Universidad Zaragoza
1542

Trabajo de Fin de Grado

ESTIMACIÓN Y VALIDACIÓN DE UN MODELO ECONOMÉTRICO PARA FILTRADO DE MENSAJES BASADO EN LA MINERÍA DE TEXTOS Y ANÁLISIS DE OPINIONES EN REDES SOCIALES

Autora

Laura Cruz Ramírez

Director

Alberto Turón Lanuza

Facultad de Economía y Empresa / Universidad de Zaragoza

Curso: 2019-2020

Índice

Introducción

El presente Trabajo de Fin de Grado nace del conocimiento como estudiante del grado de Marketing e Investigación de Mercados, de la importancia de la investigación y el análisis adecuado de los distintos mercados que existen hoy en día. La información es una herramienta muy útil para todas las empresas no obstante, no es posible sacarle provecho o utilidad sino se sabe analizar ni entender correctamente dicha información. Una adecuada búsqueda de información y análisis de la misma puede suponer un efectivo plan de marketing con el que la empresa mejore su posicionamiento y realice estrategias más eficaces a los distintos segmentos a los que se dirija.

Las técnicas de investigación de mercados se pueden clasificar según el tipo de información utilizada, la naturaleza de la información, el horizonte temporal y según los objetivos y finalidad. Estas técnicas son múltiples y diversas; la estadística es la herramienta más utilizada para el análisis de datos (Ana Garrido, 2018).

En este proyecto, el objetivo es explicar la utilidad y modo de empleo de una de tantas técnicas existentes. Considero que haber escogido la línea de Sistemas de Soporte de Decisión y Gestión del Conocimiento es una muy buena oportunidad para plasmar y ejemplificar la gran utilidad de estos métodos. A lo largo de todo el grado he cursado distintas técnicas de investigación, las cuales puedo clasificar como convencionales, desde la realización de entrevistas en profundidad o encuestas, hasta la obtención de información mediante la observación y medida. Me supone un reto motivador aprender el uso de otros procedimientos; en este caso en concreto, he tenido que familiarizarme con una sección de las herramientas y sistemas de decisión que desconocía. El núcleo de este trabajo de fin de grado es la metodología a seguir para el análisis de opiniones de manera que se sepa utilizar esa información a conveniencia de la organización. Concretamente, dicho análisis se llevará a cabo mediante la lectura de distintas opiniones expuestas en una red social. La plataforma de la que haré uso para obtener la información será Twitter.

Las relaciones sociales son las interacciones entre dos o más personas o actores y estas han existido desde siempre. Con el paso del tiempo han surgido diversas formas de interactuar, de manera que hoy en día podemos encontrar en Internet múltiples plataformas que facilitan dicha intercomunicación. Twitter[1], definido como un servicio *microblogging*, es una de las redes sociales más importantes y cuenta con más de 300 millones de usuarios. Es por ello que he seleccionado esta red para el análisis, además de porque es la única que permite el acceso por completo a todos los tweets que lanzan sus usuarios.

Desde la aparición de estos medios de comunicación han estado en un continuo auge, hasta el punto de que incluso se han convertido en una vía de comercialización para la gran mayoría de empresas. Uno de los sectores que está muy presente en las redes sociales, es el sector textil. A través sus cuentas se pueden encontrar, sus catálogos, promociones y opiniones de los usuarios. Este sector protagoniza una estrategia que se ha convertido prácticamente en la más importante, y es la estrategia de la *fast fashion* la cual está basada en el hiperconsumo mediante el constante cambio de existencias con la creación de nuevos productos (Daniela Delgado, 2008).

Así pues, queriendo aprovechar la tendencia de dicha red social junto con la otra gran tendencia de utilizar estrategias *fast fashion* he querido realizar un análisis de los tweets de los clientes de Asos[2], compañía que comercializa moda en más de 200 países y está dirigida principalmente a los jóvenes. El objetivo es realizar dicho análisis desde el punto de vista de la organización con la finalidad de explicar el procedimiento a seguir de manera que sepan usar la información con un propósito de mejorar las estrategias de marketing.

La LO 15/19999, de 13 de diciembre, de protección de datos de carácter personal otorga el derecho a la utilización de datos de carácter personal empleados en el presente trabajo, en concreto gracias al art.4 de la LO citada, el cual expresa lo siguiente: “Los datos de carácter personal sólo se podrán recoger para su tratamiento, así como someterlos a dicho tratamiento, cuando sean adecuados, pertinentes y no excesivos en relación con

[1] http://twitter.com

[2] https://www.asos.com

el ámbito y las finalidades determinadas, explícitas y legítimas para las que se hayan obtenido".

Comentado [MOU1]: Hace unos años en un TFG en el que también se analizaban tweets pusimos este párrafo. Déjalo por ahí por si acaso.

