento" dependiendo del idioma o del concepto que se propongan aprender. Por este motivo, se planteó la **segunda contribución** de esta tesis que ofrece un paso más en relación con los recursos relacionados al Análisis de Sentimientos.

Como se ha mencionado antes, en la literatura se encuentran principalmente dos enfoques para la clasificación de textos: el enfoque léxico en base a diccionarios léxicos de sentimientos y el aprendizaje automático. Este último enfoque se basa en el uso de algoritmos/métodos automáticos que aprenden a clasificar los mensajes utilizando datos previamente rotulados, llamados *datasets* de sentimiento (Tsytsarau & Palpanas, 2012). En este sentido, tras una exhaustiva revisión del estado del arte, se eligieron cuatro algoritmos/métodos, frecuentes en la literatura, con el objetivo de determinar cuál de ellos presenta mejor "rendimiento" con relación a la precisión de clasificación y el tiempo de análisis en distintos idiomas. Para llevar a cabo esa tarea, fueron utilizados 21 *datasets* en 5 idiomas distintos, recursos procedentes del estudio anterior de esta tesis. En este sentido, el objetivo principal de esta segunda contribución fue **estudiar y comparar a cuatro algoritmos frecuentes en la literatura para**

las tareas de clasificación automática de sentimientos, a fin de determinar el algoritmo con mejor rendimiento (precisión y tiempo de análisis) para los idiomas español, inglés, portugués, italiano y alemán.

Los resultados pusieron de manifiesto que, **respecto al tiempo** requerido por los algoritmos en realizar los análisis, el método *Naïve Bayes* (NB) presenta con diferencia los mejores resultados en todos los idiomas y viene seguido de los algoritmos *Support Vector Machine* (SVM), *Decisión Tree* (DT) y *Random Forest* (RF).

Respecto a la precisión de clasificación, el algoritmo SVM ocupa el primer puesto en el ranking, seguido por NB, DT y RF. Al considerar las diferencias de los resultados en clasificación y tiempo entre los algoritmos SVM y NB, se concluye que el SVM ofrece una clasificación (*accuracy* y *F-measure*) alrededor de 7,58% mejor que el NB. Sin embargo, a cambio exige un sacrificio en tiempo ya que es 113,6% más lento que el NB.

En otras palabras, el método SVM clasifica mejor, pero exige un gran sacrificio de tiempo. Por otro lado, el NB clasifica más rápido y exige un pequeño sacrificio en la calidad de la clasificación.

Otro punto por destacar es que el algoritmo SVM presenta gran dificultad para predecir correctamente los datos en el idioma italiano frente a los demás idiomas. Cabe mencionar que la calidad del *dataset* utilizado para entrenar los algoritmos es clave para conseguir buenos resultados, de modo que constituye un precedente para realizar investigaciones futuras

con grupos de entrenamientos en el idioma italiano aún más robustos a fin de corroborar este fenómeno y validarlo científicamente.

Los resultados conseguidos con este estudio revelaron el algoritmo/método que clasifica mejor (precisión) y más rápido (tiempo) frente cada uno de los 5 idiomas analizados. Estos resultados fueron consideraros como base para el diseño y elección de los recursos de Análisis de Sentimientos implementados en el sistema de apoyo a decisiones LOGOS, construido con recursos de Inteligencia Artificial (IA) que decidan de manera automática cuál algoritmo utilizar, según el idioma que se analice, para ofrecer los mejores resultados de clasificación y una mayor precisión en los informes generados por el sistema.

Los sistemas de apoyo a decisiones (*Decision Support Sistems*, DS), así como otras tecnologías tales como el procesamiento analítico en línea (OLAP), cuadro de mando integral, entre otros, constituyen sistemas de Inteligencia Empresarial *(Business Intelligence) (BI)* que buscan mejorar el flujo de trabajo y el proceso de toma de decisiones tanto a nivel táctico como estratégico dentro y fuera de las empresas (Aruldoss et al., 2014; Chen & Wang, 2010; Eckerson, 2008; Lee et al., 2016; Lee & Widener, 2016; Li et al., 2008; Zeng et al., 2012).

A este respecto, como **tercera contribución** se procedió a diseñar, construir e implementar un sistema web de Inteligencia empresarial, multilingüe y gratuito que, a través del procesamiento de datos provenientes de la web, opere como apoyo a decisiones de marketing, para

uso de las empresas, de la comunidad científica y también de la sociedad en general.

LOGOS, el sistema desarrollado como parte de esta tesis doctoral, integra los recursos de Minería de Datos, Análisis de Sentimientos e Inteligencia Artificial, que a través de una interfaz intuitiva ofrece diversos informes de apoyo a la toma de decisiones empresariales.

Concretamente, el sistema se compone por un cuadro de mando fijo que permite la búsqueda y descarga de mensajes en base a una cuenta de usuario, #hashtag o palabra cualquiera, como también la subida de ficheros de datos a la aplicación.

Para la obtención de los datos de Twitter, LOGOS se conecta automáticamente al microblog por medio de la API *Streaming* y procede con la extracción de los mensajes disponibles. Luego, aplica diferentes clases de preprocesamiento de datos provenientes de la Minería de Datos y del Procesamiento del Lenguaje Natural, para generar los informes ofrecidos por el sistema.

El acceso a los informes se da por distintas pestañas. Cada una de ellas aporta características específicas contenidas en los mensajes y ofrece al analista distintas perspectivas e interpretaciones de los datos. Además, los datos son presentados en formatos de visualización variados como listas y distintas gráficas que pueden ser configurados y ordenados de distintas maneras, para que los datos se muestren según las necesidades del analista. LOGOS también permite la descarga de cada informe como

ficheros del tipo .xls para permitir su manejo por aplicaciones de hojas de cálculo, por ejemplo. Este conjunto de herramientas ofrece una mejor visualización, comprensión y manejo de los mensajes obtenidos de Twitter, lo que estimula la obtención de *insights* provenientes de los datos.

Las nueve pestañas de informes que componen el sistema son: *Tweets, Retweets, Hashtags (top20), Users mentioned (top 20), Geo Map, Sentiment, Wordcloud, Dispositives y User Twitter Profile*. Concretamente, estos recursos en conjunto proporcionan a) la visualización y la descarga de los tuits y retuits en ficheros .xls; b) la visualización y la descarga de listados, gráfica y ranking de los hashtags más frecuentes en los mensajes, así como de los usuarios más mencionados, en solitario o por pares; c) la visualización geográfica de los mensajes y el porcentaje de mensajes geolocalizables de la muestra; d) la visualización y la descarga de los mensajes identificados como positivos o negativos mediante la aplicación del Análisis de Sentimientos por medio de algoritmos de aprendizaje automático; e) la creación de nubes de palabras con la posibilidad de determinar distintas configuraciones como, por ejemplo, el número de palabras a ser consideradas o la exclusión de "palabras vacías" consideradas irrelevantes para el análisis, entre otros; f) la visualización y la descarga del ranking de dispositivos utilizados en las publicaciones de los mensajes; g) la identificación de los perfiles en Twitter de los usuarios de la muestra, el número de tuits publicados por cada uno de ellos, su número de seguidores y su idioma; y finalmente h) la posibilidad de ejecución de todos estos análisis inicialmente en 3 idiomas distintos, (inglés, español y portugués) ampliables a más idiomas en proyectos futuros.

En definitiva, el meta sistema web desarrollado ofrece a los analistas diversos tipos de informes que, por medio de distintos listados y gráficas, revelen valiosa información respecto a las manifestaciones de los usuarios en Twitter, como sus preferencias respecto a productos/servicios entre otros muchos aspectos, que pueden resultar de gran utilidad de cara a la toma de decisiones empresariales.

Tras finalizadas las etapas de diseño y de implementación, fue puesta en marcha la fase de prueba, llevada a cabo como **última contribución** de esta tesis. Por medio de cuatro aplicaciones prácticas relacionadas con las empresas Nike y Samsung en dos idiomas distintos (español y portugués), ha quedado patente la utilidad de LOGOS en transformar el *Big Data* en *Smart Data*. Esta transformación genera información que contribuye a entender el comportamiento del consumidor desde la perspectiva del **Marketing 3.0**, e indicando a las empresas un camino a seguir fundamentado en el valor humano y con especial protagonismo del consumidor.

Los resultados pusieron de manifiesto que LOGOS cumple con su propósito al extraer y procesar los mensajes de la red social Twitter y gestionar estos datos por medio de la correcta aplicación de técnicas de Inteligencia Artificial, Minería de Datos, Procesamiento del Lenguaje Natural y del Análisis de Sentimientos automáticos. Otro punto a destacar es en relación con las distintas formas de visualización ofrecidas por LOGOS, como su diversidad de gráficas. Estos recursos aportan una

perspectiva visual holística de los datos, ofreciendo una primera lectura más clara y ágil de la información.

Relacionado con los *insights* facilitados por LOGOS, los principales resultados afirman que esos pueden ser utilizados para:

1. Identificar demandas de los usuarios y temas destacados para el desarrollo de nuevas propuestas y acciones de marketing;
2. Monitorear los distintos aspectos tanto positivos como negativos, relacionados con las campañas tanto ejecutadas como en ejecución;
3. Medir el éxito de las acciones de marketing puestas en marcha;
4. Identificar tanto los puntos fuertes como los puntos débiles en los productos y servicios ofrecidos por la marca, así como las características específicas que los hacen tener éxito o no según la percepción de los usuarios;
5. Identificar problemas con "terceros" como la falta de productos en tiendas, prácticas de precio inadecuadas y problemas con la logística y la falta de información en los equipos de ventas;
6. Gestionar de manera activa y preventiva situaciones conflictivas con los clientes;
7. Anticiparse y gestionar posibles crisis o situaciones de conflictos relacionadas a la marca, al mercado y a los consumidores;
8. Identificar a los principales competidores de la marca, sus debilidades, fortalezas, entre otros aspectos;

9. Identificar oportunidades de mercado para el desarrollo de nuevos productos y servicios por medio de la "escucha activa" del cliente en la red;
10. Identificar los usuarios influyentes i "embajadores" de la marcar para desarrollo de acciones de marketing.

En base a lo expuesto anteriormente, LOGOS ha demostrado su efectividad en las tareas propuestas por el estudio, de modo que cumple con el objetivo principal establecido para esta tesis doctoral en **"Crear un método y un sistema web gratuito de inteligencia empresarial que, a través de técnicas de Minería de Datos y Análisis de Sentimientos aplicadas a informaciones provenientes de la web, sirva en distintos idiomas, como soporte de decisiones estratégicas de marketing para las empresas contemporáneas".**

Finalmente, las implicaciones relativas a las diversas propuestas llevadas a cabo en esta tesis alcanzan a tres ámbitos principalmente: el académico, las empresas y la sociedad.

9.3. PRINCIPALES IMPLICACIONES DE LA TESIS

Las aportaciones derivadas de este trabajo alcanzan a tres ámbitos principalmente: el académico, el de las empresas y el del consumidor.

- **Principales aportaciones para el ámbito académico**: a) ofrecer a la comunidad científica un compendio de *datasets* que sirva

como recurso para futuras investigaciones relacionadas con la Minería de Datos y el Análisis de Sentimientos; b) revelar cuál es el algoritmo que mejor funciona (entre los cuatro más frecuentes en la literatura) para identificar polaridad en textos en función de varios idiomas (español, inglés, portugués, italiano y alemán); c) desarrollar un método y un sistema de Minería de Datos y Análisis de Sentimientos de código abierto.

- **Principales implicaciones para las empresas**: poner a disposición del público, en particular PyMEs (sobre todo de países en vía de desarrollo como Brasil) que cuentan con recursos muchas veces limitados, un sistema web multilingüe, gratuito e intuitivo, basado en la información proveniente de medios sociales, para generar informes que sirvan como soporte de decisiones estratégicas de marketing. Concretamente, LOGOS permite a las empresas: a) identificar demandas de los usuarios y temas destacados para el desarrollo de nuevas propuestas y acciones de marketing; b) identificar y seguir a las campañas de marketing en línea, tanto ejecutadas como en ejecución; c) obtener una retroalimentación (*feedback)* sobre el éxito de las acciones de marketing puestas en marcha; d) identificar tanto los puntos fuertes como los puntos débiles en los productos y servicios ofrecidos por la marca, así como las características específicas que los hacen tener éxito o no según la percepción de los usuarios; e) identificar problemas con "terceros" como la falta de productos en tiendas, prácticas de

precio inadecuadas y problemas con la logística y la falta de información en los equipos de ventas; f) gestionar de manera activa y preventiva situaciones conflictivas con los clientes; g) anticiparse y gestionar posibles crisis o situaciones de conflictos relacionadas a la marca, al mercado o a los consumidores; h) identificar a los principales competidores de la marca, sus debilidades y fortalezas, entre otros aspectos; i) identificar oportunidades de mercado para el desarrollo de nuevos y mejores productos y servicios por medio de la "escucha activa" del cliente en la red; j) identificar perfiles influyentes del sector; k) obtener información sobre los clientes, sobre el negocio y sobre los competidores; l) realizar seguimiento de contenidos en línea; m) contribuir a la construcción y mantenimiento de la credibilidad de la marca; e, n) identificar oportunidades para estimular el compromiso entre las partes y construir relaciones con los clientes.

- **Principales implicaciones para el consumidor**: el conocimiento que se puede obtener a partir de la aplicación de herramientas como LOGOS, resulta también muy útil desde el punto de vista de los consumidores en la medida que: a) pueden sentirse (mejor) escuchados y valorados por las empresas; b) pueden recibir productos y servicios cada vez más adaptados a sus necesidades resultado de las prácticas de seguimiento de los medios por parte de las empresas; d) pueden disfrutar de acciones de marketing cada vez más acordes y adaptadas a sus perfiles; e) pueden

sentirse más cercanos a la marca; y f) sentirse importantes a través de la participación y co-creación de las acciones de marketing.

En gran medida, los consumidores de grandes empresas ya experimentan los beneficios derivados del monitoreo de medios sociales. Sin embargo, como comentamos previamente estas prácticas son más difíciles de llevar a cabo por PyMEs, por lo que el consumidor de estas puede sentirse en menor medida escuchados por sus empresas locales.

Herramientas como LOGOS pueden contribuir a solventar o soslayar, en alguna medida, esta problemática, acercando a las PyMEs a la realidad de los medios sociales, ofreciendo la posibilidad de múltiples análisis de la información contenida en estos y mejorando las relaciones con sus clientes, contribuyendo a la supervivencia de estas empresas en Los mercados competitivos de la actualidad.

Finalmente, los resultados de esta tesis se han diseminado en las siguientes publicaciones (Tabla 46):

Tabla 46. Publicaciones de la Tesis.

Año	Medio	Publicación
2016	XXVIII Congreso de Marketing AEMARK 2016	Minería de Opiniones y Análisis de Sentimientos Multilingüe: Comparación y Análisis de Cuatro Algoritmos Automáticos al Caso de Twitter
	XXI Semiárido Anual de la Asociación de Investigadores y Estudiantes Brasileños en Cataluña (APEC)	Desvelando el Qué, Quién y Dónde: Análisis del Público de Red Bull en Twitter a Través de la Minería De Datos
2017	II Jornadas de Investigadores en Formación (JIFFI)	Estudio Comparativo por Medio del Aprendizaje de Maquina Supervisado.
	XXII Semiárido Anual de la Asociación de Investigadores y Estudiantes Brasileños en Cataluña (APEC)	Metodología y Aplicación de la Minería de Opiniones y Análisis de Sentimientos web, para el Desarrollo Empresarial de España y Brasil
2018	Revista Soft Computing (factor de impacto JCR (2.367)	A survey of multilingual human-tagged short message datasets for sentiment analysis tasks
2019	II Congreso Nacional de Investigadores en Formación (JIFFI)	LOGOS: Sistema web multi-idioma de apoyo a decisiones de Marketing

Fuente: Elaboración propia.

