UNIVERSIDADE DA CORUÑA

Facultad de Economía y Empresa

Trabajo de Fin de Grado

Técnicas Estadísticas aplicadas a Redes Sociales y Análisis de Sentimiento: el caso Volkswagen

David Abad Fernández

Tutor: Prof. Dr. Xosé Manuel Martínez Filgueira

Grado en Economía

Año 2016

Trabajo de Fin de Grado presentado en la Facultad de Economía y Empresa de la Universidad de A Coruña para la obtención del Grado en Economía

Resumen

La estadística como las matemáticas disponen de un sinfin de aplicaciones cuya utilización como herramienta de análisis económico resulta muy válida. La aplicación a la que este trabajo hace referancia no es otro que a la investigación de mercados. Son muchas las técnicas que se utilizan en este campo abarcando desde las más sencillas como realizar medias o tablas de frecuencias hasta las más complejas como clústers, análisis de la varianza o regresiones. Internet es desde hace años un gran almacén de datos de muy diverso tipo, que actualmente se encuentra en constante crecimiento. Lo mismo ocurre con las redes sociales. Desde su creación, no han hecho más que expandirse exponencialmente cada año, recogiendo en sus bases de datos multitud de información referente a los usuarios que las utilizan. Es por ello que ya muchas empresas y organizaciones se dedican cada vez a más a elaborar memorias sobre la información que se encuentra en internet y las redes sociales, con el objetivo de obtener datos susceptibles para su análisis y procesamiento.

Palabras clave: Estadística, Investigación de Mercados, Redes Sociales, Internet, Twitter, Análisis de Sentimiento.

Número de palabras: 14.698

Resumo

A estatística coma as matemáticas dispoñen dun sinfin de aplicacións cuxa utilización como ferramenta de análisis económico resulta moi válida. A aplicación á que este traballo refírese non é outra que á investigación de mercados. Son moitas as técnicas que se empregan neste campo comprendendo dende as máis sinxelas como realizar medias ou táboas de frecuencias ata as máis complexas coma clústers, análise da varianza o regresións. Internet é dende fai anos un gran almacén de datos de moi diverso tipo, que actualmente atópase en constante crecemento. O mesmo ocorre coas redes sociais. Dende a súa creación, non fixeran máis que expandirse expoñencialmente cada ano, recollendo nas súas bases de datos multitude de información referente ós usuarios que as utilizan. É por isto que xa moitas empresas e organizacións adícanse cada vez máis a elaborar memorias sobre a información que se atopa no internet e as redes sociais, co obxectivo de obter datos susceptibles para a súa análise e procesamento.

Palabras chave: Estatística, Investigación de Mercados, Redes Sociais, Internet, Twitter, Análise de Sentemento.

Número de palabras: 14.698

Abstract

Statistics, as mathematics have lots of aplications which use as an economic analyze tool results very valious. The aplication that this project is refering to is not other than the marketing research. There are amounts of techniques that are used on this area, from the easiest, like elaborating medians or frecuency tables to the most complex ones, as clusters, variance analysis or regresions. Internet is, since a couple of years ago a great database of different types of information, and nowadays is constantly growing. The same conclusion is aplicated to the Social Networks. Since their creation, It's been growing exponencially every year , taking into their databases lots of information of the users of this tool. This reason is the answer to the question of why lots of bussines and organisations are keen to develope and elaborate memories about the information that we look for on the internet and the social networks, with the objective of obtain suceptible data for their analysis and procedure.

Keywords: Statistics, Marketing Research, Social Media, Internet, Twitter, Sentiment Analysis.

Number of words: 14.698

Índice

Índice de figuras

Índice de tablas

Introducción

El tema escogido de este trabajo, nace a raíz de la importancia que tiene la información y sus múltiples vías de aprovechamiento. Ésta se encuentra disponible en multitud de formas y tamaños, siendo crucial para su procesamiento una serie de técnicas y un marco teórico apto que permita hacer un uso correcto de la misma, para un ulterior análisis y conclusiones. Su aplicación a diferentes aspectos de la economía es absolutamente crucial; empresas, estados y organizaciones deben ser capaces de manejarla, ya que la información siempre es poder. La estadística es una herramienta que permite realizar, de múltiples formas, análisis de datos e información, y su uso está absolutamente implantando tanto en la comunidad científica como en la empresarial. Es por ello que en este trabajo se expondrá su utilidad en la investigación de mercados. Se mostrarán las diversas técnicas estadísticas que emplea, concluyendo con un caso práctico en el que se analizará la información procedente de consumidores potenciales y su posible impacto sobre los resultados de una empresa.

Las técnicas para la obtención de información susceptible de análisis son muy diversas. Las opiniones y sentimientos de los consumidores son muy importantes para las empresas, ya que son unos excelentes indicadores a tener en cuenta para el posterior desarrollo de productos y campañas de publicidad. Las redes sociales cada día adquieren más usuarios y la cantidad de información personal que se encuentra en estos portales es de unas proporciones gigantescas, siendo su utilización, un medio de obtención de información para empresas y gobiernos que a día de hoy en alza. Es por ello

que se ha usado información procedente de estas redes para un caso práctico en que mostrar los usos de la estadística en referencia a la investigación de mercados. Concretamente se estudiará la repercusión que ha tenido en las redes sociales la última crisis de Volkswagen relativa al trucaje ilegal de sus motores, como "proxy" a una investigación de mercados, y en qué medida los datos obtenidos reflejan una posible relación de lo que realmente ha pasado y lo que se ha dicho en esta red social donde se realizará un estudio estadístico de las variables utilizadas en este análisis. El trabajo se inicia describiendo conceptos básicos de la Estadística (capítulo 1), y la investigación de mercados (capítulo 2) y su relación (capítulo 3). En el siguiente capítulo se describen las redes sociales y se explica su funcionamiento, acabando en el capítulo 5 con el caso aplicado, usando técnicas estadísticas e información de redes sociales para estudiar el efecto que el caso Volkswagen, pudo tener sobre su mercado.

1. La estadística y sus aplicaciones

1.1. Antecedentes

Son muchas y variadas las utilidades que las matemáticas ofrecen al hombre, abarcando un sinfín de campos que van desde las ciencias más complejas hasta aspectos simples como escoger los elementos de la cesta de la compra de un hogar tipo. A través de este amplio medio de conocimiento, se han definido otras especialidades que, a partir de las opciones que ofrecen como herramienta de trabajo, definen y ejecutan su campo de actuación. Una de ellas es la Estadística.

Las técnicas estadísticas se usan en muchos aspectos de la vida. Las encuestas realizadas en periodos electorales para obtener información previa a las elecciones y así intentar predecir el resultado de las mismas, el médico o el veterinario que investigan sobre una determinada enfermedad y sus posibles medicamentos en una población objetivo, el ingeniero o el arquitecto a la hora de elaborar sus informes de controles de calidad del producto diseñado o el economista a través de diversos índices e indicadores tales como el PIB, PNB, IDH, renta per cápita entre otros, usando la información que proporcionan para analizar la coyuntura económica.

Existen muchas definiciones de “Estadística”, Martín-Pliego (2007, p. 5) la define como “*La tecnología del método científico que proporciona para la toma de decisiones cuando éstas se adoptan en ambiente de incertidumbre, siempre que pueda ser medida en términos de probabilidad”*

La filología de esta palabra procede del latín y el italiano, concretamente de los términos *“statista”* (político, hombre de estado) y *“statisticum collegium”* (consejo de estado), se define como la exteriorización cuantitativa de las cosas de estado. De esta manera es posible realizar un análisis en retrospectiva, haciendo referencia a etapas tempranas de la historia de la humanidad, donde surgía ya la necesidad de hacer inventario de individuos y objetos, tales como: animales, personas, prendas o armas. En uno de los libros del Antiguo Testamento, bajo el nombre de "Números", se describe el censo hecho por Moisés después de la salida de Egipto: *“…hagan un censo de toda la comunidad de Israel por clanes y por familias patriarcales, anotando uno por uno los nombres de todos los varones”*[1].

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

En A Coruña, desde el año 1246, el Archivo de la Colegiata, guarda en sus muros, documentación de muy diverso carácter, siendo ésta económico (o "de fábrica"), censal y administrativo (Velo, 2009; p. 27). Se organiza en la siguiente clasificación:

- Pergaminos
- Planos y dibujos
- Documentación General
- Música y partituras

La utilidad de estos datos versaba fundamentalmente en obtener una visión mucho más amplia de los asuntos seculares, en tanto a su propia administración como a su economía, además de un estudio sociológico a través de los primitivos censos que en aquella época se realizaban.

Las aplicaciones estadísticas son muchas y muy diversas, comprendiendo desde aspectos cotidianos hasta los más complejos, además de abarcar una perspectiva histórica amplísima, que se ha desarrollado hasta nuestros días.

1.2. Conceptos clave

El término principal en el que se centra el estudio de esta materia son las "variables". Pueden agruparse en dos grupos, en tanto a su naturaleza. De este modo se distinguen:

- Variables cualitativas: Este grupo hace referencia a aquellas que, debido sus propias características, no pueden ser medidas a través de los números. Dentro de las mismas hay dos grupos:
    - Variables cualitativas nominales: Son aquellas que no permiten prelación de ninguna clase. Los colores son un ejemplo: (rojo, verde, azul, amarillo)
    - Variables cualitativas ordinales: Son aquellas que si permiten algún tipo de prelación. Las alturas de una población (alto, medio, bajo), o graduaciones militares (raso, cabo, sargento, teniente…)

- Variables cuantitativas: Son todas aquellas que pueden expresarse en términos numéricos, y por lo tanto es posible aplicarles todas las operaciones matemáticas que sean pertinentes respecto a su estudio. De esta forma se distinguen dos tipos:
    - Variable contínua: Se define como aquella que puede tomar cualquier valor comprendido dentro de un intervalo. Un ejemplo es la “altura” ya que entre dos valores puede haber algunos intermedios a los que se denominan “errores de medición”

    - Variable discreta: Dentro de un intervalo, solamente pueden tomar un espectro finito de valores. Así pues el número de hijos de una familia, o el tamaño de la plantilla de una empresa se encuentra dentro de su definición.

1.3. Clasificación de los datos

La forma más comúnmente utilizada para organizar la información existente es mediante escalas de medidas o escalas métricas. Según el concepto medido, las escalas confieren distintos tipos de intensidad. De tal forma, la ordenación se establece en cuatros grupos. La aplicación de dichas escalas se realizará en función del análisis estadístico que requiera el estudio en cuestión (Trespalacios, Bello & Vázquez, 2005).

- Escala nominal: Se utiliza para clasificar los conceptos medidos en categorías no numéricas mutuamente excluyentes, ya que una unidad o característica estudiada, solamente puede pertenecer a una de estas categorías. Es la escala más simple, y en ella se encuentran las variables cualitativas, ya que ésta nos indica si el individuo (persona, empresa…) se sitúa dentro de una determinada clase.

- Escala ordinal: En ella, las unidades que la componen, establecen un orden de prelación, en función de una determinada característica. En comparación con la escala nominal, ésta escoge dentro de un mismo grupo de variables; aquella o aquellas que más se ajustan a sus propias particularidades. El caso de las preferencias de un consumidor

dentro de un amplio espectro de productos es una realidad donde dicha escala se aplica.

- Escala de intervalos: Es posible apreciar la diferencia entre dos unidades en tanto a su magnitud. Los datos de carácter económico son un ejemplo; por tanto, salarios, precios o pensiones se englobarían dentro de este nivel categórico. Otro ejemplo sería la Escala de Likert de actitudes no comparativas[2].

- Escala de proporción o de razón: Es en este punto donde la presencia del "cero absoluto" es determinante para dar explicación a las unidades que se engloban en esta escala. De esta forma al hacer referencia en un extremo, ya que es una escala acotada inferiormente, es posible hacer un análisis en términos comparativos. La variable objeto de estudio en este caso es cuantitativa, como por ejemplo las edades y los pesos de una población o las ventas de una empresa.

1.4. Estadística y Economía

En la ciencia económica, es de obligado cumplimiento tratar con una inmensidad de datos, variables y números, fruto de las propias labores que este campo inciden. Para ello es necesario utilizar las herramientas que las matemáticas y en particular la estadística, en particular ofrecen.

Ante una sociedad cada vez más globalizada y por lo tanto cambiante, donde existe una cantidad astronómica de información, es menester la aplicación de un método de estudio que permita al científico económico, estudiar todas las interrelaciones que actualmente acontecen, pues; "*ciertas respuestas pertenecientes al pasado se entremezclan con posturas ya impregnadas de hábitos de futuro, siendo en definitiva, imposible establecer normas fijas de comportamiento y leyes inmutables que regulen las relaciones económicas" (Martín-Pliego, 2007; p. 9).*

Los indicadores económicos son una muestra de cómo los métodos estadísticos se aplican a la economía, en este caso, al espectro

[2] La escala de Likert es una escala psicométrica, utilizada en las ciencias sociales para cuestionarios de investigación, donde se expresa, mediante los cuestionarios Likert, el nivel de acuerdo o desacuerdo con determinadas afirmaciones.

macroeconómico de un estado. El Producto Interior Bruto (PIB), la tasa de paro o la renta per cápita se elaboran siguiendo estos patrones. Para su obtención se cuenta con organismos nacionales y supranacionales (INE en España y Statistisches Bundesamt en Alemania, ámbos son claros ejemplos de entidades cuyo propósito es la recopilación de información y su posterior tratamiento para la elaboración de los indicadores económicos además de informes de carácter público y gubernamental). Con este procesamiento de la información, las empresas y los países pueden extraer datos muy relevantes que revelan su comportamiento interno, y cómo éste es capaz de afectar a sus factores de producción, de tal manera que se hace posible un diagnóstico de todos aquellos parámetros relevantes, facilitando así, un método para detectar errores e ineficiencias y su consiguiente subsanación.

Esta utilidad en el ámbito más puramente económico se extiende también al mundo empresarial del cual solo mencionaremos un campo de aplicación pero con más extensión puesto que en él se inserta el ejemplo aplicado a través del cual se realiza este trabajo.

2. Investigación de Mercados

2.1. Definición

Un estudio de mercado es un proceso que consiste en la recogida, análisis e interpretación de una información relativa a los integrantes de un mercado objeto de estudio. La American Marketing Association (AMA) lo como:"*La investigación de mercados es la función que vincula al consumidor, cliente y público al vendedor a través de información - información utilizada para identificar y definir las oportunidades y los problemas de comercialización; generar, refinar y evaluar las acciones de marketing; monitorear el desempeño de la comercialización; y mejorar la comprensión de la comercialización como un proceso. La investigación de mercados especifica la información necesaria para abordar estas cuestiones, diseña el método de recogida de información, administra e implementa el proceso de recogida de datos, análisis de los resultados , y comunica los resultados y sus implicaciones"* . (Aprobado en octubre de 2004)[3]

A través de esta definición, es posible extrapolar las principales características de la investigación de mercados (Molina, 2014):

[3] Véase https://www.ama.org/AboutAMA/Pages/Definition-of-Marketing.aspx

- **Sistemático**: Emplea el método científico para llevar a cabo su propósito, bien organizado y esquematizado, empleando unos sistemas de control rigurosos.

- **Objetivo:** Es de personalidad imparcial, evitando cualquier familiariedad, obteniendo así mayor homogeneidad y unicidad de los resultados.

- **Informativo:** Su principal finalidad es la de proporcionar información que sea válida para la toma de decisiones.

- **Orientado a la toma de decisiones:** El propósito del estudio, es siempre el de reducir la incertidumbre ante un problema de decisión, de esta manera se minimizan los riesgos que implican.

