UNIVERSIDADE DA CORUÑA

Facultad de Economía y Empresa

Trabajo de Fin de Grado

Nuevas técnicas estadísticas: Text Mining en Web.

Un análisis de marcas de móviles.

Nicole Villaverde Medina

Tutor: Prof. Dr. Xosé Manuel Martínez Filgueira

Grado en Economía

Año 2017

Trabajo de Fin de Grado presentado en la Facultad de Economía y Empresa de la Universidade da Coruña para la obtención del Grado en Economía

Resumen

La aparición y desarrollo de las nuevas Tecnologías de la Información y la Comunicación (TIC) ha incrementado exponencialmente la cantidad de datos disponibles para la investigación, y la capacidad de analizarlos. En particular, la constante evolución de Internet y su papel como almacén de datos de diverso tipo, ha provocado un enorme interés en técnicas que saquen partido de esta abundancia de datos, concretamente en lo referido al análisis automático de texto. Con este tipo de análisis, y gracias a técnicas que se agrupan en la categoría genérica de *Text Mining* y que combinan organización de información con técnicas estadísticas, se hace posible la extracción de información, la cual puede ser muy relevante para distintas entidades como empresas o gobiernos.

El objetivo de este documento es presentar y describir el flujo de trabajo en el *Text Mining*, su relación con la Estadística y la posibilidad de utilizarlo partiendo de los conocimientos adquiridos en mi titulación y con una breve preparación. Esta posibilidad de utilizarlo demostraría su utilidad, ya que con mayor tiempo y preparación podría ser utilizado para abordar cuestiones más complejas.

Como parte de la presentación del trabajo con *Text Mining,* se ha realizado una aplicación práctica, relacionada con la Investigación de Mercados, en la cual, siguiendo la evolución de noticias aparecidas en un blog de tecnología entre los años 2007 y 2016, se ha visto la evolución de la imagen de Apple en sus productos de telefonía móvil, y se ha comparado con su competencia: Samsung y un conjunto amplio de otras marcas. Además, se ha visto con este análisis la importancia que, en el marketing de Apple, tiene su propia imagen de marca, mientras que el marketing de su competencia parece estar más enfocado hacia las características técnicas que pueden aportar.

Palabras clave: Estadística, Minería de Datos, Minería de Textos, Minería Web, Blog.

Número de palabras: 14.765

Resumo

A aparición e o desenvolvemento das novas Tecnoloxías da Información e a Comunicación (TIC) aumentaron exponencialmente a cantidade de datos dispoñibles para a investigación, e a capacidade de analizalos. En particular, a constante evolución da Internet e o seu papel como almacén de datos de diverso tipo, provocou unha enorme interese en técnicas que lle quiten partido a esta abundancia de datos, concretamente no referido á análise automático de texto. Con este tipo de análise, e gracias a técnicas que se agrupan na categoría xenérica de *Text Mining* e que combinan organización de información con técnicas estatísticas, faise posible a extración de información, a cal pode ser moi relevante para as distintas entidades como empresas ou gobernos.

O obxetivo deste documento é presentar e describir o fluxo de traballo no *Text Mining*, a súa relación coa Estatística e a posibilidade de utilizalo partindo dos coñecementos adquiridos na miña titulación e cunha breve preparación. Esta posibilidade de utilizalo demostraría a súa utilidade, xa que con máis tempo e preparación podería ser utilizado para abordar cuestións máis complexas.

Como parte da presentación do traballo con *Text Mining,* realizouse unha aplicación práctica, relacionada coa Investigación de Mercados, na que, seguindo a evolución de noticias aparecidas nun blog de tecnoloxía entre os anos 2007 e 2016, observouse a evolución da imaxe de Apple nos seus produtos de telefonía móbil, e comparouse coa súa competencia: Samsung e un conxunto amplo doutras marcas. Ademais, se viu con esta análise a importancia que, no marketing de Apple, ten a súa propia imaxe de marca, mentres que o marketing da súa competencia parece estar mais enfocado de cara ás características técnicas que poden aportar.

Palabras chave: Estatística, Minería de Datos, Minería de Textos, Minería Web, Blog.

Número de palabras: 14.765

Abstract

The emergence and development of new Information and Communication Technologies (TIC) has exponentially increased the amount of data available for research, and ability to analyze them. In particular, the constant evolution of the Internet and its role as data warehouse of various kinds, has caused a great interest in techniques that take advantage of this wealth of data, particularly with regard to automatic analysis of text. With this tpe of analysis, and thanks to techniques that are grouped under the generic category of *Text Mining*, and combine information organization with statistical techniques, it becomes possible to extract information, which can be very relevant to different entities as companies or govemments.

The purpose of this paper is to present and describe the workflow in the *Text Mining*, his relationship with statistics and the possibility to use it and building on the knowledge acquired in my degree and with short preparation. This ability to use demonstrate its utility, so with most time and preparation, it could be used for address more complex issues.

As part of the presentation of the work with Text Mining, it has made a practice application related to markets research, in which, following the evolution of appeared news on a blog of technology between 2007 and 2016, the evolution of the image of Apple has been seen in its mobile phone products, and has been compared to its competence: Samsung and a wide range of other brands. In addition, it has been seen with this analyzes the importance that, in the marketing of Apple, has its own brand image, while marketing your competition seems to be more focused on the technical characteristics that can give.

Keywords: Statistics, Data Mining, Text Mining, Web Mining, Blog.

Number of words: 14.765

Índice

Índice de figuras

Introducción

El tema escogido de este trabajo está motivado por el ascenso de las tecnologías de la información y la "avalancha de información" (Hair et al., 2010) que ha producido. Un fenómeno como Internet forma parte de este ascenso y de esta acumulación de información, por lo que la necesidad de aprovechar esta inmensa fuente de conocimientos se ha convertido en un importante objetivo de investigación.

La Estadística, como herramienta, permite realizar análisis de información y datos, de distintas formas, y su utilización se extiende en diferentes campos. Por ello, sus herramientas forman parte de este nuevo proceso de análisis, de manejo de grandes cantidades de información; y en este trabajo se pondrán de manifiesto las técnicas utilizadas en Estadística, qué son y cómo se integran en los nuevos procesos, completando su última parte con un caso práctico en el que se analizará la información procedente de usuarios y consumidores ,que publican sus opiniones y comentarios en un blog de tecnologías y su impacto en la evolución de diferentes empresas.

Después de una breve introducción a lo que es la Estadística en el capítulo 1, se parte del *Data Mining*, al que se le dedica el capítulo 2. Se trata de una etapa que forma parte de lo que se conoce como proceso de extracción de conocimiento a partir de datos. Dentro de ella se pueden encontrar diversos tipos de minería, entre ellos la minería de web, también conocida como *Web Mining*, la cual, como su propio nombre indica, trata de analizar información proveniente de la red.

Una de las técnicas englobada en la minería de web y explicada en el capítulo 3 es el *Text Mining*, siendo, además, la utilizada en el caso práctico. Dicha minería de textos está adquiriendo mucha importancia y llegando a muchas áreas de aplicación. En Economía, por ejemplo, se puede ver en aspectos como en la creación de indicadores económicos (Hale, A. et al., 2017), en la predicción de mercados financieros (Bollen, J., Mao, H. & Zeng, X., 2011) o en la historia del pensamiento económico (Matsuyama, N., 2016).

Posteriormente, esta técnica será utilizada en el capítulo 4 para el caso práctico, en el cual las opiniones y comentarios de los usuarios y consumidores,

Índice de tablas

Introdución

El tema escogido de este trabajo está motivado por el ascenso de las tecnologías de la información y la "avalancha de información" (Hair et al., 2010) que ha producido. Un fenómeno como Internet forma parte de este ascenso y de esta acumulación de información, por lo que la necesidad de aprovechar esta inmensa fuente de conocimientos se ha convertido en un importante objetivo de investigación.

La Estadística, como herramienta, permite realizar análisis de información y datos, de distintas formas, y su utilización se extiende en diferentes campos. Por ello, sus herramientas forman parte de este nuevo proceso de análisis, de manejo de grandes cantidades de información; y en este trabajo se pondrán de manifiesto las técnicas utilizadas en Estadística, qué son y cómo se integran en los nuevos procesos, completando su última parte con un caso práctico en el que se analizará la información procedente de usuarios y consumidores ,que publican sus opiniones y comentarios en un blog de tecnologías y su impacto en la evolución de diferentes empresas.

Después de una breve introducción a lo que es la Estadística en el capítulo 1, se parte del *Data Mining*, al que se le dedica el capítulo 2. Se trata de una etapa que forma parte de lo que se conoce como proceso de extracción de conocimiento a partir de datos. Dentro de ella se pueden encontrar diversos tipos de minería, entre ellos la minería de web, también conocida como *Web Mining*, la cual, como su propio nombre indica, trata de analizar información proveniente de la red.

Una de las técnicas englobada en la minería de web y explicada en el capítulo 3 es el *Text Mining*, siendo, además, la utilizada en el caso práctico. Dicha minería de textos está adquiriendo mucha importancia y llegando a muchas áreas de aplicación. En Economía, por ejemplo, se puede ver en aspectos como en la creación de indicadores económicos (Hale, A. et al., 2017), en la predicción de mercados financieros (Bollen, J., Mao, H. & Zeng, X., 2011) o en la historia del pensamiento económico (Matsuyama, N., 2016).

Posteriormente, esta técnica será utilizada en el capítulo 4 para el caso práctico, en el cual las opiniones y comentarios de los usuarios y consumidores,

publicadas en un blog de tecnologías, se utilizan para explicar la evolución de Apple y de sus competidores: Samsung y un conjunto de otras marcas, observando en qué medida los datos obtenidos reflejan una posible relación entre el paso de los años y las palabras de los artículos y comentarios publicados en el blog; además de estudiar si esta diferencia también depende de la marca. Seguidamente se obtienen unas conclusiones que engloban tanto dicho caso práctico como el uso que se hace del *text mining*.

De esta manera, se podrá ver cómo funciona dicha técnica, aunque en caso de disponer de más tiempo y espacio, se podría realizar un análisis mucho más profundo y complejo del tema en cuestión.

1. La Estadística

Es evidente la importancia que tiene y ha tenido la Estadística en el desarrollo de muchas áreas del saber científico. Esta circunstancia supone una dificultad seria a la hora de acotar sus características como ciencia independiente. Se pueden encontrar aplicaciones de esta disciplina en los más variados campos de la actividad humana (Economía, Medicina, Ingeniería, Física, Psicología, etc.). De esta manera, la Estadística ha llegado a ser lo que es gracias al uso que de ella han ido haciendo las diferentes ramas del conocimiento y del saber, convirtiéndose en un instrumento metodológico de las demás ciencias. (Carrasco, 2005).

Para Martin-Pliego (2007), la Estadística *"se configura como la tecnología del método científico que proporciona instrumentos para la toma de decisiones cuando éstas se adoptan en ambiente de incertidumbre, siempre que esta incertidumbre pueda ser medida en términos de probabilidad."* Por ello, la Estadística se preocupa de los métodos de recogida y descripción de datos, así como de generar técnicas para el análisis de esta información.

Etimológicamente, esta palabra procede del latín e italiano, de los términos *"statista"* (político, hombre de estado) y *"statisticum collegium"* (consejo de estado). (Martín Pliego, F.J., 1994). Esta procedencia está relacionada con el concepto de Estadística como la ciencia del Estado[1], porque en sus orígenes se utilizaba exclusivamente con fines estatales, en el sentido de que los gobiernos de las distintas naciones tenían (y tienen) la necesidad, por razones de organización, de conocer las características de su población para gestionar el pago de impuestos, el reclutamiento de soldados, el reparto de tierras o bienes, la prestación de servicios públicos, etc. Esta necesidad llevó a los gobernantes a establecer sistemas para recoger y procesar de alguna manera la información obtenida, es decir, de hacer estadísticas sobre la población.

Normalmente, los primeros estudios estadísticos que se hacían eran los censos, que son estudios descriptivos sobre todos los integrantes de una población. Además, una forma de agilizar la recogida y tratamiento de la información sobre algunas

[1] Véase "Libro de los Números Cap. 1" extraído de INE http://www.ine.es/explica/docs/historia_estadistica.pdf

características de la población era (y sigue siendo) a través de los registros, que son listados en los que los ciudadanos deben inscribirse cuando, por ejemplo, nace un hijo o hay una defunción.

Con el tiempo y el desarrollo científico surgieron alternativas a los censos: las encuestas. Se estudiana así sólo una parte de la población y se extendían sus resultados a toda la población. Para que este cambio se pudisese producir, fue necesario el desarrollo de la Teoría de la Probabilidad, de la Inferencia Estadística y del Muestreo que dió comienzo en la Edad Moderna pero que aceleró su desarrollo al entrar en el siglo XX. El desarrollo científico y filosófico ocurrido durante ese periodo también propició la aplicación de la Estadística a las ciencias sociales con fines no políticos, y además el surgimiento de nuevas técnicas y herramientas amplió las posibilidades de su uso.

1.1 Relación con la Economía

La ciencia económica, como el resto de las ciencias sociales, tiene por objeto el estudio del comportamiento, del ser humano y de la sociedad en general, en este caso desde la óptica económica, intentando descubrir las interrelaciones y diferentes actitudes de los individuos o grupos sociales ante los estímulos de carácter económico.

El científico económico se encuentra con una sociedad en continuo cambio, donde ciertas respuestas pertenecientes al pasado se entremezclan con posturas ya impregnadas de hábitos del futuro siendo, en definitiva, imposible establecer normas fijas de comportamiento y leyes inmutables que regulen las relaciones económicas (Martín-Pliego, 2007).

Además, el científico económico encuentra que sus evidencias empíricas no pueden ser obtenidas y repetidas en ningún laboratorio económico controlando las condiciones de partida.

Ante esta situación, la utilización del método estadístico como método de investigación en las ciencias sociales, no sólo es aconsejable sino que se hace imprescindible. La posibilidad de disponer de instrumentos objetivables para la

verificación estadística de las hipótesis que sobre un determinado comportamiento económico se establezcan constituye la única salida racional al proceso de investigación económica.

Por lo tanto, la Economía mantiene un vínculo de dependencia muy fuerte con la estadística. Para poder analizar cualquier realidad social o económica de interés es imprescindible el empleo de diferentes métodos estadísticos que permitan la observación del fenómeno y la recolección de datos. Con esto se pretende lograr una mayor compresión de la realidad estudiada que facilite la toma de decisiones. (Carrasco, 2005).