Objetivos

El principal objetivo de este proyecto es desarrollar un procedimiento que rastree las redes sociales con la finalidad de buscar los mensajes publicados en dichas plataformas en relación a una empresa determinada de manera que seleccione de forma automática aquellos a los que los responsables de dicha organización deberían prestar atención.

La explicación está basada en unos pasos a seguir de la herramienta de la minería de textos a través de los cuales se podrán extraer las variables a analizar de manera más profunda. Este análisis en concreto es denominado análisis de opinión y es el determinante para poder examinar la información extraída.

En un plano más específico, en el trabajo se plantean los objetivos de conocer y saber utilizar, en un proyecto de estas características, técnicas de procesamiento del lenguaje natural como la minería de textos o el análisis de opiniones, que pueden aportar información valiosa al proceso de selección anteriormente citado.

Marco teórico

Text mining

"El text mining también conocido como la minería de texto es un área de investigación del procesamiento automático de la información. Se define como el proceso de descubrimiento de patrones interesantes y nuevos conocimientos en una colección de textos, es decir, es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos." (Marcial Contreras Barrera, 2014).

En primer lugar, el primer paso es la determinación del propósito de estudio; por ejemplo, en el caso de la lectura de tweets, se puede identificar geolocalización desde donde se ha escrito el tweet, usuario que emite el tweet, cantidad de seguidores y seguidos del usuario o cantidad de *"me gusta"*, de *"favorito"* o de *"retweet"* del tweet en cuestión,

etcétera. El segundo paso es recolectar, identificar, recoger y validar información; en esta fase se busca identificar las fuentes más relevantes para el objetivo de estudio de la minería de texto de modo que se evalúe su relevancia y se realicen las anotaciones necesarias. Posteriormente, el siguiente paso a seguir es la eliminación de información no útil mediante acciones como el análisis léxico, el tratamiento y separación de palabras vacías (artículos, preposiciones, conjunciones), tratamiento de palabras compuestas entre otras muchas más. La finalidad en esta etapa del *text mining* es poder identificar las palabras clave facilitando la selección de características deseadas. El procedimiento es un tratamiento de limpieza previo al análisis. La finalidad es que el texto quede limpio en un formato en el que únicamente aparezca *palabra espacio palabra espacio*, de manera que pase a ser una variable canónica en el que todos los *tweets* aparezcan de igual forma. A continuación, se procede a la extracción y análisis de las clases, relaciones, asociaciones o secuencias para poder encontrar evidencias de conceptos y de estructuras existentes. Además, en esta fase se obtienen datos que deben ser representados de forma estructurada informáticamente para facilitar el análisis. En último lugar, en esta etapa se presentan los resultados mediante resúmenes y visualización para su información. Existe la posibilidad de almacenar la información procesada en bases de datos de manera que en un futuro se pueda proceder a su recuperación (Marcial Contreras Barrera, 2014).

Las técnicas de minería de textos son utilizadas en múltiples áreas donde los volúmenes de información hacen que sea imposible leer o estudiar todos los artículos. Su uso puede comprender ámbitos como la biomedicina, descubrimientos de fármacos o investigaciones sobre le cerebro humano hasta inteligencias del gobierno. También es una herramienta necesaria para el monitoreo en redes sociales o el análisis de sentimientos u opiniones, utilidad esta última que también concierne al presente proyecto. Para la aplicación de la minería de textos existe una amplia gama de softwares tanto de libre acceso como de tipo comercial. De este último tipo se encuentran vendedores muy conocidos como SPSS, IBM Text Analytics, Autonomy, SAS o Gate.

Métodos estadísticos de clasificación

Un procedimiento muy habitual en la toma de decisiones es la clasificación de alternativas, de manera que sea capaz de valorar la mejor opción. Dado que habitualmete existe un grado de incertidumbre en cualquier decisión, es habitual el uso de la Estadística, concretamente en el campo de la probabilidad, con el fin de cuantificar dicha

incertidumbre. La taxonomía tiene un enfoque multivariante puesto que en la toma de decisiones comúnmente son más de una variable las que se deben de tener en cuenta. Hay distintas técnicas para llevar a cabo la clasificación; entre las más utilizadas se encuentra el *análisis discriminante*.

En este análisis se estudian las técnicas de clasificación de sujetos en grupos ya definidos. Se parte de una muestra de *N* sujetos en los que se han medido *p* variables cuantitativas independientes. El problema del análisis discriminante es que no permite la inclusión de variables con distribución no normal ni variables cualitativas. Dado este problema, surge como modelo alternativo la *regresión logística* (Luis Miguel Molinero Casares, 2002).

Otro modelo de clasificación es el algoritmo de Naive Bayes.