- **Relevante:** Trata de obtener toda la información que resulte de importancia y no incluir información redundante en el proceso.

- **Fiabilidad:** Es un requisito muy importante que la información obtenida refleja la realidad en la mayor medida posible es decir, que sea exacta precisa y libre de errores.

2.2. Importancia de un Estudio de Mercado

Un gran número de empresas y organizaciones destinan un porcentaje importante de sus presupuestos anuales a la realización de investigaciones de mercados, (a través de departamentos propios o con la colaboración de entidades externas dedicadas a estos fines). Los principales motivos e inquietudes para la elaboración de estos estudios son los siguientes (BIC Galicia, 2005):

- Permite conocer el mercado en el que una empresa u organización pretende localizar su propia actividad. De esta manera la elaboración de una estrategia con la que proceder se tornará más eficaz.
- Ofrece una visión actualizada del sector de actividad; permitiendo analizar su desarrollo y tendencia en el tiempo, para así elaborar estrategias futuras.
- Es muy válida para reconocer oportunidades de negocio, o para la identificación de otras vías para el desarrollo de un proyecto o decisión empresarial.
- Es una herramienta que permite evaluar el funcionamiento de las empresas y el resultado de sus decisiones, además del impacto que tienen sobre el mercado.

2.3. Tipos de Estudio de Mercado

Dependiendo de los objetivos que se pretendan alcanzar, existen diferentes modelos a seguir para la elaboración de una investigación de estas características. De tal modo, es en este punto, necesario una identificación de los mismos, donde procede establecer al siguiente detalle (BIC Galicia, 2005):

- **Lanzamiento de una nueva línea de actividad:** Es prudente, que llegado el momento de poner en marcha una nueva idea o producto a disposición del público, se realicen todos los estudios y comprobaciones previas a su lanzamiento, para de esta forma, conocer una aproximación respecto a cuál sería su impacto. Por ello es determinante un amplio conocimiento de los factores de la demanda para configurar el diseño de la oferta, y que ésta cumpla las expectativas.

- **Entrada en un nuevo mercado:** La penetración en otros mercados no está tampoco exenta de estudio. Problemas tales como las diferencias sociales, políticas y económicas, además de los gustos y las costumbres particulares de cada región, hacen a esta herramienta particularmente útil en el momento de la toma decisiones.

- **Evaluación de las causas de las variaciones de ventas:** Tarde o temprano las empresas se enfrentan a variaciones significativas en su volumen de facturación, pudiendo ser éste, positivo o negativo. La entrada de nuevos competidores[4] o coyunturas económicas desfavorables pueden llegar a tener unas consecuencias adversas sobre los proyectos que la empresa está acometiendo, o los que en futuro desarrollará. El objetivo del estudio de mercado en estas situaciones es el de proporcionar nuevas vías de actividad que permitan paliar los efectos de las nuevas condiciones en las que se ven involucrados la empresa y sus proyectos.

- **Impacto de una campaña publicitaria:**Un gran número de empresas realizan importantes inversiones en campañas de publicidad. Se espera que ésta tenga los efectos esperados en el público objetivo. Es por ello necesario predecir el impacto que tendrá sobre el mercado de tal forma que permita (si los hay) corregir fallos, para maximizar el efecto que la campaña se espera que tenga.

[4] Las fuerzas competitivas de Porter son una herramienta muy importante para conocer el macroentorno de la empresa. Son las siguientes: 1. Amenaza de nuevos competidores, 2. Amenaza de productos o servicios sustitutivos, 3. Poder de negociación de los clientes, 4. Poder de negociación de los proveedores, 5. Competencia en el mercado.

- **Modificación de un producto:** Debido a un mercado en constante cambio, con consumidores cada vez más exigentes y la fuerte competencia en los mercados, (tanto a nivel nacional como internacional), exige que las empresas desarrollen nuevos productos acordes a las necesidades que el público requiere.

2.4. Fases de un estudio de mercado

En el momento en el que se decide realizar una investigación de estas características, es necesario definir unas pautas a seguir, que se describen a continuación (BIC Galicia, 2005):

- **Definición y establecimiento de objetivos:** Consiste en identificar un problema o necesidad. Esta es la parte más complicada ya que va a condicionar el planteamiento del propio estudio y por consiguiente las conclusiones del informe. El hecho de no realizar correctamente esta etapa, implicaría la toma errónea de decisiones y por lo tanto el diseño ineficaz de acciones para el cumplimiento de los objetivos. La regla general para la correcta definición del problema consiste en que debe aportar al investigador toda la información que necesita para abordar el problema de decisión gerencial y guiar al investigador en el transcurso del proyecto (Molina, 2014*)*. Debe expresarse con claridad la finalidad del estudio para evitar cualquier tipo de desviaciones, y además debe de ser susceptible de llevarse a cabo. Aunque en la propia investigación pueden surgir nuevas vías de trabajo adicionales o incluso la modificación de las ideas iniciales.

- **Diseño de investigación y recopilación de información:** Una vez estén claros los objetivos, es el momento de determinar qué información es necesaria para la elaboración del estudio. Es aquí donde se deben seleccionar las fuentes de información más adecuadas para el cumplimiento de este objetivo. Las fuentes de información a su vez pueden ser primarias o secundarias. Seguidamente se procede a seleccionar los diferentes tipos de investigación: Exploratoria y Descriptiva (Concluyente).

Diseño	Exploratoria	Descriptiva
Objetivo	• Proporcionar conocimiento	• Comprobar hipótesis específicas
Características	• Informal • Proceso flexible y sin estructurar • Muestra pequeña y no representativa	• Se define claramente la información requerida • Proceso de investigación formal y estructurado • Muestra grande y representativa
Resultados	• Aproximativo	• Concluyente
Consecuencia	• Suele venir acompañada de otra investigación exploratoria u otra concluyente	• Los resultados obtenidos se usan para la toma de decisiones

Tabla 1: Investigación exploratoria y descriptiva.
Fuente: Malhotra (2005) , Molina (2014)

- **Recopilación de los datos:** Se corresponde con el punto central de este epígrafe e implica un laborioso y extenso trabajo por parte del investigador. Se debe evaluar hasta qué punto la información recogida es transcendental para el estudio, o si por el contrario, conviene sustituirla por otros datos. La calidad de la información es fundamental para obtener un análisis de rigor. Si se trata de una encuesta, esta tarea se conoce como "trabajo de campo" (Trespalacios, Bello & Vázquez, 2005). En esta situación es necesario contar con personal cualificado que actúe mediante las encuestas personales (casa por casa, centros comerciales...), desde una oficina (telefónicas o computadoras), por correo (correo tradicional y encuestas de panel) o electrónicamente (por correo electrónico o usando internet) (Malhotra, 2008). Para

este fin, ha de realizarse una correcta selección, formación, supervisión y evaluación con el objetivo de minimizar los errores (Molina, 2014):

- ♦ Selección: Decisión de las características del equipo.
- ♦ Formación: Elaboración de las preguntas y forma de la entrevista
- ♦ Supervisión: Realización de los controles de muestreo pertinentes
- ♦ Evaluación: Se tienen en cuenta costes, calidad de los datos y tasas de respuesta.

- **Análisis y preparación de los datos:** La forma en la cual se organiza los datos es muy importante, ya que su objetivo es minimizar la presencia de los sesgos, y se pueda realizar una correcta toma de decisiones, para que sea lo más adecuada posible. Para el procesamiento de la información, es necesario disponer de herramientas informáticas. La aplicación de técnicas estadísticas son imprescindibles, variando, en tanto a su complejidad y sus posibilidades; son las técnicas de análisis univariable, análisis bivariable y multivariable.

- **Elaboración del informe final:** Deberá recoger de forma estructurada toda la información de la etapa anterior, además de contener todos los procedimientos que se han usado, el diseño de la investigación, los datos que fueron obtenidos, y las consiguientes conclusiones alcanzadas en el estudio. Así pues a la vista del informe, es importante que se propongan una serie de recomendaciones dirigidas a la organización o empresa para la cual se está realizando el trabajo.

2.5. Fuentes de información

La información es un elemento crucial en la toma de decisiones para las empresas, ya que permite (si es adecuada), minimizar la incertidumbre y el sesgo que pueda derivar de cualquier decisión o investigación. Proporciona las claves para una adecuada orientación al mercado y permite afianzar la posición global frente a la competencia.

Los cambios en la conducta del consumidor, las modas y la tecnología, provocan grandes transformaciones en el entorno de la empresa, ocasionando que patrones de procedimiento habituales y estrategias comúnmente utilizadas, pierdan utilidad, o la carezcan por completo. Es primordial acudir a varias fuentes de información para poder escoger un diseño acorde a las

características demandadas del estudio. Para este fin es necesario clarificar qué tipo de información se necesita y la obtención de la misma, es en este punto donde se distinguen dos tipos de fuentes de información: Fuentes de datos primarios y Fuentes de datos secundarias.

- **FUENTES DE DATOS PRIMARIOS**

Son datos que por sus características no existen y tampoco están publicados, si no que es el investigador el encargado de su obtención. Aportan información de primera mano aplicando los distintos métodos de investigación existentes, es por tanto necesario crearla o generarla de forma expresa.

El procedimiento de obtención de los datos en este caso, es mediante la propia investigación de mercados, llevado a cabo de manera específica en el tiempo durante un período determinado, a tenor de un problema, necesidad, cambio de tendencia o irregularidad que se haya visto en el mercado. Estas fuentes destacan por las dificultades que conllevan en tanto a la recopilación de la información deseada, es necesario mucho tiempo y personal cualificado para dicha tarea.

- **FUENTES DE DATOS SECUNDARIOS**

Son aquellas, donde la información recogida ya está elaborada y por lo tanto existe a disposición de los usuarios. En este punto se incluyen informes que pueden resultar de utilidad para la toma de decisiones en las empresas. Al ser una información ya existente, su disponibilidad es inmediata. Es el punto de partida de la búsqueda de información, permite en muchas ocasiones el análisis y la toma de decisiones con argumentos suficientes, sin tener que recurrir a investigaciones de mercado más lentas y caras (Trespalacios *et al*, 2015). Los datos secundarios se dividen en dos grupos diferenciados en tanto a su procedencia: Internos y Externos:

  - **Datos internos:** Son aquellos de los que dispone la propia empresa u organización, siendo estos intrínsecos a su propio funcionamiento, generados en sus departamentos o basados en la experiencia. Son las primeras fuentes a las que el investigador debe acudir por su reducido coste de oportunidad que implica su consulta; fáciles y conocidos; Cuentas Anuales (Balance, pérdidas y ganancias, memoria), Libros Contables (Libro Diario, Libro Mayor). Los esfuerzos dedicados a la generación de bases de datos que recojan esta información, facilitará siempre la toma decisiones.

- **Datos externos:** La recopilación de este tipo de información consiste en localizar el origen disponible para nuestro propósito, y posteriormente hacer una valoración del interés que merece. Debe comenzar la búsqueda en fuentes de carácter más genérico, con el fin de aproximarse paulatinamente a lo específico, con información más concreta, en lo relativo a la investigación.

3. La estadística aplicada a investigación de mercados

Para llevar a cabo el trabajo del investigador, es necesario contar con herramientas que, por una parte permitan una selección de la información, y por otra su posterior tratamiento para finalmente proceder a su análisis y alcanzar finalmente el objeto del estudio. Las ventajas reportadas son destacables. Aquí se presentan algunas de ellas: (Merino, Pintado, Sánchez & Grande; 2010).

- Permite llegar a conclusiones que no se habrían podido obtener mediante la simple observación superficial de los datos.

- Permite alcanzar interpretaciones concluyentes de la información obtenida. Las aplicaciones estadísticas conducen a observaciones objetivas.

Existe un amplio espectro de herramientas que la estadística pone a disposición de esta materia para realizar sus análisis. Seguidamente se presentarán las más utilizadas organizadas en cuatro grupos:

- Muestreo (probabilísitico y no probabilístico)
- Análisis Univariante
- Análisis Bivariante
- Análisis Multivariante

3.1. Muestreo Estadístico

En todas las investigaciones estadísticas, existe un conjunto de elementos de los que se extrae la información. Dichos elementos constituyen la "población" o "universo estadístico" (López, 1999). La forma en la que escojamos estos elementos para realizar la investigación da lugar a un censo,

en el caso de que se opere con todos los elementos de la población íntegramente. Sin embargo este proceso entraña dificultades que, en la mayoría de los casos, exigen que esta vía sea deshechada . Algunos de estos inconvenientes podrían ser su elevado coste de realización, la destrucción de elementos que podría ocasionar la toma de información o porque la población sea infinita.

Las consecuencias que estos problemas implican, obliga a tomar otras vías de investigación. La toma de solo una parte de la población estadística recibe el nombre de muestreo estadístico. La totalidad de los elementos extraídos para la obtención de información se denomina muestra. El tamaño muestral es el calificativo que se le atribuye al número total de elementos de los que se compone la muestra.

La razón por la cual se realizan labores de muestreo estadístico en los estudios de mercado es por su bajo coste, rapidez en la obtención de resultados, y además de no presentar demasiada complejidad, ya que resultaría muy caro entrevistar a todos los individuos de una población. Los procedimientos de muestreo, permiten conocer la información que se necesita en el análisis, simplemente contando la datos proporcionados por un número determinado de elementos, llamados muestra, (Merino *et al*, 2010).

El objetivo central del muestreo es la obtención de información. Es primordial seguir unas etapas, para que el objetivo del estudio permanezca siempre nítido. Estas son las fases que una investigación de este calado debe tener, para su correcto ejecución y planificación, (López, 1999).

- Planificación de los objetivos: Deben ser claros y concisos. Es importante mantener intactos los más importantes, y además deben ser lo suficientemente simples para que sean entendidos por aquellos que van a realizar el trabajo de campo.

- Delimitación de la población objetivo: Esta selección debe hacerse con suma cautela, ya que es a partir de la misma, de donde se extraerá la muestra, que habrá de ser objeto del estudio.

- Establecimiento del marco: Es necesario que exista alguna lista o mapa conceptual, que sirva como guía en el universo que se cubrirá.

- Diseño de la muestra: La población se tiene que poder dividir en unidades de muestreo. Las muestras deben de tener la capacidad de representar de una manera adecuada a sus poblaciones, es decir, que las muestras sean representativas de sus poblaciones.

- Encuesta piloto: Es muy útil, cuando se realizan encuestas de grandes dimensiones, tomar una pequeña muestra para una prueba piloto, pues permite clarificar a los encuestadores y verificar el manejo de las actividades de campo; además se pueden obtener estimaciones de algunos parámetros poblacionales.

- Trabajo de campo: Consiste en la obtención de datos que componen la muestra objeto del estudio.

- Procesamiento de los datos: Las encuestas generan gran cantidad de información. Esta labor ha de realizarse de forma automatizada, empleando lo mayor posible el uso de las TICs.

- Evaluación de resultados: Después de la obtención de los mismos, es necesario comprobar la calidad de la encuesta, antes de proceder a la difusión de sus resultados.

- Presentación de resultados: Su simple publicación, no refleja el trabajo empleado para obtenerlos. Se necesita una presentación adecuada y ordenada, que permita conocer la calidad de la investigación.

- Difusión de los resultados: Una vez finalizada la encuesta, es necesario encontrar un medio por el que se divulgue la información obtenida del análisis.