9.4. PRINCIPALES LIMITACIONES DE LA TESIS

Las diferentes propuestas de esta tesis doctoral suponen importantes avances en las investigaciones relacionadas con las técnicas de Minería de Datos y Análisis de Sentimientos automáticos en los medios sociales, y su aplicación en el desarrollo de herramientas de apoyo a las decisiones empresariales. Sin embargo, al igual que todas las investigaciones empíricas, las diferentes propuestas de esta tesis presentan limitaciones que pueden afectar a la generalización de los resultados. A este respecto, a continuación, se describen algunas de las principales limitaciones identificadas en este trabajo.

9.4.1. LIMITACIONES DERIVADAS DEL COMPARATIVO DE ALGORITMOS MULTILINGÜE

Esta limitación se relaciona con el número de tuits utilizados para medir el rendimiento de los algoritmos en cada idioma. Concretamente, debido a la novedad de los estudios en esta área de conocimiento, se ha encontrado y utilizado en el estudio un número más amplio de *datasets* validados científicamente para los idiomas inglés, español y portugués que para los idiomas italiano, alemán y francés. Es cierto que cuanto más entrenado está el algoritmo en un determinado tema, más "sabrá" sobre éste y en consecuencia mejores tienden a ser sus resultados de clasificación. En otras palabras, cuantos más tuits se utilizan para entrenar un algoritmo, mejor tiende a ser su rendimiento en la clasificación. En este caso, aunque los *datasets* utilizados para entrenar los algoritmos en los distintos idiomas se consideran suficientes, si se replica igualando el número de tuits utilizados para cada idioma, los resultados podrían diferir de los obtenidos en este estudio.

9.4.2. LIMITACIONES DEL SISTEMA CLASIFICADOR DE SENTIMIENTOS

El sistema clasificador de Minería de Datos y Análisis de Sentimientos utilizado en `RapidMiner` para evaluar los algoritmos de aprendizaje automático, fue utilizado, testado y validado científicamente (Barreira, 2013; Gonçalves & Fernandes Brito, 2013). Sin embargo, se pueden plantear otras posibilidades de tratamiento de los datos, ordenando de manera distinta los procesos de PLN, así como los parámetros de MD

ofrecidos por la herramienta. Este abanico de posibilidades puede hacer que el algoritmo aprenda mejor o peor sobre un determinado contexto, de modo que este aspecto puede suponer una limitación para el estudio. Teniendo en cuenta esta particularidad, previamente en esta se detallan las estructuras de operadores utilizadas en RapidMiner, así como las configuraciones para cada uno de los algoritmos.

9.4.3. LIMITACIONES DE PROGRAMACIÓN Y DE DISEÑO DE LOGOS

Las **limitaciones de programación**, por un lado, se relacionan sobre todo con el lenguaje de programación R utilizado en el desarrollo del sistema. Concretamente, se trata de un lenguaje de código abierto y en constante actualización. Considerando que todos los sistemas poseen sus limitaciones, a lo largo de la construcción de LOGOS fueron planteados algunos recursos especiales como gráficas, animaciones y otros informes que fueron inviables de implementarse en R. De todos modos, como ha quedado evidente en los resultados, hemos explotado al máximo los mejores recursos disponibles en R, buscando optimizar el tiempo del analista con implementaciones que generasen una presentación adecuada de los datos e informes de fácil manejo y visualización.

Las **limitaciones de diseño,** por otro lado, se relacionan con la utilización de la interfaz virtual de R llamada Shiny. Esta permitió la creación de una interfaz gráfica interesante e intuitiva, sin embargo, si la comparamos a otros lenguajes de programación como Python, por ejemplo, sería posible

implementar recursos estéticamente distintos y posiblemente ofrecer una interfaz aún más amigable al sistema LOGOS.

9.4.4. LIMITACIONES DE LA DESCARGA DE DATOS

Esta limitación está asociada a cómo el método de descarga utilizado para la obtención de las muestras de Nike y Samsung podría influir en la calidad de éstas. Aunque la herramienta R permite descargar más de 1500 tuits de una sola vez, este número es difícil de alcanzar debido a que Twitter limita la descarga de mensajes y muchas veces repite los datos que ofrece. Concretamente, esta característica, además de exigir una limpieza posterior de los mensajes repetidos, puede provocar que se "escape" algún mensaje a la hora de la descarga. Este aspecto puede suponer una limitación en la descarga de los datos que componen las muestras utilizadas en las aplicaciones prácticas de esta tesis.

9.4.5. LIMITACIONES RELACIONADAS CON LA GEOLOCALIZACIÓN DE LOS MENSAJES

En los análisis de geolocalización realizados en la aportación 4, tanto en las muestras de Nike como en las de Samsung, se ha detectado un aspecto importarte conectado a la baja adhesión de los usuarios al sistema geolocalización (<1%). Se observa que la muestra @Nike_Spain presenta un 0,14% de mensajes geolocalizables. La muestra @nikebrasil presenta apenas un 0,07% de mensajes geolocalizables, @SamsungEspana un 0,18%,

y @SamsungBrasil apenas un 0,38% de mensajes geolocalizables. Concretamente, este aspecto no representa una deficiencia de la herramienta o del estudio, sin embargo, la falta que datos que justifiquen la obtención de conclusiones robustas y significativas utilizando a estos mensajes ha sido un aspecto limitante.

9.4.6. LIMITACIONES RELACIONADAS CON LOS EMOTICONOS

Esta limitación se relaciona con la importancia de los emoticonos en el lenguaje utilizado por los clientes en los medios sociales. Según los autores Churches, Nicholls, Thiessen, Kohler y Keage (2014), los emoticonos humanizan el mensaje y aportan sentimientos de manera gráfica, lo que supone una particular ventaja sobre las palabras. Estos autores en el estudio titulado *"Emoticons in mind: An event-related potential study",* explican que el símbolo ":-)" despierta en las personas la misma reacción que tendrías al ver alguien sonreír de manera presencial. A este respecto, otros estudios tratan de la importancia de los emoticonos en las conversaciones entre empresas y clientes. Otras investigaciones recientes (Hill, 2016; Leung & Chan, 2017), afirman que las empresas que utilizan los emoticonos y emojis son consideradas más "amigables" por sus clientes y que las respuestas en las redes que van acompañadas por emoticonos, son recordadas más fácilmente por los usuarios frente a la respuestas sin este recurso visual.

En este sentido, dada la importancia que cobran los emoticonos en los mensajes publicados en la web, no haberlos considerado puede suponer una limitación al estudio.

9.4.7. PRINCIPALES LIMITACIONES DEL MEDIO SOCIAL

En líneas generales, esta tesis se centró en desarrollar un sistema web de apoyo a decisiones de marketing capaz de extraer y procesar información contenida en el medio social Twitter. Sin embargo, hoy en día, existen otros medios sociales que ganan relevancia cada día entre los internautas como, por ejemplo, Facebook e Instagram. Estas redes también pueden aportar conocimientos competitivos a las empresas. De modo que, no considerar estas redes puede suponer una limitación, dado que también son importantes fuentes de datos en la actualidad.

9.4.8. PRINCIPALES LIMITACIONES DEL ANALISIS DE SENTIMIENTOS

La comprensión del lenguaje natural no es una tarea trivial para los seres humanos y tampoco para los ordenadores, ya que las diferentes maneras de escritura y expresiones utilizadas en un mensaje influyen en cómo cada uno lo entiende. De este modo, identificar polaridad y opiniones en textos subjetivos utilizando algoritmos automáticos tiene sus limitaciones y restricciones.

De acuerdo con Liu (2012), una de las principales limitaciones del AS es la dificultad en clasificar e interpretar aspectos como la ambigüedad, ironía o sarcasmo, entre otras expresiones presentes en el lenguaje cotidiano.

Otras dificultades que interfieren en la correcta clasificación de sentimientos son a) textos con errores y sentencias sintácticamente mal formadas (lo que es bastante común en los blogs y medios sociales); b) distinguir si un texto es una opinión o un hecho; d) un texto puede referirse a más de un elemento de interés con opiniones diferentes sobre los ítems, lo que puede confundir la clasificación; e) el uso de pronombres para referenciar a "cosas" puede dificultar la identificación de sentencias que mencionen el "ítem" de interés; f) uso de términos informales y abreviaciones de palaras en Internet que son bastante usados en medios sociales, entre otros (Farías et al., 2016; González-Ibáñez, Muresan, & Wacholder, 2011; Poria et al., 2016). Por lo que estas dificultades derivadas de la dificultad de interpretación del lenguaje natural y del propio uso que los internautas realizan de los medios sociales, constituyen una importante limitación para el Análisis de Sentimientos en general y para esta tesis en particular.

9.4.9. PRINCIPALES LIMITACIONES DE LA HERRAMIENTA

En cuanto a las principales limitaciones relacionadas con la herramienta, podemos citar, el hecho de que, de momento, LOGOS esté acotado para atender apenas a los idiomas inglés, español y portugués. Esto es un

limitante ya que con más idiomas posibilitarían expandir los análisis a otros mercados.

La extracción de información proveniente de apenas un medio social, también se pude considerar una limitación. El acceso a información de otros medios como Instagram, Facebook y/o Google+, también ampliamente utilizados por los consumidores, extendería las posibilidades de análisis potenciando los resultados.

Finalmente, otra limitante de la herramienta está en la falta de mecanismos de autogestión como por ejemplo, el acopio automático y/o programado de datos. En la versión actual de LOGOS, este proceso es gestionado manualmente por el analista, que debe estar pendiente de la descarga de datos, lo que es inviable muchas veces para análisis a escala mayores.

9.5. FUTURAS LINEAS DE INVESTIGACIÓN

Para finalizar, se mencionan algunas de las posibles futuras líneas de trabajo con las que es factible ampliar y profundizar a partir de esta tesis doctoral. Destacan las siguientes:

En primer lugar, sería interesante realizar investigaciones que busquen contrastar las metodologías utilizadas en este trabajo con otras metodologías posibles. En esta línea, se indican algunas posibilidades como: a) aumentar e igualar al número de tuits de los diferentes idiomas utilizados

para entrenar a los algoritmos; b) ordenar de formas distintas los procesos de tokenización relativos al Procesamiento del Lenguaje Natural, así como los parámetros del modelo de Minería de Datos utilizados en RapidMiner y c) utilizar otras aplicaciones de MD disponibles para la elaboración del ranking de rendimiento de los algoritmos de aprendizaje automático.

Considerando las limitaciones relacionadas con el idioma, las aplicaciones prácticas llevadas a cabo en esta tesis doctoral, se han centrado en estudiar los tuits generados en los idiomas español y portugués provenientes de las cuentas de Nike y Samsung en España y Brasil por un período de doce meses (de mayo de 2016 a abril de 2017). No obstante, estas marcas están presentes en diversos países y se comunican con sus seguidores en diversos idiomas. Nike por ejemplo se comunica en las redes

en más de 55[30] países y Samsung en más de 150[31], por lo que se asume la posibilidad de extender la investigación a otros idiomas a fin de obtener resultados más amplios sobre el qué hablan y qué piensan los clientes de estas marcas en otros países.

De este modo, además de consolidar la utilidad de LOGOS como herramienta de apoyo a decisiones en estos idiomas, se revelarían otros

[30] https://www.nike.com/language_tunnel (acceso el 07/08/2019).

[31] https://www.samsung.com/es/function/ipredirection/ipredirectionLocalList/ (acceso el 07/08/2019).

casos de estudio, así como particularidades relativas a los idiomas agregados. Además de la ampliación de las aplicaciones prácticas considerando otros idiomas, sería adecuado considerar la clasificación e interpretación de los emoticonos, así como sus implicaciones para cada idioma en particular.

Por último, también se tiene en cuenta como futuras líneas de investigación, las ampliaciones e incorporaciones de nuevos recursos a LOGOS, como son:

- Incluir más idiomas a parte del inglés, español y portugués tanto para análisis como en la interfaz.
- Ampliar la posibilidad de recoger y analizar información de más medios sociales y fuentes de datos como Facebook, Instagram, Google+, blog, etc.
- Adaptar mecanismos de detección automática del idioma de cada texto y aplicación del algoritmo de análisis de sentimientos que mejor se adapte al idioma en cuestión.
- Incluir la posibilidad al analista de usar otros mecanismos y servicios de terceros de Análisis de Sentimientos como por ejemplo, Google Cloud, Meanning Cloud, Indico, entre otros.
- Incluir otros mecanismos de aprendizaje automático, como por ejemplo *Deep Learnig*, para aumentar la precisión, la robustez y la rapidez de los análisis.
- Implementar recursos de soporte a varios análisis y/o analistas de manera simultánea y de espacios de trabajo internos diferenciados

(al menos un espacio de trabajo por analista registrado en el sistema).

- Incorporar opciones de informes más interactivos, como los hipertextuales y autoreferenciados, así como gráficos automáticamente generados y exportados desde LOGOS, y de alta calidad.
- Incluir mecanismos de registros seguros mediante la incorporación de mecanismos de autenticación cifrados.
- Dotar al sistema de mecanismos para el acopio automático de datos que posibiliten la descarga programada.
- Adaptar mecanismos de autogestión como, copia de seguridad; respaldo de datos, configuraciones de análisis, los proprios informes, etc.

9.6. LECCIONES APRENDIDAS

Este apartado es el resultado de una reflexión final que pretende capturar, organizar y diseminar algunas de las lecciones aprendidas durante el proceso de esta tesis doctoral.

He aprendido mucho en mi acercamiento al mundo académico y científico, con sus publicaciones, revistas y rankings. Con los congresos, seminarios y con la estancia de investigación, aprendí a organizar, plasmar y exponer ideas. Aprendí a estar más atento a los detalles. Aprendí que los trabajos académicos deben cumplir altos niveles en cuanto a las formas, la

escritura, el enlace de la ideas, los procesos, la metodología así como en su ejecución.

También he podido construir relaciones con personas de distintos lugares y distintas áreas de conocimiento diferentes a la mía. Relaciones que espero mantenerlas de por vida. Y, ¿cómo mencionar relaciones importantes sin mencionar a mis "jefes"? Así es como suelo dirigirme a mis directores, Marisa y Antonio. Sin embargo, nunca sentí que tenía jefes y sí compañeros de camino que ya habían pasado por allí antes y por eso debía prestarles mucha atención. Que hayan estado a mi lado en este proceso ha sido todo un regalo que hizo el camino más ameno e interesante.

Por ser una tesis multidisciplinar, se me fueron exigiendo aptitudes y conocimientos muy ajenos a mi formación académica anterior. Aprendí a utilizar lenguajes de programación y sistemas gestores de bases de datos. Aprendí en la práctica a cómo construir y manejar modelos de tratamiento de datos con distintas herramientas. Sobre todo, entendí cómo todos estos y otros recursos y conceptos relacionados con la Inteligencia Artificial y provenientes de las Ciencias de la Computación, podrían trabajar para los avances de las Ciencias Económicas y Sociales y del Marketing. Aprendí que en el mundo actual, estas áreas deben de ir dadas de la mano y conversar más a menudo entre sí, ya que, como ha se ha demostrado en esta tesis, trabajan muy bien en conjunto.

Recuerdo también haber experimentado la "soledad intelectual" que una tesis exige. Es un trabajo muchas veces solitario donde los diálogos ocurren

constantemente con uno mismo. Todo un reto para mí, que *"hablo por los codos"*, y que fue superado con el ejercicio de la prudencia y la paciencia.

Si el proceso doctoral suele ser desafiador para un nativo, para un extranjero los desafíos aumentan. Lidiar con cuestiones como hablar y escribir de manera científica en otro idioma, adaptarse al entono y administrar las distancias físicas y psicológicas al mismo tiempo que escribía la tesis, me han producido cambios y me han enseñado mucho sobre mis limitaciones o, por lo menos, las que creía tener antes de eso. Aprendí que España es un país con gente muy acogedora y cautivante, dotado de instituciones académicas de alto nivel como la Universidad de Granada. Que el departamento de Comercialización e Investigación de Mercados es muy competente y por eso, me siento afortunado de poder seguir colaborando con la ciencia como miembro del grupo de investigación Marketing y Consumo.