Esta relación entre la Economía y las técnicas estadísticas se ha mantenido estable durante la mayor parte del siglo XX, ya que, aunque se ha mejorado la capacidad de cálculo y las técnicas utilizadas, se trabajaba con una cantidad de datos limitada, lo cual condicionaba las posibilidades de análisis.

Sin embargo, a finales del siglo, la expansión de la informática y las tecnologías de la información han dado lugar a una "avalancha de información" (Hair et al., 2010), que da lugar a la aparición de conceptos con data mining, relacionados con una nueva visión en el análisis de datos, y, por lo tanto, en la Estadística.

2. Data Mining

La minería de datos o *"Data Mining"* se crea por la aparición de nuevas necesidades y, especialmente, por el reconocimiento de un nuevo potencial: el valor, generalmente infrautilizado hasta ahora, de la gran cantidad de datos almacenados informáticamente en los sistemas de información de instituciones, empresas, particulares y gobiernos. Los datos pasan de ser un *"producto"* a ser una *"materia prima",* la cual hay que explotar para obtener el verdadero *"producto elaborado",* el conocimiento (Hernández et al., 2004). Un conocimiento que ha de ser especialmente valioso para la ayuda en la toma de decisiones sobre el ámbito en el que se han extraído o recopilado los datos.

Es bien cierto que la Estadística es la primera ciencia que considera los datos como su materia prima, pero las nuevas necesidades y, en particular, las nuevas características de los datos (en volumen y tipología) hacen que las disciplinas que integran lo que se conoce como *"minería de datos"* sean numerosas y heterogéneas.

2.1 Concepto de Minería de Datos

El término **minería de datos** (*data mining*) habitualmente se emplea para referirse a dos conceptos:

- Al área de la informática que estudia el análisis de datos con el fin de extraer información de interés, especialmente, en forma de conocimiento.
- A una etapa concreta de un proceso más amplio, denominado proceso de descubrimiento de conocimiento en bases de datos (KDD).

Técnicamente, el *data mining* es una fase de dicho proceso. Realmente, se trata de la fase más importante, motivo por el cual su nombre se ha extendido y se utiliza para referirse a la disciplina completa.

El término proceso de descubrimiento de conocimiento en bases de datos, más conocido como **proceso de KDD** (*Knowledge Discovery in Databases*), se utiliza para referirse al proceso de extracción automatizada de conocimiento partiendo de grandes volúmenes de datos. (Lara Torralbo, 2016).

El conocimiento extraído por el proceso de KDD ha de poseer las cuatro características siguientes:

- **No trivial**. No sirve de nada extraer conocimiento conocido por todos o que carezca de importancia.
- **Previamente desconocido**. No se aporta nada nuevo si el conocimiento extraído ya había sido descubierto anteriormente.
- **Implícito.** Se encuentra oculto en los datos.
- **Útil**. El conocimiento extraído debe servir para algo, de lo contrario no tiene ningún sentido invertir esfuerzos en extraerlo.

Siguiendo a autores como Lara Torralbo, Reyes Saldaña y García Flores, el proceso de KDD está compuesto por diferentes fases que, como la **Figura 1** muestra son las siguientes:

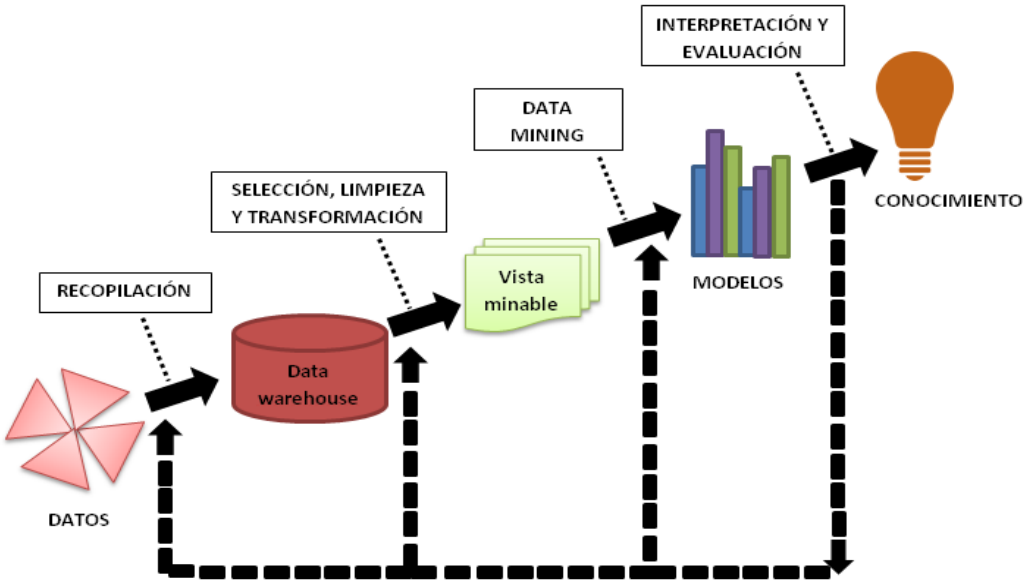

Figura 1 Proceso de KDD
Fuente: Elaboración propia a partir de (Lara, J. A., 2016)

1) **Recopilación de datos**. En esta fase, se trata de elegir un subconjunto de variables o datos, procedentes de diferentes fuentes, que se integrarán en un mismo y único repositorio de datos, denominado almacén de datos, más conocido como **data warehouse**. El resultado final de esta fase es, precisamente, ese *data warehouse.* Esto se realiza con el fin de eliminar valores redundantes e inconsistentes en los datos de varias fuentes al juntarlos dentro de una sola base de datos.

2) **Selección, limpieza y transformación de datos**. Sobre los datos recopilados en el almacén de datos no es posible realizar aún *data mining*, debido a que dichos datos pueden no estar limpios, pueden contener atributos irrelevantes, etc. Por ello, en la segunda fase del proceso de KDD se realiza una selección de los datos integrados en el *data warehouse*. Esta fase incluye operaciones básicas sobre los datos, como el filtrado para reducir ruido, crear campos explícitos con relaciones entre los atributos conocidos que puedan hacer el análisis más sencillo, etc. Dichos datos se limpian y transforman de cara a fases posteriores. El resultado de esta fase es la denominada <<vista

minable>>, que es un subconjunto limpio y transformado de los datos sobre el que ya se pueden aplicar las técnicas de *data mining* en la siguiente fase.

3) ***Data mining***. Una vez que se cuenta con una vista minable, en el siguiente paso se trata de ejecutar diferentes técnicas y algoritmos, de manera que así se puedan obtener modelos representativos.

4) **Interpretación y evaluación de modelos.** Los modelos obtenidos en la fase de *data mining* han de ser evaluados. Se trata de entender los resultados del análisis y sus implicaciones. Esto puede llevar a regresar a alguno de los pasos anteriores. Una vez comprobada la calidad de dichos modelos y de ser interpretados, a partir de ellos se obtiene el conocimiento. Este conocimiento puede implicar distintos objetivos: la meta puede ser la obtención de una descripción del sistema bajo estudio, las relaciones obtenidas pueden ser utilizadas para realizar predicciones de situaciones fuera de la base de datos, o, los resultados pueden conducir a una intervención activa en el sistema modelado.

Según algunos autores, la fase 2 se puede descomponer en varias fases. Por eso, es posible que se pueda encontrar una organización del proceso de KDD un tanto diferente a la mostrada en la **Figura 1**. Independientemente de que la fase de preparación de los datos sea una sola o se descomponga en varias, lo principal es saber que el objetivo de dicha(s) fase(s) es preparar los datos para poder realizar *data mining* con ellos en la siguiente fase.

En Witten & Frank (2000) se define la *minería de datos* como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo, debería ser automático o semi-automático (asistido) y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.

Por lo tanto, son dos los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de

información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos, etc.), y por el otro lado, usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos casos la utilidad del conocimiento minado está íntimamente relacionada con la comprensibilidad del modelo inferido, y este conocimiento va a ser obtenido a partir de diferentes herramientas que le proporciona la Estadística.

2.2 La Estadística y Data Mining

La minería de datos es un campo multidisciplinar que se ha desarrollado en paralelo o como prolongación de otras tecnologías. Por ello, la investigación y los avances en la minería de datos se nutren de los que se producen en estas áreas relacionadas.

En cuanto a la Estadística, esta disciplina ha proporcionado muchos de los conceptos, los algoritmos y técnicas que se utilizan en minería de datos, los cuales abarcan desde conceptos simples procedentes de la Estadistica Descriptiva (media, varianza,…) a propuestas más complejas procedentes de áreas como el análisis multivariante o la Teoría del Muestreo.

De esta manera, cada problema tendrá técnicas específicas para ser afrontado, lo que da lugar a un gran número de técnicas estadísticas que se pueden aplicar a problemas de *data mining*. Según Aluja (2001), algunas de ellas son:

- **Series temporales.** Partiendo de la serie de comportamiento histórica, se puede modelizar las componentes básicas de la misma, en modelos de diferente complejidad (ARIMA[2] o GARCH[3], entre otras) con la intención de realizar predicciones para el futuro (previsión de consumo de un producto, cifra de ventas, etc.).
- **Redes neuronales.** Son generalizaciones de modelos estadísticos clásicos, en los que se intenta imitar el modelo de aprendizaje de una neurona cerebral. Su novedad reside precisamente en su capacidad de aprendizaje, el hecho de usar transformaciones de las variables originales

[2] ARIMA: Modelos Autorregresivos Integrados en Medias Móviles.
[3] GARCH: es el acrónimo inglés de Heteroscedasticidad Condicional Autorregresiva Generalizada.

para predecir y la no linealidad del modelo. Esto da pié a aprender en contextos difíciles, sin tener que precisar la formulación de un modelo concreto. Su principal inconveniente es que son una caja negra para el usuario.

- **Análisis Factoriales Descriptivos.** Son técnicas de reducción d variables, agrupándolas por cercanía o mayor relación. Permiten visualizar realidades multivariantes complejas de forma más simple y, por lo tanto, exponer las regularidades estadísticas, así como eventuales discrepancias respecto de aquella e hipótesis de explicación.
- **Árboles de decisión.** Son una forma de análisis que permite, a partir de datos, la influencia de diferentes variables en una toma de decisión. La facilidad de interpretación es su principal ventaja.
- **Técnicas de <<*clustering*>>.** Son técnicas que realizan una agrupación de elementos o individuos en función de su similaridad, medida por la cercanía en unas varibales de interés.

Los campos de aplicación para estas técnicas son muy variados, y no hacen más que incrementarse debudo a la expansión de las TIC. Una de estas áreas de aplicación creciente es el análisis de datos recibidos por Internet, dando lugar al *web mining*, donde se utilizan las técnicas de *data mining* para optimizar las interacciones a través de la web.

Además, los datos a analizar pueden ser textos, lo que da lugar al *Text Mining*, que es un campo actual de investigación para la presentación de la información encontrada en la web. Dicha área de investigación será analizada posteriormente en este trabajo.

2.3 Minería de Web y textos

En principio, la minería de datos puede aplicarse a cualquier tipo de información, adaptándose sus técnicas a las diferentes situaciones que se pueden encontrar. En concreto, en este trabajo, se hará hincapié en el análisis de datos no estructurados provenientes de la web o de otros tipos de repositorios de documentos.

La World Wide Web es el repositorio de información más grande y diverso de los existentes en la actualidad. Su origen data de 1990. El código inicial fue escrito por Berners-Lee en el Laboratorio de Física de Altas Energías (CERN) en Suiza (Hernández et al., 2004, p. 546). Como él mismo afirmó*: "el principal objetivo de la web fue tener un espacio de información compartido a través del cual máquinas y personas pudieran comunicarse"*. Estaba especialmente interesado en que se pudieran comunicar máquinas y *software* de diferentes tipos. Para ello, desarrolló un identificador de recursos universal (*Uniform Resource Locator*, URL) para poder referirse a cualquier documento (u otro tipo de recurso) en el universo de información. Asimismo, en lugar del protocolo de transferencia de archivos utilizado en ese momento para el intercambio de información, creó a partir de él un protocolo de transferencia de hipertexto (*Hypertext Transfer Protocol*, HTTP) más rápido que el primero y un lenguaje de marcas para hipertexto (*Hypertext Markup Language*, HTML).

Actualmente, Internet es el medio más popular e interactivo de difundir información. Pero esta situación hace que a menudo los usuarios tengan una sobrecarga de información. Según Kosala & Blockeel (2000) algunos de los problemas que se presentan cuando se interactúa con la web son:

- **Encontrar información relevante**: cuando un usuario utiliza servicios de búsqueda para encontrar una información específica en la web, normalmente se introduce una pregunta con las palabras clave y obtiene como respuesta una lista de páginas ordenadas según su similitud con la pregunta. Sin embargo, estas herramientas de búsqueda tienen, por lo general, una precisión bastante baja debido a la irrelevancia de muchos de los resultados de la búsqueda. A esto se une la limitada memoria que las hace incapaces de indexar toda la información disponible en la web, por lo que se hace incluso más necesario encontrar la información relevante a la pregunta.
- **Crear un nuevo conocimiento**: la relevancia de la información obtenida en las consultas a la web es un problema estrechamente relacionado con el de crear nuevo conocimiento a partir de la información disponible en la web, es decir, una vez obtenidos los datos tras el proceso de búsqueda probablemente se quiera extraer coincidencias, resúmenes, patrones, regularidades y, al fin y al cabo, conocimiento a partir de esos datos. Se puede decir, que si encontrar

información en la web es un proceso orientado a la recuperación, la obtención de conocimiento útil es un proceso orientado a la minería de datos.

- **Personalización de la información**: a menudo se asocia este problema con la presentación y el tipo de la información, ya que los diferentes usuarios suelen tener gustos distintos a la hora de preferir ciertos contenidos y presentaciones cuando interactúan con la web. Muy relacionado con este problema está el de aprender de los usuarios, es decir, saber qué es lo que los usuarios hacen y quieren. Esto permite personalizar la información incluso para un usuario individual (diseño de portales web, de herramientas software, filtros de correo, etc.).