3.2. Procedimientos de muestreo

La primera cuestión que debe preguntarse el investigador es decidir qué técnica de muestreo empleará en el análisis. Existen dos grupos claramente diferenciados. El muestreo no probabilístico y el probabilístico:

- **MUESTREO NO PROBABILÍSTICO**

Es la práxis habitual para conseguir la información de una población objetivo, intentar obtener la información muestral sin demasiados costes (López, 1999). En este tipo de muestreo no existe un método que permita determinar la probabilidad de seleccionar un elemento que forme parte de la muestra. Por ello las estimaciones son difícilmente extrapolables a la población de interés (Merino *et al*, 2010). Es muy frecuente su uso debido a la carencia de un marco poblacional. Existen diferentes tipos de muestreo probabilístico, atendiendo a su uso y características:

- Muestreo intencional u opinático: Se realiza a través de informantes profesionales del campo que ocupa la encuesta (López, 1999).

- Muestreo aplicando criterio: El investigador es quién elige a los elementos de la muestra. Su argumento de decisión está basado en sus propios conocimientos del campo estudiado (Trespalacios *et al*, 2005).

- Muestreo por cuotas: Se puede considerar como un muestreo por criterio en dos etapas (Malhotra, 2008). La primera consiste en desarrollar categorías de control para elementos de la población, y la segunda en selección de los elementos que se incluirán en la muestra, acordes a unos criterios.

- Muestreo por bola de nieve: Aplicado en aquellos casos donde el investigador no consigue identificar a los elementos que puedan proporcionar información del tema objeto del estudio(Trespalacios *et al*, 2005).

• MUESTREO PROBABILÍSTICO

Se utiliza el muestreo probabilístico para medir el grado de representatividad de la muestra. En este caso se podrá establecer la probabilidad de obtener cada una de las muestras que sea posible seleccionar (López, 1999); a través de un tipo de procedimiento de muestreo, cuando sea aleatoria la selección de las distintas muestras comprendidas en el referido espacio muestral, ocurriendo en condiciones de azar.

Dentro del muestreo probabilístico se contemplan el muestreo aleatorio simple, el muestreo aleatorio sistemático, el muestreo aleatorio estratificado y el muestreo por conglomerados.

- MUESTREO ALEATORIO SIMPLE

Este apartado se divide entre muestreo aleatorio simple con reposición y sin reposición. Si no se especifica este muestreo se entiende siempre como muestreo aleatorio simple sin reposición, también llamado muestreo irrestríctamente aleatorio (López, 1999).

El muestreo aleatorio simple sin reposición, es un procedimiento con probabilidades iguales, que consiste en obtener elementos muestrales aleatoriamente sin devolverlos a su población original, no interviniendo su orden de colocación. Los elementos extraídos de forma repetida son sucesos imposibles en este caso.

Por lo que respecta al muestreo aleatorio simple con reposición, los elementos seleccionados se devuelven a la población original, por lo se puede repetir la extracción de elementos muestrales.

- MUESTREO ALEATORIO SISTEMÁTICO

Se considera una población de tamaño N, agrupando sus elementos en n zonas de tamaño k (N=nk). El procedimiento de extracción de una muestra de tamaño n es eligiendo una unidad al azar en la primera zona, y para seleccionar las n-1 unidades restantes, se seleccionará en función de dónde se ha extraído la primera. Por ejemplo; si se ha seleccionado en la primera zona es la segunda unidad, se extraerán n-1 restantes para la muestra, tomando la segunda unidad de cada zona (López, 1999).

- MUESTREO ALEATORIO ESTRATIFICADO

En el muestreo aleatorio estratificado una población heterogénea de N $\{u_i\}_{i=1,2,\dots,N}$ unidades, se subdivide en L subpoblaciones lo más homogéneas posibles denominados estratos $\{u_{hi}\}_{i=1,2,\dots N_h}^{h=1,2,\dots,L}$ de tamaños $N_1, N_2, \dots, N_L$ (López, 1999). Para obtener la muestra estratificado de tamaño n, se selecciona n_h elementos (h=1,2,..., L), de cada uno de los L estratos de los que se divide la población de manera independiente (Blázquez, 2014). Una decisión importante es la afijación, es decir, el reparto de la muestra por estratos. Hay tres tipos fundamentales de afijación: simple, proporcional y óptima.

- MUESTREO POR CONGLOMERADOS

Las unidades muestrales no son el elemento individual de la población, son grupos de elementos mutuamente excluyentes llamados conglomerados. Los casos más habituales de este tipo de muestreo son la selección aleatoria de familias de una población para realizar un estudio de los individuos que pertenecen a dichas familias, un ejemplo muy común son los estudios que realizan los ingenieros agrónomos en las diferentes cultivos de una explotación. Salientan en este punto, dos tipos de muestreo por conglomerados: Monoetápico y Bietápico.

3.3. Análisis Univariante

Las técnicas de análisis de datos univariantes analizan cada variable de forma individual. Es el más sencillo de todos. Se emplea para describir las características muestrales o como mera introducción a análisis más complejos. Permiten además una visión general de las características de los datos (Díaz y Martín-Consuegra, 2014).

Las técnicas utilizadas tienen un carácter descriptivo. Uno de los primeros tipos de análisis que se realizan en este campo son las distribuciones de frecuencias de una variable. La información que facilita son el número de respuestas asociadas con distintos valores de la variable analizada (Díaz y Martín-Consuegra, 2014). Una tabla de frecuencias es un instrumento fácil de comprender, pero en ocasionas genera demasiado volumen de información, por lo que es imprescindible resumirla en una serie de estadísticos descriptivos. Lo más utilizados se agrupan en las medidas de tendencia central, de dispersión y de deformación.

- **Medidas de posición**: Son las características básicas de la información de algunos valores sintéticos (Martin-Pliego, 2007).

Media aritmética: Se define como el sumatorio de todas los valores de una distribución, multiplicado por el número total de datos (Martín-Pliego, 2007). Sin embargo la media no informa de manera cuantitativa de la disparidad entre los datos, pudiendo surgir valores extremos en la distribución que ocasionen errores en el análisis.

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i n_i}{N}$$

Mediana: Es aquel valor, que divide a la distribución de datos por la mitad, es decir, el mismo número frecuencias a su derecha que a su izquierda.

Moda: Indica en una distribución, cuál es el dato que más se repite. Es una buena medida para las variables nominales, siendo la única aplicable en este apartado.

- **Medidas de dispersión**: Coeficientes estadísticos que indican si los datos están cercanos o alejados respecto a un promedio (Escuder, Murgui; 1995).

Rango: Es la diferencia cuantificable entre el mayor y el menor valor de la distribución:

$$Ra = X_n - X_1$$

Varianza: Se define como la media aritmética de los cuadrados de las desviaciones de los valores de la variable a la media aritmética. Es decir, indicará la mayor o menor dispersión de los datos respecto a la media. Si ésta es muy grande, la media no será representativa. La varianza no viene expresada en las mismas unidades de medida que la distribución, ya que están al cuadrado, lo cual dificulta su interpretación; por lo tanto se recurre a la Desviación Típica, que es la raíz cuadrada de la varianza, de esta manera se expresará en las mismas unidades que la distribución (Martín-Pliego, 2007)

$S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \frac{n_i}{N}$ Varianza

$S = \sqrt{S^2}$ Desviación Típica

- **Medidas de forma:** Son una serie de tipologías que se establecen para clasificar las distribuciones en función de su forma, es decir; según su representación gráfica.

Asimetría: Consiste en un indicador que muestra el nivel de simetría (o asimetría), que presenta una distribución. Se pretende con esto evitar la representación gráfica.

Curtosis: Estas medidas pretenden estudiar la distribución de frecuencias de la zona central de la distribución. El hecho de que exista una mayor o menos concentración de frecuencias alrededor de la media y en la zona central de la distribución, ocasionará un mayor o un menor apuntamiento de la distribución; de esta forma estas medidas se llaman también de apuntamiento (Martín-Pliego, 2007).

3.4. Análisis Bivariante

Es frecuente que el investigador, por las características de su estudio, le interese conocer la relación existente entre dos variables. En función del tipo de análisis requerido, se utilizará una herramienta u otra, acordes con las especificaciones propuestas.

- Coeficiente de correlación lineal

Se utiliza para conocer la relación entre dos variables. Para ello se recurre al coeficiente de correlación lineal:

$$r = \frac{S_{xy}}{S_x S_y}$$

Es un coeficiente acotado, entre los valores 1 y -1; indicando el primero, correlación perfecta positiva, y el segundo correlación perfecta

negativa. Ahora bien la correlación entre ámbas variables puede ser nula (r=0) ya que la covarianza se puede anullar sin cumplir la condición de independencia (Martín-Pliego, 2007).

- Regresión Simple

La regresión tiene por objeto , poner de manifiesto la estructura de dependencia que mejor explique el comportamiento de la variable Y sobre la variable X. La regresión será lineal cuando cuando la curva de regresión obtenida sea una recta.

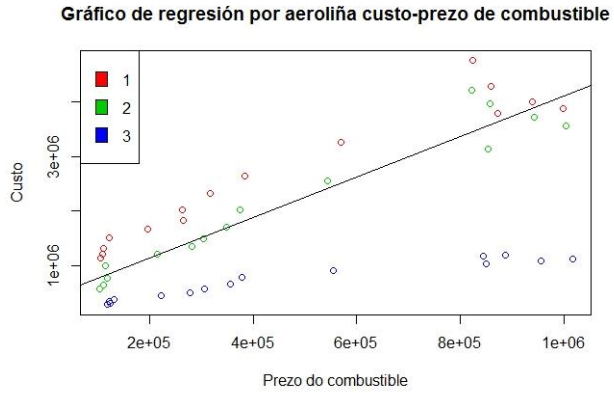

Figura 1: Regresión de tres aerolíneas
Fuente: elaboración propia a partir de datos de la FAA.[5]

- Contraste Chi-Cuadrado

Es un contraste de hipótesis realizado para variables cualitativas, donde se pretende averiguar si las variables analizadas son independientes. Este es el estadístico utilizado (Merino *et al*, 2010):

$$X^2 = \sum_{j=1}^{c} \sum_{i=1}^{f} \frac{(n_{ij} - n_i \frac{n_j}{n})^2}{n_i \frac{n_j}{n}}$$

- Análisis de la varianza

Es una técnica estadística que consiste en analizar las diferencias entre las medias de dos o más poblaciones. Explota los datos provenientes de situaciones experimentales, pero también es extrapolable a los obtenidos mediante encuestas (Díaz y Martín-Consuegra, 2014).

[5] Análisis realizado a partir de datos proporcionados por la FAA (http://www.faa.gov/), de 3 aerolíneas donde se analizan el precio del combustible por aerolínea y los costes totales de cada aerolínea.

3.5. Análisis Multivariante

Cuando se dispone de gran cantidad de datos y éstos no se pueden entender directamente y el análisis clásico de medias, varianzas correlaciones etc... no pueden proporcionar una visión en conjunto, se emplean estas técnicas. Existen varias técnicas:

- Análisis de regresión múltiple

Los modelos de regresión lineal están basados, en el concepto de dependencia de las variables, las cuales, una de ellas será interpretada como "variable explicada" o "variable dependiente", mientras que el resto se denominan "explicativas" o " independientes". La relación entre aquellas no es determinista, y además se incluye en el modelo un grado de aleatoriedad que se denomina "término de perturbación" (Mallou, 2003).

- Análisis conjunto

Es una técnica estadística que permite explicar una variable de respuesta (o dependiente) de tipo ordinal, en función de dos o más variables de carácter nominal (atributos o factores) (Mallou, 2003). El objetivo principal es averiguar cuáles son los atributos más significativos, que se emplean en el análisis de elementos (Vergara, 2015).

- Análisis multivariante de la varianza

Se trata de una herramienta usada para contrastar simultáneamente la igualdad de las medias $\mu_1, \ldots, \mu_k$, lo cual implica una gran ventaja, que permite omitir los contrastes para todas las parejas posibles de medias, a través del modelo "t" de student (Mallou, 2003). De esta manera se reducen operaciones y errores acumulativos.

- Análisis discriminante

Consiste en otra herramienta de clasificación y asignación de un elemento del grupo del que se conocen unos determinados atributos (Mallou, 2003). Su objetivo consiste en identificar las diferencias existentes entre los grupos, de tal manera que se compruebe que el elemento seleccionado forme parte de alguno de los grupos.

- Análisis factorial

Técnica consistente en simplificar las numerosas relaciones existentes en un grupo de variables cuantitativas. Para conseguirlo, busca factores que la relacionan las variables, que de por si no están relacionadas (Mallou, 2003).

- Análisis Cluster

Conjunto de técnicas cuyo propósito es formar grupos a partir de un conjunto de elementos. Los grupos deben de estar compuestos con elementos que sean similares (homogeneidad interna), y al mismo tiempo diferentes entre los diversos grupos (heterogeneidad de los grupos), (Díaz y Martín-Consuegra, 2014).

- Análisis de correspondencias

Método de reducción de la variabiliad en el que se busca reorganizar las categorías de 2 o más atributos, teniendo en cuenta la relación existente entre ellos. Existen dos versiones fundamentales de este análisis; el Análisis de Correspondencias Simples (ACS), aplicado a tablas de contigencia que se obtienen mediante el cruce de dos variables nominales, y el Análisis de Correspondencias Múltiples (ACM), el cuál es una extensión del anterior al caso de dos o más variables nominales (Mallou, 2003).

4. Investigación de mercados on-line

Internet se está convirtiendo en una herramienta que, cada vez más, muchos investigadores usan como fuente de elaboración de sus trabajos. Existen a disposición del público multitud de bases de datos de diversa índole, muchas de ellas de acceso gratuito. Un ejemplo son los organismos nacionales de procesamiento estadístico, tales como el INE en España, o el Census en Estados Unidos. Aquí se encuentra multitud de información, como datos referentes a economía, demografía o mercado laboral.

Las ventajas de realizar una investigación por internet, son muchas, haciendo hincapié en la facilidad de obtención de los datos. La amplia oferta de organismos que los proporciona posibilitan extender el ámbito del propio estudio u obtener diferentes puntos de vista sobre un mismo tema. Además no hay límites geográficos para la investigación, desaparecen las fronteras, y se reducen costes.

La recogida de información en internet, es muy útil a la hora de elaborar encuestas. A día de hoy, donde cada vez más consumidores disponen de acceso a la web, se aprecia, a efectos prácticos, la versatilidad de este tipo de herramientas en la investigación de mercados, siendo una práctica con tendencia al alza (Merino *et al*, 2010).

4.1. Las redes sociales

Otra alternativa, dentro del marco de obtención de información a través de internet, es en foros y redes sociales. En estos medios, existe mucha información referente a opiniones, comentarios y datos relativos a los propios usuarios, tales como, su nombre (o nick de usuario), procedencia, fecha de nacimiento, e incluso información de sus propios familiares (caso de Facebook, donde en cada perfil, se indica, a discreción del propio usuario, el número de familiares y el nombre de cada uno, además de su rasgo de parentesco). La utilidad que tienen estas nuevas fuentes de información, cada vez más explotadas, es la de omitir el paso de encuestar a los consumidores sobre un tema en concreto; ya que en la mayoría de los casos, es información que éstos han proporcionado con anterioridad en las plataformas sociales. La información obtenida a través de este tipo de canales puede resultar de mucha utilidad en diferentes estudios de marketing y, particularmente, en investigación de mercados.

El mayor inconveniente a priori para explotar este tipo de datos, es conocer cuáles pueden ser útiles en la investigación, y, de ser así, qué procedimientos hay que seguir para procesar la información de una manera viable. No todas las redes sociales pueden ser aptas para los experimentos que se pretendan llevar a cabo. Es necesario saber qué rango de edades pueden tener por término medio, los usuarios de estos portales, y quiénes participan en ellas. Por lo tanto, el establecimiento de un perfil básico de usuario se torna muy útil y necesario en el momento de emprender investigaciones de estas características.