En definitiva, el camino que uno recorre desde que se propone escribir una tesis hasta el momento del producto final es fascinante. Por tratarse de un proyecto a largo plazo, personas inmediatistas como yo, aprenden que esta característica debe de ser domada. Que el intelecto necesita su tiempo para crear y madurar las ideas, y que estas, se seguirán transformando con el paso del tiempo. Se trata de un proceso lento y constante. De modo que aprendí a esperar y respetar el tiempo que exigen las cosas. Aprendí también a planear, porque algo que exige tiempo también exige planificación y control. Es muy fácil "bajar la guardia" y relajarse en proyectos muy largos. La planificación te ayuda a seguir avanzando.

Finalmente, aprendí el valor y la importancia de un proyecto como éste para la ciencia, para la sociedad y para uno mismo, lo que es muy alentador, me llena como académico, como profesional del marketing y principalmente como ser humano.

REFERENCIAS BIBLIOGRÁFICAS

Abbasoglu, M. A., Gedk, B., & Ferhatosmanoglu, H. (2014). Aggregate profile clustering for streaming analytics. *Computer Journal*, *58*(9), 2092–2108. https://doi.org/10.1093/comjnl/bxv023

Abdelfattah, M., Galal, D., Hassan, N., Elzanfaly, D. S., & Tallent, G. (2016). A Sentiment Analysis Tool for Determining the Promotional Success of Fashion Images on Instagram. *Sustainable Vital Technologies in Engineering & Informatics*, 7.

Abdulla, N. A., Ahmed, N. A., Shehab, M. A., & Al-Ayyoub, M. (2013). Arabic Sentiment Analysis: Lexicon-based and Corpus-based. *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.

Accenture. (2018). *INNOVATING IN THE NEW*. Retrieved from https://www.accenture.com/_acnmedia/pdf-89/accenture-fiscal-2018-annual-report.pdf

Aggarwal, C. C. (2018). Opinion Mining and Sentiment Analysis. In *Machine Learning for Text* (pp. 413–434). https://doi.org/10.1007/978-3-319-73531-3_13

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*. Retrieved from http://www.andrew.cmu.edu/user/ngm/15-721/summaries/

Al-Kabi, M., Al-Ayyoub, M., Alsmadi, I., & Wahsheh, H. (2016). A prototype for a standard Arabic sentiment analysis corpus. *International Arab Journal of Information Technology*, *13*(1A), 163–170.

Al-Twairesh, N., Al-Khalifa, H., & Al-Salman, A. (2015). Subjectivity and sentiment analysis of Arabic: Trends and challenges. *IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, *2014*(June), 148–155. https://doi.org/10.1109/AICCSA.2014.7073192

Alaei, A. R., Becken, S., & Stantic, B. (2019a, February 1). Sentiment Analysis in Tourism: Capitalizing on Big Data. *Journal of Travel Research*, Vol. 58, pp. 175–191. https://doi.org/10.1177/0047287517747753

Alaei, A. R., Becken, S., & Stantic, B. (2019b, February 14). Sentiment Analysis in Tourism: Capitalizing on Big Data. *Journal of Travel Research*, Vol. 58, pp. 175–191. https://doi.org/10.1177/0047287517747753

Alharbi, A. S. M., & de Doncker, E. (2019). Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. *Cognitive Systems Research*, *54*, 50–61. https://doi.org/10.1016/j.cogsys.2018.10.001

Alsaeedi, A. (2019). EFTSA: Evaluation Framework for Twitter Sentiment Analysis. *Journal of Software*, 24–35. https://doi.org/10.17706/jsw.14.1.24-35

Alves, H., Fernandes, C., & Raposo, M. (2016). Value co-creation: Concept and contexts of application and study. *Journal of Business Research*, *69*(5), 1626–1633. https://doi.org/10.1016/j.jbusres.2015.10.029

Ananda, A. S., Hernández-García, Á., & Lamberti, L. (2016). N-REL: A comprehensive framework of social media marketing strategic actions for marketing organizations. *Journal of Innovation & Knowledge*, *1*(3), 170–180. https://doi.org/10.1016/j.jik.2016.01.003

Anandarajan, M., Hill, C., & Nolan, T. (2018a). *Learning-Based Sentiment Analysis Using RapidMiner*. https://doi.org/10.1007/978-3-319-95663-3_15

Anandarajan, M., Hill, C., & Nolan, T. (2018b). *Modeling Text Sentiment: Learning and Lexicon Models*. https://doi.org/10.1007/978-3-319-95663-3_10

Anderson, Carl. (2019). *Business Intelligence*. https://doi.org/10.1007/978-3-319-97556-6_6

Anderson, Cris. (2006). *A Cauda Longa: do mercado de massa para o mercado de nicho* (Elsevier, Ed.). Rio de Janeiro.

Ankit, & Saleena, N. (2018). An Ensemble Classification System for Twitter Sentiment Analysis. *Procedia Computer Science*, *132*, 937–946. https://doi.org/10.1016/j.procs.2018.05.109

Araujo, & Kollat, J. (2018). Communicating effectively about CSR on Twitter. *Internet Research*, *28*(2), 419–431. https://doi.org/10.1108/intr-04-2017-0172

Araujo, M., Pereira, A., Reis, J., & Benevenuto, F. (2016). *An Evaluation of Machine Translation for Multilingual Sentence-level Sentiment Analysis*. 1140–1145. https://doi.org/10.1145/2851613.2851817

Arnott, D., Lizama, F., & Song, Y. (2017). Patterns of business intelligence systems use in organizations. *Decision Support Systems*, *97*, 58–68. https://doi.org/10.1016/j.dss.2017.03.005

Arnott, D., & Pervan, G. (2014). A critical analysis of decision support systems research revisited: The rise of design science. *Journal of Information Technology*, *29*(4), 269–

293. https://doi.org/10.1057/jit.2014.16

Arnott, D., & Pervan, G. (2016). A critical analysis of decision support systems research. *Formulating Research Methods for Information Systems: Volume 2*, 127–168. https://doi.org/10.1057/9781137509888_5

Aruldoss, M., Lakshmi Travis, M., & Prasanna Venkatesan, V. (2014). A survey on recent research in business intelligence. *Journal of Enterprise Information Management*, *27*(6), 831–866. https://doi.org/10.1108/JEIM-06-2013-0029

Aryo Prakoso, A., Winantesa Yananta, B., Fitra Setyawan, A., & Muljono. (2018). A Lexicon-Based Sentiment Analysis for Amazon Web Review. *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, ISemantic 2018*, 503–508. https://doi.org/10.1109/ISEMANTIC.2018.8549812

Azevedo, A. (2017). Data Mining and Knowledge Discovery in Databases. In *Encyclopedia of Information Science and Technology, Fourth Edition* (pp. 1907–1918). https://doi.org/10.4018/978-1-5225-2255-3.ch166

Azorín-Richarte, D., Orduna-Malea, E., & Ontalba-Ruipérez, J.-A. (2016). Redes de conectividad entre empresas tecnológicas a través de un análisis métrico longitudinal de menciones de usuario en Twitter. *Revista Española de Documentación Científica*, *39*(3), 1–20. https://doi.org/10.3989/redc.2016.3.1316

Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. A. (2016). The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. *Journal of Marketing Research*, *53*(3), 297–318. https://doi.org/10.1509/jmr.14.0380

Baca-Gomez, Y. R., Martinez, A., Rosso, P., Estrada, H., & Farias, D. I. H. (2016). Web service SWePT: A hybrid opinion mining approach. *Journal of Universal Computer Science*, *22*(5), 671–690.

Bakharia, A., & Dawson, S. (2011). SNAPP: a bird's-eye view of temporal participant interaction. *LAK '11 Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, 168–173. https://doi.org/10.1145/2090116.2090144

Balog, K., Mishne, G., & Rijke, M. De. (2006). Why are they excited? Identifying and explaining spikes in blog mood levels. *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (EACL '06)*, 207–210. Retrieved from http://dl.acm.org/citation.cfm?id=1609010

Bandaru, S., Ng, A. H. C., & Deb, K. (2017). Data mining methods for knowledge discovery

in multi-objective optimization: Part A - Survey. In *Expert Systems with Applications* (Vol. 70). https://doi.org/10.1016/j.eswa.2016.10.015

Barbosa, L., & Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics*, 36–44. Retrieved from http://dl.acm.org/citation.cfm?id=1944571

Bari, A., & Saatcioglu, G. (2018). Emotion Artificial Intelligence Derived from Ensemble Learning. *Proceedings - 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications and 12th IEEE International Conference on Big Data Science and Engineering, Trustcom/BigDataSE 2018*, 1763–1770. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00266

Barreira, R. G. (2013). *Análise de Sentimentos com RapidMiner*.

Basile, P., Basile, V., Nissim, M., & Novielli, N. (2015). Deep Tweets: from Entity Linking to Sentiment Analysis. *Second Italian Conference on Computational Linguistics CLiC-It*, 41–45. Retrieved from https://iris.unito.it/retrieve/handle/2318/1532924/75495/Accademia_University_Press_978-88-99200-62-6.pdf#page=271

Basile, P., & Novielli, N. (2014). UNIBA at EVALITA 2014-SENTIPOLC Task : Predicting tweet sentiment polarity combining micro-blogging , lexicon and semantic features. *Evalita - Evaluation of NLP and Speech Tools for Italian*, 58–63. Retrieved from http://s3.amazonaws.com/academia.edu.documents/43743624/UNIBA_at_EVALITA_2014-SENTIPOLC_Task_Pre20160315-15932-1od5t2n.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1477044332&Signature=0bYP13oEDVMFCAz/zmNH4NcvH6A=&response-content-disposition=inlin

Basile, V., Bolioli, A., Nissim, M., Patti, V., & Rosso, P. (2014). Overview of the Evalita 2014 SENTIment POLarity Classification Task. *4th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Evalita 2014*, 50–57. https://doi.org/<10.12871/clicit201429>. <hal-01228925>

Bauwelinck, N., Jacobs, G., Hoste, V., & Lefever, E. (2019). LT3 at SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 436–440. Retrieved from https://www.aclweb.org/anthology/papers/S/S19/S19-2077/

Beineke, P., Hastie, T., Manning, C., & Vaithyanathan, S. (2004). Exploring Sentiment Summarization. *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text Theories and Applications*, *07*, 1–4. Retrieved from

http://www.aaai.org/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-003.pdf

Benítez, R., Escudero, G., & Kanaan, S. (2013). *Inteligencia Artificial Avanzada*. 1–214.

Benkhelifa, R., & Laallam, F. Z. (2018). *Opinion Extraction and Classification of Real-Time YouTube Cooking Recipes Comments*. https://doi.org/10.1007/978-3-319-74690-6_39

Berners-Lee, T., Fischetti, M., & Foreword By-Dertouzos, M. L. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor* (1st ed.). HarperBusiness.

Berry, M. J. A., & Linoff, G. S. (2011). Data mining techniques for Marketing, Sales, and Costumer Relationship Management. In *SIGMOD Record*. https://doi.org/http://doi.acm.org/10.1145/235968.280351

Berson, A., Smith, J. S., & Thearling, K. (2000). *Building Data Mining Applications for CRM*. McGraw-Hill.

Betz, C., Burkhalter, M., & Jung, R. (2019). *Prerequisites for Value Co-Creation in Business Ecosystems*. Retrieved from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1455&context=amcis2019

Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment Analysis: Measuring Opinions. *Procedia Computer Science*, *45*, 808–814. https://doi.org/10.1016/J.PROCS.2015.03.159

Bharti, A., & Lecturer, G. (2016). Custome Churn Management. *UNI JOURNAL OF RESEARCH*, *3*, 14–22.

Bhattacharjya, J., Ellison, A., & Tripathi, S. (2016). An exploration of logistics-related customer service provision on Twitter. *International Journal of Physical Distribution & Logistics Management*, *46*(6/7), 659–680. https://doi.org/10.1108/IJPDLM-01-2015-0007

Bhattacharya, S., Srinivasan, P., & Polgreen, P. (2014). Engagement with Health Agencies on Twitter. *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0112235

Bhuta, S., Doshi, A., Doshi, U., & Narvekar, M. (2014). A Review of Techniques for Sentiment Analysis Of Twitter Data. *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 583–591. https://doi.org/10.1109/ICICICT.2014.6781346

Bigsby, K. G., Ohlmann, J. W., & Zhao, K. (2019). The turf is always greener: Predicting decommitments in college football recruiting using Twitter data. *Decision Support*

Systems, *116*, 1–12. https://doi.org/10.1016/j.dss.2018.10.003

Bijmolt, T. H. A., Leeflang, P. S. H., Block, F., Eisenbeiss, M., Hardie, B. G. S., Lemmens, A., & Saffert, P. (2010). Analytics for customer engagement. *Journal of Service Research*, *13*(3), 341–356. https://doi.org/10.1177/1094670510375603

Borau, K., Ullrich, C., Feng, J., & Shen, R. (2009). Microblogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence. *International Conference on Web-Based Learning "ICWL."*

Bosco, C., Patti, V., & Bolioli, A. (2015). Developing Corpora for Sentiment Analysis : The Case of Irony and Senti – TUT (Extended Abstract). *Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 4158–4162. https://doi.org/10.1109/MIS.2013.28

Brand, W. (2013). *Big Data for Dummies* (1st ed.; I. John Wiley & Sons, Ed.). Hoboken, New Jersey.

Bravo-Marquez, F., Mendoza, M., & Poblete, B. (2013). Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis. *Second International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '13*, 1–9. https://doi.org/10.1145/2502069.2502071

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bruseberg, A., & Mcdonagh-Philp, D. (2002). Focus groups to support the industrial/product designer: a review based on current literature and designers' feedback. In *Applied Ergonomics* (Vol. 33). Retrieved from https://tecfa.unige.ch/tecfa/maltt/cosys-1/textes/bruseberg02.pdf

Bruwer, J., & Johnson, R. (2010). Place-based marketing and regional branding strategy perspectives in the California wine industry. *Journal of Consumer Marketing*, *27*(1), 5–16. https://doi.org/10.1108/07363761011012903

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Cai, Y., Lau, R. Y. K., Liao, S. S. Y., Li, C., Leung, H. F., & Ma, L. C. K. (2014). Object typicality for effective Web of Things recommendations. *Decision Support Systems*, *63*, 52–63. https://doi.org/10.1016/j.dss.2013.09.008

Campos, A. C., Mendes, J., do Valle, P. O., & Scott, N. (2018, March 4). Co-creation of tourist experiences: A literature review. *Current Issues in Tourism*, Vol. 21, pp. 369–400. https://doi.org/10.1080/13683500.2015.1081158

Canales, P. R., & Hernández, A. F. (2013). La Aparición de Nuevos Consumidores. In *Marketing en Una Nueva Era* (pp. 88–105). Ibergaceta Publicaciones S.L.

Carvalho, L. M. (2010). LEGITIMAÇÃO INSTITUCIONAL DO JORNALISMO INFORMATIVO NAS MÍDIAS SOCIAIS DIGITAIS: estratégias emergentes no conteúdo de Zero Hora no Twitter. *Legitimação Institucional Do Jornalismo Informativo Nas Mídias Sociais Digitais: Estratégias Emergentes No Conteúdo de Zero Hora No Twitter*. Dissertação de Mestrado.

Castellucci, G., Croce, D., Cao, D. De, & Basili, R. (2014). A Multiple Kernel Approach for Twitter Sentiment Analysis in Italian. *Evaluation of NLP and Speech Tools for Italian (Evalita´14)*, 98–103. Retrieved from http://www.fileli.unipi.it/projects/clic/proceedings/vol2/clicit2014217.pdf

Chafale, D., & Pimpalkar, A. (2014). Review on Developing Corpora for Sentiment Analysis Using Plutchik's Wheel of Emotions with Fuzzy Logic. *International Journal of Computer Sciences and Engineering (IJCSE)*, *2*(10), 14–18.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2015). *Shiny: web application framework for R* (p. 106). p. 106. Retrieved from https://shiny.rstudio.com/

Chao, M. C. H., & Florenthal, B. (2016). A comparison of global companies' performance on Twitter and Weibo. *International Journal of Business Environment*, *8*(3), 242. https://doi.org/10.1504/IJBE.2016.079691

Chattamvelli, R. (2015a). *Data Mining Methods.* Alpha Science International Ltd.