La enorme cantidad de información disponible hace de la web un área fértil para la minería de datos cuyas técnicas pueden resolver los problemas que se acaban de mencionar.

Sin embargo, a diferencia de las bases de datos relacionales que poseen una estructura bien definida, la web es poco estructurada por naturaleza. Esto significa que muchas de las técnicas de minería de datos no pueden aplicarse directamente, deben modificarse o, incluso, deben definirse nuevas técnicas. De hecho, tradicionalmente, la minería de datos se ha aplicado a las bases de datos, ya que era un formato de fácil procesamiento por los computadores, mientras que la información en la web reside en documentos enfocados al consumo humano tales como páginas personales, publicitarias, información general o catálogos de productos.

Se puede definir entonces la **Minería Web** como el uso de técnicas de minería de datos para descubrir y extraer información automáticamente desde el World Wide Web. (Etzioni, 1996).

Minar la web no es un problema sencillo, debido a que muchos de los datos son no estructurados o semi-estructurados, que muchas páginas web contienen datos multimedia (imágenes, texto, vídeo y/o audio), y a que estos datos pueden residir en diversos servidores o en archivos (como los que contienen los logs). Otros aspectos que dificultan la minería web son cómo determinar a qué páginas se debe acceder y cómo seleccionar la información que va a ser útil para extraer conocimiento. Toda esta diversidad hace que la minería web se organice en torno a tres categorías (Hernández et al., 2004):

- Minería del contenido, para encontrar patrones de los datos de las páginas web.
- Minería de la estructura, entendiendo por estructura los hipervínculos y URLs.
- Minería del uso que hace el usuario de las páginas web (navegación).

Dentro de estas tres categorías, en este trabajo se va a centrar en la primera de ellas, la minería del contenido, ya que nuestro objetivo es estudiar como obtener información de texto, en particular del texto de una web y con ella extraer conocimiento útil, en particular desde el punto de vista de la investigación de mercados, el área escogida para el ejemplo aplicado que se realizará.

3. Text Mining

Desde el comienzo de la historia, el hombre utiliza la escritura para preservar y transmitir sus conocimientos. Con el tiempo, el hombre fue creando normas sobre la manera de registrar y preservar el conocimiento, en particular, la definición de alfabetos que permitieron sistematizar el texto escrito. El texto pasó entonces a ser el proceso formal de transmisión de conocimientos entre las personas.

Más recientemente, con el advenimiento de la sociedad de la información, la cantidad de información textual generada y disponible para cada uno de nosotros ha crecido exponencialmente. Por lo tanto, existe una brecha cada vez mayor entre la cantidad de datos y la información producida y la capacidad de leer, asimilar y convertirlos en conocimiento. Por otro lado, las soluciones de gestión de la información que ofrece la Tecnología de la Información y la Comunicación (TIC) están orientadas al tratamiento de datos estructurados, no siendo tan eficaz para la gestión de contenido textual[4].

La creciente cantidad de datos textuales que se producen, almacenan y se difunden, debido al uso masivo e intensivo de los medios informáticos, muy especialmente en la actividad económica e institucional, hace necesario la utilización de métodos y algoritmos con capacidad para tratar y analizar datos lingüísticos que comportan imprecisiones, vaguedad y, en parte, incertidumbre (Pavone, 2015).

[4] Véase https://run.unl.pt/handle/10362/6244

Hoy en día las soluciones encontradas ya no se basan sólo en las herramientas estadísticas, sino que son el resultado de un enfoque multidisciplinar que combina con igual importancia, herramientas informáticas, estadísticas y lingüísticas, principalmente en el área de investigación conocida como *Text Mining* (Bolasco et al., 2005).

3.1 Concepto de Minería de Textos

El *text mining* o **minería de textos,** es un área de investigación emergente que apunta a la extracción de la información significativa presente en textos escritos en lenguaje natural. Con esto se entiende cualquier aplicación de métodos que analizan automáticamente los datos textuales con el fin de obtener, a partir de fuentes no estructuradas, el conocimiento utilizable. Intrínsecamente interdisciplinario, el *text mining* proviene de ámbitos cercanos a él, como el *data mining* (Feldman & Sanger, 2007). En este contexto, la minería de textos se ha convertido en una herramienta indispensable para la extracción de conocimiento significativo desde datos textuales.

Por lo tanto, el objetivo de la minería de textos es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado. Por no estructurado nos referimos a texto libre, generalmente en lenguaje natural, aunque también podría ser código fuente y otro tipo de información textual.

Aggarwal y Zhai (2012) definen este concepto como "*el retorno de documentos/información relacionada con el conjunto de palabras clave que el usuario entiende que describen la respuesta que espera"*. Es decir, el usuario define un conjunto de términos relacionados con la información que desea y espera el retorno de información relacionada con estos mismos términos. Por lo tanto, se percibe que está buscando la similitud entre un conjunto de palabras y el conjunto de documentos existentes.

Miner et al. (2012) mencionan la misma definición explicada anteriormente pero añaden un detalle que puede hacer toda la diferencia en una investigación por similitud: la utilización de un diccionario de sinónimos durante el proceso. Claramente, si está asociado al conjunto de palabras clave definidas por el usuario, está otro

conjunto de palabras que son sinónimos de las primeras referidas, posiblemente aumentará el número de resultados obtenidos.

Este proceso se ha convertido en un método popular para el análisis y la compresión de grandes conjuntos de datos no susceptibles a las técnicas tradicionales de investigación cualitativa.[5]

3.2 Text Mining y Data Mining

De manera análoga a la minería de datos la minería de textos busca extraer información útil desde fuentes de datos a través de la identificación y exploración de patrones interesantes. En el caso de la minería de textos, sin embargo, las fuentes de datos son colecciones de documentos y se encuentran patrones interesantes no entre los registros de base de datos formalizados, sino datos textuales no estructurados en los documentos de estas colecciones.

Ciertamente, la minería de textos deriva gran parte de su inspiración y dirección de investigación sobre el *data mining*. Por lo tanto, no es sorprendente encontrar que la minería de textos y los sistemas de minería de datos evidencian muchas similitudes arquitectónicas de alto nivel.

Dado que la minería de datos supone que los datos ya se han almacenado en un formato estructurado, gran parte de su enfoque de pre-procesamiento recae en dos tareas críticas: Hacer limpieza y normalizar datos y crear un gran número de combinaciones de tablas. Por el contrario, para los sistemas de minería de texto, las operaciones de pre-procesamiento se centran en la identificación y extracción de entidades representativas de documentos en lenguaje natural. Estas operaciones de pre-procesamiento son responsables de transformar los datos no estructurados almacenados en las colecciones de documentos en un formato intermedio estructurado más explícitamente, lo cual es una preocupación que no es relevante para la mayoría de los sistemas de minería de datos. (Feldman & Sanger, 2007).

[5] Véase http://journal.code4lib.org/articles/11626

3.3 La colección de documentos y el documento

Un elemento clave de la minería de textos es su enfoque en la colección de documentos (Miner et al., 2012). En su forma más simple, una colección de documentos puede ser cualquier agrupación de documentos basados en texto.

Sin embargo, prácticamente hablando, la mayoría de las soluciones de minería de textos están dirigidas a descubrir patrones en colecciones de documentos muy grandes. El número de documentos en dichas colecciones puede oscilar entre los muchos miles y las decenas de millones.

Las colecciones de documentos pueden ser estáticas, en cuyo caso el complemento inicial de los documentos permanece sin cambios, o dinámico, que es un término aplicado a las colecciones de documentos caracterizadas por la inclusión de documentos nuevos o actualizados a lo largo del tiempo. Las colecciones de documentos extremadamente grandes, así como colecciones de documentos con tasas muy altas de cambio de documentos, pueden plantear desafíos de optimización de rendimiento para varios componentes de un sistema de minería de texto.

Los sistemas de minería de texto, sin embargo, no suelen ejecutar sus algoritmos de descubrimiento de conocimiento en colecciones de documentos no preparados. Un énfasis considerable en la minería de textos se dedica a lo que se conoce comúnmente como operaciones de pre-procesamiento.

También es un elemento básico en la minería de texto el documento. Para fines prácticos, un documento puede definirse como una unidad de datos textuales discretos dentro de una colección que por lo general, pero no necesariamente, se correlaciona con algún documento del mundo real, como un informe comercial, un memorándum legal, un correo electrónico, un trabajo de investigación, un manuscrito, un artículo, un comunicado de prensa o una noticia. (Feldman & Sanger, 2007)

3.4 Fases del Text Mining

Generalmente, un trabajo de *text mining* incluye las siguientes fases (De la Calle, G., 2014, p. 40-41) (Sakaji, G., 2016, p. 20-22):

- **Recuperación de información o selección de documentos.** Ya que todo proceso de minería de texto trabja sobre un corpus textual, la primera tarea consiste en determinar, construir o definir el conjunto de textos relevantes para el proceso de minería de texto en cuestión. Ese conjunto de textos puede ser obtenido de varios tipos de fuentes como artículos, periódicos, noticias, etc. En este trabajo han sido obtenidos a través de un blog de telefonía móvil.

  Dos de los principales problema que surgen durante esta etapa son la detección de documentos irrelevantes y de documentos que, siendo relevantes, no han sido incluidos en la colección. Sin embargo, el triunfo en esta etapa depende en gran parte de la correcta caracterización realizada previamente sobre los documentos que se han recuperado.

- **Fase de pre-procesamiento de textos o extracción de información**. Su objetivo es detectar y extraer información relevante a partir de textos escritos en lenguaje natural. Esto se realiza con el fin de mejorar la calidad de los datos disponibles y organizarlos de una forma más fácil para el posterior análisis. Es decir, se trata de crear una representación estructurada intermedia útil para la siguiente etapa, partiendo de los textos (datos no estructurados) contenidos en los documentos del corpus.

  Los mecanismos más utilizados para el pre-procesamiento de textos son la eliminación de *stopwords* y *stemming*.

  - *Stopwords*, son términos como preposiciones, artículos o pronombres, considerados de escaso valor semántico, por lo que al poseer una frecuencia muy elevada en los documentos se eliminan. De esta forma se evitan los problemas de ruido documental.[6]

[6] Véase https://www.upf.edu/hipertextnet/numero-5/pln.html

- *Stemming,* es un proceso por el que se cortan las palabras de los documentos antes de indexarlos[7], con el fin de identificar palabras de la misma raíz y eliminar afijos derivativos.[8]

- **Fase de Data Mining o de descubrimiento.** Consiste en descubrir el conocimiento que está oculto en el corpus original. Para ello, los distintos métodos tratan de establecer relaciones y de encontrar patrones entre los diferentes elementos extraídos en la anterior fase de pre-procesamiento de textos.

  Con esta fase se obtiene la evaluación que comprende la interpretación de los patrones extraídos, con el fin de comprobar si se han alcanzado los objetivos trazados.

El *text mining* es un instrumento típico de las aplicaciones destinadas a empresas e instituciones (Bolasco, et al., 2005), que, al tener que interactuar con grandes masas de materiales de texto tienen el problema de seleccionar, dentro de estas enormes fuentes, los datos de interés para extraer información capaz de producir valor. Las operaciones típicas de *Text Mining* son: la categorización de textos, el agrupamiento de texto, extracción de entidad / conceptos, *sentiment analysis*, *document summarization*, y *entity relation modeling*.

3.5 Herramientas para el Text Mining

Las herramientas para la minería de textos se pueden desarrollar ad hoc para cada caso concreto. No obstante, existe un importante abanico de sistemas de text mining preparados para su aplicación.

En este trabajo se utilizará el *paquete tm* de R en el desarrollo del caso práctico. Dicho paquete será el utilizado para analizar los textos extraídos artículos y comentarios de un blog de telefonía móvil, en las que se analizarán dos marcas, Apple y Samsung, y el conjunto de artículos englobados en la categoría “Otras”.

[7] Se refiere a ordenar una serie de datos o informaciones de acuerdo a un criterio común a todos ellos, para facilitar su consulta y análisis.

[8] Véase https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

4. Caso práctico: Text Mining en Web.

"*Uno de los retos más complicados a la hora de abordar el caso Apple Vs Samsung es intentar no caer en la simplificación de "buenos y malos", en abstraerse de las filias y fobias alrededor de las dos marcas y de las dos plataformas e intentar ser coherente con la admiración que uno siente por los innovadores en tecnología y lo que se viene criticando desde hace años contra las patentes software*." [9]

Una vez expuesta la metodología del *text mining*, y su forma de trabajar, se va a realizar un caso práctico en el que se muestra la aplicación de lo expuesto hasta ahora.

El objetivo va a ser analizar la imagen de dos marcas de telefonía móvil, Apple y Samsung, en un periodo que abarca los años 2007 (año de aparición del primer iPhone en el mercado) hasta 2016. Se han tomado como fuente los comentarios posteados en el blog XatakaMóvil durante dichas fechas. Posteriormente se han procesado mediante un algorítmo de análisis de textos para realizar con los datos obtenidos un análisis estadístico y obtener unas conclusiones finales dentro del marco teórico de la investigación de mercados.

En este caso se encuadraría dentro de la Investigación de Mercados porque ésta es un área especialmente interesada en temas de analítica web y recolección de información a partir de mensajes textuales.

4.1 La Investigación de Mercados

La **investigación de mercados** *es la función que vincula al consumidor, cliente y público con el vendedor a través de la información. Esta información se utiliza para*

[9] Véase: https://www.xataka.com/analisis/no-no-fue-un-gran-dia-para-los-innovadores

identificar y definir las oportunidades y los problemas de marketing y a través de ella generar, redefinir, evaluar y controlar la ejecución de las acciones de marketing. (American Marketing Association, AMA[10]).

La investigación de mercados proporciona información pertinente y actualizada de los diferentes agentes que actúan en él. Por tanto, su finalidad es la obtención de información útil para la toma de decisiones. De hecho, no se debe considerar a la investigación como la solución a problemas empresariales sino un instrumento más que permita minimizar riesgos y, en consecuencia, las decisiones puedan ser más acertadas.

En el siglo XXI, las empresas ya son conscientes de la importancia que tiene actualizar constantemente la información sobre sus clientes; y, al mismo tiempo, valorar si el término <<constantemente>> equivale a todos los días, todas las semanas, todos los meses o todos los años. ¿Qué información puede obtener una empresa sobre el mercado, clientes, competidores…? La información puede ser muy variada, desde noticias en prensa, visitas a un establecimiento, flujos de visitantes en una ciudad, seguimiento de una web o repeticiones de la compra de un producto, entre otras (Esteban y Molina, 2014).