Regresión logística

Como he comentado anteriormente, esta técnica es similar la anterior pero cuenta con ciertas diferencias que le dotan de ciertas ventajas a diferencia del análisis discriminante. En este modelo la variable dependiente o respuesta presenta dos categorías, la ocurrencia y no ocurrencia del acontecimiento definido por la variable, codificándose con los valores uno y cero, respectivamente. Respecto a las variables independientes o explicativas, no existe ninguna restricción, éstas pueden ser cuantitativas o cualitativas. En este último caso se puede crear una variable ficticia por cada uno de los niveles, dándole valor 1 o 0 según s está o no presente el correspondiente rasgo (lógicamente, sólo una de las variables ficticias tomará valor 1).

Así pues, con una variable dependiente definida, el modelo de regresión logística expresa la ocurrencia o no del acontecimiento en términos de probabilidad. Se utiliza la función logística para estimar dicha probabilidad mediante la siguiente formulación:

$$Prob(Y\imath = 1 \mid X) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i X_i}}$$

La variable respuesta Y solo puede tomar dos valores:

$$Y\imath = 1 \; con \; pobabilidad \; p$$

$$Y\imath = 0 \; con \; probabilidad \; 1 - p$$

Donde el valor 1 denota la ocurrencia del acontecimiento y el valor 0 de la no ocurrencia del acontecimiento. La función de probabilidad de este modelo, conocido también como modelo *logit*, garantiza que el resultado de la estimación esté acotado entre 0 y 1. Sin embargo, hay que establecer un umbral a partir del cual el resultado se considere con una probabilidad baja o con una probabilidad alta. De esta manera cuando el resultado no sea ni 1 ni 0, se generalizará su probabilidad en función de si está por encima o por debajo del umbral. Habitualmente, el umbral establecido es 0'5, de esta forma, a todo resultado que sea mayor a 0'5 se le asignará el valor 1, y por lo tanto significará la ocurrencia del acontecimiento. De lo contrario, si el resultado es menor a 0'5 se le asignará el valor 0, por consiguiente, la no ocurrencia del acontecimiento.

Análisis de sentimientos u opiniones

El análisis de sentimientos también conocido como la minería de opinión, según el Director de Marketing de la compañía Bitext[3], empresa especializada en servicios y tecnologías de la información, *"es el proceso por el que determinamos si una frase o acto de habla contiene opinión positiva o negativa sobre una entidad concreta o concepto."* (Manuel Delgado, 2015).

A pesar de no usarse sólo en el ámbito comercial y de las redes sociales, el análisis de sentimientos u opiniones está muy ligado a este sector puesto que permite conocer de manera mucho más sencilla las opiniones que vienen dadas en un texto con un gran volumen de valoraciones o comentarios. Cuestión de gran interés para las compañías puesto que supone la opinión sincera y real del consumidor a diferencia de las encuestas o entrevistas donde el cliente se puede sentir en la obligación de no hablar mal de la marca o con restricción a ellos. Además, el coste de este método se menor que la realización de las otras técnicas de investigación, siendo el número de opiniones mayor que al que se puede alcanzar con dichas técnicas.

El lenguaje natural que es usado en las redes sociales, en su mayor parte, supone un problema para los sistemas de análisis de texto sencillos, puesto que no distinguen oraciones que contengan ironía, comparaciones, condicionales o el uso del lenguaje

[3] https://www.bitext.com

informal. Es por esto el éxito del análisis de sentimientos u opiniones y su continua evolución ya que acentúan cada vez más los múltiples significados de las oraciones con la intención de incrementar la capacidad de otorgar información realmente útil.

Dentro de este campo, se puede clasificar en dos tipos:

- **Detección de polaridad**. Principalmente, es la técnica que decreta si una opinión es positiva o negativa, pero también se puede asignar un valor numérico dentro de un rango determinado de manera que se obtiene una valoración de la opinión más concreta. Se basa sobre todo en diccionarios semánticos. Su inconveniente es que depende mucho de la calidad, tamaño y dominio de los datos de entrenamiento.
- **Análisis del sentimiento basado en características**. Es la capacidad de señalar las diferentes características de un producto formuladas en un comentario. La finalidad es determinar qué tipo de orientaciones definen las opiniones de los consumidores. De esta manera se identifican conjuntos de características en las opiniones de los usuarios de la red social o foro de reseñas, para así proporcionar un informe especificado sobre la polaridad de cada característica (Peñalver Martínez, 2015). Así pues, cabe destacar los siguientes términos:
  - *Corpus*: Se le denomina corpus al conjunto de comentarios. En este caso, el corpus del presente trabajo es la totalidad de todos los *tweets*.
  - *Documento*: Se considera documento a cada comentario, es decir, cada *tweet*.
  - *Término*: En este caso, cada término es cada palabra que aparece en el *tweet*.

En consecuencia, el proceso necesita enfoques basados tanto en el corpus como en diccionarios. Generalmente, el diccionario no da un dominio o un contexto dependiente; por ello es necesario el corpus. Por otro lado, en el enfoque basado en el corpus es difícil encontrar un conjunto muy grande de palabras de opinión; por este motivo el diccionario es útil en esta situación. En esta práctica se necesitan corpus, diccionario y enfoques manuales (Bing Liu, 2011).