Posteriormente, es necesaria la aplicación de un método, que permita extraer eficazmente grandes cantidades de información en el menor tiempo posible, ya que a disposición del público en general no hay bases de datos que facilitan la obtención de información sobre este tipo de plataformas; por lo que se torna imprescindible acceder de lleno y directamente a estos portales para conseguirla. Existen diversas herramientas para la extracción de datos de las redes sociales, (algunas de ellas son gratuitas y está disponibles en la web, por ejemplo, TwitterCounter o Twitteranalyzer). Su mayor problema radica, en que la información que se puede obtener a través de estas aplicaciones puede resultar escasa, por varios motivos:

1. Normalmente realizan un filtrado de la información muy pobre y aleatorio, lo cual afecta a la veracidad y precisión del trabajo del investigador, y hace inevitable la aparición de multitud de sesgos en la información obtenida.

2. A efectos legales, el derecho a difundir la información que dichas redes sociales tienen es potestad casi exclusiva de las mismas.

Aunque la información que allí se encuentra es de carácter público, existe determinados datos de naturaleza sensible, cuya difusión supondría un verdadero problema, a tenor las legislaturas que cada país contemple en estos casos.

3. Muchas de estas herramientas están "limitadas", ya que algunas disponen de versiones de pago, y ni tan siquiera abonando una subscripción a estas aplicaciones, se garantizaría en ningún caso que realmente sea de utilidad para nuestros propósitos.

La conclusión subyacente es que usando este tipo de herramientas, su capacidad como instrumento útil para uso comercial o académico, en la mayoría de los casos es muy deficiente. Sin embargo, ya existen empresas, que ofrecen en el mercado servicios para solventar estos problemas pudiendo así realizar un análisis más objetivo de nuestro tema.

A su vez, Twitter dispone de aplicaciones propias que permiten la extracción de datos de sus servidores, de libre acceso y gratuitas. Son las "API rest" y "Streaming API". No obstante, dichas herramientas también tienen sus limitaciones, además de usar una interfaz, de por sí árida y complicada, muy poco accesible para el usuario medio de internet, con unos conocimientos de programación y "social media" escasos.

Por una vía o por otra, la extracción de información de las redes sociales es, hoy en día una realidad, y su aprovechamiento como herramienta para la elaboración de estudios de mercado, análisis de palabras, psicología y otras disciplinas, es un servicio cada vez más demandado y en alza en los últimos años.

4.2. Investigación de Mercados y redes sociales

Se han realizado muchos avances en este campo del márketing, para la obtención de información susceptible de ser empleada en estos estudios, aunque de momento, no ofrezca las facilidades, que otras fuentes disponen (Merino *et al*, 2010). Sin embargo el interés que suscita este tipo de investigaciones es doble (Rambocas & Gama; 2013):

- Primeramente, estudiosos del márketing han reconocido, la profunda influencia que las redes sociales ejercen sobre el consumidor, destacando de esta manera el trabajo de Gruen et al (2006), citado en Rambocas & Gama (2013), y sus comportamientos de compra. A la vez del gran reto que han supuesto para los investigadores la explotación de este tipo de información; Rambocas & Gama (2013) cita a Stanton *et al, (*2001)

- En segundo lugar, cada vez es más sencillo obtener los datos para nuestras investigaciones. Es posible conseguir gran cantidad de información en tiempo real y sin contaminar por la presencia de un buscador externo.

Una vez estén claros los objetivos del análisis y los datos recolectados correctamente, llegó el momento, de emprender todas aquellas acciones destinadas a su posterior procesamiento, con el objetivo de reflejar en el estudio, las conclusiones, virtudes o desventajas, además de los inconvenientes que la propuesta presenta. Para ello es necesario poner en práctica técnicas informáticas de minería de datos, procesamiento de lenguaje natural y análisis de sentimiento, que serán muy útiles para los propósitos de los investigadores, a la vez que imprescindibles.

4.3. Data mining y análisis de sentimiento

En este punto es el momento en el que se procesan, todos los datos que se han obtenido a través de las diversas fuentes nombradas anteriormente. Existen varias técnicas en esta materia, que facilitan el tratamiento y posterior análisis de los datos.

El concepto de minería de datos, (o "data mining") hace referencia al tratamiento de grandes cantidades de información con la finalidad de extraer conclusiones de los mismos, en tanto a su comportamiento bajo determinadas circunstancias, además de buscar ciertos patrones en los mismos.

El análisis de sentimiento, es una herramienta de la "minería de datos" que recurre al procesamiento del lenguaje natural (Bugeja, s.f.), lingüístico computacional y analítico para identificar y extraer contenido de interés de un conjunto de información textual. Además puede analizar la información extraída sin incurrir en retrasos y de manera sistemática (Rambocas & Gama, 2013) . Su utilidad en la investigación de mercados es tal, que proporciona la capacidad al investigador para aprender sobre los sentimientos que expresan los consumidores (por un determinado producto o campaña, por ejemplo). La metodología a emplear se detalla en la siguiente figura:

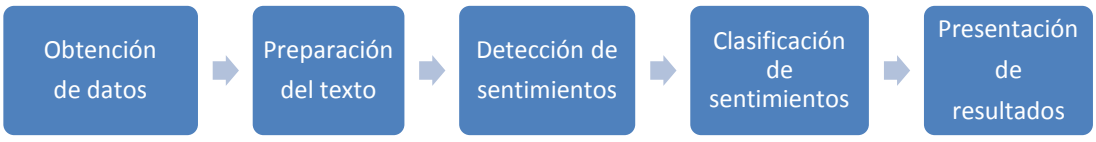

Figura 2: Proceso del análisis de sentimiento
Fuente: Elaboración propia a partir Rambocas & Gama (2013)

- Obtención de los datos: Se trata de una fase primordial, ya que su procedencia marcará de manera significativa las etapas posteriores. En el caso de recurrir a internet, y a las redes sociales en concreto, normalmente la información se encontrará en un estado desordenado y desagregado en múltiples portales (Rambocas &Gama, 2013), donde se expresan multitud de opiniones.

- Preparación del texto: Una vez se disponga del grueso de datos para realizar el análisis, es necesario tratarlo. Los textos extraídos han de ser formateados a disposición del investigador. Los programas informáticos son una opción muy interesante en este caso, sobre todo si se está tratando con cantidades muy importantes de información.

- Detección de sentimientos: Se escogerá un método, en función de las características de la información recogida, que permite detectar todas aquellas peculiaridades que se pretendan obtener. El desarrollo de un algoritmo matemático para programación informático es una opción muy productiva.

- Clasificación de sentimientos: Una vez obtenido el algoritmo o el método que permita la detección, se torna imprescindible asignarle cómo éste clasificará la información que va a procesar. Normalmente, se asocia a "malo", "bueno", "positivo", "negativo" o "neutro", pero las posibilidades en este punto son muy variadas.

- Presentación de resultados: El propósito principal, es convertir todos los datos procesados, en información legible y útil para su análisis y establecimiento de conclusiones. El uso de herramientas gráficas como diagramas de barras, gráficos de línea y tablas de frecuencia, son herramientas óptimas para dar formato al análisis que se pretende realizar.

Los usos, que más se han publicado de análisis de sentimiento, han sido sobre la atención que ha suscitado a consumidores sobre un producto en concreto y, la extracción de opiniones de algún tipo de producto específico como teléfonos, películas, hoteles, restaurantes y empresas en general (Rambocas & Gama, 2013). Además una parte de las mediciones de mercado, respecto a la publicidad de marca pueden ser medidos en las redes sociales, ofreciendo el análisis de sentimiento, una alternativa poco costosa para medir la efectividad de las decisiones que se toman en el márketing (de Groot, 2012). Las aplicaciones en un ambiente empresarial, están hoy en día en alza y cada vez más empresas, del ámbito de la publicidad y el márketing, ofrecen estos servicios. Dichos servicios incluyen:

- ✓ Rastreo de usuarios y no usuarios e índices de productos y servicios.
- ✓ Monitorización de las actividades que enfrenta la empresa con el fin de prevenir efectos virales.
- ✓ Evaluar rumores del mercado, la competencia, y las últimas tendencias entre los consumidores.
- ✓ Medición de la respuesta del público a una determinada acción o evento relacionado con la actividad de la empresa.

Sin embargo, el análisis de sentimiento, no solamente tiene aplicaciones comerciales, si no que se emplea en otras áreas. La creciente actividad y usuarios que reúnen las redes sociales (Pereira, 2014) y su uso para aprocechamiento comercial, han suscitado el surgimiento de determinadas prácticas y conductas, de dudosa legalidad, o que podrían ser potencialmente peligrosas para la seguridad de las personas, países u organizaciones. El gobierno de los Estados Unidos, es muy activo en estos campos. Según el diario New York Times (2006), destinan cada año alrededor de 2,4 millones de dólares a la investigación y desarrollo de equipos informáticos que permiten monitorear cualquier actividad que resulte potencialmente peligrosa para la seguridad nacional, por parte de cualquier fuente, y sea cual sea su procedencia. En este aspecto, resalta la actividad de la ultrasecreta NSA de Estados Unidos (Agencia para la Seguridad Nacional). Se creó en el año 1952 por el presidente Harry S. Truman, y no fue hasta el año 1970 cuando se reveló su existencia. Su función principal consiste en la seguridad de la información. En los últimos años ha tomado mucha relevancia a raíz de los documentos filtrados por el ciudadano estadounidense Edward Snowden, donde reflejaban multitud de acciones de espionaje a millones de personas, organizaciones y países en todo el mundo.

Los partidos políticos también se benefician de la información que proporciona este tipo de estudios. Pueden realizar sondeos en multitud medios donde se faciliten opiniones relativas a sus ofertas políticas o campañas electorales, para complementar por esta vía la opinión de los votantes y el público en general. Algunos ejemplos de este tipo de actividades las han desarrollado empresas como "Crimson Hexagon" con sede en Massachusetts (EE.UU). Su análisis se basó en el sentimiento del público acerca de los derrames de crudo acontecidos en Golfo de México. A su vez la compañía británica "Linguamatics" analizó 130.000 cuentas de la red social "Twitter", para realizar un sondeo acerca de la opinión pública sobre las elecciones del Reino Unido. Su análisis fue sorprendente, ya que los resultados fueron muy similares a las encuestas políticas tradicionales, y predijo, dentro de un porcentaje de votos, que el Partido Conservador ganaría finalmente las elecciones.

Estos estudios confirman la utilidad que puede llegar a alcanzar las aplicaciones del análisis de sentimiento aplicadas a las TICs en multitud de áreas, siendo esta nueva vía muy reciente.

5. Caso práctico: La crisis de los motores trucados de Volkswagen desde la perspectiva de Twitter.

El presente análisis tiene como objetivo realizar un estudio de la crisis que ha tenido la empresa alemana de vehículos utilitarios Volkswagen respecto al trucaje ilegal de sus motores, instalando inhibidores de emisiones de CO2, con el objetivo de burlar los controles para el test de gases de efecto invernadero. Se han tomado como fuente los comentarios posteados en Twitter desde el 17 de septiembre al 31 de diciembre de 2015. Posteriormente se han procesado mediante un algorítmo de análisis de textos para con los datos obtenidos realizar un análisis estadístico y obtener unas conclusiones finales dentro del marco teórico de la investigación de mercados.

Este estudio abarca un nicho, el del análisis de las reddes sociales cuya explotación hoy en día está en alza. Por eso para analizar los posibles efectos de este escándalo sobre el mercado de productos de Volkswagen se realizará un estudio de los comentarios y opiniones que ha suscitado en una red social; Twitter. Buscando la manera de comparar estos resultados con la evolución del mercado. Sin embargo no se ha podido disponer de datos diarios o semanales del mercado de Volkswagen por lo que se han sustuido por un "proxy", los valores de su cotización en la bolsa de Frankfurt, ya que las reacciones del público podrían afectar a ámbos mercados de forma paralela. Se procederá por lo tanto en este caso práctico a analizar la evolución en Twitter del escándolo Volkswagen comparándola con el valor de las cotizaciones.

5.1. Cómo llegó Volkswagen a una de las peores crisis de su historia

El 18 de septiembre de 2015, la Agencia de Protección Ambiental (EPA) de EE.UU, después de detectar irregularidades en los test de emisiones de CO2 para los vehículos de turismo, acusó a la firma germana de instalar premeditadamente determinados aparatos electrónicos e informáticos que controlaban la expulsión de partículas contaminantes, que actuaban en el

momento que los autos realizaban las inspecciones pertinentes de acuerdo con la legislación.

¿Cómo se descubrió el fraude? a través del trabajo de un ecologista llamado Peter Mock director de "International Council for Clean Transportation" un grupo dedicado a preservar el medio ambiente, orientando sus acciones al control de los medios de transporte. Se pretendía en un principio, demostrar que los test de emisiones en Europa son mucho más laxos que en Estados Unidos. Para ello se comprobaría si los vehículos emitían menos emisiones cuando se realizaban los controles allí. Se tomaron como muestra varios modelos de turísmos: Volkswagen Jetta, Volkswagen Passat y BMW X5. Además se instalaron unos medidores de emisiones en los maleteros de estos coches, y los resultados fueron sorprendentes. Esperaban que los vehículos pasasen la prueba con mejores tasas de partículas contaminantes, ya que en Estados Unidos los controles son más estrictos. Mientras que la realidad demostró que el BMW X5 cumplía perfectamente con las leyes, los turismos de la marca Volkswagen llegaron a superar esta tasa hasta 35 veces por encima de lo permitido[6].

En base a estas gravísimas conclusiones del estudio de Mock, La EPA entra en escena, haciendo sus propias indagaciones, identificando a más de 482.000 vehículos afectados en Estados Unidos[7]. El día 23 de septiembre de ese mismo año se produce la asunción de culpa de Volkswagen, produciéndose la dimisión de Martin Winterkorn, (Nieto, 2015), sucedido posteriormente el día 23 de septiembre por Mathias Müller. Se reconoció públicamente que se había estado alterando los motores producidos en las plantas de fabricación de la Volkswagen, admitiendo que la cifra final de vehículos afectados podría rondar los 11 millones, aunque podría ser mayor, según reconocieron. Las primeras medidas que se tomaran por parte de la compañía alemana fueron las de realizar una dotación contable por valor 6.500 millones de euros para hacer frente al pago de futuras indemnizaciones. Por su parte el "Bundeskagabinett" el Gabinete Federal de Alemania, exigió a la compañía total transparencia con estos hechos. Como es habitual en estos casos, las acciones de Volkswagen, se desplomaron en la bolsa de Frankfurt.

5.2. La red social Twitter

Twitter es una red social basada en mensajes cortos de hasta 140 caracteres. Detrás de estos mensajes se articula todo lo demás. Cuando un usuario entra en Twitter, podrá encontrar un listado cronológico de Tweets que han publicado otros usuarios a los que siguen. La información creada en Twitter es pública por defecto, y la convierte en la red social más usada para la

[6] Véase: http://cincodias.com/cincodias/2015/09/22/empresas/1442919977_921118.html

[7] Véase: http://www.bbc.com/mundo/noticias/2015/09/150922_volkswagen_escandalo_trampa_perdidas_ac

obtención de información y análisis en diversos ámbitos, desde empresas que quieren saber más sobre sus clientes hasta investigaciones científicas (Artero y Marcos, 2014).