Esta reflexión sobre lo que significa <<constantemente>>, se vuelve pertinente ante la aparición de internet y la avalancha de información que con ella ha llegado. Informacion que con frecuencia está colocada ahí por consumidores, y que para una marca puede ser una fuente de información.

4.1.1 La Investigación On-line

Internet ya se ha convertido en una de las herramientas más importantes de la investigación de mercados. Lo que hasta hace tan sólo unos años era una promesa, hoy se considera como una alternativa viable en muchos de los estudios comerciales que se realizan.

[10] Véase https://www.ama.org/Pages/default.aspx

Más que una alternativa a la investigación de mercados tradicional, la investigación on-line se debe considerar como un apoyo más, del que se puede sacar mucho provecho si se utiliza en las situaciones adecuadas.

En Internet hay mucha más información de la que ofrecen ciertos organismos como el Instituto Nacional de Estadística en España, ya que los propios consumidores son los que generan más datos a través de foros, redes sociales y blogs. Sin embargo, este tipo de información no está estructurada y es más difícil de explotar, aunque en los últimos años se han hecho grandes avances en esta área (*Text Mining*, análisis semántico, análisis de redes, etc.). Incluso en las redes sociales, se puede seguir la actividad de un individuo, y aprender cuáles son sus intereses, sus hábitos o sus preferencias.

Por este motivo la Investigación de Mercados es una de las áreas más interesadas en el campo del análisis automático de textos, interés que sólo puede aumentar, dadi el incremento exponencial que está teniendo la información on-line.

4.1.2 La Investigación de Mercados y el análisis automatizado de textos

"*Los blogs, páginas web y sobre todo lugares como Facebook o Twitter son las nuevas fuentes de información de las empresas, y están sustituyendo de forma progresiva los canales tradicionales. Los consumidores tienden a ser más francos en Internet, por lo que sus opiniones ofrecen mucha mejor información de sus actitudes y sus posibles reacciones*", explica el vicepresidente Senior de Análisis Estratégico en SPSS, Colin Shearer.

El problema es que esta información es textual. No está codificada ni es numérica para poder ser analizada con las herramientas de análisis estadístico clásicas. Además de esto, su volumen es tan cambiante y descomunal, que realizar esta tarea de forma manual es inviable.

Por ello, han surgido herramientas de análisis automatizado de textos (lo que en epígrafes anteriores se ha denominado *text mining*), que son capaces de cuantificar

los términos y los contextos de las manifestaciones, declaraciones o conversaciones de los individuos en la red. El análisis de contenido permite identificar conceptos claves en cualquier texto, en base al análisis de tipo morfológico y sintáctico, lógico, semántico y estadístico (Merino, 2015).

Por lo tanto, para realizar un análisis de contenido, las técnicas aplicadas se basan en los siguientes análisis.

El **análisis morfológico y sintáctico** permite eliminar las ambigüedades presentes en el texto a analizar, clasificando cada palabra desde un punto de vista gramatical y reduciéndola a su entrada en un diccionario.

El **análisis lógico** consiste en la identificación automática del rol funcional desarrollado en el texto de cada una de las palabras (sujeto, complemento de lugar, de tiempo o de modo, complemento objeto, etc.), permitiendo comprender quien hace cada cosa, cuándo, cómo y dónde.

El **análisis semántico** recoge, además, el significado profundo de cada palabra. Proyectar esta información en el tiempo permite medir con eficacia el grado de satisfacción de los clientes en una empresa de frente a sus iniciativas comerciales.

Por último, el **análisis estadístico** asigna los textos <<conceptualizados>> a categorías temáticas predefinidas y personalizadas (categorización).

4.2 Desarrollo del trabajo aplicado

Para realizar el análisis, se han tomado los artículos escritos en el blog de móviles XatakaMovil[11], bajo las categorías Apple y Samsung, que van desde el 1 de Enero de 2007 a 31 de Diciembre de 2016. El objetivo es realizar un estudio comparativo entre estas dos marcas, siguiendo los textos y los comentarios de sus noticias. Además, dado que estas marcas son líderes del mercado, se han analizado también las noticias agrupadas en la categoría de Otras, como representativa de una evolución general del mercado. El espacio temporal seleccionado comprende desde el año de lanzamiento

[11] Véase https://www.xatakamovil.com/

del primer iPhone hasta el último año completo que se puede incluir en el análisis. Sin embargo, la totalidad de los años sólo se ha utilizado en el análisis previo. Para el análisis de *text mining* que se va a llevar a cabo, se han seleccionado tres años en concreto, 2007, 2011 y 2016, el inicial, un año intermedio y el final, ya que tres años con suficiente distancia facilitarán la observación de diferencias entre ellos y el análisis no será así tan largo.

4.2.1 Herramientas utilizadas

Dadas las características del trabajo, se ha recurrido a software libre para realizar todo el proceso, concretamente se ha utilizado el software estadístico R, usando Rstudio como interface y manejándolo a partir de plantillas ya preparadas.

R[12] es un conjunto integrado de servicios de software para la manipulación de datos, cálculo y representación gráfica, el cual incluye, entre otros, un conjunto de operadores para los cálculos en matrices, instalaciones gráficas para análisis de datos y de visualización, un simple y eficaz lenguaje de programación, etc. Está diseñado en torno a un cierto lenguaje informático, y permite a los usuarios añadir funcionalidad adicional mediante la definición de nuevas funciones.

Debido a la extensibilidad y flexibilidad de R, se ha mantenido constantemente como un programa popular para los datos y han aparecido en la actualidad difeentes aplicaciones de minería de texto como el paquete *tm*[13], que es el usado en este trabajo.

Por otro lado, Rstudio[14] es un entorno de desarrollo integrado popular y eficiente para trabajar con R, particularmente útil para aquellos que comienzan su trabajo con este lenguaje. Además de la consola básica, muchas herramientas que pueden reducir en gran medida la curva de aprendizaje.

En este trabajo se utilizará dicha herramienta para el análisis de los datos de Apple y Samsung.

[12] Véase https://cran.r-project.org/doc/contrib/rdebuts_es.pdf
[13] Véase https://cran.r-project.org/web/packages/tm/
[14] Véase https://www.rstudio.com/

4.2.2 Obtención del texto

La primera fase en cualquier proceso de text mining es la recuperación de información, la cual en este caso ha consistido en descargar el conjunto de noticias del blog encuadrado en las categorías seleccionadas. Para ello se recurrió a técnicas de "*Web Scraping*" (Munzert, Rubba, Meibner & Nyhuis, 2015) para la extracción de la información.

Según Glez-Peña et al. (2013, p. 789) el *Web Scraping* generalmente se define como un "*proceso de extracción y combinación de contenidos de interés desde la Web de forma sistemática"*. En dicho proceso, un agente de software imita la interaccion de navegación entre los servidores web y el humano en un recorrido Web convencional. De esta forma, el agente software, paso a paso, va accediendo a tantos sitios web como sean necesarios, analizando su contenido para encontrar y extraer los datos de interés y las estructuras de los contenidos según se desee.

Para este trabajo se ha utilizado como herramienta para Web Scraping los comandos básicos de R para la lectura de textos, ya que una página web puede ser considerada como un fichero de texto normal almacenada en un ordenador remoto.

El proceso de obtención del texto se ha llevado a cabo a través de cuatro pasos:

1º) Estudio de las características de la web para decidir qué páginas se adaptaban a lo que se pretendía en el estudio. En este análisis se han identificado tres categorías del blog: **Apple**, **Samsung** y **Otras**, así como la estructura de sus páginas, y de las páginas de noticias relacionadas con ellas.

2º) Posteriormente, se ha descargado el texto "en bruto" de la página del blog, con su código. El proceso utilizaba la información anterior para identificar las categorías, y apartir de la página inicial de cada una de ellas, que tiene diez noticias, se buscó el enlace a las diez noticias anteriores y a la siguiente página de categoría, para poder descargarlas, y así sucesivamente hasta llegar al año 2007 en cada una de esas categorías.

3º) A continuación, se detectaron dentro de ese código las partes que identifican la información que pudiera ser relevante para este estudio: nombre del artículo, fecha, *tags*, texto del artículo y texto de los comentarios. Esto se ha conseguido identificando diferentes tipos de etiquetas que separaban partes de la página del código HTML. Por ejemplo, el inicio del texto del artículo se identificaba con la etiqueta ** ,** y el inicio del texto de los comentarios con **<AML.Comments.config.data>**.

4º) Siguiendo las identificaciones detectadas en el paso anterior, se procedió a extraer la información para cada artículo y sus comentarios, almacenándolos como una colección de documentos con la que posteriormente se trabajaría.

En total se han extraído 1.567 artículos de Apple, 2.168 de Samsung y 1.921 de Otras. En cuanto a los comentarios, un total de 1.429 de Apple, 2.008 de Samsung y 1.703 de Otras. Sin embargo, el análisis se realizará únicamente comparando los años 2007, 2011 y 2016, de los cuales se han extraído 387 artículos de Apple, 518 de Samsung y 577 de Otras, mientras que los comentarios han sido en Apple 329, en Samsung 405 y en Otras 431.

4.2.3 Preparación del texto

Esta es la segunda fase del proceso de *text mining*, la fase de pre-procesamiento de textos o extracción de información. Teniendo ya los datos organizados como una colección de documentos, siendo cada uno de ellos un fichero de texto tipo “texto elemental”, identificados con la extensión .txt, se ha llevado a cabo la limpieza de los mismos a través de diferentes pasos[15]. Para ello se utiliza a través de R un paquete llamado *tm* que está diseñado para analizar una colección de documentos de texto, conocido como corpus.

Los pasos específicos para este proceso y su código se ha adaptado del artículo publicado por Maceli (2016) y los pasos que se han seguido son los siguientes:

[15] Véase http://journal.code4lib.org/articles/11626

1º) En el primer paso se ha **elaborado el corpus de documentos** a partir de los textos bajados en la fase anterior de obtención del texto. Entendiendo como un corpus la forma específica en la que *tm* organiza la colección de documentos para que el software pueda actuar sobre él.

2º) A continuación, se realiza la **limpieza de dichos textos**, buscando la eliminación de palabras genéricas que no aporten significado y homogeneizar las diferentes formas de las que sí nos interesan:

- Se pone todo en minúsculas.
- Eliminación de los espacios en blanco.
- Se realiza el *stemming*: se ha homogeneizado las palabras de todos los textos de manera que aparezcan de una sola forma las variedades de cada palabra (masculinos y femeninos, singulares y plurales, tiempos verbales a infinitivo, etc.).

3º) **Se eliminan los *stopwords***, quitando palabras que no son significativas como “el”, “la”, “estas”...

4º) Después **se han quitado los números y símbolos de puntuación**.

De esta forma se eliminan cacteres extraños, palabras no deseadas y especios en blanco excesivos, para asegurarse de que las palabras se cuentan correctamente.

Finalmente, con los textos limpios se elaboran las matrices de términos de los documentos (Document Term Matrix), que serán el objeto sobre el cual se aplicarán los análisis estadísticos con los que se realizará la tercera fase del text mining, la fase de *data mining* o descubrimiento, en la que se obtendrá la información sobre marcas que eran el objetivo de esta aplicación práctica.

4.3 Análisis de los datos

4.3.1 Conocimiento previo de los datos

En este primer análisis lo que se tratará es de ver la evolución de los artículos y sus comentarios por años y marcas. Se buscará con él un "proxy" del interés por las marcas, suponiendo que un mayor interés por una marca provoca un mayor número de noticias sobre ella.

Comenzando con el análisis más global, se ha obtenido los siguientes gráficos de barras donde se representan el número de artículos publicados por marca y el número medio de comentarios por marca, respectivamente.

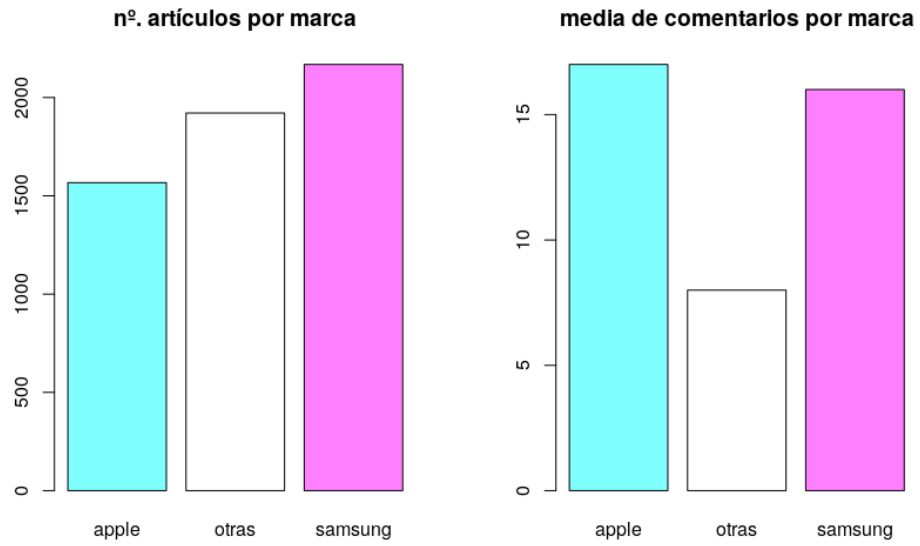

Figura 2 Nº de artículos y nº medio de comentarios por marca
Fuente: Elaboración propia a partir de datos de XatakaMóvil.

Se observa que Apple cuenta con un número medio de comentarios mayor que Samsung y Otras. Dicho número medio en Apple es 26, mientras que en Samsung es 22 y en Otras solo 12. Esto quiere decir que los usuarios del blog han publicado más comentarios sobre temas relacionados con Apple. Sin embargo, Samsung es el más publicado y Apple el que menos. Esta diferencia entre los dos indicadores puede muy bien deberse a la capacidad de Apple para ser el foco de atención y para diferenciarse, es decir, a su capacidad para "dar que hablar".

Comparando lo mismo, pero por años, se obtiene que en el año 2010 es cuando más comentarios se publican, 31 comentarios de media, seguido del 2011 con 30 (Figura 15).