Aplicaciones

En este apartado se pretende explicar algunas aplicaciones que ejemplifiquen la utilidad del *text mining* y la gran repercusión que ha tenido en distintos ámbitos, desde el sector científico y farmacéutico hasta el sector comercial o de videojuegos.

Filtrado de correo no deseado (*SPAM*)

Comenzaré comentando uno de las aplicaciones de este método de la minería de texto más conocido, que es la clasificación del *spam*. *Spam* es el correo electrónico no solicitado que se envía a un gran número de destinatarios con fines publicitarios. El hecho de que es un envío de forma masiva supone en la mayoría de los casos una molestia para el usuario del correo electrónico. Dada esta situación, han surgido varias soluciones entre ellas la consideración de este como correo basura mediante el uso de un filtro anti-*spam* basado en el algoritmo de Naive Bayes. Básicamente, este algoritmo utiliza una base de datos en los que aparecen palabras *spam* y palabras no *spam*. Haciendo uso del *text mining*, es decir, analizando el contenido del correo, obtiene los datos con los que determinará si el contenido es o no *spam*. Las principales características de dicho correo es que la dirección no resulta familiar para el usuario además de que no suele aparecer un correo electrónico al que reenviar. No obstante, la característica que más destaca, es el contenido publicitario del mensaje, en el que aparecen listados de productos en venta, ofertas o anuncios de sitios web. (Pedro Echevarría Briones, 2009).

Filtrado de referencias bibliográficas en el área médica

Otro ejemplo que he considerado interesante es el uso de la minería de textos en el área médica. Concretamente, la empresa de salud natural *Swanson*[4] analizó artículos de la base MEDLINE, una de las bases de datos más conocidas y utilizadas en el ámbito de la medicina, que contiene referencias bibliográficas. El análisis de *Swanson* se basó en hallar la relación entre la migraña y la deficiencia de magnesio. Existen múltiples casos en este ámbito, como puede ser el descubrimiento de fármacos o la toxicología predictiva (Marcial Contreras Barrera, 2014).

Aplicaciones al marketing

Para finalizar este apartado, comentaré otro uso de estas técnicas, esta vez aplicadas al marketing. La minería de datos supone hoy en día, para la toma de decisiones en el mercado, supone una herramienta muy útil y estratégica. Cada vez son más las empresas que usan este tipo de técnicas, puesto que estos sistemas favorecen la comprensión del comportamiento del consumidor de modo que se diseñen estrategias de

[4] https://swansoneurope.com

promoción acorde a las necesidades de los clientes. Siempre tratando de lograr el objetivo inicial de toda empresa, que es aumentar su rentabilidad. Gracias al *text mining* en el marketing se han facilitado cuestiones como:

- **Segmentación del mercado**. Se identifican características comunes de los clientes de la empresa, de manera que se facilita la agrupación de dichos clientes en distintos segmentos.
- **Detección del riesgo de pérdida de clientes**. Existe la posibilidad de predecir qué clientes muestran mayor predisposición a dejar la compañía o escoger al competidor.
- **Marketing directo**. Se determina qué perspectivas se incluyen en una lista de correo con el fin de obtener la mayor tasa de respuesta posible.
- **Marketing interactivo**. Existe la predicción de lo que un usuario considera interesante, analizando lo que éste busca en un sitio web.
- **Análisis de la cesta de la compra**. La finalidad es extraer información de qué productos o servicios habitualmente son comprados a la vez.
- **Análisis de tendencias**. Se revela la diferencia de hábitos de un consumidor en función de la temporada.

El uso de estos sistemas inteligentes puede suponer un significativo incremento en las ventas de la empresa además de que también ayuda al cliente a sugerirle u orientarle que producto puede interesarle más de manera que se afianza su compra (Diana Arteaga, 2018)

Metodología

En el presente proyecto el objetivo principal, como ya se ha comentado anteriormente es plasmar la utilidad y el gran provecho que puede suponer para una compañía el uso de la minería de texto, concretamente para el sector del marketing y la investigación de mercados, el análisis de sentimientos u opiniones. Se va a ampliar al caso de un análisis de opiniones de usuarios de Twitter sobre la marcada de ropa ASOS.

Planteamiento del problema y definición del objetivo

La finalidad es que desde el equipo de marketing de la multinacional ASOS sepan cual sería el procedimiento a llevar a cabo en caso de que quieran conocer qué

valoraciones tanto positivas como negativas plasman sus clientes en la red social de Twitter. Si bien es cierto que la información obtenida puede enfocarse en la toma de decisiones de nuevas estrategias, también se puede tomar un enfoque de mejora de las posibles debilidades que puedan percibir sus usuarios.

La red social en la que me centraré será Twitter, mi intención es extraer de manera automática los tweets que tienen relación con la compañía y seleccionar aquellos que merecen la atención del responsable del área comercial y de atención al cliente de la empresa.