Chattamvelli, R. (2015b). *Data Mining Methods*. Retrieved from http://cds.cern.ch/record/2216975

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, *54*(8), 88. https://doi.org/10.1145/1978542.1978562

Chaurasia, V., & Pal, S. (2014). Data Mining Techniques : To Predict and Resolve Breast Cancer Survivability. *International Journal of Computer Science and Mobile Computing*, *3*(1), 10–22.

Chen, H., & Zimbra, D. (2010). AI and opinion mining. *IEEE Intelligent Systems*, *25*(3), 74–76. https://doi.org/http://doi.org/10.1109/MIS.2010.75

Chen, & Huang, T. R. (2019). Christians and buddhists are comparably happy on twitter: A large-scale linguistic analysis of religious differences in social, cognitive, and emotional tendencies. *Frontiers in Psychology*, *10*(FEB), 113. https://doi.org/10.3389/fpsyg.2019.00113

Chen, M., Han, J., & Yu, P. S. (1996). Data Mining : An Overview from Database Perspective. *IEEE Transactions on Knowledge*. Retrieved from http://ieeexplore.ieee.org/abstract/document/553155/

Chen, M. K., & Wang, S. C. (2010). The use of a hybrid fuzzy-Delphi-AHP approach to develop global business intelligence for information service firms. *Expert Systems with Applications*, *37*(11), 7394–7407. https://doi.org/10.1016/j.eswa.2010.04.033

Chen, Y. S. (2010). The drivers of green brand equity: Green brand image, green satisfaction, and green trust. *Journal of Business Ethics*, *93*(2), 307–319. https://doi.org/10.1007/s10551-009-0223-9

Chisholm, A., & Hofmann, M. (2016). *Text Mining and Visualisation - Case Studies Using Open-Source Tools* (CRC Press Taylor & Francis Group, Ed.). Retrieved from http://www.crcpress.com

Chowdhury, G. G. (2003). Natural Language Processing. *Annual Review of Information Science and Technology*, *37*(37), 51–89. https://doi.org/10.1002/aris.1440370103

Churches, O., Nicholls, M., Thiessen, M., Kohler, M., & Keage, H. (2014). Emoticons in mind: An event-related potential study. *Social Neuroscience*, *9*(2), 196–202. https://doi.org/10.1080/17470919.2013.873737

Ciribeli, J. P., & Paiva, V. H. P. (2011). Redes e mídias sociais na internet: realidades e perspectivas de um mundo conectado. *Revista Mediação*, *13*(12). Retrieved from http://www.fumec.br/revistas/index.php/mediacao/article/view/509

Coletta, L. F. S., Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. J. (2014). Combining classification and clustering for tweet sentiment analysis. *Brazilian Conference on Intelligent Systems, (BRACIS´14)*, 210–215. https://doi.org/10.1109/BRACIS.2014.46

Cordero-Guzmán, D., & Rodríguez-López, G. (2017). Business intelligence: a strategy for the management of productive enterprises. In *Ciencia UNEMI* (Vol. 10, pp. 40–48).

Cotelo, J. M., Cruz, F. L., Enríquez, F., & Troyano, J. A. (2016). Tweet categorization by combining content and structural knowledge. *Information Fusion*, *31*, 54–64. https://doi.org/10.1016/j.inffus.2016.01.002

Crisci, A., Grasso, V., Nesi, P., Pantaleo, G., Paoli, I., & Zaza, I. (2018). Predicting TV programme audience by using twitter based metrics. *Multimedia Tools and Applications*, *77*(10), 12203–12232. https://doi.org/10.1007/s11042-017-4880-x

Culnan, M. J., McHugh, P., & Zubillaga, J. I. (2010). How large US companies can use Twitter and other social media to gain business value. *MIS Quarterly Executive*, *9*(4),

243–259.

Cumbreras, M. Á. G., Cámara, E. M., Román, J. V., & Morera, J. G. (2016). TASS 2015 - The Evolution of the Spanish Opinion Mining Systems. *Procesamiento de Lenguaje Natural*, *56*, 33–40.

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (2017). Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. *ACM SIGIR Forum*, *51*(2), 148–159. https://doi.org/10.1145/3130348.3130362

Da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. J. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, *66*, 170–179. https://doi.org/10.1016/j.dss.2014.07.003

Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. *Pacific Finance Association Annual Conference 2001, Undefined*. Retrieved from https://scholar.google.com/scholar?hl=es&as_sdt=0%2C5&q=S.+Das+and+M.+Chen%2C+"Yahoo%21+for+Amazon%3A+Extracting+market+sentiment+from+stock+message+boards%2C"+in+Proceedings+of+the+Asia+Pacific+Finance+Association+Annual+Conference+%28APF

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th International Conference on World Wide Web*, 519–528. https://doi.org/10.1145/775152.775226

David, S., & Shwartz, S. S. (2014). Understanding Machine Learning: From Theory to Algorithms. In *Understanding Machine Learning: From Theory to Algorithms*. https://doi.org/10.1017/CBO9781107298019

Deepashri, & Kamath, A. (2017). Survey on Techniques of Data Mining and its Applications. *International Journal of Emerging Research in Management &Technology*, *ISSN*(62), 2278–9359. Retrieved from https://www.ermt.net/docs/papers/Special_Issue/2017/ICETE/33p.pdf

Delibašić, B., Hernández, J. E., Papathanasiou, J., Dargam, F., Zaraté, P., Ribeiro, R., … Linden, I. (2015). Decision Support Systems V - Big Data Analytics for Decision Making: First International Conference, ICDSST 2015 Belgrade, Serbia, May 27-29, 2015 Proceedings. *Lecture Notes in Business Information Processing*, *216*. https://doi.org/10.1007/978-3-319-18533-0

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1–30. Retrieved from http://dl.acm.org/citation.cfm?id=1248547.1248548

Dickinson, B., Ganger, M., & Hu, W. (2015). Dimensionality Reduction of Distributed Vector Word Representations and Emoticon Stemming for Sentiment Analysis. *Journal of Data Analysis and Information Processing*, *3*, 153–162. https://doi.org/10.4236/jdaip.2015.34015

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9

Ding, X., Liu, B., & Yu, P. S. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. *International Conference on Web Search and Web Data Mining (WSDM'08)*, 231–239. https://doi.org/10.1145/1341531.1341561

Dosciatti, M. M., Ferreira, L. P. C., & Paraiso, E. C. (2013). Identificando emoções em textos em português do brasil usando máquina de vetores de suporte em solução multiclasse. *X Encontro Nacional de Inteligência Artificial e Computacional*, 1–12. Retrieved from http://www.ppgia.pucpr.br/~paraiso/mineracaodeemocoes/publicacoes.php

Duarte, A., Brito, A. V., & Medeiros, F. P. A. (2009). Desenvolvimento de um Método para Utilização de Redes Sociais na Internet como Ferramentas de Apoio ao Ensino e Aprendizagem. *Anais Do XX Simpósio Brasileiro de Informática Na Educação*.

Ducange, P., Fazzolari, M., Petrocchi, M., & Vecchio, M. (2019). An effective Decision Support System for social media listening based on cross-source sentiment analysis models. *Engineering Applications of Artificial Intelligence*, *78*, 71–85. https://doi.org/10.1016/j.engappai.2018.10.014

Duncan, B., & Zhang, Y. (2015). Neural Networks for Sentiment Analysis on Twitter. *14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI´CC´15)*, 275–278. https://doi.org/10.1109/ICCI-CC.2015.7259397

Eckerson, W. W. (2008). Performance Dashboards Measuring, Monitoring, and Managing Your Business. In *Business* (Vol. 351). Retrieved from http://books.google.com/books?hl=en&lr=&id=daiXfV1jcakC&oi=fnd&pg=PR7&dq=Performance+dashboards:+measuring,+monitoring,+and+managing+your+business&ots=AIC7mCjG35&sig=iSdPKxxpnFqlfW40QLfsoBNYRoY

Einwiller, S. A., & Steilen, S. (2015). Handling complaints on social network sites - An analysis of complaints and complaint responses on Facebook and Twitter pages of large US companies. *Public Relations Review*, *41*(2), 195–204. https://doi.org/10.1016/j.pubrev.2014.11.012

Elbashir, M. Z., Collier, P. A., & Davern, M. J. (2008). Measuring the effects of business intelligence systems: The relationship between business process and organizational

performance. *International Journal of Accounting Information Systems*, *9*(3), 135–153. https://doi.org/10.1016/J.ACCINF.2008.03.001

Elbashir, M. Z., Collier, P. A., Sutton, S. G., Davern, M. J., & Leech, S. A. (2013). Enhancing the Business Value of Business Intelligence: The Role of Shared Knowledge and Assimilation. *Journal of Information Systems*, *27*(2), 87–105. https://doi.org/10.2308/isys-50563

Epstein, M. J. (2018). *Making Sustainability Work: Best practices in managing and measuring corporate social, environmental and economic impacts*. https://doi.org/10.4324/9781351280129

Esuli, A., & Sebastiani, F. (2006). Determining term subjectivity and term orientation for opinion mining. *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, *2*(1), 193–200. Retrieved from http://acl.ldc.upenn.edu/eacl2006/main/papers/13_1_esulisebastiani_192.pdf

Esuli, A., & Sebastiani, F. (2010). Machines that learn how to code open-ended survey data. *International Journal of Market Research*, *52*(6), 775–800. https://doi.org/10.2501/S147078531020165X

Ewing, M. T., Wagstaff, P. E., & Powell, I. H. (2013). Brand rivalry and community conflict. *Journal of Business Research*, *66*(1), 4–12. https://doi.org/10.1016/j.jbusres.2011.07.017

Fan, S., Lau, R. Y. K., & Zhao, J. L. (2015). Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix. *Big Data Research*, *2*(1), 28–32. https://doi.org/10.1016/j.bdr.2015.02.006

Farías, Delia Irazú Hernańdez, Patti, V., & Rosso, P. (2016). Irony Detection in Twitter: The Role of Affective Content. *ACM Transactions on Internet Technology*, *16*(3), 1–24. https://doi.org/10.1145/2930663

Farías, Delia Irazú Hernández, Patti, V., & Rosso, P. (2018). ValenTO at SemEval-2018 Task 3: Exploring the Role of Affective Content for Detecting Irony in English Tweets. *Proceedings of The 12th International Workshop on Semantic Evaluation*, 643–648. Retrieved from http://www.aclweb.org/anthology/S18-1105

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding Topic Signals in Large-Scale Text. *Conference on Human Factors in Computing Systems - CHI'16*, 4647–4657. https://doi.org/10.1145/2858036.2858535

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37.

https://doi.org/10.1609/aimag.v17i3.1230

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. https://doi.org/10.1145/240455.240464

Femina, B. T., & Sudheep, E. M. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia Computer Science*, *46*(Icict 2014), 725–731. https://doi.org/10.1016/j.procs.2015.02.136

Ferreira, T., Pedrosa, I., & Bernardino, J. (2017). *Business Intelligence for E-commerce: Survey and Research Directions*. https://doi.org/10.1007/978-3-319-56535-4_22

Flavián, C., Guinalíu, M., & Torres, E. (2005). The influence of corporate image on consumer trust: A comparative analysis in traditional versus internet banking. *Internet Research*, Vol. 15, pp. 447–470. https://doi.org/10.1108/10662240510615191

Fleiss, J. L. (1971). *Measuring Nominal Scale Agreement Among Many Raters*. *76*(5), 378–382. Retrieved from http://www.wpic.pitt.edu/research/biometrics/Publications/Biometrics Archives PDF/395-1971 Fleiss0001.pdf

Flekova, L., Preoţiuc-Pietro, D., & Ruppert, E. (2015). *Analysing domain suitability of a sentiment lexicon by identifying distributionally bipolar words*. 77–84. https://doi.org/10.18653/v1/w15-2911

Foley, É., & Guillemette, M. G. (2010). What is Business Intelligence? *International Journal of Business Intelligence Research*, *1*(4), 1–28. https://doi.org/10.4018/jbir.2010100101

Fong, N. M., Fang, Z., & Luo, X. (2014). Geo-Conquesting: Competitive Locational Targeting of Mobile Promotions. *Ssrn*, *LII*(October), 726–735. https://doi.org/10.2139/ssrn.2439398

Fresno, N. (2014). Is a picture worth a thousand words? The role of memory in audio description. *Across Languages and Cultures*, *15*(1), 111–129. https://doi.org/10.1556/Acr.15.2014.1.6

Friedland, L., Joseph, K., Swire-Thompson, B., Grinberg, N., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Friedman, M. (1940). A Comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, *11*, 86–92.

https://doi.org/10.1214/aoms/1177731944

Gandini, A. (2016). *The Reputation Economy: Understanding Knowledge Work in Digital Society*. Retrieved from https://books.google.com.br/books?hl=es&lr=&id=2ItPDAAAQBAJ&oi=fnd&pg=PR5&dq=reputation+economy&ots=KQvP6dd21s&sig=-yxuLe_zr6WGpSmqvuf__PmAdIs#v=onepage&q=reputation economy&f=false

García-Peñalvo, F. J., & Conde-González, M. A. (2017). Statistical Implicative Analysis Approximation to KDD and Data Mining: A Systematic and Mapping Review in Knowledge Discovery Database Framework. *Ninth International Conference on Advances in Databases, Knowledge, and Data Applications*, (c), 70–77.

García, S., Molina, D., Herrera, F., & Lozano, M. (2007). Tests no paramétricos de comparaciones múltiples con algoritmo de control en el análisis de algoritmos evolutivos: Un caso de estudio con los resultados de la sesión especial en optimización continua. *Actas de Las I Jornadas Sobre Algoritmos Evolutivos y Metaheurísticas*, 219–227.

Gardener, M. (2012). *Beginning R: the statistical programming language*. Retrieved from https://books.google.es/books?hl=pt-BR&lr=&id=iJoKYSWCubEC&oi=fnd&pg=PR21&dq=Beginning+R:+the+statistical+programming+language&ots=5845Exazcy&sig=KRsZXENtCDuiZaA2_VN6YWf41XM&redir_esc=y#v=onepage&q=Beginning R%3A the statistical programming language&f=fal

Gaspar, R., Pedro, C., Panagiotopoulos, P., & Seibt, B. (2016). Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior*, *56*, 179–191. https://doi.org/10.1016/j.chb.2015.11.040

George, A., Barathi Ganesh, H. B., Anand Kumar, M., & Soman, K. P. (2018). TeamCEN at SemEval-2018 Task 1: Global Vectors Representation in Emotion Detection. *Proceedings of The 12th International Workshop on Semantic Evaluation*, 334–338. Retrieved from https://nlp.stanford.edu/projects/

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Reyes, A., & Barnden, J. (2015). SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. *9th International Workshop on Semantic Evaluation (SemEval´15)*, 470–478. Retrieved from http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval080.pdf

Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report, Stanford*, (1), 12. https://doi.org/10.1016/j.sedgeo.2006.07.004

Go, A., Huang, L., & Bhayani, R. (2009). Twitter Sentiment Analysis. *CS224N - Final Project Report*, 17. https://doi.org/10.1007/978-3-642-35176-1_32

Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D., & Gelbukh, A. (2016). Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. *Sensors (Switzerland)*, *16*(9), 1374. https://doi.org/10.3390/s16091374

Gonçalves, R., & Fernandes Brito, P. (2013). Análise de sentimento usando Support Vector Machine. *XV Encoinfo - Encontro de Computaçao e Informática Do Tocantins*, 58–67. Retrieved from http://ulbra-to.br/encoinfo/index.php/encoinfo/encoinfo-2013/paper/view/20/255

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: a closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, 581–586. Association for Computational Linguistics.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Retrieved from https://books.google.com.br/books?hl=pt-BR&lr=&id=omivDQAAQBAJ&oi=fnd&pg=PR5&dq=%5BGoodfellow+et+al.+2016%5D+Goodfellow,+I.,+Bengio,+Y.,+e+Courville,+A.+(2016).+Deep+learning+(adaptive+computation+and+machine+learning+series).+Cambridge:+The+MIT+Press.&ot

Greene, S., & Resnik, P. (2009). More thanWords: Syntactic Packaging and Implicit Sentiment. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL* (pp. 503–511). https://doi.org/10.3115/1620754.1620758

Grégoire, Y., Salle, A., & Tripp, T. M. (2015). Managing social media crises with your customers: The good, the bad, and the ugly. *Business Horizons*, *58*(2), 173–182. https://doi.org/10.1016/j.bushor.2014.11.001

Gupte, A., Joshi, S., Gadgul, P., & Kadam, A. (2014). Comparative Study of Classification Algorithms used in Sentiment Analysis. *(IJCSIT) International Journal of Computer Science and Information Technologies*, *5*(5), 6261–6264. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.660.5055&rep=rep1&type=pdf

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*.