Twitter tiene una serie de términos que identifican cada acción que tiene lugar dentro de esta plataforma (Artero y Marcos, 2014):

- **Tweet:** Es el mensaje que publica el usuario y está limitado a 140 caracteres.
- **Follower:** Es un usuario que decide suscribirse al contenido que publica otro.
- **Hacer Follow:** Acción de seguir o suscribirse al usuario.
- **Friend:** Usuario de Twitter que sigue al usuario
- **Retweet:** Copia de un Tweet de otro usuario en el perfil propio. Puede añadir este último algún comentario al respecto además del Tweet original.
- **Mensaje directo:** Mensaje privado que solo puede ver el usuario al que va destinado.
- **Timeline de usuario:** Listado cronológico de Tweets publicados por el usuario.
- **Hashtag:** Etiquetas a las que pertenecen o pueden pertenecer los Twets. Van precedidos por el símbolo #.
- **Trending topic:** Tendencia sobre lo que se está hablando dentro de la red.

¿Cómo se puede extraer la información de Twitter? a través de dos programas: el API Rest y Streaming API (Russell, 2014). La diferencia que existe entre ellos, es que el primero permite acceder a la funcionalidad interna de Twitter por un sistema en forma de caja negra, mientras que el segundo se pone en contacto con la arquitectura de esta red social, de manera externa estableciendo una conexión, que terminará cuando el usuario lo estime. Las limitaciones de estos programas, son fundamentalmente la imposibilidad de acceder a un histórico de los datos. Además de estos dos métodos, existen otras posibilidades de extracción.

Aplicación	Limitación temporal	Limitación tamaño
Streaming API	Solo tiempo real	-
API Rest	NO	3.200 últimos tweets

Tabla 2: Las aplicaciones para extracción dde información de Twitter
Fuente: Elaboración propia a partir de (Artero y Marcos, 2014).

¿Qué información se puede extraer de Twitter? (Artero y Marcos, 2014):

- **Quién** escribió el Tweet, junto con sus datos públicos; nombre completo, localización, lenguaje, etc. Un estudio de la demografía de Twitter afirma que se trata de personas adultas de unos 35 años; Artero & Marcos (2014) cita a Dreyer (2011).

- **Qué** contenido muestra cada Tweet publicado por el usuario.

- **Cuándo.** Fecha y hora de publicación.

- **Dónde.** Coordenadas geográficas desde donde se publica el Tweet. Aunque esta información solo está incluida en el 2,02 % de los Tweets; Aretero & Marcos (2014) cita a Leetaru (2011).

5.3. Presentación de los datos

Para el presente análisis, se han tomado los comentarios realizados en Twitter, con referencia a Volkswagen desde el 17 de septiembre de 2015 a 31 de diciembre de 2015. Además se han recogido los "Tweets" del mismo período del año 2014, con el objetivo de realizar un estudio comparativo en una situación de supuesta normalidad. El espacio temporal seleccionado comprende el inicio y el posterior desarrollo de la crisis hasta final de año. También se ha recogido la cotización de las acciones de esta compañía para el período de 2015, con el objetivo de visualizar posibles relaciones en la bolsa de Frankfurt.

5.3.1. Obtención del texto

Primeramente es necesario un método de extracción de la información. Como los datos que se pretenden obtener pertenecen al total histórico, y ninguna de las aplicaciones disponibles para el público (y de forma gratuita) permite realizar ninguna indagación en este ámbito, (para uso académico y comercial), fue necesario el desarrollo de un sistema que pudiese acceder a la arquitectura web de Twitter. Por ello, y a través de la propia aplicación y sin el requisito tan siquiera de crear una cuenta en esta red social, se recurrió a técnicas de "Web Scraping" (Malik, Kumar & Rizvi, 2011) para la extracción de la información. Estos métodos consisten en obtener datos de páginas web a través de aplicaciones informáticas de software. Se han empleado dos herramientas para esta finalidad, que se detallan a continuación:

- Aplicación web de Twitter "búsqueda avanzada de Twitter"
- RStudio y R (procesamiento y análisis de datos estadístico)

A través de la primera se seleccionó como parámetro, que los tweets deberían estar escritos en lengua inglesa. El motivo principal de escoger este idioma es que Volkswagen es un producto global por lo que es más interesante reccoger un efecto más amplio, que únicamente se puede aproximar usando los Tweets en inglés como fuente, reforzado además por el hecho de que el 51% de los comentarios están en inglés. Según el Instituto Cervantes (2015), el inglés es la lengua más usada en internet, concretamente cuenta con 463.790.410 usuarios en la red, alcanzando un porcentaje, sobre el total de un 29,10%.

Segidamente es neceseario crear un filtrado, que permita solamente obtener aquellos "Tweets" que se relacionen con Volkswagen. Por ello se ha elaborado la búsqueda mediante etiquetas, es decir; las menciones o comentarios que se han realizado a raíz de la relación que tengan con la empresa alemana. Estas etiquetas o "hashtags" se han seleccionado con el fin de solamente encontrar menciones que hagan referencia únicamente a Volkswagen a través de sus "hastags oficiales". Se detallan a continuación:

- #volkswagen
- #vw

De esta manera se obtendrá mediante filtrado todos los comentarios realizados que contengan estas dos últimas, desechando de esta forma todos los demás.

El siguiente paso es delimitar la búsqueda dentro del espectro temporal. Esto es sencillo, ya que simplemente hay que añadir el rango de fechas que interesa analizar, (17/09/2015-31/12/2015). Una vez introducidos todos los parámetros en los patrones de búsqueda, se procede a realizar el filtrado. Una vez hecho esto, la primera información que se obtendrá, no es otra que los tweets realizados en esas fechas y con esos hashtags, aunque no se proporcionan todos de una vez. El sistema para obtener el total, es conseguir en el navegador web que se utilice, la posibilidad de cargar todos los datos. En la mayor parte de las ocasiones no es así, dadas las limitaciones del navegador, imposibilita extraerla a un soporte manejable. Por ello, se han de realizar repetidas búsquedas en espacios temporales muy cortos. Esto supone una de las limitaciones de este tipo de búsqueda, ya que es un proceso bastante lento. Posteriormente se formatean los datos a unos archivos de texto. De esta forma se han obtenido 8 ficheros pertenecientes a los meses y los años que se realizó la búsqueda; (correspondientes al período septiembre-diciembre 2015 y 2014). Concretamente se han extraído un total de 80.562 comentarios de 2015 y 35.209 en 2014, dando un total de 115.771 tweets obtenidos para el análisis.

5.3.2. Preparación del texto

Una vez los datos se encuentren contenidos en los ficheros (se han escogido archivos de procesamiento de texto tipo "texto elemental" identificados con la extensión .txt) es necesario realizar un nuevo filtrado, ya que no es posible tratarlos para su análisis sin antes organizarlos, identificando y separando cada uno de los Tweets individuales. Para ello se ha utilizado un "script" de R que permite diferenciar automáticamente el inicio y el final de cada Tweet.

A continuación se detalla un ejemplo de cómo se encuentran los datos hasta este punto:

Lee M Kelsall @LeeMKelsall 3 nov. 2015
My Father in Law's Beetle is Mega !!!!
#beetle #volkswagen #volkswagenbeetle
https://instagram.com/p/9pFQjlsODo3Z3EJ9xUzywgNjLvVzh0tUPEZOM0/ …
0 retweets 2 me gusta
Responder Retwittear
Me gusta 2
Más
Lavada Chim @Lavadaxs96 3 nov. 2015
#Volkswagen also lied about its gas-powered cars http://dlvr.it/Cdv17B
1 retweet 0 me gusta
Responder Retwittear 1
Me gusta
Más

Se observa que los Tweets empiezan de esta forma "*Lee mKelsall @LeeMKsall 3 nov.2015";* es decir indicando el nombre del usuario, seguido del nombre de la cuenta precedido por un "@", a continuación de la fecha de publicación. Otro dato importante es que el contenido del mensaje finaliza siempre en todos los casos en la palabra "Más". Teniendo en cuenta estas dos características, y mediante el Script de R, se preparó el texto, obteniendo a partir de él los Tweets individualizados. En algunas ocasiones, no ha funcionado de forma correcta, y no todo se procesó adecuadamente, lo que obligó a realizar ese trabajo a mano. A través de otro Script, se extrajeron del texto procesado anteriormente las variables que podrían interesar para el estudio, que se citan a continuación:

- "fecha"
- "usuario"
- "nombre de la cuenta"
- "texto del tweet"
- "nº de retweets"
- "nº de me gusta"

Se recogió toda esta información, formando una base de datos en formato.csv (coma separated value), es decir, "datos separados por comas". A continuación, se realizará un nuevo filtrado, en este caso por palabras clave, constituyendo esta parte el análisis de sentimiento.

5.3.3. Detección y clasificación de sentimientos

Se ha empleado para este fin un léxico de opinión. Es un listado que contiene palabras "positivas" y "negativas" (Minqing Hu & Bing Liu 2004)[8]. Su funcionamiento consiste en clasificar en estos dos grupos todas aquellas que coincidían en uno de los listados. Esta clasificación se realizó para cada tweet lo que servió para clasificar en función del tipo de palabra más frecuente en cada uno de ellos. Por la tanto se obtienen las siguientes variables:

- "número de tweets positivos"
- "número de tweets negatios"
- "número de tweets neutros"

Las variables anteriores se han organizado en diferentes bases de datos en función del espectro temporal requerido en cada caso; practicándose el desglose en días y semanas.

5.3.4. Presentación de resultados

Una vez obtenidas todas las variables, se decidirá qué herramientas se usará para proceder a su análisis. El hecho de que los datos se encuentren distribuidos en series temporales, sugiere que se empleen gráficos de estadística descriptiva para observar sus patrones de una forma más visual. Concretamente se han utilizado gráficos de línea y de barras. De esta forma, se apreciará las tendencias y posibles puntos de inflexion, además de momentos influyentes que han tenido las variables, y extrapolándolas (en algunos casos) para la explicación de las demás variables. Se han tomado datos absolutos además de tasas de variación y porcentajes para la elaboración de los mismos.

Uno de los datos muy importantes que nos permite obtener el análisis consiste en qué palabras son las más repetidas, y cuál es el peso que tienen sobre el resto. Para ello, se recurre a un gráfico llamado nube de palabras. Mediante su utilización se puede observar de una manera muy sencilla, cuáles

[8] El listado al que se hace referencia ha sido elaborado por el profesor Bing Liu de la Univerisad de Illinois en Chicago (UIC) profesor de Ciencias de la Computación. El link es el siguiente: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

son las palabras que más trascendencia han tenido en el conjunto del período analizado. Permite tener una aproximación bastante efectiva de qué sentimientos y reacciones se han desarrollado en relación a Volkswagen durante la crisis. Y no solo eso, además pueden servir de ayuda para identificar determinadas actuaciones (positivas o negativas) que ha desarrollado la empresa, o incluso identificar por las opiniones de los usuarios, a sus segmentos de mercado. Las nubes de palabras en este trabajo se han realizado para la totalidad de los períodos en cuestión, tanto en 2015 como 2014.

Finalmente una de las cuestiones que ha suscitado el análisis gráfico es, si se podría apreciar alguna relación (correlación) de las variables seleccionadas con algún agente externo a Twitter, o incluso dentro de las propias variables recogidas íntegramente de esta red social. La variable externa recopilada en este caso, ha sido la cotización de Volkswagen durante el período que abarca desde el 17/09/2015 al 31/21/2015. Su obtención ha sido muy sencilla, ya que actualmente existen a disposición del público, multitud de portales que disponen de información al ámbito bursátil. Concretamente se han recopilado estos datos de XETRA (Exchange Electronic Trading), que es la plataforma electrónica de negociación de la bolsa de Alemania de Frankfurt (DAX). Para este fin se ha recurrido a una comparación mediante regresión.

5.4. Análisis de los datos

En este epígrafe se procederá al análisis gráfico a través de las variables obtenidas en las etapas anteriores. La razón por la cual se han obtenido datos del mismo período en dos años consecutivos, no es otra que la de realizar un estudio comparativo entre ámbos.

En el primer período, que corresponde a partir del 17 de septiembre de 2014 hasta el final de ese mismo año, la opinión de los usuarios que en Twitter se ve reflejada puede concluirse como bastante positiva. Para empezar, se parte de un gráfico de líneas, donde se reflejan por día, los tweets positivos, negativos y finalmente los neutros:

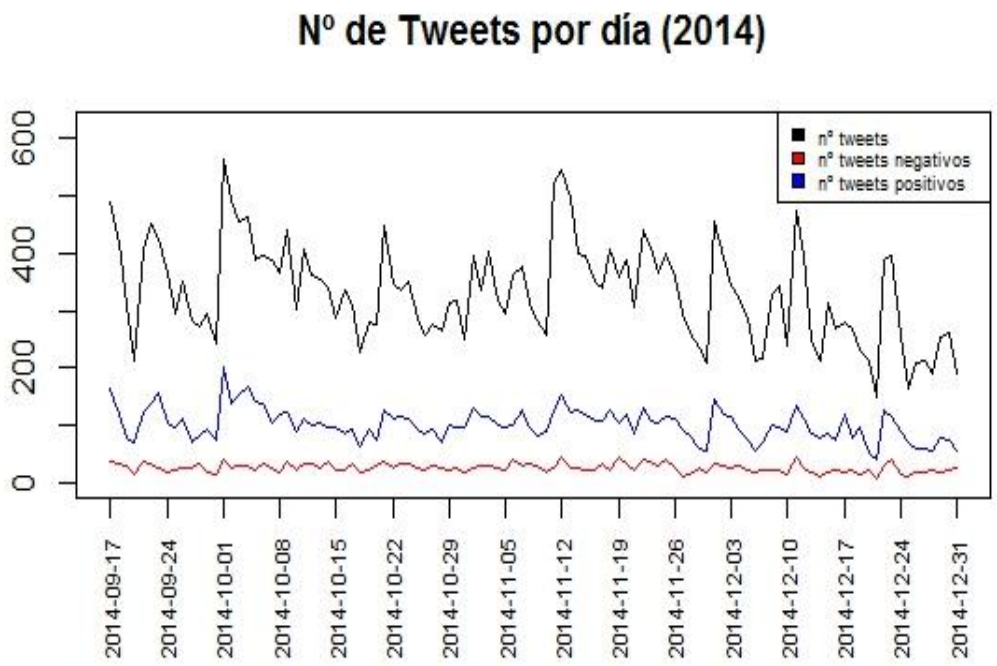

Figura 3: Caso Volkswagen, Tweets por día. (2014)
Fuente: Elaboración propia a partir de los datos recogidos en Twitter.

La tendencia es clara en todo el período. Los tweets positivos superan a los negativos siempre. Se puede observar a partir de los datos que la percepción que tienen los usuarios de esta red social sobre Volkswagen es más positiva que negativa. El hecho de que la línea de positivos tenga cierta tendencia a los "dientes de sierra" es explicada por el hecho de tratarse de cifras absolutas y no relativas. Es por ello que cuando el número total de comentarios (representado en negro, donde se recogen los Tweets positivos, negativos y neutros) presenta algún pico o caída, también lo presenta los positivos, esto quiere decir que cuando se publican más comentarios, aumentan significativamente los positivos más que los negativos, siendo los negativos, poco variables en relación al número total de tweets, junto con los positivos. El volumen de publicaciones asciende en este período a 35.209. No llegando a superar en ningún, el volumen total de comentarios a los 600 por día, estableciendo en este punto, su máximo. Por otra parte el mínimo de comentarios realizados por día es de menos de 200. Así mismo no hay demasiados hechos destacables que hagan una especial incidencia en la opinión de los usuarios de Twitter, los picos de comentarios positivos, normalmente en este caso, se deben en gran medida al trabajo de los "Community Managers" de Volkswagen y la incidencia positiva que hayan podido tener los diferentes Tweets de usuarios aislados, que pudieran tener cierta relevacia, o una combinación de ámbos factores.