Determinación de la población y variables a estudiar

Así pues, la población objeto del estudio se centrará en todos los *tweets* publicados en la red que incluyan términos 'asos' o 'asos_teayuda' ya sea a modo de mención al usuario, de hashtag o como texto simple, puesto que el uso de estos términos es un indicio de que el *tweet* puede estar referido a la compañía.

Las variables en las que me voy a fijar serán:

- *Verificado*: En esta red social existen cuentas verificadas, esto es a modo de comprobación de que la cuenta pertenece a quien dice ser. Esto suele darse en el caso de usuarios con cierta popularidad con el fin de que nadie pueda hacerse pasar por dichos usuarios. Considero esta variable es importante puesto que refleja el nivel de influencia que pueda tener respecto al comentario que haga, ya sea positivo o negativo. En el programa esta variable aparece con el nombre de *verified.*
- *Retweet*: Se le llama *retweet* cuando un usuario copia el tweet de otro en su perfil, teniendo la posibilidad de añadir algún comentario más respecto al original. Un tweet con muchos retweets va a significar un comentario con bastante liderazgo lo cual resulta interesante para la empresa, conocer qué tipos de comentarios son los que más popularidad tienen en dicha red social. En el programa esta variable aparece con el nombre de *retweet_count*.
- *Tweet*: Un comentario breve publicado en Twitter (Diccionario de Cambridge). Esta variable es la que más información proporciona puesto que de aquí se extraen las palabras con valores positivos y negativos, es decir, a partir del texto que aparezca en el *tweet* se procede al análisis de opiniones con la detección de

polaridad. En el *tweet* esta variable aparece con el nombre de *text*. Cabe destacar que esta variable no se analiza como tal en este trabajo; es la más importante, puesto que a partir del *tweet* se extrae la información más relevante, pero la función de esta variable es obtener, a partir del texto que contiene, las variables derivadas de ella, que se a continuación.

- *Puntuación*: Esta variable aparece como *score* y es el resultado de la suma de puntuaciones que yo he decidido adjudicar a distintas palabras clave. Más adelante se detalla el criterio a seguir de las puntuaciones.
- *Análisis de opinión*: Las opiniones son subjetivas; por lo tanto en este tipo de programas existe lo denominado resumen de opiniones, en los cuales se generalizan las opiniones en lo que expresan, ya sea enfado, sorpresa, alegría, miedo… Esta variable clasifica el *tweet* en una de las seis emociones básicas: enfado, alegría, tristeza, miedo, sorpresa o disgusto.
- *Análisis de polaridad*: Además de lo comentado anteriormente del resumen de opiniones, se puede percibir la orientación de la opinión contenida en el mensaje, es decir, a través de análisis semántico del texto se puede llegar a distinguir si una opinión es positiva, negativa o si está exenta de opinión es decir, neutral.

Determinación de las palabras clave

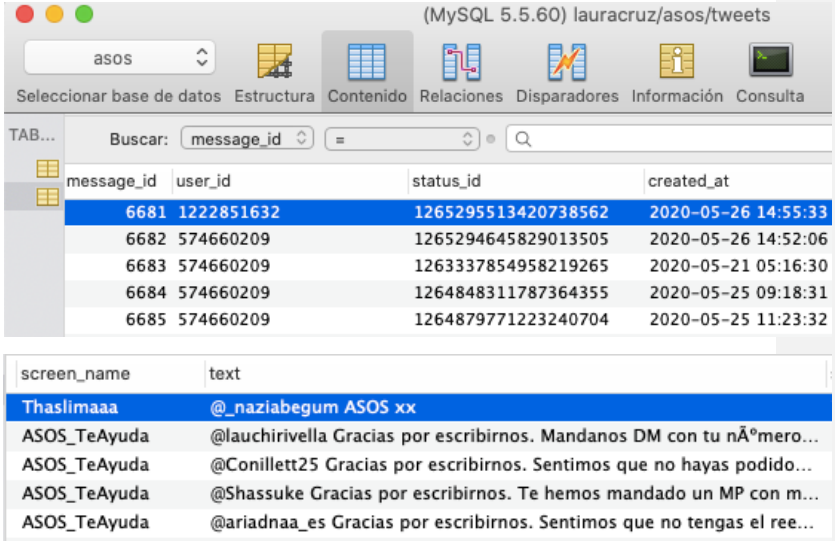
Para proceder al análisis de opiniones, hay que identificar distintas palabras clave que puedan aparecer en el comentario, de forma que posteriormente se le asignará una puntuación distinta a cada palabra en función de su importancia, y se pueda crear una variable en el que aparezcan las distintas puntuaciones de cada *tweet*. Tras analizar los *tweets* que se han recogido en este proyecto, he considerado que de las palabras que más se repiten y a las que se debe prestar especial atención son:

- bien, bueno, buena, bonito, bonita, precioso, preciosa.
- gusta, encanta, favorito, favorita, feliz.
- feo, fea, horrible, caro, cara.
- gracias, por favor, devolucion, pedido, codigo.
- estafa, estafadores, timo, timan, dinero, harto, harta.