Han, J., Kamber, M., & Jian, P. (2012). Data mining: Concepts and techniques. In *Elsevier* (Vol. 7).

Hangya, V., & Farkas, R. (2016). A comparative empirical study on social media sentiment analysis over various genres and languages. *Artificial Intelligence Review*,

1–21. https://doi.org/10.1007/s10462-016-9489-3

Haque, T. U., Saber, N. N., & Shah, F. M. (2018). Sentiment analysis on large scale Amazon product reviews. *2018 IEEE International Conference on Innovative Research and Development, ICIRD 2018*, 1–6. https://doi.org/10.1109/ICIRD.2018.8376299

Haro-de-Rosario, A., Sáez-Martín, A., & Caba-Pérez, M. del C. (2018). Using social media to enhance citizen engagement with local government: Twitter or Facebook? *New Media and Society*, *20*(1), 29–49. https://doi.org/10.1177/1461444816645652

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, 174–181. https://doi.org/10.3115/979617.979640

Haykin, S. (2004). Neural networks: a comprehensive foundation. In *McMaster University Hamilton, Ontario, Canada* (Vol. 2). https://doi.org/10.1017/S0269888998214044

Hemsley, B., Palmer, S., Dann, S., & Balandin, S. (2018). Using Twitter to access the human right of communication for people who use Augmentative and Alternative Communication (AAC). *International Journal of Speech-Language Pathology*, *20*(1), 50–58. https://doi.org/10.1080/17549507.2017.1413137

Hernández Farías, D. I., Laganà, I., Patti, V., & Bosco, C. (2017). Towards an Italian lexicon for polarity classification (polarITA): A comparative analysis of lexical resources for sentiment analysis. *CEUR Workshop Proceedings*, *2006*(December 2017), 11–12. https://doi.org/10.4000/books.aaccademia.2417

Herschel, R. T. (2019). Business Intelligence. In *Advanced Methodologies and Technologies in Business Operations and Management* (p. 11). https://doi.org/10.4018/978-1-5225-7362-3.ch043

Hill, J. (2016). The Impact of Emojis and Emoticons on Online Consumer Reviews , Perceived Company Response Quality , Brand Relationship , and Purchase Intent . *University of South Florida Scholar Commons Graduate*, *4*(November), 85. Retrieved from http://scholarcommons.usf.edu/etdhttp://scholarcommons.usf.edu/etd/6513

Hodes, R. L., Cook, E. W., & Lang, P. J. (1985). Individual differences in autonomic response: conditioned association or conditioned fear? *Psychophysiology*, *22*(5), 545–560. https://doi.org/10.1111/j.1469-8986.1985.tb01649.x

Hootsuite. (2019). *2019 Digital Yearbook: headline internet, social media and mobile use*

data for every country in the world. Retrieved from https://pt.slideshare.net/wearesocial/2018-digital-yearbook-86862930

Hopke, J. E., & Simis, M. (2017). Discourse over a contested technology on Twitter: A case study of hydraulic fracturing. *Public Understanding of Science*, *26*(1), 105–120. https://doi.org/10.1177/0963662515607725

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, 168–177. https://doi.org/http://dx.doi.org/10.1145/1014052.1014073

Hu, X., Tang, L., Tang, J., & Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. *Sixth ACM International Conference on Web Search and Data Mining (WSDM´13)*, 537–546. https://doi.org/10.1145/2433396.2433465

Hurtado, L.-F., & Pla, F. (2014). ELiRF-UPV en TASS 2014: Análisis de Sentimientos, Detección de Tópicos y Análisis de Sentimientos de Aspectos en Twitter. *Procesamiento Del Lenguaje Natural*, 1–7.

Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, *30*(4), 330–338. https://doi.org/10.1016/j.jksues.2016.04.002

IAB. (2017). *Estudio Anual Redes Sociales 2017*. Retrieved from http://iabspain.es/wp-content/uploads/iab_estudioredessociales_2017_vreducida.pdf

Ilhan, B. E., Kübler, R. V., & Pauwels, K. H. (2018). Battle of the Brand Fans: Impact of Brand Attack and Defense on Social Media. *Journal of Interactive Marketing*, *43*, 33–51. https://doi.org/10.1016/j.intmar.2018.01.003

Ingenbleek, P. T. M., & van der Lans, I. A. (2013). Relating price strategies and price-setting practices. *European Journal of Marketing*, *47*(1), 27–48. https://doi.org/10.1108/03090561311285448

Inmon, W. H., & Linstedt, D. (2014). *Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault*. https://doi.org/10.1016/B978-0-12-802044-9/00046-5

Istanbulluoglu, D. (2017). Complaint handling on social media: The impact of multiple response times on consumer satisfaction. *Computers in Human Behavior*, *74*, 72–82. https://doi.org/10.1016/j.chb.2017.04.016

Jakob, N., & Gurevych, I. (2010). Extracting opinion targets in a single-and cross-domain setting with conditional random fields. *Proceedings of the 2010 Conference on*

Empirical Methods in Natural Language Processing, (October), 1035–1045. Retrieved from http://portal.acm.org/citation.cfm?id=1870759

Jiang, Li, Z., & Ye, X. (2019). Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartography and Geographic Information Science*, *46*(3), 228–242. https://doi.org/10.1080/15230406.2018.1434834

Jiang, M., Kumar, S., Subrahmanian, V. S., & Faloutsos, C. (2017). *KDD 2017 Tutorial: Data-Driven Approaches towards Malicious Behavior Modeling*. (17), 3–6. https://doi.org/10.475/123

Jindal, N., & Liu, B. (2006). Identifying comparative sentences in text documents. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*, (April), 244. https://doi.org/10.1145/1148170.1148215

Joachims, T. (1999). Making large-scale SVM learning practical. *Universität Dortmund*, 41–56. https://doi.org/10.1109/ICEMI.2009.5274151

Johny, M. C. P., & Scholar, P. G. (2017). Customer Churn Prediction : A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, *8*(5), 2178–2181.

Jukes, S. (2019). Crossing the Line between News and the Business of News: Exploring Journalists' Use of Twitter. *Media and Communication*, *7*(1), 248. https://doi.org/10.17645/mac.v7i1.1772

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: an introduction to natural language processing* (S. Russell & P. Norvig, Eds.). Retrieved from http://books.google.com/books?id=fZmj5UNK8AQC&pgis=1

Kamps, J., Marx, M., Mokken, R. J., & Rijke, M. De. (2004). Using WordNet to measure semantic orientations of adjectives. *FNWI: Institute for Logic, Language and Computation (ILLC)*, 1–5. Retrieved from http://dare.uva.nl/document/2/37038

Kandadai, V., Yang, H., Jiang, L., Yang, C., Fleisher, L., & Winston, F. K. (2016). Measuring Health Information Dissemination and Identifying Target Interest Communities on Twitter: Methods Development and Case Study of the @SafetyMD Network. *JMIR Research Protocols*, *5*(2). https://doi.org/10.2196/resprot.4203

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/http://dx.doi.org/10.1016/j.bushor.2009.09.003

Karpathy, A. (2017). Convolutional neural networks for visual recognition. Retrieved from http://cs231n.github.io/convolutional-networks/

Katarya, R., & Yadav, A. (2018). A comparative study of genetic algorithm in sentiment analysis. *Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018*, (Icisc), 136–141. https://doi.org/10.1109/ICISC.2018.8399051

Keh, H. T., & Pang, J. (2010). Customer Reactions to Service Separation. *Journal of Marketing*, *74*(2), 55–70. https://doi.org/10.1509/jmkg.74.2.55

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *7*(1), 100–107. https://doi.org/10.1109/2945.981847

Kenyon-Dean, K., Ahmed, E., Fujimoto, S., Georges-Filteau, J., Glasz, C., Kaur, B., … Ruths, D. (2018). *Sentiment Analysis: It's Complicated!* 1886–1895. https://doi.org/10.18653/v1/n18-1171

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, 1367–1374. https://doi.org/10.3115/1220355.1220555

Kim, Song, Y. I., & Rim, H. C. (2016). Opinion retrieval for twitter using extrinsic information. *Journal of Universal Computer Science*, *22*(5), 608–629. Retrieved from https://pdfs.semanticscholar.org/1fa7/95421ecb82e1e6d5266b597daf909b78eb30.pdf

Kohl, C., Knigge, M., Baader, G., Böhm, M., & Krcmar, H. (2018). Anticipating acceptance of emerging technologies using twitter: the case of self-driving cars. *Journal of Business Economics*, *88*(5), 617–642. https://doi.org/10.1007/s11573-018-0897-5

Kotler, P., & Armstrong, G. (2013). Principles of Marketing 15th Global Edition. In *Pearson Education Limited* (Vol. 6). Retrieved from https://www.bookdepository.com/Principles-Marketing-Global-Edition-Dr-Philip-T-Kotler/9781292220178%0Ahttps://books.google.com/books?id=TahuAwAAQBAJ&pgis=1

Kotler, P., Kartajaya, H., & Setiawan, I. (2010). *Marketing 3.0 : from products to customers to the human spirit*. Wiley.

Kotsenas, A. L., Arce, M., Aase, L., Timimi, F. K., Young, C., & Wald, J. T. (2018). The Strategic Imperative for the Use of Social Media in Health Care. *Journal of the American College of Radiology*, *15*(1), 155–161.

https://doi.org/10.1016/j.jacr.2017.09.027

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. *Departmental Papers (ASC)*, 1–12. Retrieved from http://repository.upenn.edu/asc_papers

Kubat, M., Bratko, I., & Michalski, R. (1998). *A Review of Machine Learning Methods*. https://doi.org/10.1017/CBO9781107415324.004

Kumar, A., & Jaiswal, A. (2019). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency Computation*. https://doi.org/10.1002/cpe.5107

Kuo, Y. F., & Feng, L. H. (2013). Relationships among community interaction characteristics, perceived benefits, community commitment, and oppositional brand loyalty in online brand communities. *International Journal of Information Management*, *33*(6), 948–962. https://doi.org/10.1016/j.ijinfomgt.2013.08.005

Lafaye de Micheaux, P., Drouilhet, R., & Liquet, B. (2013). The R Software: Fundamentals of Programming and Statistical Analysis. In *Statistics and Computing*. https://doi.org/10.1007/978-1-4614-9020-3

Lahuerta-Otero, E., & Cordero-Gutiérrez, R. (2016). Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter. *Computers in Human Behavior*, *64*, 575–583. https://doi.org/10.1016/j.chb.2016.07.035

Langley, P. (1992). An Analysis of Bayesian Classifiers. *Aaai*, (90), 223–228.

Laroche, M., Habibi, M. R., & Richard, M. O. (2013). To be or not to be in social media: How brand loyalty is affected by social media? *International Journal of Information Management*, *33*(1), 76–82. https://doi.org/10.1016/j.ijinfomgt.2012.07.003

Laroche, M., Habibi, M. R., Richard, M. O., & Sankaranarayanan, R. (2012). The effects of social media based brand communities on brand community markers, value creation practices, brand trust and brand loyalty. *Computers in Human Behavior*, *28*(5), 1755–1767. https://doi.org/10.1016/j.chb.2012.04.016

Larose, D. T. (2014). *Discovering knowledge in data : An introduction to data mining* (Vol. 2; W. John, Ed.). https://doi.org/http://dx.doi.org/10.1002/9781118874059

Lau, R. Y. K., Li, C., & Liao, S. S. Y. (2014). Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. *Decision Support Systems*, *65*(C), 80–94. https://doi.org/10.1016/j.dss.2014.05.005

Lee, C., & Longo, V. (2016). Dietary restriction with and without caloric restriction for healthy aging. *F1000Research*, *5*(0), 1–7.

https://doi.org/10.12688/f1000research.7136.1

Lee, H.-M., Long, J., & Visinescu, L. (2016). The Relationship between a Business Simulator, Constructivist Practices, and Motivation toward Developing Business Intelligence Skills. *Interdisciplinary Journal of Information, Knowledge, and Management*, *15*, 593–609. Retrieved from https://www.informingscience.org/Publications/3602?Search=business intelligence

Lee, I. (2018). Social media analytics for enterprises: Typology, methods, and processes. *Business Horizons*, *61*(2), 199–210. https://doi.org/10.1016/j.bushor.2017.11.002

Lee, M. T., & Widener, S. K. (2016). The Performance Effects of Using Business Intelligence Systems for Exploitation and Exploration Learning. *Journal of Information Systems*, *30*(3), 1–31. https://doi.org/10.2308/isys-51298

Lee, S. W., Song, Y. I., Lee, J. T., Han, K. S., & Rim, H. C. (2012). A new generative opinion retrieval model integrating multiple ranking factors. *Journal of Intelligent Information Systems*, *38*(2), 487–505. https://doi.org/10.1007/s10844-011-0164-5

Leung, H. C., & Chan, W. T. Y. (2017). Using emoji effectively in marketing: an empirical study. *Digital & Social Media Marketing*, *5*(1), 76–96. Retrieved from https://www.ingentaconnect.com/content/hsp/jdsmm/2017/00000005/00000001/art00007

Lévy, P. (2004). Inteligencia colectiva. In La Découverte (Essais) (Ed.), *Por una antropología del ciberespacio*.

Li, Ott, M., Cardie, C., & Hovy, E. (2014). Towards a General Rule for Identifying Deceptive Opinion Spam. *Acl-2014*, 1566–1576. https://doi.org/10.3115/v1/P14-1147

Li, Q., Wang, C., Liu, R., Wang, L., Zeng, D. D., & Leischow, S. J. (2018). Understanding Users' Vaping Experiences from Social Media: Initial Study Using Sentiment Opinion Summarization Techniques. *Journal of Medical Internet Research*, *20*(8), e252. https://doi.org/10.2196/jmir.9373

Li, S.-T., & Tsai, F.-C. (2013). A fuzzy conceptualization model for text mining with application in opinion polarity classification. *Knowledge-Based Systems*, *39*, 23–33. https://doi.org/10.1016/j.knosys.2012.10.005

Li, S. T., Shue, L. Y., & Lee, S. F. (2008). Business intelligence approach to supporting strategy-making of ISP service management. *Expert Systems with Applications*, *35*(3), 739–754. https://doi.org/10.1016/j.eswa.2007.07.049

Li, Y. H., & Jain, A. K. (1998). Classification of text documents. *The Computer Journal*, *41*(8), 537–546. https://doi.org/10.1093/comjnl/41.8.537

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*, *2*(December), 18–22.