Si se repite el análisis, (pero esta vez en datos relativos); es decir representando gráficamente los porcentajes de los comentarios, tanto positivos

como negativos, las conclusiones son idénticas. Para ello se ha recurrido de nuevo a las líneas temporales, por días. Se aprecia que el porcentaje positivo respecto del negativo es claramente superior, encontrandose los primeros entre el 20% y 40%, llegando incluso a sobrepasarlo. Mientras que los negativos en ningún caso supera el 20%, no alcanzando ni tan siquiera a rozarlo (Fig. 12). También se ha hecho una medición a través de gráficos de barras, donde se han tomado datos semanales en lugar de datos diarios, ya que debido a la gran cantidad de días que se recogen, no sería práctico debido a la infinidad de barras que se generarían, además de ser confuso.

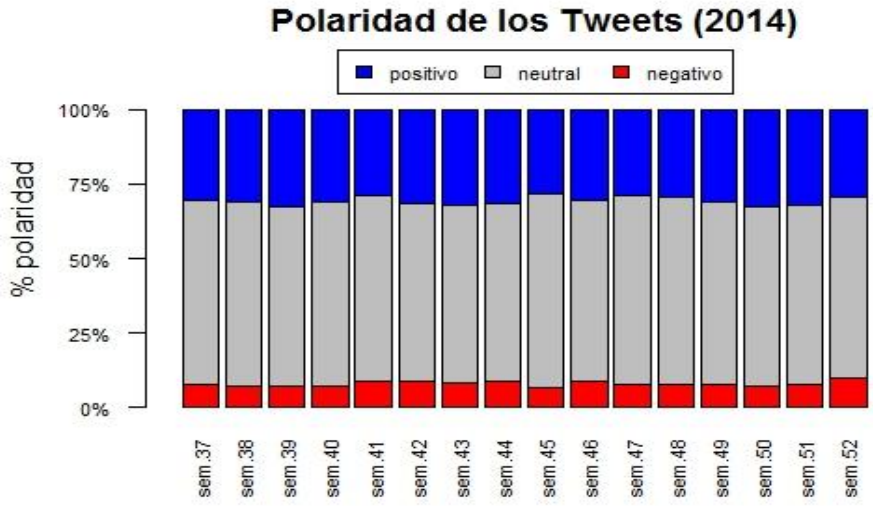

Figura 4: Porcentaje de polaridad de Tweets (2014)
Fuente: Elaboración propia a partir de datos recogidos en Twitter.

Nuevamente, el análisis semanal, y en porcentaje de Tweets positivos y negativos, se concluye que no ha habido demasiadas variaciones de los datos que se representan, indicando una prevalencia de las publicaciones positivas sobre las negativas en porcentaje sobre el total. Además se ha incluido una nueva variable para este tipo de gráfico, los tweets neutrales en porcentaje sobre el total. En una primera pasada, es significativamente superior a los negativos y positivos, siendo su explicación muy simple. Al no existir grandes cambios, y no encontrarse Volkswagen en un contexto adverso en tanto a la propia imagen de marca que ofrece la compañía, ni tampoco indicar, como se ha visto anteriormente, que existan malas opiniones en general , a la vista de los datos obtenidos de Twitter, el alto porcentaje de comentarios neutros se debe en gran medida a las publicaciones y menciones que han hecho los usuarios aislados e independientes sobre la empresa. Éstos se constituyen por ejemplo de fotos subidas de los propios usuarios mostrando algo relacionado con Volskswagen, tales como sus propios utilitarios, como menciones inocuas a la referida impresa, a ojos del programa que ha procesado las publicaciones.

En este punto, es posible conocer los sentimientos y opiniones que ha suscitado la empresa Volskwagen en la red social Twitter. Se ha podido observar que son predominantemente positivos y neutros, siendo poco

representativo el volumen de negativos. La siguiente cuestión del análisis es qué se ha dicho, qué palabras han usado los usuarios para referirse a este tema en cuestión, y en qué medida destacan unas sobre otras, además de la propia naturelaza de las mismas, en tanto a qué motivos relacionados con Volkswagen, hacen referencia. Para ello se dispone de un nuevo tipo de gráfico: la nube de palabras.

Para su elaboración se han hecho varios cortes temporales, con el fin de averiguar qué es lo que de decía exactamente en cada período y con qué frecuencia. Las nubes se han realizado a partir de los datos por semanas, del número de palabras positivas, negativas y neutras. En un primer momento se ha seleccionado para, todo el período de los meses de 2014, es decir; todas las semanas desde septiembre a diciembre.

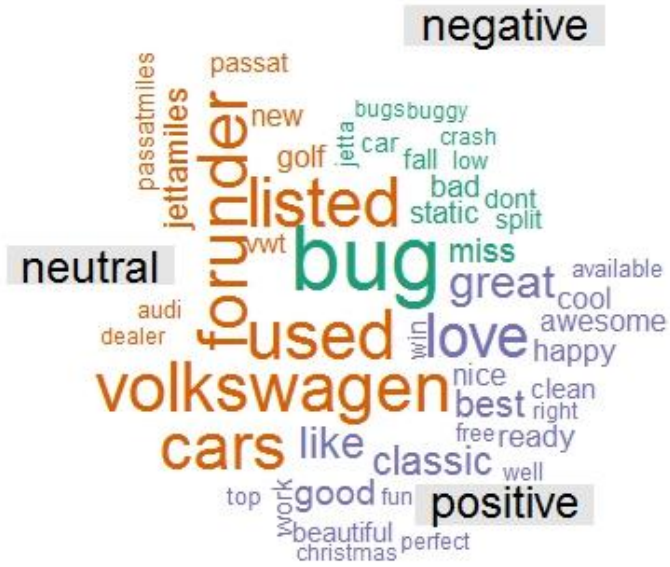

Figura 5: Nube de palabras, 50 más repetidas (2014). Total periodo analizado.
Fuente: Elaboración propia a partir de datos de Twitter

Se han seleccionado las 50 palabras que más se repetían. En este tipo de gráfico las más grandes son las que más frecuencia tienen. Nuevamente, el número de palabras positivas y neutras, es claramente superior a las negativas. Se detecta la escasa significancia de las palabras negativas. Analizando su significado, es posible extraer varias conclusiones. Primeramente las palabras positivas y neutras, que hacen referencia a los aspectos más positivos que pueda ocasionar Volkswagen; en tanto a su linea de productos, como a los sentimientos que ellos desprenden. Las palabras que más se repiten en este caso son "love", "best", "great", "volkswagen". Los usuarios tienen una buena opinión en general y lo expresan de forma muy abierta con estas palabras que son muy significativas. Tanto usuarios que sean

clientes, como los que no lo sean, si Twitter se tratase de un mercado, parece que a nivel global se encuentra contento y satisfecho. La impresión que reflejan estas opiniones, es predominante positiva. En lo referente a los aspectos negativos, las conclusiones que permiten extraer son que los descontentos, se deben a los clásicos problemas que puede enfrentar el utiliario común, o simplemente, los relacionados con un trato en ocasiones deficiente al público al que va destinado. Observando la nube, destaca la palabra "bug" entre las negativas, que significa "error" en inglés, aunque no se ahonda mucho más en la naturaleza del mismo, no se observa que ninguna de las palabras listadas puedan dar pistas sobre la procedencia de los mismos. Esto indica que su procedencia será muy variada, y las causas estén muy diluidas dentro de los datos extraídos, y por lo tanto aportando muy poca significancia.

Los demás cortes temporales se han hecho cada cinco semanas. Las similitudes con la nube que contiene las de todo el período son muy grandes, ya que las tendencias son muy similares, sin grandes cambios. Sin embargo se ha realizado una nueva nube, que solo contiene 10 palabras, (las más repetidas), que recoge el período total. Lo que muestra es que de diez palabras, solamente una es mala, siendo ésta "bug", vieja conocida en este tipo de gráficos, 6 palabras son neutras y 3 son positivas, y muy significativas; "love", "great" y "like". Nuevamente se reprueba la baja significatividad de las palabras negativas, concluyendo así, que la red social Twitter, a través de las opiniones de sus usuarios ha tratado a Volkswagen de una manera bastante afectiva en el período comprendido entre septiembre y diciembre de 2014.

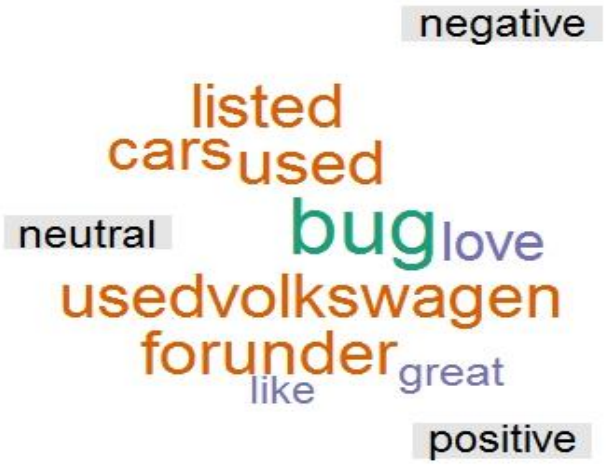

Figura 6: Nube de palabras, 10 más repetidas (2014). Total periodo analizado.
Fuente: Elaboración propia a partir de datos de Twitter.

El siguiente punto pretende analizar el período correspondiente al espacio que abarca de septiembre a diciembre de 2015. Para ello se usará el

mismo tipo de técnicas que en 2014, aunque añadiendo otra variable, la cotización de la acción de Volkswagen.

El gráfico de líneas, que representa en valores absolutos el número de tweets por días tanto positivos como negativos, además de añadir el total, muestra una forma muy distinta que el del periodo anterior. Para empezar se observa un pico muy grande que comienza a ascender desde el 18 de septiembre. Aumenta el número total de comentarios, tanto positivos como negativos. No obstante es desde este momento donde éstos dos últimos se cortan en el gráfico, y los negativos toman una ventaja muy importante, algo inédito teniendo en cuenta los datos anteriores. El punto máximo de comentarios, tanto positivos como negativos es el día 23 de septiembre, coincidiendo con la asunción de culpa por parte de Volkswagen, y la dimisión de Martin Winterkorn. La situación comienza a normalizarse en la segunda quincena de octubre dónde se aprecia que el número de comentarios negativos se acerca a los positivos, produciendose en algunas ocasiones que el número de positivos supere a los negativos. A excepción de algunos picos de comentarios acaecidos durante finales del mes de octubre y principios de noviembre, la evolución que toman ámbos anteriores es muy similar, ya que dejan de producirse puntos álgidos, y su número es muy parejo. Es a partir de la primera quincena de diciembre cuando al fin, los positivos superan a negativos, retornando definitivamente a la tendencia original, aunque con restos de esta crisis, que se aprecian sobre todo en el volumen de comentarios. Concretamente en este período se han extraído 80.562 Tweets para su análisis.

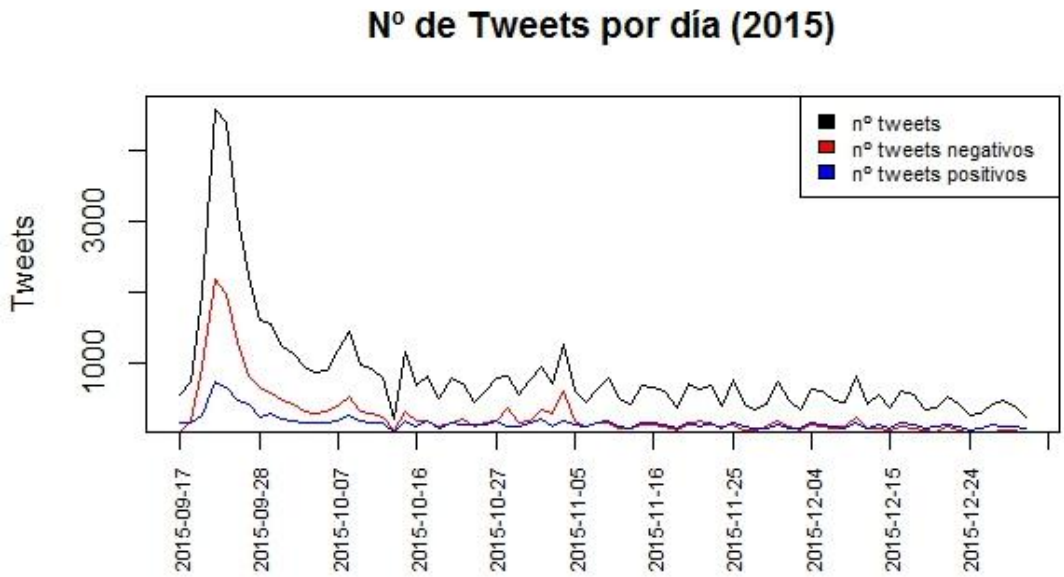

Figura 7: Caso Volkswagen, Tweets por día. (2015).
Fuente: Elaboración propia a partir de datos de Twitter.

El análisis en términos porcentuales del volumen de Tweets revela resultados paralelos. Es significativamente superior el porcentaje de comentarios negativos frente a positivos desde el inicio de la crisis hasta la primera quincena de diciembre. Los máximos han sido en algunas ocasiones superiores al 50% del total, algo nuevamente inusual e inédito. Posteriormente, a final de año se vuelve a retomar el volumen normal de publicaciones, destacando un repunte porcentualmente superior de comentarios positivos frente a negativos.

Mediante un gráfico de barras, es más sencillo observar como los tweets se han distribuido porcentualmente sobre el total. Además es a partir de este punto donde se ha incluido en el análisis la nueva variable para analizar: la cotización de Volkswagen.

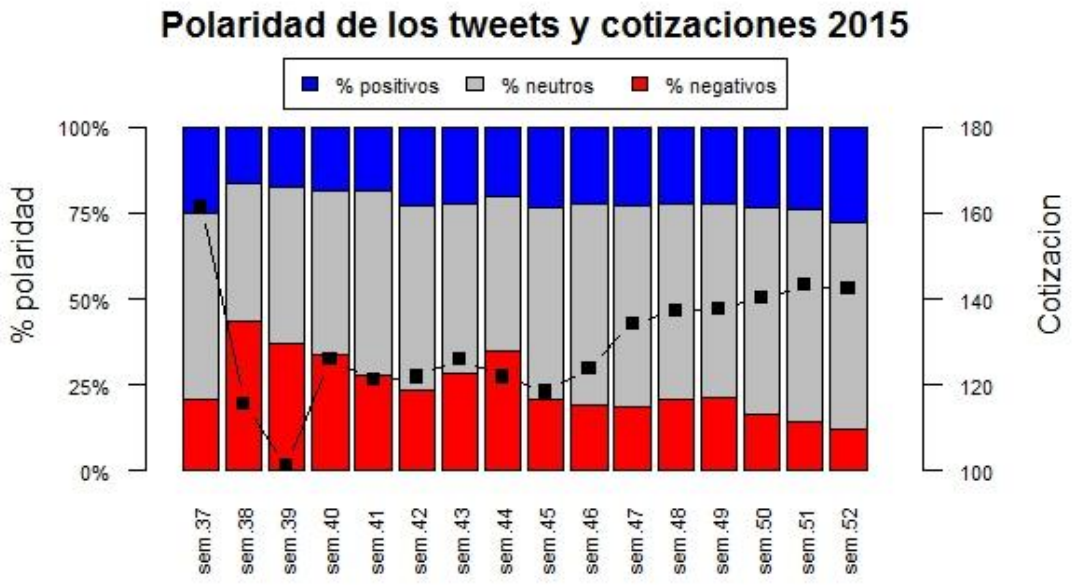

Figura 8: Porcentaje de polaridad de Tweets y cotizaciones (2015)
Fuente: Elaboración propia a partir de datos recogidos en Twitter y XETRA.