Recogida de datos

La recogida de datos, como he mencionado anteriormente, ha consistido en la la extracción de tweets mediante la API de Twitter y seleccionado todos los mensajes que contuviesen las palabras 'asos' o asos_teayuda'. Esta extracción de datos se realizó durante el periodo comprendido desde el 18 de Mayo de 2020 hasta el 12 de Junio de 2020. Los datos fueron almacenados en una base de datos creada en el laboratorio de datos del Grupo Decisión Multicriterio Zaragoza[5], grupo de investigación adscrito a la Facultad de Economía y Empresa de la Universidad de Zaragoza.

A continuación se muestra la manera en que aparecen los datos:



Obtención de variables derivadas

En primer lugar se procedió a la limpieza y estructuración de los textos necesaria para facilitar su análisis. Se eliminaron todos los caracteres no estándares, como tildes, eñes, emoticonos… De esta manera, toda forma que aparece en un formato no estándar, lo sustituye, por ejemplo, donde aparece una "ñ" aparecerá una "n", o donde aparece una

[5] http://gdmz.unizar.es

"ó" aparecerá una "o". También se eliminaron los signos de puntuación, las filas de espacio, los tabuladores… y se convirtió todo en minúsculas.

De la variable *text* se identificaron aquellos mensajes que contenían las palabras consideradas clave, a las que se adjudicó una mayor o menor puntuación en función de su importancia.

Al tratarse de una empresa de moda, realmente los *tweets* que resulten interesantes para la empresa serán aquellos que expresen alguna insatisfacción o queja por parte del usuario o todo lo contrario, aquellos que expresen una satisfacción o alegría del cliente sobre el producto o los productos recibidos.

Así pues, la puntuación asignada a las palabras anteriormente seleccionadas fue la siguiente:

- Puntuación de 1: bien, bueno, buena, bonito, bonita, precioso, preciosa.
- Puntuación de 2: gusta, encanta, favorito, favorita, feliz.
- Puntuación de 3: feo, fea, horrible, caro, cara.
- Puntuación de 4: gracias, por favor, devolucion, pedido, codigo.
- Puntuación de 5: estafa, estafadores, timo, timan, dinero, harto, harta.

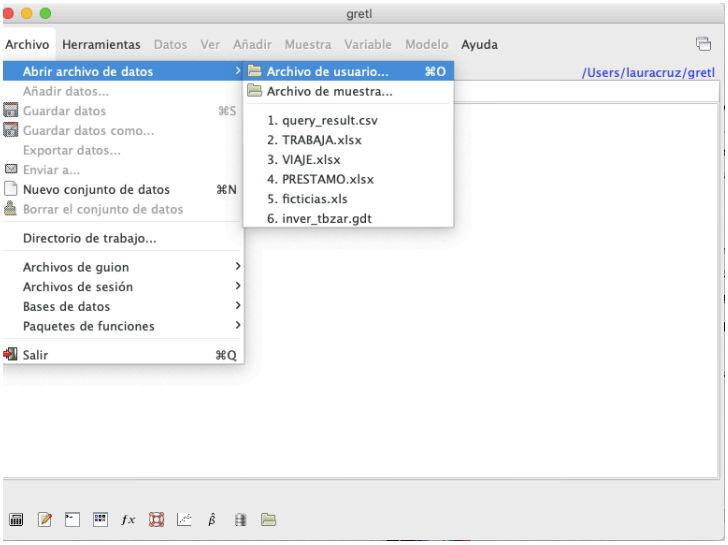
He adjudicado la puntuación de manera que cuanto mayor puntuación tenga el *tweet*, más urge prestarle atención puesto que mayor es la probabilidad de que implique la insatisfacción de algún cliente. Esta variable derivada llamada *score* es el resultado de la suma de puntuaciones adjudicadas a las distintas palabras que aparecen en cada *tweet*. Cabe destacar que si una palabra aparece más de una vez no se sumará la puntuación más de una vez. El criterio de puntuación es que a cuanto mayor insatisfacción exprese el cliente vía Twitter, mayor puntuación tenga el comentario.