Sin embargo, en el año 2007 es cuando menos comentarios se han escrito. Esto puede darse por distintos motivos. Uno de ellos es que en ese momento el blog aun no era tan popular, pero también puede influir la situación de Apple, la marca globalmente más comentada, ya que en Junio de dicho año es cuando se lanza el primer iPhone de Apple y "*El modelo de negocio para el primer año del iPhone fue un desastre*", explica Tony Fadell, uno de los desarrolladores de Apple del dispositivo. "*Pivotamos y lo descubrimos en el segundo año*".[16] Esta situación empezaría a cambiar en 2010, año en el que las dos áreas donde Apple ve crecimiento astronómico están en las líneas de iPhone y iPad[17], con la presentación del iPhone 4s (El terminal de Apple fue presentado en sociedad el pasado mes de octubre y desde ese momento sus cifras de venta han sido meteóricas y nunca vistas por una empresa acostumbrada a vender sus productos en cantidades masivas[18]) y con el lanzamiento de su primer iPad[19], y por lo tanto, momento en el que se disparan los comentarios en este blog.

Para conocer la evolución de las marcas en el blog, se ha obtenido el siguiente gráfico de líneas, que explica el número de artículos por marca y año.

[16] Véase http://www.eleconomista.es/tecnologia-gadgets/noticias/8464605/06/17/Hace-10-anos-salio-a-la-venta-el-primer-iPhone-la-piedra-angular-de-Apple.html
[17] Véase http://www.macworld.com/article/1156506/applefin.html
[18] Véase https://www.adslzone.net/article7961-el-iphone-4s-fue-el-telefono-mas-vendido-en-el-ultimo-trimestre-de-2011.html
[19] Véase https://www.applesfera.com/apple/ipad-historia-de-un-tablet-primera-parte

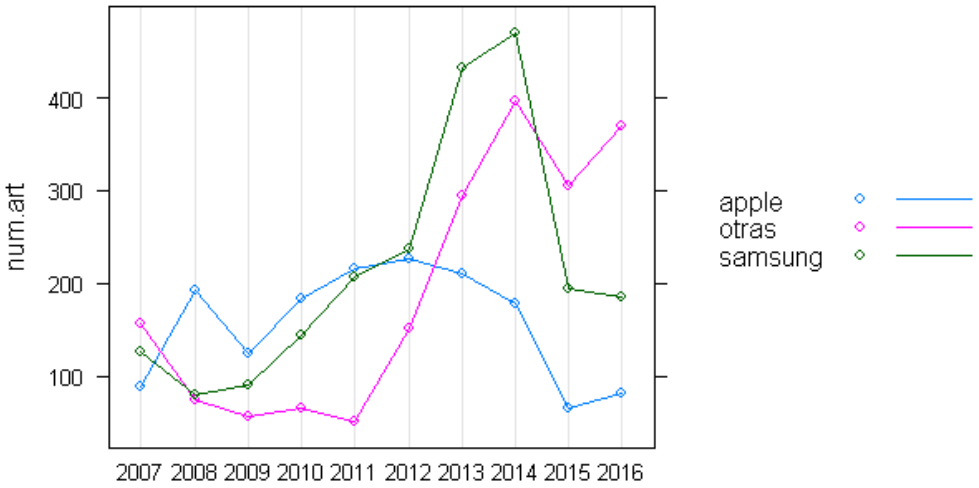

Figura 3 Nº de artículos por marca y año
Fuente: Elaboración propia a partir de datos de XatakaMóvil.

Aunque en el 2007 Apple haya comenzado con un menor número de artículos, durante el periodo 2008-2011 es la marca con mayor número de artículos en XatakaMovil. A partir del año 2012, comienza a disminuir dicho número hasta ser desde 2013 a 2016 la marca con menos artículos en el blog.

Sin embargo, Samsung y Otras, exceptuando algunos años, crecen casi en paralelo. El número de artículos de Samsung se dispara desde el año 2008 hasta el 2014. Esto podría darse debido que a partir de 2010, dicha marca lanza al mercado la serie de smartphones Galaxy con el sistema operativo Android. Desde entonces, Samsung no deja de crecer y evolucionar[20]. A pesar de ello, esta marca deja el puesto de mayor número de artículos publicados sobre ella en el año 2014, pasando de publicarse 470 artículos en dicho año a 195 el año siguiente.

Cabe destacar que en el año 2015, las tres variables analizadas sufren una reducción en el número de artículos publicados. En ese año, según el informe elaborado por la Comisión Nacional de los Mercados y la Competencia (CNMC), el sector ha estado marcado por las guerras de precios y el agravamiento de la crisis económica.[21]

[20] Véase http://www.samsung.com/es/aboutsamsung/samsungelectronics/history/history/
[21] Véase https://cincodias.elpais.com/cincodias/2015/11/11/empresas/1447269007_460682.html

En el último año analizado, Otras es quien más peso tiene en cuanto a número de artículos publicados. Dentro de esta categoría se encuentran marcas como Huawei, ZTE, Lenovo o Nokia, entre otras. Más rápido de lo que se pensaba, marcas chinas como Huawei, Lenovo y Xiaomi aumentaron su cuota de mercado en todo el mundo al ganarse la confianza del público[22]. Precisamente en 2011 se presentaron en España los productos de Huawei y ZTE iniciando la progresión de este tipo de marcas que se observa en la gráfica. En esta línea, otras marcas como Meizu, Xiaomi u OnePlus llegaron al mercado nacional hace poco más de un año y ya forman parte del catálogo de teléfonos inteligentes más vendidos.[23] Este tema será analizado posteriormente con mayor detalle.

4.3.2 Aplicación de *text mining:* análisis de frecuencias

Después de un análisis previo, del cual se obtuvo una visión general de la evolución de las marcas, se va a comenzar a utilizar *text mining* para profundizar en ese conocimiento comentado con anterioridad, y tratar de descubrir características que hablen sobre esa evolución.

Se ha comenzado el análisis con Apple, ya que es la marca de referencia, y en función del análisis previo de número de artículos y comentarios, Samsung y la categoría Otras se comentarán a partir de esta referencia, viendo similitudes y diferencias con el análisis de Apple.

Como técnicas de análisis se han utilizado frecuencias de ocurrencia de las palabras, separadas por marca, año y tipo de documento, visualizándolas mediante nubes de palabras, y, en ocasiones, diagramas de barras.

Las nubes de palabras se han formado con las 30 palabras más frecuentes en cada agrupación de documentos. En general, las palabras más repetidas son las que en las nubes figuran en color gris, centradas y en un tamaño mayor al resto. Por lo tanto, las frecuencias están escaladas por colores y tamaño.

[22] Véase http://www.abc.es/economia/abci-marcas-chinas-moviles-conquistan-mercado-mundial-bajos-precios-201602172116_noticia.html

[23] Véase http://economia.elpais.com/economia/2014/06/12/actualidad/1402594208_091445.html

Para obtener las nubes se han quitado las palabras “Apple”, “iPhone”, “iPad”, “iPod”, “Samsung” y “Galaxy”. Así se obtienen unos resultados más claros, ya que si se dejasen dichas palabras, tendrían como palabras más repetidas éstas dentro de la propia marca, lo cual es una obviedad.

APPLE

2007

Figura 4 Nubes de palabras de artículos y comentarios (Apple, 2007)
Fuente: Elaboración propia a partir de datos de XatakaMóvil.

Las palabras que más se repiten en los artículos son “teléfono” y “móvil”, dos formas de decir lo mismo, que se refieren al producto en sí. Hay otras palabras como “lanzamiento”, “noticia”, “hecho”, “oficial” y “dato”, que podrían estar refiriéndose a la salida del primer iPhone ocurrida en ese año.

Al aparecer palabras como “España” y “Europa”, y ya que el blog analizado es español, igual esto se puede dirigir a que se publicaron artículos respecto a las zonas geográficas donde saldría a la venta dicho producto. También palabras como son “operadora”, “compañía” y “Vodafone” podrían estar relacionadas. Según The Guardian, los dos móviles más esperados, el iPhone y el Google Phone, ya andan buscando novias para distribuir su teléfono en el mercado europeo. Apple está sopesando la opción de Vodafone, que parece que ha sido el más rápido y al tratarse de un operador que tiene presencia en varios países parte con ventaja.[24]

[24] Véase https://m.xataka.com/moviles/iphone-con-vodafone-y-google-phone-con-o2#c206427

Una palabra importante en estas nubes es “poder”, que se refiere al verbo, y es una palabra que no tendrá importancia en otras nubes, con lo que se tendría otra muestra del lanzamiento, y la descripción de “lo que puede hacer” el dispositivo.

Respecto a los comentarios**,** se observa que se habla más sobre “contrato”, “libre”, “producto”, “mercado”, lo cual puede reflejar intereses más específicos de los usuarios. Una diferencia importante con las referencias dadas por los artículos.

Otra diferencia importante es que los comentarios no se utiliza tanto la palabra “teléfono”, sino que los usuarios se dirigen más con “móvil” al ahora de referirse al iPhone. Esto indica una pauta lógica de textos menos técnicos que los elaborados por los gestores del blog.

2011

Figura 5 Nubes de palabras de artículos y comentarios (Apple, 2011)
Fuente: Elaboración propia a partir de datos de XatakaMóvil

Esto muestra dos cambios, que se consideran relevantes: Apple daba mucha importancia en su marketing a su sistema operativo[25] (“iOS”) y lo que podía hacer (“aplicación”), y el debate estaba dando entrada a su competidor, “Samsung”, con su sistema operativo (“Android”), diferente al de los dispositivos iPhone, que aparece destacado en los comentarios, pero incluso se ve en la nube de noticias. Debate que se desprende de la aparición de palabras como “mejor” y “bien” en los comentarios.

En ambas nubes se publica sobre las características de los dispositivos. Se encuentran palabras como “pantalla”, “sistema”, “terminal”, aunque aún no de manera tan destacada como en la siguiente revisión.

[25] Véase https://wiboomedia.com/marketing-app-ios/

2016

Figura 6 Nubes de palabras de artículos y comentarios (Apple, 2016)
Fuente: Elaboración propia a partir de datos de XatakaMóvil

Por último, en el año 2016, se produce un cambio notable: en los artículos aparecen destacadas palabras como "pantalla", "pulgada", "sistema", "procesador", mientras que en los comentarios están "pantalla", "precio", "tamaño", "cámara", "diseño". En estas nubes el marketing parece que se desplaza hacia las características técnicas, perdiendo relevancia aspectos exclusivos de nubes anteriores como la propia marca o el sistema operativo de la misma.

Sin embargo, las más destacadas en los artículos son "modelo" y "plus". Debido a que en dicho año, Apple lanza al mercado sus dos últimos modelos de iPhone, con un marketing en el que tiene mucha importancia su "pantalla" que, tanto en artículos como en comentarios, aparece bastante repetida.

Samsung sigue apareciendo, lo que lo confirmaría como el principal rival para Apple. Además, en los comentarios se habla de "Note", marca de Samsung que se explicará con mayor detalle en la nube de Samsung de 2016.

Al menos en los artículos, aparece la expresión "Smartphone". Pero en los comentarios no ocurre lo mismo. Los usuarios siguen utilizando "móvil" como palabra principal a la hora de referirse a los dispositivos.

Por otra parte, "iOS" y "Android" siguen apareciendo, aunque con menor frecuencia que el año analizado anteriormente. Esto indica que es un tema importante para publicar y comentar.

SAMSUNG

2007

Figura 7 Nube de palabras de artículos y comentarios (Samsung, 2007)
Fuente: Elaboración propia a partir de datos de XatakaMóvil

En el año 2007, se obtienen dos nubes en las que predomina “móvil”, y si se observan los diagramas de barras (Figura 16), se ve que esta palabra tiene una frecuencia muy elevada, en comparación con el resto de palabras.

En los artículos se tiende a tratar temas relacionados con las características técnicas de los dispositivos (“pantalla”, “cámara” y “megapíxel”), mientras que en los comentarios los usuarios tratan temas relacionados con marcas de móviles como “Nokia”, en aquel momento un competidor importante, y, otra vez, “Vodafone” como compañía telefónica. En aspectos técnicos, a los usuarios parece interesarles más “Internet”, “WiFi” y “GPS”.

Como hecho relevante para Samsung, figura el lanzamiento al mercado el modelo G800. Se trata del primer móvil del mundo en presumir de una resolución de 5 megapíxeles y zoom óptico 3x.[26] Lo que podría explicar que los textos de Samsung tengan referencias a aspectos técnicos.

[26] Véase http://blogs.elcorreo.com/elartilugio/2007/12/19/samsung-g800-erase-movil-una-camara-pegado/

2011

Figura 8 Nubes de palabras de artículos y comentarios (Samsung, 2011)
Fuente: Elaboración propia a partir de datos de XatakaMóvil

Este año ocurre algo parecido a lo que ocurría en Apple con la importancia del sistema operativo. En la nube de Apple del 2011 predominaba "iOS", y en esta "Android", tanto en artículos como comentarios.

Otra palabra que cabe destacar es "Nexus", el modelo de referencia para Samsung en ese año, y que podría relacionarse con "nuevo", "dispositivo" y "mercado".[27]

Tanto en artículos como comentarios, se siguen publicando palabras relacionadas con las características técnicas de los móviles, características que se diferencian de las nubes de Apple. Entre ellas están "megapíxel", "resolución", "memoria", "sistema", "precio", "pantalla", "pulgada", etc.

En ambas nubes también aparecen marcas de móviles y modelos distintos. Forman parte de ellas, "Apple", "Nokia", "Google" y modelos como "Nexus" de LG y "iPhone". La impresión es que la competencia percibida para Samsung es mayor que para Apple.

[27] Véase https://elandroidelibre.elespanol.com/2011/12/review-y-analisis-a-fondo-del-samsung-galaxynexus-pure-google-puro-android.html

2016

Figura 9 Nubes de palabras de artículos y comentarios (Samsung, 2016)
Fuente: Elaboración propia a partir de datos de XatakaMóvil

A la hora de analizar las nubes de palabras para este año, predomina un modelo de móviles entre todas las palabras, tanto para comentarios como artículos. Se trata del modelo “Note”. Lo relativo a esta palabra se podría descomponer en dos partes:

- El 19 de agosto de 2016 comienzan las ventas del Samsung Galaxy Note 7 en mercados determinados como Estados Unidos[28]. Ya desde su presentación, se sitúa como “el nuevo smartphone a batir en el mercado”[29].