La posibilidad de poder añadir esta nueva variable de estudio, sobre el mismo gráfico de barras donde se añade en términos porcentuales el volumen de publicaciones, resulta una introducción del análisis final, que consiste en una regresión. El comportamiento de las cotizaciones resulta curioso si se compara con la cantidad de comentarios publicados y la naturaleza de los mismos. Cuando se origina la crisis, su volumen aumenta de forma muy significativa respecto al mismo período el año anterior. Además durante las primeras semanas, los tweets negativos superan a los positivos. La cotización se desploma enormemente durante estas fechas, pudiendose apreciar una cierta relación inversa entre el porcentaje de publicaciones negativas y los valores de cotización. Cuando este porcentaje aumenta, las acciones caen, y

su comportamiento más errático ocurre en el período donde los comentarios negativos son mayores, si se comparan con los positivos. El valor de las cotizaciones retoma una relativa calma en tanto a su oscilación a partir del momento en que las opiniones negativas empiezan a decaer en términos porcentuales, nuevamente es un dato muy singular, si se relaciona con opiniones de una red social. ¿Existe una relación entre ámbas variables?

Nuevamente se retoma para el análisis un gráfico de barras, esta vez añadiendo la tasa de variación de las cotizaciones. Las variaciones han sido muy negativas a principios de período, coincidiendo con un siginificativamente mayor porcentaje de comentarios negativos. En la mayoría de los casos, el análisis gráfico parece indicar una vinculación entre el porcentaje de publicaciones negativas y la tasa de variación de las cotizaciones. Es muy negativa cuando aumenta el porcentaje de tweets desfavorables, sobre todo al inicio. Posteriormente la evolución de ámbas variables parece indicar a una vuelta a las conclusiones que se han llegado a través de los gráficos anteriores, y no es otra que el retorno a una normalidad de las variables. Las variaciones se tornan mucho menos erráticas y los comentarios negativos pierden posiciones en valores porcentuales a favor de los neutros y positivos.

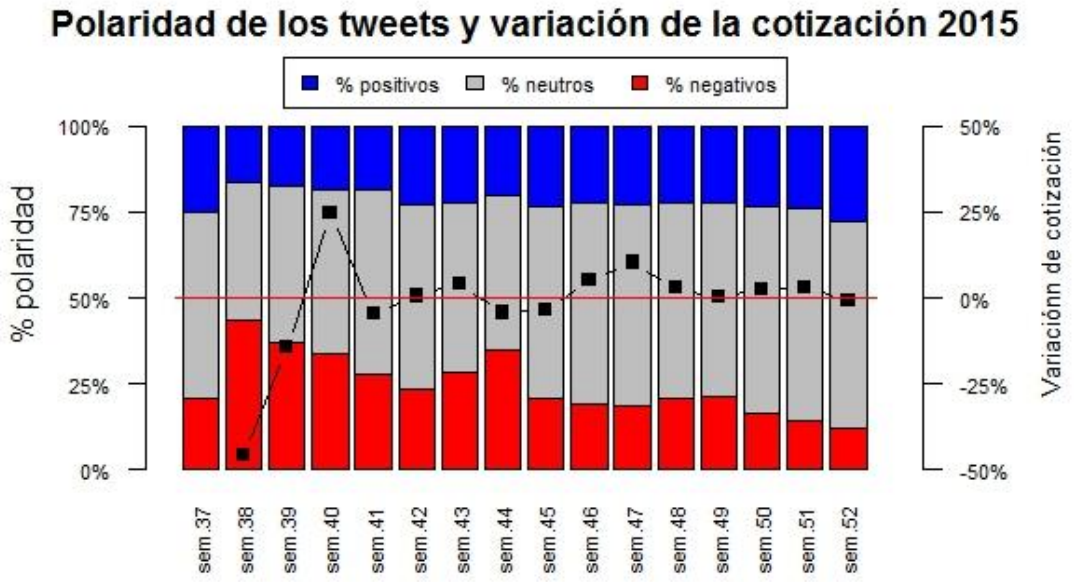

Figura 9: Porcentaje de polaridad de Tweets y tasa de variación de la cotización de Volkswagen (2015).
Fuente: Elaboración propia a partir de datos recogidos en Twiter y XETRA.

Las siguientes cuestiones, son; ¿cuáles han sido las opiniones de los usuarios?, ¿se relacionan con el contexto de la crisis, o han sido efecto de la casualidad?. Nuevamente las nubes de palabras vuelven a tomar relevancia para dar respuesta a estas preguntas. Para ello se ha tomado como primera

aproximación el total del período analizado en 2015. Resalta la presencia de las palabras negativas, y no solo eso ya que su significado hace referencia a los motivos por los cuales esta crisis ha sucedido. Nuevamente se han tomado las cincuenta palabras más repetidas, donde su tamaño indica su menor o mayor frecuencia.

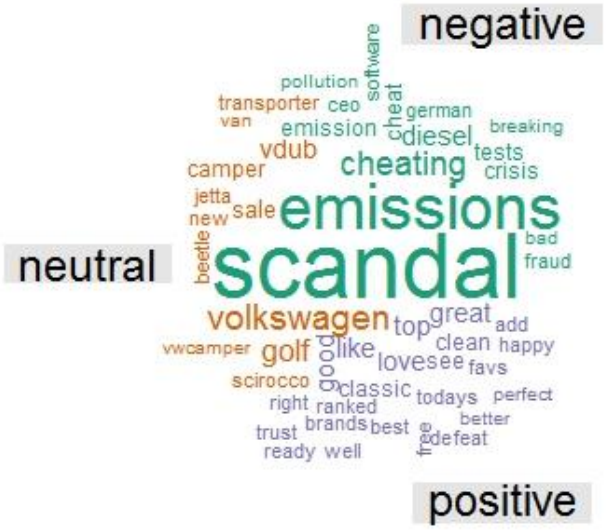

Figura 10: Nube de palabras, 50 más repetidas (2015). Total, periodo analizado.
Fuente: Elaboración propia a partir de datos recogidos en Twitter.

Las palabras que más destacan son "scandal", "emissions" y "cheating", su significado en castellano es "escándalo", "emisiones" y "trucaje". Si tuviera que definirse esta crisis en tres palabras, estás resultarían muy válidas. Esto prueba que las oscilaciones de comentarios negativos frente a positivos en Twitter, y su volumen total, no se debe al simple azar, si no que ha tenido una causa claramente externa, la "crisis de los motores trucados de Volkswagen". Los palabras neutras y positivas, siguen en una línea similar a las del mismo período de 2014, aunque se aprecia un mayor número y variedad, a consecuencia del aumento del número de Tweets. El hecho de que este público, exprese opiniones tan graves sobre la compañía germana, implica un daño muy fuerte a la imagen de marca que han tenido desde siempre. "un coche fiable, un coche alemán". Los utilitarios alemanes siempre han tenido muy buena fama, no sin merecerla en muchas ocasiones, no obstante, el hecho de una gran empresa automobilística germana incurra en este tipo de fraudes, no solo afectará a Volkswagen, ya que éste es un grupo de muchas empresas, entre las que se incluyen, Audi, Bentley, Ducatti, Lamborghini, Porsche o SEAT.

Nuevamente se recurre a otra nube de palabras, pero en este caso solamente se incluirán diez. Los resultados son muy diferentes a los del mismo período de 2014. Se aprecia una significancia mayor de las palabras negativas

frente a las neutras, en tanto a su frecuencia, ya que aunque no sean porcentualmente superiores a las neutras, se repiten mucho más y por eso son más grandes y resalta mucho más. La palabra "scandal" aparece muy grande, siendo un buen calificativo general para el caso analizado.

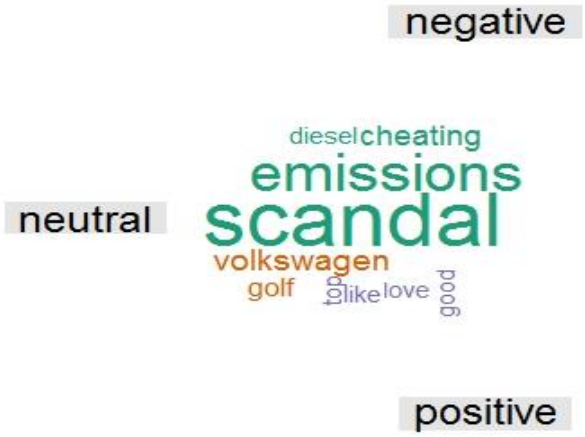

Figura 11: Nube de palabras, 10 más repetidas (2015). Total periodo analizado.
Fuente: Elaboración propia a partir de datos de Twitter.

Para concluir el análisis se retomará una cuestión que surgió a través del análisis gráfico; ¿guarda alguna relación la tasa de variación de la cotización con el porcentaje de comentarios negativos? Podría ser atrevido establecer una afirmación tajante desde el principio, por ello las aplicaciones estadísticas proponen una forma de observar la relación entre ámbas variables y cómo de fuerte ha sido; para este propósito se empleará una regresión; donde la variable explicada será la variación de la cotización y la explicativa el porcentaje de tweets negativos. Para este propóstito se han recogido los datos diarios de ámbas variables, para disponer de un mayor número de datos. Los resultados han sido los siguientes:

```
Call:
## lm(formula = Yb ~ X1b)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.663  -5.396   2.200   5.598   9.792
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.5990     6.7254   2.022  0.06605 .
## X1b          -1.1741     0.2722  -4.314  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.582 on 12 degrees of freedom
## Multiple R-squared:  0.6079, Adjusted R-squared:  0.5753
## F-statistic: 18.61 on 1 and 12 DF,  p-value: 0.001007
```

La primera impresión es que se trata de un buen R^2, para tratarse de datos diarios, con mayores oscilaciones que datos de intervalos temporales más largos. Además la explicativa escogida "Porcentaje de Negativos" se muestra muy significativa. La conclusión es que en este caso el porcentaje de negativos se relaciona negativamente con la variación de la cotización, es decir; se observa a la vista de la regresión que un aumento del porcentaje de Tweets negativos, implica un descenso de la variación de la cotización de Volkswagen. Aumentos unitarios en la variable explicativa implican un descenso de la explicada en -1,1741. A pesar de tratarse de un período de tiempo corto en una situación especial hace que estos resultados no sean aplicables a períodos diferentes. De esta forma no se podría asegurar que la relación obsevada se pueda extrapolar a períodos diferentes o a situaciones más generales.

Conclusiones

La información, bien entendida e interpretada es siempre crucial. Las empresas en concreto están necesariamente obligadas a hacer uso de ella. Para este fin resulta preciso la utilización de todas las técnicas disponibles para su análisis y procesamiento. Se ha visto en los últimos años que las redes sociales han experimentado un auge sin precedentes; tanto es así que grandes como pequeñas coporaciones disponen de cuentas en dichas plataformas, y además también existen otras que su función principal es la de gestionar los perfiles de empresas o asesorarlas en este aspecto. Se han convertido en instrumentos donde es posible (hoy en día) realizar negocio; anuncios, promociones de productos o mejoras de la imagen de marca; estas son pues, algunas de las posibilidades que dichas plataformas ofrecen al mundo empresarial.

Esta cantidad de información, unida a la transmitida por individuos, en muchos casos potenciales clientes, convierte a estas redes en una nueva fuente para la obtención de datos. En el ámbito del conocimiento del mercado y los consumidores, la investigación de mercados proporciona las herramientas necesareas para conocer más a fondo en el ámbito en el que las empresas pretendan ejercer sus actividades. Es en este punto donde la Estadística (a través de sus aplicaciones) facilita todo un abanico de técnicas que se implementan para conseguir y explicar muchas de las conclusiones a las que finalmente se llega cuando se realiza una investigación de mercados.

En este trabajo se ha mostrado como la Estadística provee herramientas para el aprovechamiento de esta nueva fuente, no sólo con aplicaciones tradicionales como gráficas o métodos descriptivos, sino también con técnias elementales de "text mining", más novedosas, que permiten extraer y presentar la información a partir de textos.

Se ha mostrado además cómo se puede manejar de forma simple la información procedente de redes sociales, información que puede resultar de utilidad para las empresas por varios motivos, uno de ellos es -sin duda- la posibilidad de extraer grandes cantidades de información respecto a las

opiniones que profesan los consumidores, sobre un tema en concreto, además de la posibilidad de transformarlos en "medidas de sentimientos" que orienten sobre la situación de un mercado en particular, que concretamente en este caso ha sido la evolución de la crisis de Volkswagen a través de Twitter, en la que se ha observado un paralelísmo entre la evolución de la cotización en bolsa de la empresa y la de los sentimientos expresados a través de las críticas y opiniones en Twitter

Otra conclusión obtenida en el análisis ha sido la gran diferencia entre los términos más mencionados durante la época de crisis y el mismo período del año anterior. En ésta se muestra como el principal término negativo "bug" se relaciona con problemas más cotidianos de una marca de coches, además de ser un número relativamente pequeño de comentarios negativos. En contraste, en la coyuntura de crisis el sentimiento negativo se focaliza en términos relativos al escándalo, produciédose un mayor porcentaje de comentarios negativos que durante la época de normalidad.

Otro de los aspectos a destacar en la realizada labor exploratoria, fue la posibilidad de extarer, de forma gratuíta y sin necesidad de realizar las clásicas encuestas (que hubieran llevado mucho tiempo), más de 115.000 opiniones. Una posibilidad interesante sería continuar con el estudio estableciendo una desagregación por edades y procedencias geográficas de los usuarios que escribieron los comentarios, ya que de esta forma sería viable establecer perfiles variados respecto de las opiniones.

En definitiva en Twitter se encuentran, datos personales, opiniones y procedencias geográficas, todos son ingredientes que tanto organizaciones como empresas desean conocer para que de esta forma, poder satisfacer las necesidades del mercado de una manera más eficiente.

Este análisis no pretende demostrar si existe una superioridad (en cuanto a las posibilidades que ofrece esta nueva vía), sobre las encuestas y medios de obtención de información tradicionales, además que las limitaciones que para el investigador académico presenta; por lo tanto, el trabajo se presenta como un canal más para la obtención de información, como un complemento a las técnicas más asentadas en este caso, ya que la cantidad de datos que se alojan en aquellas plataformas es inmensa y puede resultar, por ejemplo, muy atractiva para complementar a los "focus groups", es decir; las entrevistas en grupo, ya que las redes sociales son uno de los mayores medios de intercambio de opiniones y discusión más importantes que existen actualmente.

Finalmente cabe resaltar la relevancia que cada vez más aportan los datos que se encuentran disponibles en internet, pudiendo ser muy variada la gestión respecto a los usos tanto implícitos como explícitos de los mismos. No obstante ¿hasta qué punto resulta lícito obtener las opiniones y datos personales para su aprovechamiento comercial sin que se avise previamente? y sobre todo ¿ a algún usuario de Twitter le habrá (a su criterio) ocasionado algún perjuicio el hecho de que se obtuviesen sus opiniones y algunos datos personales para la elaboración de este trabajo académico?

Sin duda en este nuevo microcosmos quedan elementos tanto para analizar como para aprender y sobre todo, por explotar.

Bibliografía

Artero, M. y Marcos, R. (2014). *Extracción, análisis y visualización de información social desde Twitter.* (Trabajo fin de grado, Universidad Complutense de Madrid). Recuperado de: http://eprints.ucm.es/26486/.