Estimación del modelo

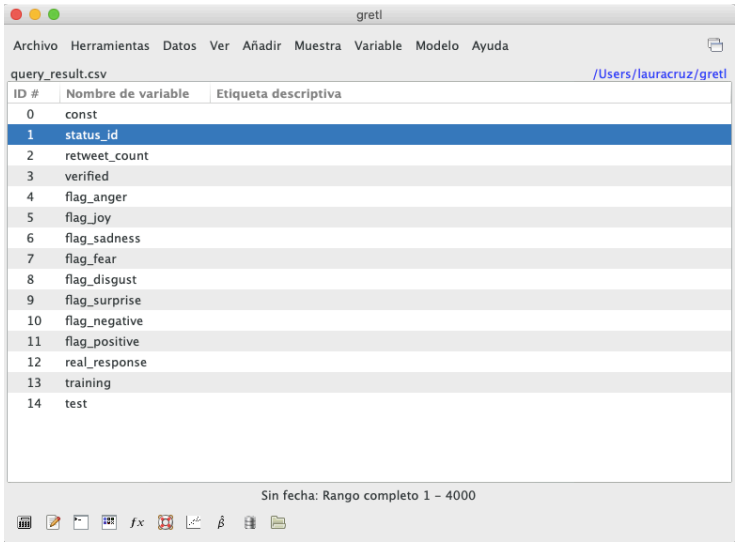
Para la estimación del modelo he utilizado el programa de Gretl , software el cual permite el análisis estadístico y la estimación de modelos econométricos, como es el caso de este proyecto. Se pueden importar datos de distintos formatos como Excel, ASCII, Stata o archivos CSV. En este caso, he importado los datos desde el programa Sequel Pro en un formato CSV.

Dada la muestra de 4.000 *tweets* el procedimiento para comprobar la validez del modelo será, en primer lugar ajustar el modelo y hallar los coeficientes con Gretl con una muestra aleatoria de 2.000 *tweets* y posteriormente con los otros 2.000 comentarios restantes, validarlo. Así pues los pasos a seguir con Gretl, son los siguientes:

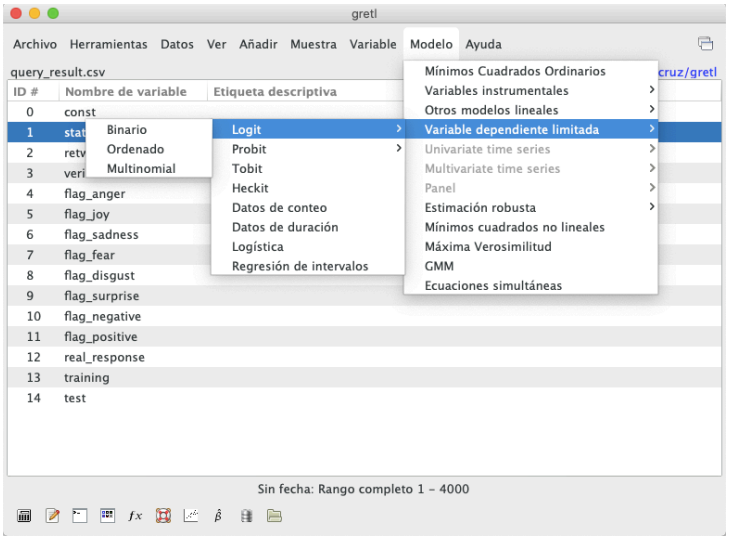
1. Abrir el archivo de datos mediante la secuencia **Archivo / Abrir Datos / Archivo de usuario:**



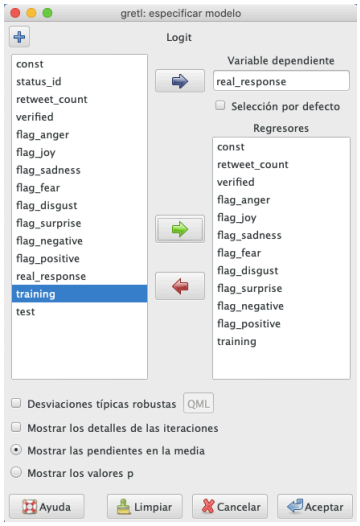
Una vez el archivo esté abierto, se podrá ver de la siguiente manera:



2. Estimar el modelo a partir de la secuencia: **Modelo / Variable dependiente limitada / Logit / Binario**



3. Especificar el modelo Logit para calcular la ocurrencia del acontecimiento: En este caso la variable respuesta es real_response y las variables alternativas son *retweet_coun*, *verified* y las *flag* correspondientes a las variables derivadas del *tweet*. Se selecciona también la variable *training* para que el programa escoja solo los 2.000 comentarios que se han separado.