- El 2 de septiembre de 2016. Samsung suspende las ventas del Galaxy Note 7 hasta nuevo aviso. Pide a todos los usuarios que devuelvan inmediatamente sus dispositivos a las tiendas debido a un defecto de fabricación y/o diseño de las baterías que causaba la explosión de las mismas[30]. Por ello, en la nube de palabras obtenida, a través de los comentarios que realizaron los usuarios en el blog, hay palabras con alta frecuencia como “problema”, “batería”, “caso”, “Note”, “modelo”, “calidad”, “fabricante”, “mAh”, etc.

En los artículos comienza a utilizarse la palabra “terminal” para referirse a los móviles, aunque en los comentarios sigue reinando “móvil”.

[28] Véase https://hipertextual.com/2016/10/samsung-galaxy-note-7-cronologia
[29] Véase https://hipertextual.com/2016/08/galaxy-note-7
[30] Véase https://hipertextual.com/2016/10/samsung-galaxy-note-7-cronologia

Se sigue dando el patrón de artículos referidos a características de los móviles y los comentarios ponen de manifiesto la existencia de más marcas que las que se mencionan en los artículos. Predominan también, en general, las palabras “pantalla” y “cámara”.

OTRAS

2007

Figura 10 Nubes de palabras de artículos y comentarios (Otras, 2007)
Fuente: Elaboración propia a partir de datos de XatakaMóvil

Como en los casos anteriores, la palabra que destaca es “móvil”. Sin embargo, en aspectos generales las nubes tienen más similitudes con Samsung que con Apple.

En los artículos, destacan las características técnicas de los dispositivos. Se encuentran palabras como “cámara”, “megapíxel”, “bluetooth”, “memoria”, “conectividad”, entre otras. Pero también figuran palabras que podrían estar relacionadas con el aspecto de los Smartphone, “modelo”, “diseño”, “color”, “teclado”, “pantalla”.

Sin embargo, en los comentarios, los usuarios aun posteando sobre las características de los móviles, mencionan ciertos aspectos que en los artículos es difícil de ver. Aquí se encuentran palabras como “barato”, “pobre”, “mejor”, “bien”, relacionadas con las opiniones que tienen los usuarios respecto a dichos dispositivos, y que es una característica claramente diferenciada de los mensajes en otras marcas: más “pragmatismo”.

Destaca la marca Nokia, en aquel momento la "otra marca" más potente del mercado. En ese momento salía a venta el Nokia N95. "*Probablemente, era el mejor móvil que podías adquirir en aquel momento, y desde luego el único con GPS integrado*".[31] Por ello puede ser que figure "GPS" en el grupo de las 30 palabras más repetidas.

2011

Figura 11 Nubes de palabras de artículos y comentarios (Otras, 2011)
Fuente: Elaboración propia a partir de datos de XatakaMóvil

En los artículos, a diferencia de los comentarios, se habla más sobre "Android", relacionado con palabras como "sistema" y "operativo", lo que confirmaría que este año es el de la relevancia de los sistemas operativos. En los comentarios, los usuarios se dirigen a los dispositivos móviles con el nombre de "móvil", "teléfono" y "terminal", mientras que en los artículos incluyen también la palabra "dispositivo".

Se publican artículos relacionados con marcas como son "Acer" y "Windows". En los comentarios, a diferencia, se habla sobre "Nokia", "HTC" y se repite "Windows".

A plena vista, en las nubes se ve que, en cuanto a las características, a los usuarios les interesa todo lo relacionado con la "pantalla", lo que en los artículos se repite con bastante frecuencia, y "precio".

Un hecho importante, en contraste con 2007, es la aparición de la palabra "China", que estaría relacionada con "fabricante" y "mercado".

[31] Véase https://www.xataka.com/espacionokia/18-telefonos-nokia-que-nos-muestran-la-alucinante-evolucion-de-la-telefonia-movil

2016

Figura 12 Nubes de palabras de artículos y comentarios (Otras, 2016)
Fuente: Elaboración propia a partir de datos de XatakaMóvil

En este año, se obtienen unas nubes muy diferentes entre artículos y comentarios, pues en ambas se observa una variedad de palabras relevantes en los artículos, mientras los comentarios se refieren a menos términos y más habituales.

Como se puede ver, la tecnología aplicada a los dispositivos móviles sigue avanzando. Por ello, se encuentran palabras que en las nubes anteriores no habían aparecido, como “GHz”.

Además, destacan los nombres de fabricantes chinos, como “Huawei” o “Xiaomi”, que presentan móviles bastante elegantes en su diseño y con especificaciones bastante sorprendentes si se tiene en cuenta su bajo “precio”, palabra bastante repetida por los usuarios en los comentarios del blog.

Respecto a esto, en los comentarios se postea sobre “marcas” producidas en “China” como son “Xiaomi” y “Redmi”, y dentro de ésta última se habla de uno de sus “modelos”, “Note”.[32]

Se siguen tratando los temas relacionados con “pantalla”, “cámara”, “megapíxel”, “pulgada”, aunque aparecen palabras que en las ocasiones anteriores no habían salido como son “mAh” (miliamperios/hora) relacionada con la palabra “batería”, “sensor”, “led”, “ram” (relacionada con “memoria”), etc.

[32] Véase https://www.xatakamovil.com/otras/xiaomi-redmi-note-4-diseno-potencia-y-autonomia-por-menos-de-120-euros

4.3.3 Aplicación de *text mining*: técnicas multivariantes

En esta tercera parte del análisis se usarán técnicas estadísticas específicas de variables cualitativas: medidas de asociación y análisis de correspondencias. Se aplicarán sobre dos variables, la primera recoge y combina las 30 palabras más frecuentes en cada combinación de marca, año y tipo de documento (directorio), mientras que la segunda indica a qué directorio pertenece un elemento.

Para poder llevar a cabo este análisis, previamente se ha realizado un paso intermedio en el cual se han transformado las matrices de palabras en tablas de contingencia para poder realizar contrastes de independencia, análisis de correspondencias y medidas de asociación. A continuación se procede a analizar las diferentes asociaciones que hay entre palabras y marcas, y, palabras y años, tanto para comentarios como para artículos. Para ello, se ha utilizado la V de Cramer.

La V de Cramer es una medida de asociación, que se utiliza para mediar la fuerza de la asociación entre una variable nominal con otra variable nominal, o con una variable ordinal. Ambas variables pueden tener más de dos categorías[33]

Se ha calculado dicha medida para el directorio (artículos y comentarios), únicamente para los artículos, y únicamente para los comentarios, y se obtuvo lo siguiente:

```
#V de Cramer directorio #V de Cramer artículos #V de Cramer comentarios
 CramerV(t.exercicio1)   CramerV(t.exercicio1)  CramerV(t.exercicio1)

  ## [1] 0.1454079      ## [1] 0.2104848        ## [1] 0.120911
```

Figura 13 V de Cramer para directorio, artículos y comentarios
Fuente: Elaboración propia a partir de datos de XatakaMóvil

El resultado más elevado se obtiene en los artículos. Para su interpretación, se sigue la siguiente tabla:

[33] Véase http://groups.chass.utoronto.ca/pol242/Labs/LM-3A/LM-3A_content.htm

LEVEL OF ASSOCIATION	Verbal Description	COMMENTS
0.00	No Relationship	Knowing the independent variable does not help in predicting the dependent variable.
.00 to .15	Very Weak	Not generally acceptable
.15 to .20	Weak	Minimally acceptable
.20 to .25	Moderate	Acceptable
.25 to .30	Moderately Strong	Desirable
.30 to .35	Strong	Very Desirable
.35 to .40	Very Strong	Extremely Desirable
.40 to .50	Worrisomely Strong	Either an extremely good relationship or the two variables are measuring the same concept
.50 to .99	Redundant	The two variables are probably measuring the same concept.
1.00	Perfect Relationship.	If we the know the independent variable, we can perfectly predict the dependent variable.

Tabla 1 Interpretación de V de Cramer
Fuente: POL242 LAB Manual: Exercise 3A
(http://groups.chass.utoronto.ca/pol242/Labs/LM-3A/LM-3A_content.htm)

Por lo tanto, el valor de la V de Cramer es 0.21 para artículos y, al estar comprendido entre 0.20 y 0.25, se entiende que la relación entre palabras y marcas y palabras y años es moderada. Dependen moderadamente unas variables con otras. Sin embargo, el valor obtenido en los comentarios es de 0.12, por lo que la relación entre las variables es muy débil. Incluso el valor de la V de Cramer en el directorio es baja, 0.14, por lo que la relación entre las variables es débil. Debido a estos resultados, no es aceptable el análisis con comentarios y con el directorio donde se combinan tanto artículos como comentarios. Por ello, se realizará únicamente el análisis a partir de los artículos publicados en el blog.

El análisis de correspondencias es una técnica estadística multivariante de reducción de dimensiones, aplicado a variables cualitativas. Su objetivo es coger una situación compleja con 2 o más variables, y multiples categorías, y reducirla a algo más interpretable, reorganizando sus asociaciónes y creando nuevas variables, llamadas dimensiones o ejes. Pero este proceso de reorganización también puede ser visto como una forma de estudiar tablas de contingencia complejas transformándolas en algo más simple (Greenacre, M., 2008), que se puede analizar gráficamente.

En esta fase del análisis la idea principal es explicar las dos dimensiones que figuran en los gráficos, obtenidas a partir de la información extraida de los artículos, y

aquellas categorías más relevantes, que aparecen apartadas del centro o tienen más contribución a la formación de las dimensiones estudiadas.

En un primer momento, se obtuvieron las **Figuras 17, 18** y **19**, pero se ha escogido la **Figura 15**, ya que es el conjunto de los anteriores pero filtrado, de manera que se han eliminado todas las palabras que tienen contribución menor a 1, así solo se presentan las palabras relevantes y se facilita el análisis.

Observando los autovalores de la F**igura 14**, y en particular el porcentaje acumulado de variabilidad explicada, Cumulative % of var., se ve que con 2 factores se explica el 68.067%, por lo tanto se considera que 2 factores son suficientes.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Variance	0.142	0.100	0.043	0.025	0.019
% of var.	39.950	28.117	12.256	7.194	5.329
Cumulative % of var.	39.950	68.067	80.322	87.516	92.845

Figura 14 Variabilidad explicada
Fuente: Elaboración propia en R a partir de datos de XatakaMóvil

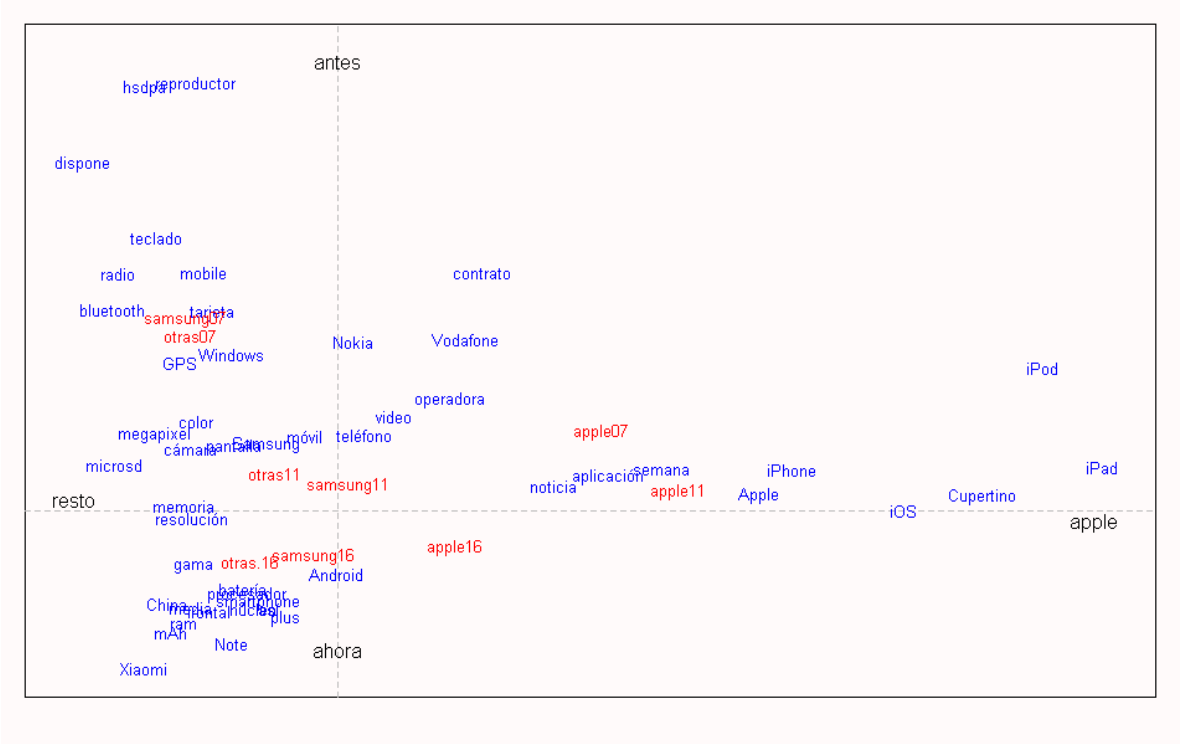

Figura 15 Gráfico de correspondencias
Fuente: Elaboración propia a partir de datos de XatakaMóvil

En la **Figura 15** se ven de forma más clara las disparidades entre marcas y palabras. En cuanto a su interpretación, se realiza por ejes.

Interpretación del primer eje:

1º) Las contribuciones de cada categoría a la formación del eje

Se observa que los artículos que más influyen son los de Apple en el año 2011 (60.65% de contribución) en el lado positivo (coordenada=0.879) y Otras en el año 2016 (13.33% de contribución) en el lado negativo (coordenada= - 0.224). Además, en la **Figura 18** se observa claramente la agrupacion de los 3 directorios de Apple en una parte del gráfico y el resto en la parte opuesta.

Por lo tanto, se sitúan en un extremo los artículos que tienen relación con Apple frente a los que hablan tanto de Apple como de Samsung, englobados en la categoría Otras y a los que se hace referencia en el gráfico como resto.