Blázquez Resino, Juan José. (2014). Procedimientos de muestreo y tamaño de la muestra. En Águeda Esteban Talaya, Arturo Molina Collado, (coord.). *Investigación de Mercados* (pp. 117-142). Díaz Sánchez, Estrella y Martín Consuegra, David. (2014). Técnicas de análisis de datos. En Águeda Esteban Talaya, Arturo Molina Collado, (coord.). *Investigación de Mercados* (pp. 175-188). Molina Collado, Arturo. (2014). Introducción a la Investigación de Mercados. En Águeda Esteban Talaya, Arturo Molina Collado, (coord.). *Investigación de Mercados* (pp.13-27). Madrid: ESIC.

Bugeja, R.,(sin fecha). *Twitter Sentiment Analysis for Marketing Research.* (Tesis, University of Malta). Recuperado de http://staff.um.edu.mt/cabe2/supervising/undergraduate/overview/rachel_bugeja.pdf

C.E.E.I GALICIA, S.A (BIC GALICIA), (2010). *Cómo realizar un Estudio de Mercado.* Recuperado de http://www.bicgalicia.es/dotnetbic/Portals/0/banner/ARCHIVOS/Manuales%20Pymes/3RealizarEstudodeMercado_C.pdf

de Groot, R. (2012). *Data Mining for Tweet Sentiment Classification: Twitter Sentiment Analysis.* Saarbrücken*:* LAP Lambert*.*

Escuder, R. Y Santiago Murgui, J, (1995). *Estadística aplicada. Economía y ciencias sociales.* Valencia: Tirant Lo Blanch.

Instituto Cervantes. (2015). *El Español una lengua viva*. Recuperado de http://www.cervantes.es/imagenes/File/prensa/El%20espaol%20una%20lengua%20viva.pdf

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

Malhotra. N. K. (2008). *Investigación de Mercados* (5ª Ed). México: Pearson Educación.

Malik, Sanjay Kumar, & S. A. M. Rizvi. (2011). *Information extraction using web usage mining, web scrapping and semantic annotation.* Computational Intelligence and Communication Networks (CICN). New Dehi: Indraphasta University New Delhi. Recuperado de http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6112910

Mallou, J. (2003). *Análisis Multivariable para las Ciencias Sociales* (1ª Edición). Madrid: Pearson Education.

Martín-Pliego, Fco. J. (2007). *Introducción a la Estadística Económica y Empresarial: Teoría y Práctica* (3ª Edición) (p. 5). Madrid: AC.

Merino, Mª J., Pintado, T., Sánchez, J., Grande, I y Estévez, M. (2010). *Introducción a la Investigación de Mercados*. Madrid: ESIC.

Minking, H. & Bing, L., (2004). Mining and sumarizing Customer Reviews. En Procedings of the ACM SIGKDD, *International Conference on Knowledge Discovery and Data Mining,* (KDD-2004), Aug 22-25, 2004, Seattle. Recuperado de http://delivery.acm.org/10.1145/1020000/1014073/p168-hu.pdf?ip=193.144.61.240&id=1014073&acc=ACTIVE%20SERVICE&key=DD1EC5BCF38B3699%2ED6A91496E324B249%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=628567868&CFTOKEN=38318364&__acm__=1465578679_26404c3b7db79b0e91de14c13500f327

Nieto, M., (2015). *Engagement de contenidos: Social Data y prensa on-line. El Caso de Volkswagen* (Social Bussines Analytics/2015). Madrid: Instituto de Ingeniería del Conocimiento.

Pereira, I. S., (2014). *A era de um mercado social: a relaçao entre o Twitter e o mercado accionista.* (Tesis de máster, Universidade Nova de Lisboa). Recuperado en: https://run.unl.pt/bitstream/10362/14539/1/TGI0026.pdf

Pérez, C. (1999). *Técnicas de Muestreo Estadístico* (1ª Edición). Madrid: RA-MA

Rambocas, M., Gama, J.,(2013). Marketing Research: The Role of Sentiment Analysis. *Research Work in Progress*, (489). Recuperado de http://wps.fep.up.pt/wps/wp489.pdf

Russell, M., (2014). *Mining the social web.* (2ª Ed.). Sebastopol CA, EE.UU: O´Reilly Media.

Trespalacios, J.A.,Vázquez, R. y Bello, L., (2005). *Investigación de Mercados: Métodos de recogida y análisis de la información para la toma de decisiones en Márketing.* Madrid: Thomson.

Velo, I., (2009). *Archivo de la Colegiata: Clasificación e Inventario guía* (p. 27). A Coruña: Pernas.

Vergara, A., (2015). *Investigación de Mercados: Estudio del Público Objetivo de Smartphone.* (Trabajo fin de grado, Universitat de Barcelona). Recuperado de http://diposit.ub.edu/dspace/bitstream/2445/66710/1/TFG-ADE-Vergara-Alejandro-juliol15.pdf

-Hemerotecas:

Cinco días: http://cincodias.com/

BBC: http://www.bbc.com/

Anexo

Anexo A: 2014

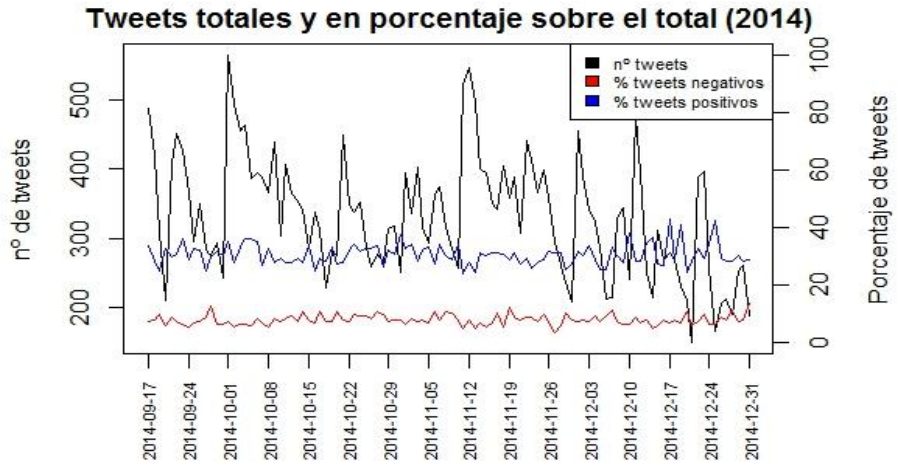

Figura 12: Tweets totales y en porcentaje sobre el total. (2014)

Fuente: Elaboración Propia a partir de datos de Twitter.

Figura 13: Nube de palabras, semanas 37-41 (2014)

Fuente: Elaboración propia a partir de datos de Twitter.

Figura 14: Nube de palabras, semanas 42-46 (2014)
Fuente: Elaboración propia a partir de datos de Twitter.

Figura 15: Nube de palabras, semanas 47-52 (2014)
Fuente: Elaboración propia a partir de datos de Twitter.

Anexo B: 2015

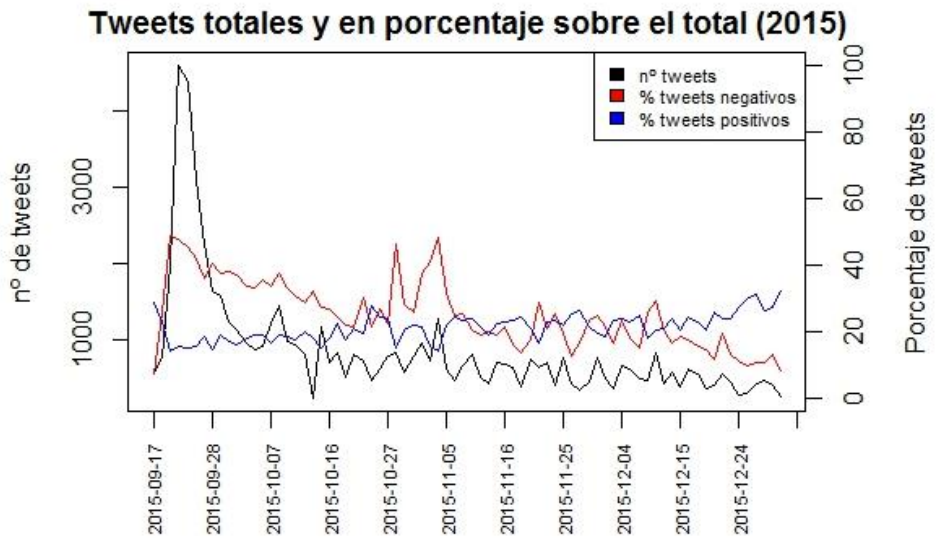

Figura 16: Tweets Totales y porcentaje sobre el total (2015)
Fuente: Elaboración propia a partir de los datos de Twitter.

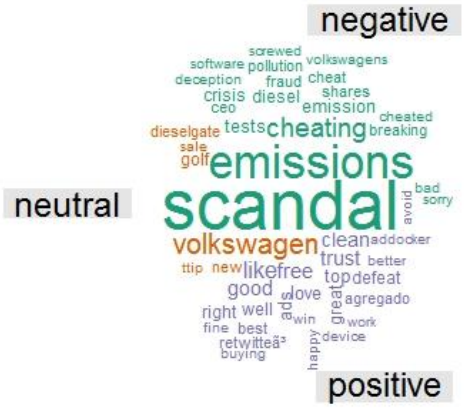

Figura 17: Nube de palabras, semanas 37-38. (2015)
Fuente: Elaboración propia a partir de datos de Twitter.

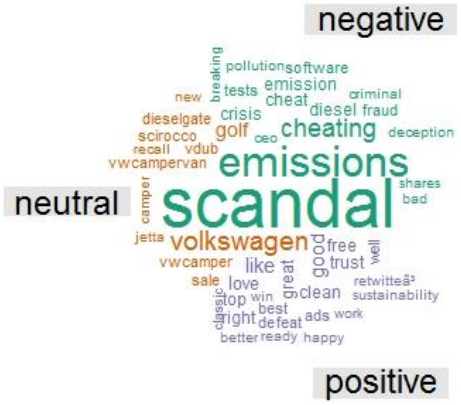

Figura 18: Nube de palabras, semanas 37-41. (2015)
Fuente: Elaboración propia a partir de datos de Twiter

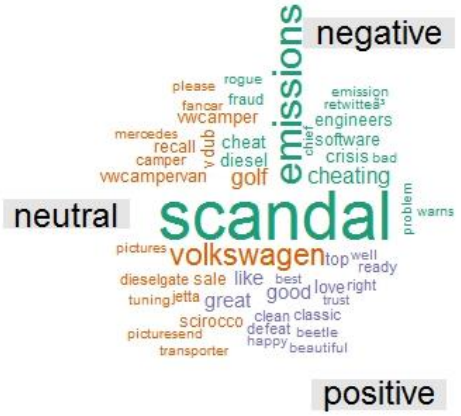

Figura 19: Nube de palabras, semanas 42-46. (2015)
Fuente: Elaboración propia a partir de datos de Twitter.

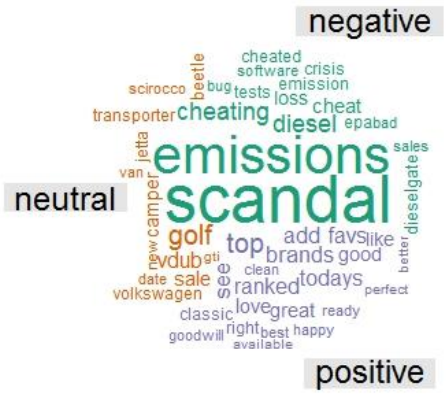

Figura 20: Nube de palabras, semanas 45-46. (2015)
Fuente: Elaboración Propia a partir de datos de Twitter.

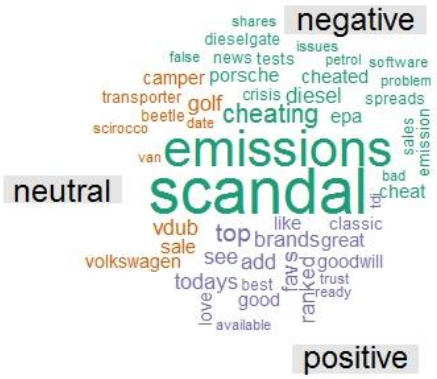

Figura 21: Nube de palabras, semanas 44-45. (2015).
Fuente: Elaboración propia a partir de datos de Twitter.

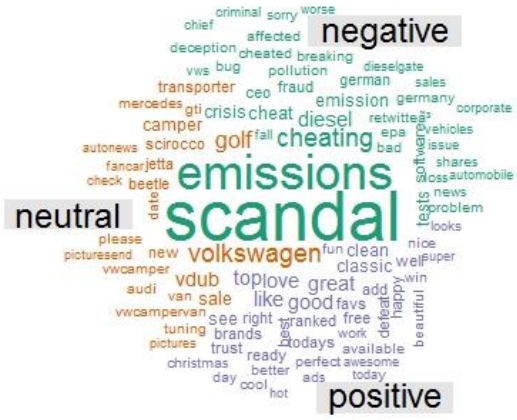

Figura 22: Nube total, 100 palabras (2015)
Fuente: Elaboración propia a partir de datos de Twitter.

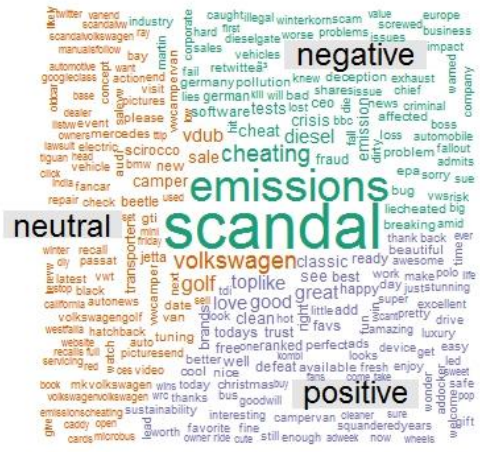

Figura 23. Nube total, 500 palabras (2015)
Fuente: Elaboración propia a partir de datos de Twitter.

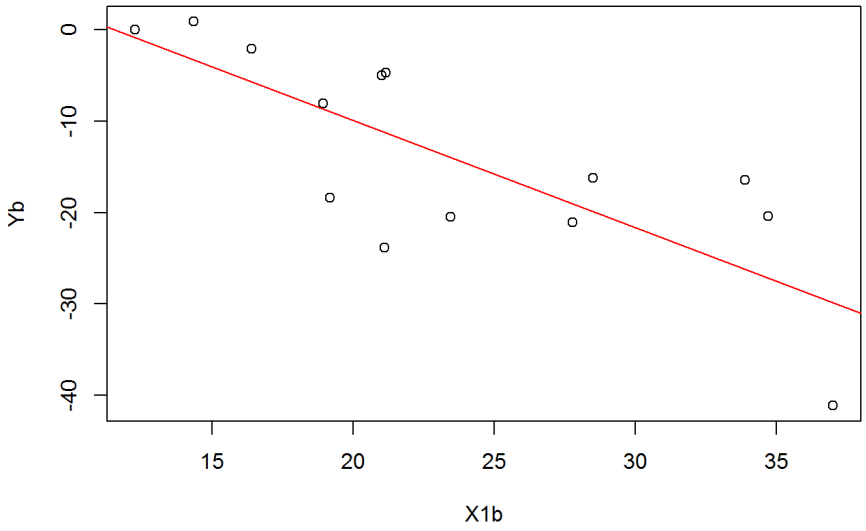

Figura 24: Gráfico de regresión. Variación de la cotización y porcentaje de Tweets negativos.
Fuente: Elaboración propia a partir de datos de Twitter.