Lim, E. E., Chen, H. H., & Chen, G. G. (2013). Business intelligence and analytics: Research directions. *ACM Transactions on Management Information Systems*, *3*(4), 1–10. https://doi.org/10.1145/2407740.2407741

Liozu, S. M., & Hinterhuber, A. (2013). Pricing orientation, pricing capabilities, and firm performance. *Management Decision*, *51*(3), 594–614. https://doi.org/10.1108/00251741311309670

Liu, B. (2007). Web data mining: exploring hyperlinks, contents, and usage data. In *Data-centric systems and applications*. https://doi.org/http://dx.doi.org/10.1007/978-3-642-19460-3

Liu, B. (2012). Sentiment Analysis and Opinion Mining. In *Synthesis lectures on human language technologies* (Vol. 5). https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, B. (2017). Many Facets of Sentiment Analysis. In *A Practical Guide to Sentiment Analysis* (pp. 11–39). https://doi.org/10.1007/978-3-319-55394-8_2

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. *Proceedings of the 14th International Conference on World Wide Web*, 342–351. https://doi.org/http://dx.doi.org/10.1145/1060745.1060797

Lochter, J. V., Pires, P. R., Bossolani, C., Yamakami, A., & Almeida, T. A. (2018). Evaluating the impact of corpora used to train distributed text representation models for noisy and short texts. *Proceedings of the International Joint Conference on Neural Networks*, *2018-July*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489355

López, M. C. L. de A., García, B. C., & Fernández, J. G. (2018). Estrategias de gestión de los clubes de golf de la Comunidad de Madrid en Twitter. *Cuadernos.Info*, (42), 71–84. https://doi.org/10.7764/CDI.42.1304

López, Morales, R., & Cavero, O. (2018). Digital marketing strategy through social networks in the context of Ecuadorian SMES. *CienciAmerica*, *7*(2), 18. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=6553438

Loshin, D. (2013). Business Intelligence: The Savvy Manager's Guide. In *Morgan Kauf*. https://doi.org/10.1016/B978-0-12-385889-4.00001-6

Lu, & Seah, Z. Y. (2018). Social Media Influencers and Consumer Online Engagement Management. In *Digital Marketing and Consumer Engagement*. https://doi.org/10.4018/978-1-5225-5187-4.ch070

Lu, X., Ba, S., Huang, L., & Feng, Y. (2013). Promotional marketing or word-of-mouth? Evidence from online restaurant reviews. *Information Systems Research*, *24*(3), 596–612. https://doi.org/10.1287/isre.1120.0454

Luo, Z., Osborne, M., & Wang, T. (2015). An effective approach to tweets opinion retrieval. *World Wide Web*, *18*(3), 545–566. https://doi.org/10.1007/s11280-013-0268-7

Lusch, R. F., Liu, Y., & Chen, Y. (2010). The phase transition of markets and organizations: The new intelligence and entrepreneurial frontier. *IEEE Intelligent Systems*, *25*(1), 71–75. https://doi.org/10.1109/MIS.2010.27

Lv, H., Yu, G., & Wu, G. (2018). Relationships among customer loyalty, customer satisfaction, corporate image and behavioural intention on social media for a corporation. *International Journal of Information Technology and Management*, *17*(3), 170. https://doi.org/10.1504/IJITM.2018.092879

Ma, S., Sun, X., Lin, J., & Ren, X. (2018). *A Hierarchical End-to-End Model for Jointly Improving Text Summarization and Sentiment Classification*. Retrieved from http://snap.stanford.edu/data/web-Amazon.html

Magnani, M., Montesi, D., & Rossi, L. (2011). Conversation retrieval for microblogging sites. *Information Retrieval*, *15*(3–4), 354–372. https://doi.org/10.1007/s10791-012-9189-9

Maipradit, R., Hata, H., & Matsumoto, K. (2019). Sentiment Classification using N-gram IDF and Automated Machine Learning. *IEEE Software*, 1–1. https://doi.org/10.1109/MS.2019.2919573

Maks, I., Izquierdo, R., Frontini, F., Agerri, R., Azpeitia, A., & Vossen, P. (2014). Generating Polarity Lexicons with WordNet propagation in five languages. *European Language Resources Association (ELRA)*, 1155–1161. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/847_Paper.pdf

Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, Vol. 30, pp. 3514–3554. https://doi.org/10.1108/IJCHM-07-2017-0461

Marolt, M., Zimmermann, H. D., & Pucihar, A. (2018). Exploratory study of social CRM use in SMEs. *Engineering Economics*, *29*(4), 468–477.

https://doi.org/10.5755/j01.ee.29.4.20246

Maroto, C. (2016). *Big data 2016 "IN A NUTSHELL."* https://doi.org/10.1089/big.2013.0037.Published

Martín-Valdivia, M. T., Martínez-Cámara, E., Perea-Ortega, J. M., & Ureña-López, L. A. (2013). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, *40*(10), 3934–3942. https://doi.org/10.1016/j.eswa.2012.12.084

Martínez-López, F. J., Anaya-Sánchez, R., Aguilar-Illescas, R., & Molinillo, S. (2015). *Online Brand Communities: Using the Social Web for Branding and Marketing*. https://doi.org/10.1007/978-3-319-24826-4

McCue, C. (2007). *Data Mining and Predictive Analysis: Inteligence Gathering an Crime Analysis*. Elsevier Inc.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Merchanteí, P. (2015). Big data. El nuevo mantra. *En Nuevas Tecnologías En La Investigación de Mercados, Investigación y Marketing N. 128*, 20–23.

Michalski, R. S. (1983). A theory of methodology of inductive learning. In *Machine Learning: An Artificial Intelligence Approach* (pp. 111–161). https://doi.org/10.1016/0004-3702(83)90016-4

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mishra, D., & Kumar, R. (2017). Knowledge Discovery in Databases (KDD): A Comparative Evaluation of Scientific Databases. *Asian Journal of Information Science*. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=22316108&AN=127115910&h=Qx99BKbNIKoC6wVL212BYgfwONHcKrOE%2BA9G6n%2FiFnD5FcehbRcjD3GDzFS8ve91E7EWXhaC0BsULhe1BVgtNg%3D%3D&crl=c

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Higher Education.

Moen, Ø., Havro, L. J., & Bjering, E. (2017). Online consumers reviews: Examining the moderating effects of product type and product popularity on the review impact on sales. *Cogent Business and Management*, *4*(1), 1–20. https://doi.org/10.1080/23311975.2017.1368114

Moens, M.-F., Li, J., & Chua, T.-S. (2014). Mining user generated content and its applications. In *Mining User Generated Content* (pp. 3–17). https://doi.org/US7618719

Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, *2*(August), 599–608. https://doi.org/10.3115/1699571.1699591

Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *Seventh International Workshop on Semantic Evaluation Exercises (SemEval´13)*, *2*, 321–327.

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2016). Stance and Sentiment in Tweets. *ACM Transactions on Embedded Computing Systems*, *0*(0), 22. https://doi.org/0000001.0000001

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management*, *51*(4), 480–499. https://doi.org/10.1016/j.ipm.2014.09.003

Mohanty, A., & Das, S. (2017). Impact of Customer Relationship Management (CRM) on Telecom sector in. *International Journal of Scientific Research in Science and Technology*, *3*(8), 431–437.

Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., & Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, *40*(18), 7250–7257. https://doi.org/10.1016/j.eswa.2013.06.076

Momtazi, S. (2012). Fine-grained German sentiment analysis on social media. *9th Intl. Conf. on Language Resources and Evaluation*, 1215–1220. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/999_Paper.pdf

Monard, M., & Baranauskas, J. (2003). Conceitos sobre aprendizado de máquina. In *Sistemas Inteligentes-Fundamentos e Aplicações* (pp. 39–56). Retrieved from http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf

Montesi, M., & Navarrete, T. (2008). Classifying web genres in context: A case study documenting the web genres used by a software engineer. *Information Processing and Management*, *44*(4), 1410–1430. https://doi.org/10.1016/j.ipm.2008.02.001

Montoyo, A., Martínez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments.

Decision Support Systems, *53*(4), 675–679. https://doi.org/10.1016/j.dss.2012.05.022

Moreno, A., & Lara, J. (2017). Análisis de actividad de un srvicio de teleasistencia social mediante Big DaAta y Data Mining. *Centros de Estudios Financieros*, *6*, 88–102.

Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002). Mining product reputations on the web. *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD´02)*, 341–349. https://doi.org/10.1145/775094.775098

Moss, L. T., & Atre, S. (2003). Business Intelligence Roadmap: The Complete Project Lifecycle for Decision- Support Applications. In *Communication*. Retrieved from https://books.google.com.br/books?hl=pt-BR&lr=&id=ZV8jeV4a9_AC&oi=fnd&pg=PR7&dq=Topic%2BOverview%2B+%22Business%2BIntelligence%22&ots=LumALeDQC5&sig=RpgTD-KDcjcVTpIlxvJxYockvA4#v=onepage&q=Topic%2BOverview%2B%22Business%2BIntelligence%22&f=false

Mount, M., & Martinez, M. G. (2014). Social Media: A Tool for Open Innovation. *California Management Review*, *56*(4), 124–143. https://doi.org/10.1525/cmr.2014.56.4.124

Muenchen, R. A. (2014). *The Popularity of Data Analysis Software*. (April). Retrieved from http://r4stats.com/articles/popularity/

Mukherjee, S., & Bhattacharyya, P. (2012). Sentiment Analysis in Twitter with Lightweight Discourse Analysis. *Coling*, 1847–1864. Retrieved from http://cse.iitk.ac.in/users/cs671/2013/submissions/rkjha/hw3/hw3.pdf

Mukherjee, Subhabrata, Malu, A., Balamurali, A. R., & Bhattacharyya, P. (2012). TwiSent: A Multistage System for Analyzing Sentiment. *Conference on Information and Knowledge Management (CIKM´12)*, 2531–2534. https://doi.org/10.1145/2396761.2398684

Müller, R. M., Linders, S., & Pires, L. F. (2010). Business Intelligence and Service-oriented Architecture: A Delphi Study. *Information Systems Management*, *27*(2), 168–187. https://doi.org/10.1080/10580531003685238

Naiar, F., Bourouis, S., Bouguila, N., & Belghith, S. (2018). A Fixed-Point Estimation Algorithm for Learning the Multivariate GGMM: Application to Human Action Recognition. *Canadian Conference on Electrical and Computer Engineering*, *2018-May*, 1–4. https://doi.org/10.1109/CCECE.2018.8447761

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., & Wilson, T. (2013).

SemEval-2013 Task 2: Sentiment Analysis in Twitter. *International Workshop on Semantic Evaluation (SemEval´13)*, *2*, 312–320.

Narr, S., Hülfenhaus, M., & Albayrak, S. (2012). Language-independent Twitter sentiment analysis. *Knowledge Discovery and Machine Learning (KDML´12)*, 12–14. Retrieved from http://www.dai-labor.de/fileadmin/Files/Publikationen/Buchdatei/narr-twittersentiment-KDML-LWA-2012.pdf

Nascimento, P., Aguas, R., Lima, D. de, Kong, X., Osiek, B., Xexéo, G., & Souza, J. de. (2015). Análise de sentimento de tweets com foco em notícias. *Revista Eletrônica de Sistemas de Informação*, *14*(2), 12. https://doi.org/10.5329/RESI

Nasukawa, T., & Yi, j. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd International Conference*. Retrieved from https://dl.acm.org/citation.cfm?id=945658

Negash, Solomon, & Gray, P. (2008). Business Intelligence. In *Handbook on Decision Support Systems 2* (pp. 175–193). https://doi.org/10.1007/978-3-540-48716-6_9

Negash, Solomun. (2004). Business Intelligence. *Communications of the Association for Information Systems*, *Vol. 13*(February), Article 15. https://doi.org/10.1002/9781118915240.ch7

Nelson, G. S. (2010). Business Intelligence 2 . 0 : Are we there yet ? *SAS Global Formun 2010*, 1–10. https://doi.org/10.1.1.176.2806

Nisbet, R., Miner, G., & Yale, K. (2018). Deep Learning. In *Handbook of Statistical Analysis and Data Mining Applications* (pp. 741–751). https://doi.org/10.1016/B978-0-12-416632-5.00019-0

Noureen, R., Qamar, U., Khan, F. H., & Muhammad, I. (2018). InstaSent: A novel framework for sentiment analysis based on instagram selfies. *Advances in Intelligent Systems and Computing*, *868*, 323–336. https://doi.org/10.1007/978-3-030-01054-6_23

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications and Strategies*, *65*(1), 17–37. https://doi.org/Encontrar DOI http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1008839

Obaidat, I., Mohawesh, R., Al-Ayyoub, M., AL-Smadi, M., & Jararweh, Y. (2015). Enhancing the Determination of Aspect Categories and Their Polarities in Arabic Reviews Using Lexicon-Based Approaches. *Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT´15)*, 1–6.

https://doi.org/10.1109/AEECT.2015.7360595

Okada, M., Yanagimoto, H., & Hashimoto, K. (2019). Sentiment Classification with Gated CNN for Customer Reviews. *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing, ISAI-NLP 2018 - Proceedings*. https://doi.org/10.1109/iSAI-NLP.2018.8692959

Olbrich, R., & Holsing, C. (2011). Modeling Consumer Purchasing Behavior in Social Shopping Communities with Clickstream Data. *International Journal of Electronic Commerce*, *16*(2), 15–40. https://doi.org/10.2753/JEC1086-4415160202

Olson, D. L., & Delen, D. (2008). Advanced data mining techniques. In *Control* (1st ed.). https://doi.org/10.1007/978-3-540-76917-0

Olvera-Lobo, M. D., Castillo-Rodríguez, C., & Gutiérrez-Artacho, J. (2018). Spanish SME use of Web 2.0 tools and web localisation processes. *International Conferences on WWW/Internet, ICWI 2018 and Applied Computing 2018*, 35–42. Retrieved from http://digibug.ugr.es/bitstream/handle/10481/53559/Spanish SME use of Web 2.0 tools and web localisation processes.pdf?sequence=1&isAllowed=y

Orellana-Rodriguez, C., & Keane, M. T. (2018, August). Attention to news and its dissemination on Twitter: A survey. *Computer Science Review*, Vol. 29, pp. 74–94. https://doi.org/10.1016/j.cosrev.2018.07.001

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, *31*(1), 527–541. https://doi.org/10.1016/j.chb.2013.05.024

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annua Meeting of the Association for Computational Linguistics*, 1–11. Retrieved from http://arxiv.org/abs/1107.4557

Pan, Y., Gao, H., Lin, H., Liu, Z., Tang, L., & Li, S. (2018). Identification of Bacteriophage Virion Proteins Using Multinomial Naïve Bayes with g-Gap Feature Tree. *International Journal of Molecular Sciences Article*. https://doi.org/10.3390/ijms19061779

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summation based on minimum cuts. *Proceedings of 42nd Annual Meeting on Association for Computational Linguistics (ACL´04)*, 271–279. https://doi.org/10.3115/1218955.1218990

Pang, Bo, & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

https://doi.org/http://dx.doi.org/10.1561/1500000011

Pang, Bo, Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, 79–86. https://doi.org/10.3115/1118693.1118704

Parenteau, J., Sallam, R. L., Howson, C., Tapadinhas, J., Schlegel, K., & Oestreich, T. W. (2016). *Magic Quadrant for Business Intelligence and Analytics Platforms*. Retrieved from https://www.gartner.com/doc/reprints?id=1-2XXKCD7&ct=160204&st=sb

Park, Kang, J., Choi, D., & Han, J. (2018). Understanding customers' hotel revisiting behaviour: a sentiment analysis of online feedback reviews. *Current Issues in Tourism*. https://doi.org/10.1080/13683500.2018.1549025

Park, S., Lee, K., & Song, J. (2011). Contrasting Opposing Views of News Articles on Contentious Issues. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11)*, *1*, 340–349.

Peleja, F., & Lisboa, U. N. De. (2015). *Learning Sentiment Based Ranked-Lexicons*. 435–440.