Respecto a las palabras, claramente se ve que “iPod”, “iPad”, “Cupertino”, “iOS”, “Apple”, “iPhone” y “aplicación” están asociadas con el lado positivo donde se encuentran los artículos de Apple. Mientras que en con el lado negativo, referido a Otras, están relacionadas palabras como “megapíxel”, “microsd”, “cámara” y “memoria”.

2º) Las contribuciones de cada eje a las categorías

Si se consideran los cosenos al cuadrado, que se refiere a lo que cada eje contribuye a cada categoría, se tienen los artículos de Apple del año 2011 y los referidos a Otras del 16, otra vez, mientras que este eje contribuye poco a los artículos sobre Samsung en el año 2011.

Sin embargo, para las palabras destacan “aplicación”, “Apple”, “memoria”, “resolución”, “semana” y “Cupertino”, como las que el eje de la Dimensión 1, referido a artículos de Apple por un lado y el resto por el otro, contribuye en mayor medida.

Interpretación del segundo eje:

1º) Las contribuciones de cada categoría a la formación de los ejes.

Se observa que los artículos que más influyen son los de Samsung del año 2007 (34.75% de contribución) en el lado positivo (coordenada= 0.832) y Otras en el año 2007, también, (32.58% de contribución) en el lado positivo igualmente (coordenada=0.761). Ambas categorías tienen similar contribución a la hora de formar los ejes.

Por lo tanto, se sitúan ambas en un extremo donde los artículos tienen relación con el antes y en el otro extremo estarían los que tienen relación el ahora, donde se encuentran los artículos de Otras en el año 2016 con una contribución de 18.40%. Esta asociación se ve tambien en la grafica 3, donde el año 2007 se desmarca claramente de la zona central y deja en el otro extremo el año 2016. Se observa también la cercania a este último del año 2011.

Respecto a las palabras, se obtiene que “reproductor”, “bluetooth”, “dispone”, “hdspa”, “teclado”, “mobile”, “tarjeta” y “móvil”, están relacionadas con los artículos de Samsung y Otras para el año 2007. Dichas palabras forman parte del lado positivo. Si ahora se mira el lado negativo, se aprecia que hay palabras que tienen una contribución relativamente alta. Éstas serían “ram”, “mAh”, “Smartphone”, “Xiaomi”, “procesador”, “Note”, “China”, “batería”, entre otras.

2º) Las contribuciones de cada eje a las categorías.

Si se consideran los cosenos al cuadrado, que es lo que cada eje contribuye a cada categoría, se tienen los artículos de Samsung de 2007 y a los referidos a Otras en el mismo año, mientras que este eje contribuye poco a los artículos sobre Apple en el año 2011.

Sin embargo, este eje de la Dimensión 2 contribuye a explicar en mayor medida a palabras como “tarjeta”, “reproductor”, “teclado”, “hdspa”, “Smartphone” y “dispone”, entre otras.

Por todo esto, se puede concluir con que hay dos distinciones:

En cuanto a las marcas, se observa que se dividen las relaciones por una parte en Apple y por la otra, en Samsung y Otras. Debido a esto, en el gráfico se observa que las palabras a la derecha forman parte de “el mundo Apple”, lo que parece mostrar lo distintivo que es su valor de marca. Sin embargo, del otro lado se observa que los intereses en las marcas agrupadas en Otras y Samsung son claramente técnicos. En el lado izquierdo del gráfico predominan las propias características técnicas de los móviles. Además, la marca Samsung es bastante neutral ya que sus artículos de los años 2011 y 2016 tienden al centro, por lo que no están muy relacionados con las demás variables.

Un detalle importante puede observarse en la **Figura 18,** en ella la posición de Apple en 2016 está muy cercana al centro, y separada de Apple 2007 y 2011. Esta posición coincide con lo observado en las frecuencias, que daba a Apple en 2016 un cambio de estrategia, con características más técnicas que en los años anteriores.

En cuanto al espacio temporal, su aspecto predominante es la evolución técnica. En el año 2007 se tratan temas relacionados con las características técnicas básicas de un móvil como son el reproductor, el teclado, la radio, bluetooth… Este tipo de características eran muy utilizadas en aquel periodo de tiempo. Hoy en día se deja de lado la radio y el reproductor y se cambian por aplicaciones de música como por ejemplo Spotify. El teclado se queda apartado desde que aparecen los móviles táctiles, y el bluetooth también se aparta para dejar paso a aplicaciones, como WhatsApp, a través de las cuales los usuarios pueden compartir información y archivos.

En el otro extremo, el año 2016, trataría de reunir palabras como “Note”, “mAh” y “batería”, que explicarían el comportamiento de Samsung durante este año y el problema que había surgido con su modelo Samsung Galaxy Note 7 debido a un defecto de fabricación y diseño de las baterías. Sin embargo en 2016 también figuran

palabras como “China” y “Xiaomi”, donde se refleja la relevancia que están adquiriendo los móviles asiáticos, como ocurre en las nubes de palabras de la categoría Otras.

Ambas partes aparecen muy agrupadas en el gráfico, esto da a entender que están bastante relacionadas entre sí. Aparecen agrupadas porque las relaciona el año 2016.

Conclusiones

La creciente utilización de las Tecnologías de la Información y de la Comunicación (TIC), y dentro de ellas el crecimiento de Internet, ha dado lugar a la aparición de nuevas y más amplias fuentes de datos; y por lo tanto, a nuevas necesidades en su tratamiento. Los datos pasan a ser una materia prima abundante, que se necesita explotar para obtener un conocimiento, el cual será valioso en la ayuda de toma de decisiones sobre el ámbito en el que se han extraido los datos.

Para poder llegar a dicho conocimiento, ha sido necesario desarrollar diferentes metodologías y entre las más importantes está las que se agrupan bajo el concepto de *Data Mining*, proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos, Es en este punto donde se demuestra que la Estadística ayuda a llegar a ese conocimiento final, ya que provee herramientas y técnicas para poder explicar las distintas conclusiones que se obtienen cuando se realiza un análisis.

Concretamente, dentro de este ámbito de la Estadística y la minería de datos, han surgido nuevas técnicas que sirven como soporte y como medio de análisis de la información. Entre ellas, se destaca la minería de textos, también conocida como *Text Mining*, la cual permite extraer y presentar la información a partir de textos y que ha sido la técnica aplicada en este trabajo.

Dentro del marco de la minería de textos, se ha mostrado cómo se puede manejar de forma sencilla la información procedente de Internet. Concretamente, los datos analizados se han extraído de un blog. Este tipo de comunidad virtual, como son también las redes sociales y foros, es muy utilizada en las estrategias de organismos, gobiernos o empresas que quieren llegar al público. En ellas se enfocan diversos temas y se comparte información común.

En este caso, el blog utilizado, XatakaMóvil, ha sido la base de la que se han extraído los datos. Se ha podido comprobar que hay dos partes: una se trata de la persona o entidad encargada de gestionar y publicar los artículos en el blog, y la otra se trata de los usuarios que han posteado, de manera que dejan sus comentarios y opiniones relativos al tema en cuestión. A este respecto se observa una primera conclusión: la diferencia entre textos de artículos y comentarios. En ellos, se distingue

el diferente nivel de lenguaje utilizado, observándose que los usuarios usan más la palabra “móvil” en los comentarios, frente a “terminal” y “dispositvo” como palabras utilizadas en los artículos de forma más técnica. También se observan diferentes intereses: la imagen de las empresas en las noticias (“imagen”, “noticia”, “lanzamiento”, “ventas”, “mercado”) e intereses específicos de usuarios (“precio”, “pantalla”, “modelo”, “seguro”, “compañía”).

Dentro de la industria de los dispositivos móviles, existen dos grandes marcas fácilmente reconocibles: Apple y Samsung. Cada una de ellas ha llegado a su posición actual con una estrategia distinta y claramente marcada para su enfoque en el mercado, y a pesar de ello, a ambas le ha funcionado particularmente bien. Además, se debe tener en cuenta que la relación entre las compañías es especial, ya que al mismo tiempo que son competidoras, también son socias y tienen acuerdos muy importantes en lo relacionado al suministro de componentes.

Al escoger únicamente tres años, se ha podido realizar un análisis mucho más concreto de la evolución del mercado de los dispositivos móviles y de dichas marcas.

Una de las principales conclusiones es que los resultados obtenidos de Apple son muy distintos a los obtenidos para Samsung y la categoría Otras. Esto enmarca a la compañía de Cupertino como una marca muy distinguida frente al resto. Comercializa muy pocos productos y a precios elevados. Sin embargo, sus productos tienen ese “algo” que, para muchos consumidores ”es Apple”, y que otros dispositivos no tienen. Dentro de esta categoría se han encontrado palabras referidas al “mundo Apple” como son “iPad”, “iPhone”, “Mac”, entre otras.

Por el otro lado, se encuentran Samsung y Otras, las cuales tienen un mayor número de similitudes entre sí. Cabe destacar, sobre todo, que ambas comparten el mismo sistema operativo, pero también comparten una imagen basada en características técnicas: en los artículos se observan como destacan característica técnicas, desde las más genéricas, como “procesador”, “batería”, “reproductor”, “cámara” o “pantalla”, a otras más claramente técnicas como “bluetooth”, “microsd” o “mAh”, incluso palabras referidas al aspecto como “gama”, “color”, “diseño” o “teclado”.

Sumado a esto, se obtuvo un detalle importante en el que la posición de Apple en el año 2016 cambia bastante en relación a los otros años. Dicha posición coincide con lo observado en las frecuencias, donde se ve que Apple realiza un cambio de estrategia enfocando su imagen en características más técnicas que en los años anteriores.

Se obtuvo tambien una imagen de la evolución temporal, representada por la evolución de las características técnicas de los móviles, pero tambien por la evolución de las marcas chinas en el mercado. Ya a partir del 2011 comienzan a aparecer en las nubes palabras como “China” y “coreano”, y debido a la crisis económica, estas marcas han sabido ver la puerta grande de su crecimiento en España. Siguiendo esta línea, en el año 2016 se encuentran marcas como “Huawei”, “Xiaomi”, “Meizu” y “Redmi”. Esto indica que la telefonía china ha entrado en España con fuerza, aunque ya en años anteriores se vendieran móviles chinos. El año 2016, por lo tanto, está marcado en mayor medida por dichas marcas ya que entre varias cosas, prometen móviles con una buena relación calidad-precio.

En definitiva, este tipo de detalles procedentes de los artículos y comentarios de los usuarios, son ingredientes que tanto empresas como organizaciones podrían desear conocer para que así, se puedan satisfacer y cubrir las necesidades de dicho mercado de forma más eficiente. Por ello, cabe acentuar la importancia que pueden aportar los datos disponibles en Internet y el uso que de ellos se puede realizar.

Una conclusión genérica, que merece la pena citar, es lo que se ha conseguido con la propia realización del trabajo mostrando como se puede, con los conocimientos obtenidos en el grado de Economia y un poco de preparación, entrar en el mundo del *Text Mining,* lo que nos abre puertas y nuevas posibilidades tanto en el mundo de la investigación como en el profesional.

Finalmente, concluir con que a pesar de que la tecnología ha facilitado, de forma excepcional, la vida de muchas personas, organizaciones y empresas, no hay que quitarle importancia a la protección de datos, ya que dicha protección y la privacidad de datos en Internet están en manos de los consumidores, y más hoy en día que la información publicada en la Web ha aumentado enormemente y seguirá creciendo. Se debe tener especial cuidado con ello, ya que, por ejemplo, las empresas

utilizan la información que los usuarios suben a la red para generar estructuras de poder y la información privada ya no lo es tanto. En este sentido, ¿se puede estar seguro de que nuestros datos personales están suficientemente protegidos?

Bibliografía

Aggarwal, Charu C & Zhai, C. (2012). Mining text data. Recuperado de http://charuaggarwal.net/text-content.pdf

Aluja, T. (2001). La Minería de Datos, entre la Estadística y la Inteligencia Artificial. Universitat Politècnica de Catalunya. Recuperado de https://www.idescat.cat/sort/questiio/questiiopdf/25.3.4.Aluja.pdf

Bolasco, S. Et al. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi. (Tésis Università degli Studi di Roma .La Sapienza) .Recuperado de http://www.labstat.it/home/wpcontent/uploads/2015/03/BOLASCO.pdf

Carrasco Arroyo, S. (2005). Una aproximación a la estadística desde las Ciencias Sociales. Recuperado de http://www.uv.es/~carrascs/PDF/aproximacion%20estadistica.pdf

De la Calle Velasco, G. (2014). Modelo basado en técnicas de procesamiento de lenguaje natural para extraer y anotar información de publcaciones científicas. (Tésis Doctoral, Universidad Politécnica de Madrid). Recuperado de http://oa.upm.es/30856/1/GUILLERMO_DE_LA_CALLE_VELASCO.pdf

Esteban Talaya, A y Molina Collado, A. (2014) Investigación de mercados. Madrid: ESIC Editorial.

Etzioni, O. (1996) "The World Wide Web: quagmire or gold mine?" Communications of the ACM, 39(11).

Feldman, R. & Sanger, F. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

Glez Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M. & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. Recuperado de https://academic.oup.com/bib/article-pdf/15/5/788/17488715/bbt026.pdf

Greenacre, M. (2008). La práctica del análisis de correspondencias. Bilbao: Fundación BBVA.

Hair, J.F., Black, W.C., Babin, B.J. & Anderson, R.E. (2010). Multivariate Data Analysis. (7th Edition). Prentice Hall, Upper Saddle River, New Jersey.

Hernández Orallo, J., Ramírez Quintana, M.J., Ferri Ramírez, C. (2004). Introducción a la Minería de Datos. Pearson Education, S.A. Madrid.

Instituto Nacional de Estadística (INE). (s. f.). *Historia de la Estadística*. Recuperado de http://www.ine.es/explica/docs/historia_estadistica.pdf

Kosala, R.; Blockeel, H. (2000). "Web Mining Research: ASurvey", ACM SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, vol. 2, pp: 1-15.

Lara Torralbo, J. A. (2016). Business intelligence. Madrid: CEF.

Maceli, M. (2016). Introduction to Text Mining with R for Information Professionals. *Code 4 lib journal*. Recuperado de http://journal.code4lib.org/articles/11626

Martín-Pliego, Fco. J. (2007). Introducción a la Estadística Económica y Empresarial: Teoría y Práctica. (3ª Edición). Madrid: AC.

Merino Sanz, M. J. (2015). Introducción a la investigación de mercados. (2ª Edición). Madrid: ESIC Editorial.

Miner, G. et al. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Elsevier Inc.

Munzert, S., Rubba, C., Meibner, P., Nyuis, D. (2015). Automated Data Collection with R: A practical Guide to Web Scraping and Text Mining. United Kingdom: John Wiley & Sons. Ltd.

Paradis, E. (2003). R para principiantes. Recuperado de https://cran.r-project.org/doc/contrib/rdebuts_es.pdf

Pavone, P. (2015). Un Método de Text Mining para la categorización Fuzzy de documentos. (Tésis doctoral, Universidad de Málaga). Recuperado de http://riuma.uma.es/xmlui/handle/10630/9756

Reyes Saldaña, J. F. y García Flores, R. (2005). El proceso de descubrimiento de conocimiento en bases de datos. Recuperado de http://eprints.uanl.mx/10160/1/26_el_proceso.pdf

Sakaji Kido, G. (2016). Extração de tópicos com modularidade no Twitter. (Proyecto Final de Curso, Universidade Estatudual de Londrina). Recuperado de http://www.uel.br/cce/dc/wp-content/uploads/TCC-GUILHERME_SAKAJI_KIDO-BCC-UEL-2015.pdf

Witten & Frank Clark, P.; Boswell, R. (2000). Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers.

Anexo

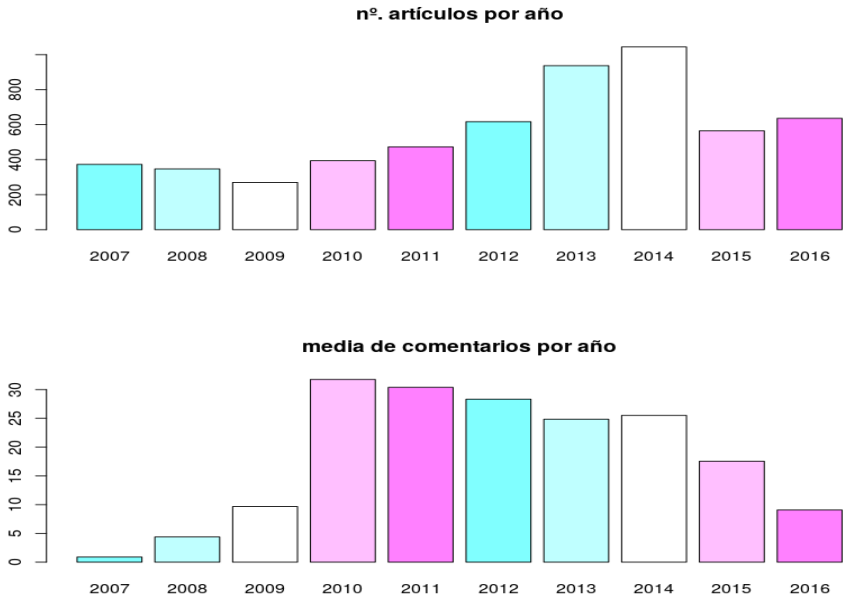

Figura 16 Nº de artículos y media de comentarios por año
Fuente: Elaboración propia a partir de datos de XatakaMóvil.

```
table(datos.artigos$anoartigo)

##
## 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
##  373  347  270  394  473  617  937 1044  565  636
```

Tabla 2 Nº de artículos por año
Fuente: Elaboración propia a partir de los datos de XatakaMóvil

```
table(datos.artigos$marca)

##
##   apple   otras samsung
##    1567    1921    2168
```

Tabla 3 Nº de artículos por marca
Fuente: Elaboración propia a partir de datos de XatakaMóvil

```
table(datos.artigos$anoartigo,datos.artigos$marca)

##
##        apple otras samsung
##   2007    89   157     127
##   2008   193    75      79
##   2009   124    56      90
##   2010   183    66     145
##   2011   216    51     206
##   2012   227   152     238
##   2013   210   294     433
##   2014   178   396     470
##   2015    65   305     195
##   2016    82   369     185
```

Tabla 4 Nº de artículos por marca y año
Fuente: Elaboración propia a partir de datos de XatakaMóvil.

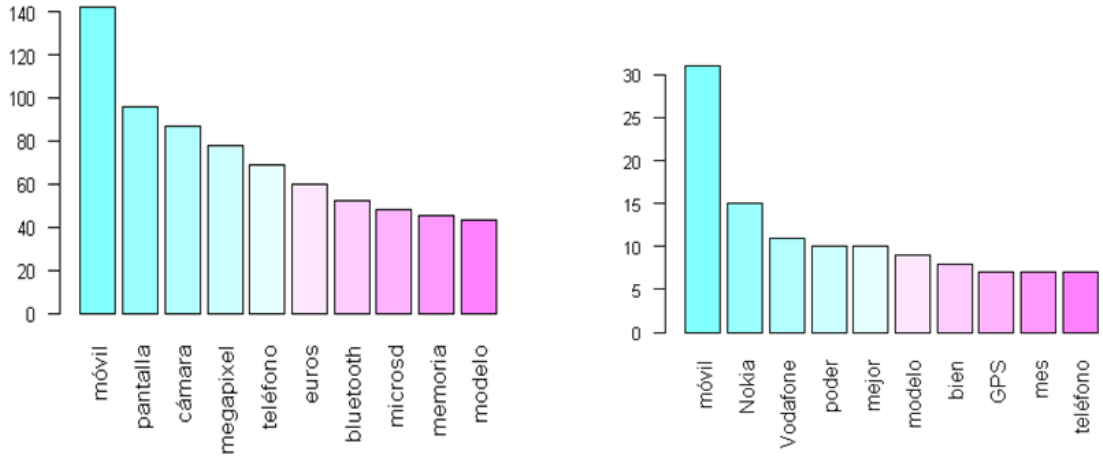

Figura 17 Diagrama de barras de artículos y comentarios, (Samsung, 2007)
Fuente: Elaboración propia a partir de XatakaMóvil

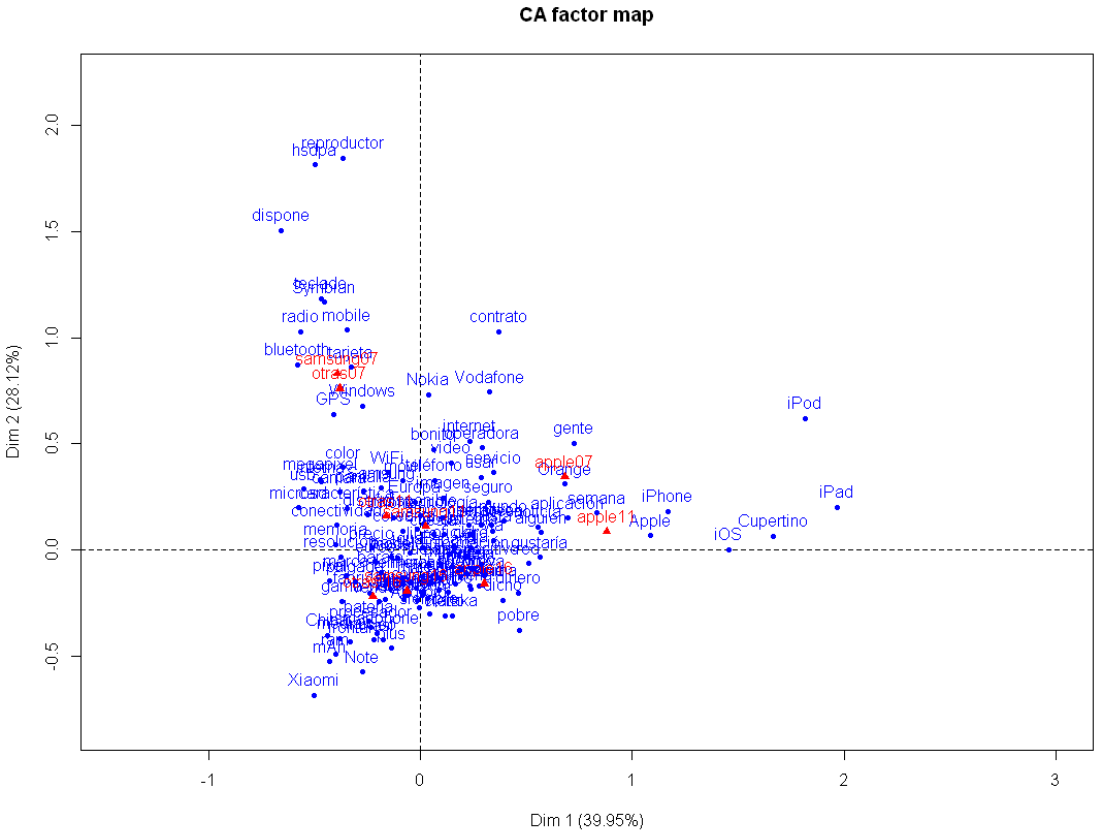

Figura 18Análisis de correspondecia entre palabras y marcas por año
Fuente: Elaboración propia a partir de datos de XatakaMóvil

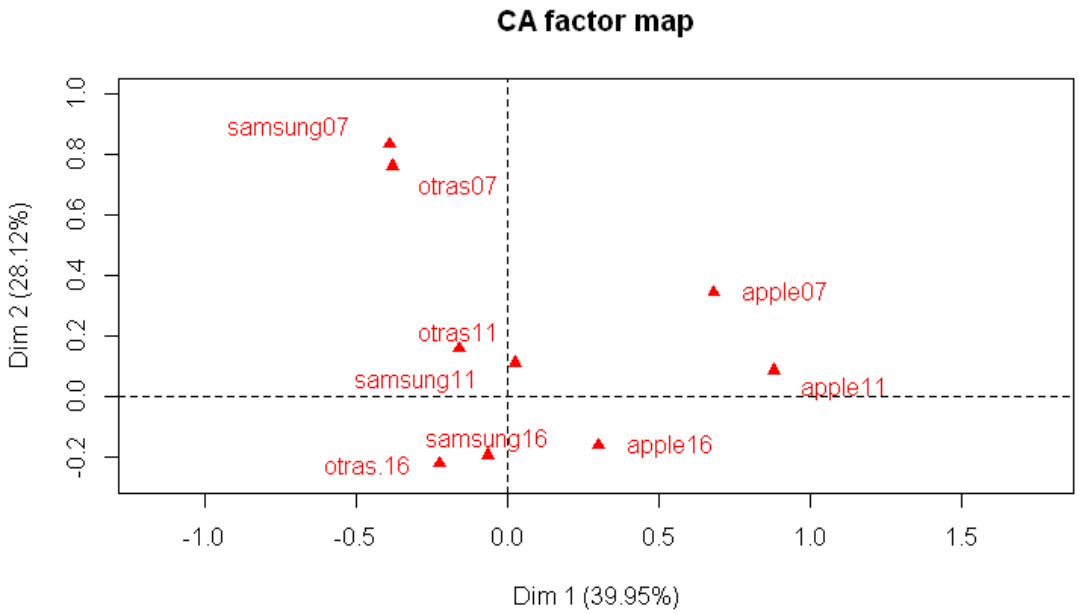

Figura 19 Análisis de correspondencia de marcas por año
Fuente: Elaboración propia a partir de datos de XatakaMóvil

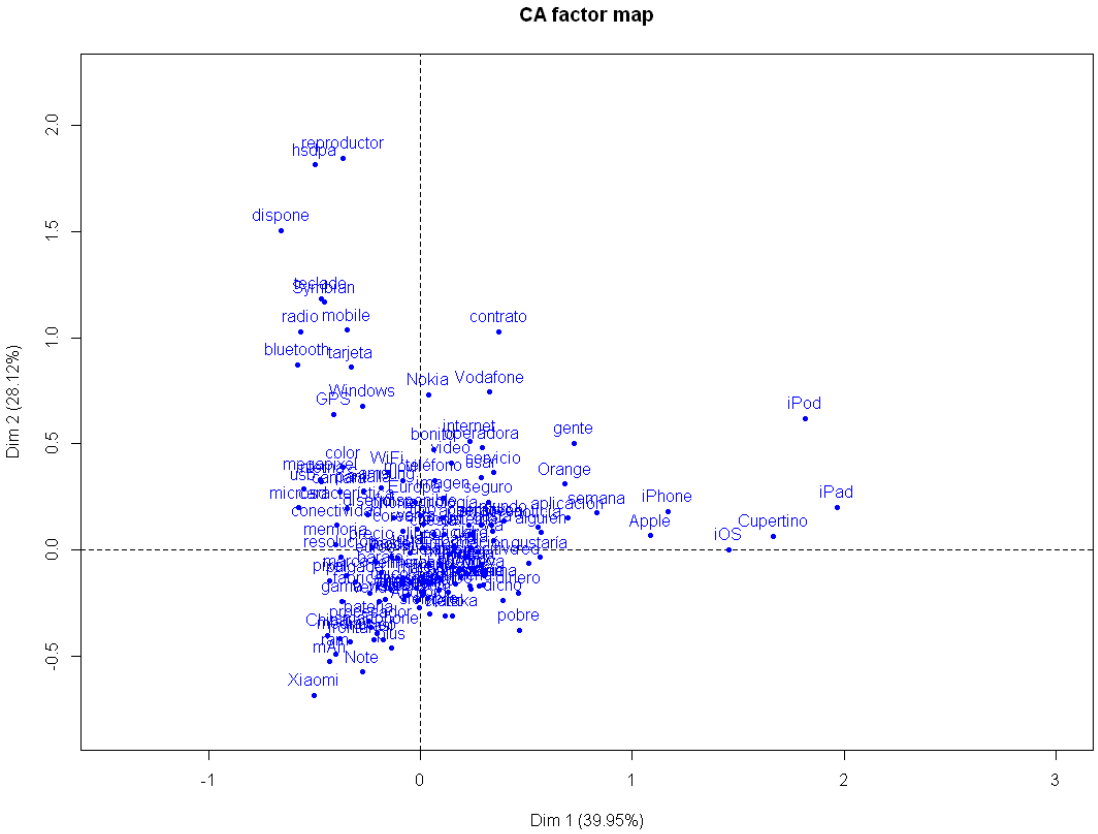

Figura 20 Análisis de correspondencia de palabras
Fuente: Elaboración propia a partir de datos de XatakaMóvil