Pemmaraju, N., Thompson, M. A., & Qazilbash, M. (2017, October 1). Disease-specific hashtags and the creation of Twitter medical communities in hematology and oncology. *Seminars in Hematology*, Vol. 54, pp. 189–192. https://doi.org/10.1053/j.seminhematol.2017.08.004

Pemmaraju, N., Utengen, A., Gupta, V., Kiladjian, J. J., Mesa, R., & Thompson, M. A. (2017, December 6). Rare Cancers and Social Media: Analysis of Twitter Metrics in the First 2 Years of a Rare-Disease Community for Myeloproliferative Neoplasms on Social Media—#MPNSM. *Current Hematologic Malignancy Reports*, Vol. 12, pp. 598–604. https://doi.org/10.1007/s11899-017-0421-y

Peng, H., Ma, Y., Poria, S., Li, Y., & Cambria, E. (2019). *Phonetic-enriched Text Representation for Chinese Sentiment Analysis with Reinforcement Learning*. Retrieved from http://arxiv.org/abs/1901.07880

Perea-Ortega, J. M., & Balahur, A. (2014). Experiments on feature replacements for polarity classification of Spanish tweets. *TASS 2014: Workshop on Sentiment Analysis at SEPLN*, 0–6. Retrieved from http://www.sngularmeaning.team/TASS2014/papers/3.JRC.pdf

Pérez, M., Carreras, D., & Bustamante, M. (2018). Uso e impacto de las redes sociales en las estrategias de marketing de las PyME´s. *Revista de Investigación Académica Sin*

Frontera: División de Ciencias Económicas y Sociales, *19*(19). Retrieved from http://revistainvestigacionacademicasinfrontera.com/sistema/index.php/RDIASF/article/view/47

Petasis, G., Spiliotopoulos, D., Tsirakis, N., & Tsantilas, P. (2014). Large-scale Sentiment Analysis for Reputation Management. *Hellenic Conference on Artificial Intelligence*, (July), 327–340. Retrieved from http://www.researchgate.net/publication/259716667_Large-scale_Sentiment_Analysis_for_Reputation_Management

Peters, A., Crane, D., & Costello, J. (2019). A comparison of students' twitter use in a postsecondary course delivered on campus and online. *Education and Information Technologies*, 1–18. https://doi.org/10.1007/s10639-019-09888-1

Peters, M. D., Wieder, B., Sutton, S. G., & Wakefield, J. (2016). Business intelligence systems use in performance measurement capabilities: Implications for enhanced competitive advantage. *International Journal of Accounting Information Systems*, *21*, 1–17. https://doi.org/10.1016/j.accinf.2016.03.001

Petrini, M., & Pozzebon, M. (2009). Managing sustainability with the support of business intelligence: Integrating socio-environmental indicators and organisational context. *Journal of Strategic Information Systems*, *18*(4), 178–191. https://doi.org/10.1016/j.jsis.2009.06.001

Pilászy, I. (2005). Text categorization and support vector machines. *The Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, *1*, 1–10.

Pino Romero, C. del, & Castello-Martinez, A. (2017). *La estrategia publicitaria basada en influencers. El caso de SmartGirl by Samsung*. Retrieved from http://bit.ly/datos-MWC17.

Pond, P., & Lewis, J. (2019). Riots and Twitter: connective politics, social media and framing discourses in the digital public sphere. *Information Communication and Society*, *22*(2), 213–231. https://doi.org/10.1080/1369118X.2017.1366539

Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1601–1612. Retrieved from https://arxiv.org/pdf/1610.08815.pdf

Porter, R., Nisbet, R., Miner, L. A., & Miner, G. (2018). Using Customer Churn Data to Develop and Select a Best Predictive Model for Client Defection Using STATISTICA Data Miner 13 64-bit for Windows 10. In *Handbook of Statistical Analysis and Data Mining Applications* (Second Edi). https://doi.org/10.1016/b978-0-12-416632-

5.00039-6

Posegga, O., & Jungherr, A. (2019). Characterizing Political Talk on Twitter: A Comparison Between Public Agenda, Media Agendas, and the Twitter Agenda with Regard to Topics and Dynamics. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2019.312

Prieto, R. R. (2016). A case of reflexive and critical interaction in a campaign about Twitter in the course of Legal Theory. *International Journal of Educational Research and Innovation*, *7*, 186–201.

Quijote, T. A., Zamoras, A. D., & Ceniza, A. (2019). Bias detection in Philippine political news articles using SentiWordNet and inverse reinforcement model. *IOP Conference Series: Materials Science and Engineering*, *482*(1), 012036. https://doi.org/10.1088/1757-899X/482/1/012036

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1023/A:1022643204877

Ramaswamy, V., & Ozcan, K. (2018). What is co-creation? An interactional creation framework and its implications for value creation. *Journal of Business Research*, *84*, 196–205. https://doi.org/10.1016/J.JBUSRES.2017.11.027

Ramos-Serrano, M., Gómez, J. D. F., & Pineda, A. (2018). Follow the closing of the campaign on streaming': The use of Twitter by Spanish political parties during the 2014 European elections. *New Media and Society*, *20*(1), 122–140. https://doi.org/10.1177/1461444816660730

Rasmussen, N. H., Goldy, P. S., & Solli, P. O. (2002). *Financial Business Intelligence: Trends, Technology, Software Selection, and Implementation*. Retrieved from https://books.google.es/books?id=CWxl3pHu754C&dq=Rasmussen,+N.H.,+Goldy,+P.S.,+Solli,+P.O.+(2002).+"Financial+BI"&lr=&hl=pt-BR&source=gbs_navlinks_s

Rath, L. T. (2011). The Effects of Twitter in an Online Learning Environment. *Library Publications and Presentations*. Retrieved from https://digitalcommons.brockport.edu/drakepubs/4

Rathore, A. S., Arjaria, S., Khandelwal, S., Thorat, S., & Kulkarni, V. (2019). Movie rating system using sentiment analysis. *Advances in Intelligent Systems and Computing*, *742*, 85–98. https://doi.org/10.1007/978-981-13-0589-4_9

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46. https://doi.org/10.1016/j.knosys.2015.06.015

Recupero, D. R., Dragoni, M., Buscaldi, D., Alam, M., & Cambria, E. (2018). Workshop on Sentic Computing, Sentiment Analysis, Opinion Mining, and Emotion Detection. *Proceedings of EMSASW2018 -4th*. Retrieved from https://gplsi.dlsi.ua.es/sepln15/en/node/36

Reginato, L., & Nascimento, A. M. (2007). Um estudo de caso envolvendo Business Intelligence como instrumento de apoio à controladoria. *Revista Contabilidade & Finanças*, *18*(spe), 69–83. https://doi.org/10.1590/S1519-70772007000300007

Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, *385–386*, 213–224. https://doi.org/10.1016/j.ins.2017.01.015

Rezende, D. A. (2005). *Engenharia de software e sistemas de informação* (3.rd; S. M. de Oliveira, Ed.). Rio de Janeiro: Brasport.

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1), 1–29. https://doi.org/10.1140/epjds/s13688-016-0085-1

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 105–112. https://doi.org/10.3115/1119355.1119369

Rinaldo, S. B., Tapp, S., & Laverie, D. A. (2011). Learning by Tweeting. *Journal of Marketing Education*, *33*(2), 193–203. https://doi.org/10.1177/0273475311410852

Robert, B., & Brown, E. B. (2004). *Agile Project Management with Scrum by Ken Schwaber*. Retrieved from https://books.google.com.br/books?hl=es&lr=&id=6pZCAwAAQBAJ&oi=fnd&pg=PT9&dq=Agile+Project+Management+with+Scrum,+Ken+Schwaber,+Microsoft+Press,+January+2004,+163pp,+ISBN+0-7356-1993-X&ots=kbuWO_9seS&sig=ivnhlb4bS3DG5Iyu_MqjYmiUipE

Rockcontent. (2018). *Social Media Trends 2018*. Retrieved from https://cdn2.hubspot.net/hubfs/355484/Ebooks MKTC/Social Media.pdf?utm_source=hs_automation&utm_medium=email&utm_content=39460531&_hsenc=p2ANqtz--VqnwN2T5oTUszCh5hBoLpdvzJc5qPXxtY5FGOBUYUhyUjLDW2AQsSLvS3oar50hHlJLgXBwTEkXe38ExOgUpvS6nWyg&_hsmi=39460531

Rodríguez A., R., & Santamaría P., C. (2012). Análisis del uso de las redes sociales en Internet: Facebook y Twitter en las Universidades españolas. *Revista*

ICONO14.Revista Científica de Comunicación y Tecnologías Emergentes, *10*(2), 228–246.

Rodríguez Sala, J. J., Santamaría Sala, L., Rabasa Dolado, A., & Martínez Bonastre, O. (2003). *Introducción a la programación. teoría y práctica*.

Romain, V. (2019). Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos. Retrieved July 11, 2019, from https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducción-a-la-clasificación-y-principales-algoritmos-dadee99c9407

Román, J. V., Morera, J. G., Cámara, E. M., & Zafra, S. M. J. (2015). TASS 2014 - The challenge of aspect-based sentiment analysis. *Procesamiento de Lenguaje Natural*, *54*, 61–68.

Roncal, I. S. V., & Urizar, X. S. (2014). Looking for features for supervised tweet polarity classification. *TASS 2014: Workshop on Sentiment Analysis at SEPLN*. Retrieved from http://www.daedalus.es/TASS2014/papers/8.Elhuyar.pdf%5Cnhttps://www.researchgate.net/profile/Inaki_San_Vicente/publication/266385260_Looking_for_Features_for_Supervised_Tweet_Polarity_Classification/links/542ebf980cf29bbc126f53c6.pdf

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., & Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in Twitter. *9th International Workshop on Semantic Evaluation (SemEval´15)*, 451–463. Retrieved from http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval078.pdf

Roventini, A., Alonge, A., Calzolari, N., Magnini, B., & Bertagna, F. (2000). ItalWordNet: a Large Semantic Database for Italian. *LREC*. Retrieved from https://www.semanticscholar.org/paper/ItalWordNet%3A-a-Large-Semantic-Database-for-Italian-Roventini-Alonge/244c9b4a20869d2cce51dc3bcf73f1dc23a78fb6

Rud, O. P. (2009). *BI success factors: Tools for aligning your business in the global economy*. John Wiley & Sons.

Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, *38*(12), 14799–14804. https://doi.org/10.1016/j.eswa.2011.05.070

Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google , GitHub, and More*. " O'Reilly Media, Inc."

Rybalko, S., & Seltzer, T. (2010). Dialogic communication in 140 characters or less: How Fortune 500 companies engage stakeholders using Twitter. *Public Relations Review*, *36*(4), 336–341. https://doi.org/10.1016/j.pubrev.2010.08.004

Sabanovic, A., & Søilen, K. (2012). Customers expectations and needs in the Business Intelligence software market. *Journal of Intelligence Studies in Business*, *2*(1), 5–20. Retrieved from https://ojs.hh.se/index.php/JISIB/article/view/27

Sabou, M., Aroyo, L., Bontcheva, K., Bozzon, A., & Qarout, R. K. (2018). Semantic web and human computation: The status of an emerging field. *Semantic Web*, *9*(3), 1–12. https://doi.org/10.3233/sw-180292

Safko, L., & Brake, D. K. (2010). *A Biblia da Midia Social: táticas, ferramentes e estratégias para construir e transformar negocios*. EDGARD BLU.

Sagot, Benoît, & Fišer, D. (2008). Building a free French wordnet from multilingual resources. *Proceedings of OntoLex*, 14–19. Retrieved from https://hal.inria.fr/inria-00614708/

Sagot, Benoit, Fišer, D., & others. (2012). Automatic Extension of WOLF. *GWC2012 - International Global Wordnet Conference*. Retrieved from https://hal.inria.fr/hal-00655774/

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM13)*, 9–21. Retrieved from http://www.di.unito.it/patti/essem13/index.html%5Cnhttp://ceur-ws.org/Vol-1096/paper1.pdf

Saif, H., He, Y., & Alani, H. (2012). Semantic Sentiment Analysis of Twitter. *The 11th International Semantic Web Conference (ISWC´12)*, 508–524. https://doi.org/10.1007/978-3-642-35176-1_32

Saif, H., He, Y., Fernandez, M., & Alani, H. (2014a). Adapting Sentiment Lexicons Using Contextual Semantics for Sentiment Analysis of Twitter. *European Semantic Web Conference (ESWC´14)*, *8798*, 54–63. https://doi.org/10.1007/978-3-319-11955-7

Saif, H., He, Y., Fernandez, M., & Alani, H. (2014b). Semantic Patterns for Sentiment Analysis of Twitter. *Proceedings of the 13th International Semantic Web Conference - Part II (ISWC´14)*, *8797*, 324–340. https://doi.org/10.1007/978-3-319-11915-1_21

Salton, G. (1975). *A theory of indexing*. Retrieved from https://books.google.com.br/books?hl=es&lr=&id=FsmZHrywfB0C&oi=fnd&pg=PP2&dq=%22A+theory+of+indexing%22&ots=CtTukMiGyd&sig=t86-yU12YNfNr0R2Ks6k7Fq5mT4

Salton, G., & Yang, C. S. (1973, April 1). On the specification of term values in automatic indexing. *Journal of Documentation*, Vol. 29, pp. 351–372. https://doi.org/10.1108/eb026562

Salton, G., Yang, C. S., & Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, *26*(1), 33–44. https://doi.org/10.1002/asi.4630260106

Salvatore, M. F., Terrebonne, J., Fields, V., Nodurft, D., Runfalo, C., Latimer, B., & Ingram, D. K. (2015). Initiation of calorie restriction in middle-aged male rats attenuates aging-related motoric decline and bradykinesia without increased striatal dopamine. *Neurobiology of Aging*, *37*, 192–207. https://doi.org/10.1016/j.neurobiolaging.2015.10.006

Sanchez-Perez, M. A., Markov, I., Gómez-Adorno, H., & Sidorov, G. (2017). Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same Spanish news corpus. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10456 LNCS*, 145–151. https://doi.org/10.1007/978-3-319-65813-1_15

Sarkar, K., & Bhowmick, M. (2018). Sentiment Polarity Detection in Bengali Tweets Using Multinomial Naïve Bayes and Support Vector Machines. *2017 IEEE Calcutta Conference, CALCON 2017 - Proceedings*, *2018-Janua*, 31–35. https://doi.org/10.1109/CALCON.2017.8280690

Sarvabhotla, K., Pingali, P., & Varma, V. (2011a). Sentiment classification: A lexical similarity based approach for extracting subjectivity in documents. *Information Retrieval*, *14*(3), 337–353. https://doi.org/10.1007/s10791-010-9161-5

Sarvabhotla, K., Pingali, P., & Varma, V. (2011b). Sentiment classification a lexical similarity based approach for extracting subjectivity in documents. *Information Retrieval*, *14*(3), 337–353.

Savoy, J. (2012). Authorship attribution based on specific vocabulary. *ACM Transactions on Information Systems*, *30*(2), 1–30. https://doi.org/10.1145/2180868.2180874

Schildt, H. (1997). *C Completo e Total* (3rd ed.). https://doi.org/10.2307/506804

Schölkopf, B, & Smola, A. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Retrieved from https://books.google.es/books?hl=es&lr=&id=y8ORL3DWt4sC&oi=fnd&pg=PR13&dq=A.+J.+Smola+and+B.+Schölkopf.+Learning+with+Kernels.+The+MIT+Press,+Cambridge,+MA,+2002&ots=bKxV4vR3FA&sig=UuuLjyQ9Uru6mS-5lzrIT1V6Po0

Schölkopf, Bernhard, & Burges, C. J. C. (1999). *Advances in kernel methods: support vector learning*. Retrieved from https://books.google.es/books?hl=es&lr=&id=_NYamXKkNM8C&oi=fnd&pg=PR7&dq=Sch%25C2%25A8olkopf+B.,+Burges+C.J.C.,+and+Smola+A.J.+1999a.+(Eds.)+Advances+in+Kernel+Methods%25E2%2580%2594Support+Vector+Learning.+MIT+Press,+Cambridge,+MA.&ots=RyeQ7y3Lv5&sig=Hf-

Scholz, T., Conrad, S., & Hillekamps, L. (2012). Opinion mining on a german corpus of a media response analysis. *International Conference on Text, Speech and Dialogue*, (1), 39–46. https://doi.org/10.1007/978-3-642-32790-2_4

Segal, M. R. (2003). *UC San Francisco Recent Work Title Machine Learning Benchmarks and Random Forest Regression Publication Date Machine Learning Benchmarks and Random Forest Regression*. Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt35x3v9t4/qt35x3v9t4.pdf?t=kro5qw

Seki, Y., Kando, N., & Aono, M. (2009). Multilingual opinion holder identification using author and authority viewpoints. *Information Processing and Management*, *45*(2), 189–199. https://doi.org/10.1016/j.ipm.2008.11.004

Seller, M. L., & Laurindo, F. J. B. (2018). Comunidade de marca ou boca a boca eletrônico: qual o objetivo da presença de empresas em mídias sociais? *Gestão & Produção*, *25*(1), 191–203. https://doi.org/10.1590/0104-530x2244-16

Serrano-Cobos, J. (2014). *Big data* y analítica web. Estudiar las corrientes y pescar en un océano de datos. *El Profesional de la Informacion*, *23*(6), 561–566. https://doi.org/10.3145/epi.2014.nov.01

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, *311*(August), 18–38. https://doi.org/10.1016/j.ins.2015.03.040

Shalunts, G., Backfried, G., & Prinz, K. (2014). Sentiment Analysis of German Social Media Data for Natural Disasters. *11th International Conference on Information Systems for Crisis Response and Management (ISCRAM´14)*, (May), 752–756. Retrieved from http://saifmohammad.com/WebDocs/SA-prop.pdf%5Cnhttp://www.jhuapl.edu/techdigest/TD/td3001/Fink.pdf

Shamma, D. A., Kennedy, L., & Churchill, E. F. (2009). Tweet the debates: understanding community annotation of uncollected sources. *The First SIGMM Workshop on Social Media (WSM '09)*, 1–8. https://doi.org/10.1145/1631144.1631148

Sharma, A., Sharma, R., Sharma, V. K., & Shrivatava, V. (2014). Application of Data Mining – A Survey Paper. *(IJCSIT) International Journal of Computer Science and*

Information Technologies, *5*(2), 2023–2025.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2011). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. Retrieved from https://books.google.be/books?id=zjXlIAETPKwC

Sidarenka, U., & Stede, M. (2016). *Generating Sentiment Lexicons for German Twitter*. 80–90. Retrieved from https://github.com/WladimirSidorenko/SentiLex.

Silva, Carvalho, P., & Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. In Springer Berlin Heidelber (Ed.), *International Conference on Computational Processing of the Portuguese Language* (Vol. 7243, pp. 218–228). https://doi.org/10.1007/978-3-642-28885-2_25

Smith, Kendall, M., Knighton, D., & Wright, T. (2018). Rise of the Brand Ambassador: Social Stake, Corporate Social Responsibility and Influence among the Social Media Influencers. *Communication Management Review*, *3*(01), PRELIMINARY. https://doi.org/10.22522/cmr20180127

Smith, L. I. (2002). A Tutorial on Principal Components Analysis Introduction. *Statistics*, *51*, 52. https://doi.org/10.1080/03610928808829796

Soler, F. (2017). Preliminares para la comprensión del concepto Logos en el Comentario a Juan de Orígenes. *Cuadernos de Teología*. https://doi.org/10.22199/S07198175.2017.0001.00002

Souto, M. C. P. de, Lorena, A. C., Delbem, A. C. B., & Carvalho, A. C. P. L. F. de. (2003). Técnicas de aprendizado de máquina para problemas de biologia molecular. *Sociedade Brasileira de Computação*, 1–45. Retrieved from http://www.cin.ufpe.br/~mcps/ENIA2003/jaia2003-14-08.pdf

Souza, M., Vieira, R., Chishman, R., & Alves, I. M. (2011). Construction of a Portuguese opinion lexicon from multiple resources. *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, 59–66.

Spencer, J., & Uchyigit, G. (2012). Sentimentor: Sentiment Analysis of Twitter Data. *The 1st International Workshop on Sentiment Discovery from Affective Data (SDAD´12)*, 56–66. Retrieved from http://ceur-ws.org/Vol-917/SDAD2012.pdf

Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011). Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. *Conference on Empirical Methods in Natural Language Processing (EMNLP´11)*, 53–63.

Steiner-Correa, F. (2017). Dando Alas a Red Bull: Una aplicación de la minería de

opiniones, para conocer qué piensan y de qué hablan los seguidores de la marca en Twitter. In *Libro Conmemorativo del X Aniversario del Máster en Marketing y Comportamiento del Consumidor* (1st ed., pp. 336–348). Retrieved from https://drive.google.com/…/0BwjEt53pHUCudTl1UWstMEs4WmM/view

Steiner-Correa, F., Viedma-del-Jesus, M. I., & Lopez-Herrera, A. G. (2018). A survey of multilingual human-tagged short message datasets for sentiment analysis tasks. *Soft Computing*, *22*(24), 8227–8242. https://doi.org/10.1007/s00500-017-2766-5

Suliman, H. (2010). Finding a Place for Twitter in Higher Education. *ELearn Magazine*. Retrieved from http://elearnmag.acm.org/archive.cfm?aid=1821980

Sung, Y., Kim, Y., Kwon, O., & Moon, J. (2010). An explorative study of Korean consumer participation in virtual brand communities in social network sites. *Journal of Global Marketing*, *23*(5), 430–445. https://doi.org/10.1080/08911762.2010.521115

Swani, K., Brown, B. P., & Milne, G. R. (2014). Should tweets differ for B2B and B2C? An analysis of Fortune 500 companies' Twitter communications. *Industrial Marketing Management*, *43*(5), 873–881. https://doi.org/10.1016/j.indmarman.2014.04.012

Teetor, P. (2011). *R Cookbook* (First; Adam Zaremba, Ed.). O'Reilly Media, Inc.

Telles, A. (2010). *A REVOLUÇÃO DAS MÍDIAS SOCIAIS: estratégias de marketing digital para você e sua empresa terem sucesso nas mídias sociais* (M. Books, Ed.). São Paulo.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173. https://doi.org/10.1002/asi.21662

Toldos, M., & Castro, M. (2013). El efecto de las dimensiones de personalidad de marca en la intención de compra de marcas de lujo en México y Brasil. *Global Conference on Business and Finance Proceedings. Vol 8*, 837–842.

Tong, R. (2001). An operational system for detecting and tracking opinions in on-line discussion. *Working Notes of the ACM SIGIR 2001*.

Toprak, C., Jakob, N., & Gurevych, I. (2010). Sentence and Expression Level Annotation of Opinions in User-Generated Discourse. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, *1*(July), 575–584. Retrieved from http://www.aclweb.org/anthology/P10-1059

Torres, C. (2009). A Bíblia do marketing digital: tudo o que você queria saber sobre marketing e publicidade na internet e não tinha a quem perguntar. *São Paulo: Novatec*, 15–83.

Tsakalidis, A., Papadopoulos, S., & Kompatsiaris, I. (2014). An Ensemble Model for Cross-Domain Polarity Classification on Twitter. *Conference on Web Information Systems Engineering - Part II (WISE´14)*, 168–177. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-11746-1_12

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, *24*, 478–514. https://doi.org/http://doi.org/10.1007/s10618-011-0238-6

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, (July), 417–424. https://doi.org/10.3115/1073083.1073153

Twitter, I. (2019). The Story of a Tweet.

Ujjwal, K. (2016). An intuitive explanation of convolutional neural networks. Retrieved from https://ujjwalkarn.me/2016/08/11/ intuitive-explanation-convnets/

Valdivia, A., Luzón, M. V., & Herrera, F. (2017). Sentiment Analysis in TripAdvisor. *IEEE Intelligent Systems*, *32*(4), 72–77. https://doi.org/10.1109/MIS.2017.3121555

van Zoonen, W., & van der Meer, T. G. L. A. (2016). Social media research: The application of supervised machine learning in organizational communication research. *Computers in Human Behavior*, *63*, 132–141. https://doi.org/10.1016/j.chb.2016.05.028

Vapnik, V., Golowich, S. E., & Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, 281–287. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.3139

Vapnik, V. N., & Chervonenkis, A. Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of Complexity: Festschrift for Alexey Chervonenkis* (pp. 11–30). https://doi.org/10.1007/978-3-319-21852-6_3

Vela, M., & Riera, V. (2013). La comunicación" on-line": nuevos consumidores, fuentes de información, medios, canales, estrategias y tácticas. *Dialnet.Unirioja.Es*. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=4495848

Viedma-del-Jesus, M. I. (2016). *Proyecto docente de investigación*. Universidad de Granada.

Vijaya, J., & Sivasankar, E. (2017). An efficient system for customer churn prediction

through particle swarm optimization based feature selection model with simulated annealing. *Cluster Computing*, 1–12. https://doi.org/10.1007/s10586-017-1172-1

Vilares, D., Doval, Y., Alonso, M. A., & Gómez-Rodríguez, C. (2014). LyS at TASS 2014: A Prototype for Extracting and Analysing Aspects from Spanish tweets. *TASS 2014: Workshop on Sentiment Analysis at SEPLN*.

Vilares, D., Doval, Y., Alonso, M. A., & Gómez-Rodríguez, C. (2015). LyS at TASS 2015: Deep learning experiments for sentiment analysis on Spanish tweets. *TASS 2015: Workshop on Sentiment Analysis at SEPLN*, *1397*, 47–52.

Vomfell, L., Härdle, W. K., & Lessmann, S. (2018). Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems*, *113*, 73–85. https://doi.org/10.1016/j.dss.2018.07.003

Wamba, S. F., & Carter, L. (2016). Social Media Tools Adoption and Use by SMEs: An Empirical Study. *Igi-Global*. Retrieved from http://fossowambasamuel.com/wp-content/uploads/2016/10/Fosso-Wamba-and-Lemuria-Social-MediaUseSMEsEmpirical-StudyAdoption.pdf

Wang, D., Zhu, S., & Li, T. (2013). SumView: A Web-based engine for summarizing product reviews and customer opinions. *Expert Systems with Applications*, *40*(1), 27–33. https://doi.org/10.1016/j.eswa.2012.05.070

Wang, J. (2009). *Encyclopedia of Data Warehousing and Mining, Second Edition*. https://doi.org/10.4018/978-1-60566-010-3

Wang, L., Niu, J., & Yu, S. (2019). SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/tkde.2019.2913641

Watson, H. J. (2010). BI Training Solutions : As Close as Your Conference Room. *Business Intelligence Journal*, *15*(2), 1–57. https://doi.org/1547-2825

Watson, H. J., & Wixom, B. H. (2007). The Current State of Business Intelligence. *Computer*, (October 2007), 96–99. https://doi.org/10.1109/MC.2007.331

Wei Di, Neel Sundaresan, Robinson Piramuthu, A. B. (2014). Is a picture really worth a thousand words?:-on the role of images in e-commerce. *Proceedings of the 7th ACM International Conference on Web Search and Data Mining - WSDM '14*, 633–641. https://doi.org/10.1145/2556195.2556226

Weiler, S., Matt, C., & Hess, T. (2019). *Understanding User Uncertainty during the Implementation of Self-Service Business Intelligence: A Thematic Analysis*.

Retrieved from https://scholarspace.manoa.hawaii.edu/handle/10125/60023

Weiss, S., & Kulikowski, C. (1991). *Computer systems that learn*. Retrieved from http://www.citeulike.org/group/1778/article/932135

Wieder, B., & Ossimitz, M. L. (2015). The Impact of Business Intelligence on the Quality of Decision Making - A Mediation Model. *Procedia Computer Science*, *64*, 1163–1171. https://doi.org/10.1016/j.procs.2015.08.599

Wiegand, M., & Klakow, D. (2012). Generalization Methods for In-Domain and Cross-Domain Opinion Holder Extraction. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (Eacl)*, 325–335.

Wiertz, C., & De Ruyter, K. (2007). Beyond the call of duty: Why customers contribute to firm-hosted commercial online communities. *Organization Studies*, *28*(3), 347–376. https://doi.org/10.1177/0170840607076003

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, *35*(3), 399–433. https://doi.org/10.1162/coli.08-012-R1-06-90

Witten, I., Frank, E., Hall, M., & Pal, C. (2016). *Data Mining: Practical machine learning tools and techniques*. Retrieved from https://books.google.de/books?hl=de&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=data+mining+practile&ots=8ICMpkmwvf&sig=luXgfdBm_spquY6JDVeBViMCCjU

Xavier, G. C., & Nunes, M. L. E. (2014). *A Rede social e as organizações empresariais-vantagens e riscos do uso das redes sociais pelas empresas*. Retrieved from http://www.ambito-juridico.com.br/site/?n_link=revista_artigos_leitura&artigo_id=14127

Xie, S., Wang, G., Lin, S., & Yu, P. S. (2012). Review spam detection via temporal pattern discovery. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 823–831. https://doi.org/10.1145/2339530.2339662

Xiong, F., Nelson, J., & Bodle, K. (2017). The adoption of new technology by listed companies: the case of Twitter. *Technology Analysis & Strategic Management*, 1–14. https://doi.org/10.1080/09537325.2017.1385759

Yang, M., Liang, Y., Zhao, W., Xu, W., Zhu, J., & Qu, Q. (2018). Task-oriented keyphrase extraction from social media. *Multimedia Tools and Applications*, *77*(3), 3171–3187. https://doi.org/10.1007/s11042-017-5041-y

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). *Sentiment Analyzer: Extracting*

Sentiments about a Given Topic using Natural Language Processing Techniques. Retrieved from http://suraj.lums.edu.pk/~cs631s05/Papers/sentiment_analysis.pdf

Young, J. A. (2017). Facebook, Twitter, and Blogs: The Adoption and Utilization of Social Media in Nonprofit Human Service Organizations. *Human Service Organizations Management, Leadership and Governance*, *41*(1), 44–57. https://doi.org/10.1080/23303131.2016.1192574

Yu, L. C., Wu, J. L., Chang, P. C., & Chu, H. S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, *41*(April), 89–97. https://doi.org/10.1016/j.knosys.2013.01.001

Yu, Y., & Wang, X. (2015). World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets. *Computers in Human Behavior*, *48*, 392–400. https://doi.org/10.1016/j.chb.2015.01.075

Zabin, J., & Jefferies, A. (2008). *Social media monitoring and analysis: Generating consumer insights from online conversation*.

Zaglia, M. E. (2013). Brand communities embedded in social networks. *Journal of Business Research*, *66*(2), 216–223. https://doi.org/10.1016/j.jbusres.2012.07.015

Zeng, L., Li, L., & Duan, L. (2012). Business intelligence in enterprise computing environment. *Information Technology and Management*, *13*(4), 297–310. https://doi.org/10.1007/s10799-012-0123-z

Zhan, M., Tu, R., & Yu, Q. (2018). Understanding readers: Conducting sentiment analysis of instagram captions. *ACM International Conference Proceeding Series*, 33–40. https://doi.org/10.1145/3297156.3297270

Zhang, L., & Liu, B. (2017). Sentiment Analysis in Social Media, Aspect Extraction for. In *Encyclopedia of Social Network Analysis and Mining* (pp. 1–11). https://doi.org/10.1007/978-1-4614-7163-9_110207-1

Zhang, L., Wang, S., & Liu, B. (2018). Deep Learning for Sentiment Analysis : A Survey. *Iley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4). Retrieved from http://arxiv.org/abs/1801.07883

Zhang, S., Zhang, X., Chan, J., & Rosso, P. (2019). Irony detection via sentiment-based transfer learning. *Information Processing and Management*, *56*(5), 1633–1644. https://doi.org/10.1016/j.ipm.2019.04.006

Zhang, Y., Song, D., Zhang, P., Li, X., & Wang, P. (2019). A quantum-inspired sentiment

representation model for twitter sentiment analysis. *Applied Intelligence*, 1–16. https://doi.org/10.1007/s10489-019-01441-4

Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research*, *41*(12), 1955–1959. https://doi.org/10.1016/j.asr.2007.07.020

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The State-of-the-Art in Twitter Sentiment Analysis. *ACM Transactions on Management Information Systems*, *9*(2), 1–29. https://doi.org/10.1145/3185045