4. Obtener los coeficientes del modelo:

Archivo Editar Contrastes Guardar Gráficos Análisis LaTeX

```
Nota: Prob(real_response = 0 | flag_surprise = 1) = 1
Quitando  flag_surprise

Modelo 1: Logit, usando las observaciones 3-3999 (n = 2000)
Se han quitado las observaciones ausentes o incompletas: 1997
Variable dependiente: real_response
Desviaciones típicas basadas en el Hessiano
Omitidas debido a colinealidad exacta: training

                coeficiente  Desv. típica      z       Pendiente
  ------------------------------------------------------------------
  const           -1.99259       0.538370   -3.701
  retweet_count   -0.226401      0.105746   -2.141     -0.00265979
  verified         0.285175      0.664369    0.4292     0.00384991
  flag_anger       1.36217       0.227845    5.979      0.0292390
  flag_joy         1.18328       0.476474    2.483      0.0254645
  flag_sadness   -16.9840     5808.88       -0.002924  -0.0125032
  flag_fear        0.106538      0.743206    0.1433     0.00131664
  flag_negative    0.0164790     0.593983    0.02774    0.000194987
  flag_positive   -0.667987      0.537203   -1.243     -0.0104368

Media de la vble. dep.  0.076500   D.T. de la vble. dep.   0.265863
R-cuadrado de McFadden  0.064622   R-cuadrado corregido    0.047964
Log-verosimilitud      -505.3601   Criterio de Akaike      1028.720
Criterio de Schwarz     1079.128   Crit. de Hannan-Quinn   1047.229

Número de casos 'correctamente predichos' = 1847 (92.3%)
f(beta'x) en la media de las variables independientes = 0.012
Contraste de razón de verosimilitudes: Chi-cuadrado(8) = 69.8271 [0.0000]

             Predicho
                0      1
  Observado 0  1847      0
            1   153      0

Sin considerar la constante, el valor p más alto fue el de la variable 6 (flag_sadness)
```

Resultados

Validación del modelo

Para proceder a la comprobación del modelo, como he comentado anteriormente se ha separado la muestra en dos partes, de las cuales 2.000 *tweets* se han utilizado para la estimación del modelo y los otros 2.000 restantes para la comprobación del buen funcionamiento del mismo. En la siguiente matriz, se puede observar el número de *tweets* que yo he distinguido entre 0 y 1, es decir, entre *no me interesa* y *me interesa* y el número de *tweets* que dentro de estos comentarios el programa ha clasificado como interesantes y no interesantes.

		Predicho	
		0	1
Observado	0	1797	147
	1	51	5

Así pues, podemos observar como el modelo ha resultado ser bastante efectivo puesto que dentro de los de los que yo he marcado como 0 que es un total de 1848, el programa ha acertado 1797, lo que implica un porcentaje de 97.24% de aciertos. Esto implica un porcentaje muy alto de aciertos, lo que determina que el funcionamiento del modelo es efectivo.

Conclusiones

Un buen análisis de la información es elemental para el crecimiento de una empresa, el mayor activo de una organización es el cliente, y de su satisfacción dependerá gran parte del éxito de una compañía. El análisis de redes sociales es una estrategia que va teniendo cada día más demanda debido al gran uso que tienen éstas en las últimas décadas además de que siguen en continuo crecimiento.

El presente trabajo ha aplicado una metodología efectiva para poder obtener una información que resulta verdaderamente útil a la hora de tomar decisiones respecto a estrategias comerciales, estrategias en las que el marketing debe de tener presente en todo momento la percepción del consumidor sobre la marca y los atributos más relevantes de ésta que la organización hace llegar al cliente.

En la lectura de *tweets* he podido descubrir en poco tiempo gracias a la técnica aplicada que existen varías aspectos que la compañía podría reforzar. He comprobado como una gran parte de insatisfacción de los clientes viene dada por la dificultad de contactar con el servicio de atención al cliente, por lo cual considero que sería conveniente mejorar el servicio del chat online desde la página web de manera que el cliente se encuentre con una respuesta lo antes posible. Además, otra de las quejas más comunes es que el servicio de paquetería no habilita las fechas correctas por lo que sería recomendable prestar especial atención a este aspecto.

Así pues, finalizo el presente trabajo con la satisfacción de haber conseguido los objetivos propuestos mediante la aplicación de técnicas de minería de textos en los que se extrae automáticamente la importancia de los mensajes y mediante las técnicas de análisis de opinión en el que se ha identificado los estados de ánimo que reflejan los *tweets*. El porcentaje de aciertos ha sido alto por lo que estas técnicas han demostrado ser muy efectivas.

Agradecimientos

Gracias al Grupo Decisión Multicriterio Zaragoza por permitirme utilizar el software que está desarrollando para las labores de minería de textos y análisis de opiniones, así como por alojar en su laboratorio la base de datos que ha dado soporte a este trabajo.

Gracias a mi tutor Alberto Turón que me ha guiado y ayudado en todo momento durante la elaboración del presente trabajo.

Bibliografía

Contreras Barrera, Marcial. Minería de texto: una visión actual. (2014). https://www.redalyc.org/pdf/285/28540279005.pdf

Delgado, Daniela. Fast Fashion: Estratégia para conquista do mercado globalizado. (2008). http://www.revistas.udesc.br/index.php/modapalavra/article/view/7598/5101

Molinero Casares, Luis Miguel. Métodos estadísticos de clasificación (2002). https://www.seh-lelha.org/metodos-estadisticos-clasificacion/

Delgado, Manuel. ¿Qué es el análisis de sentimiento? (2015). https://manueldelgado.com/que-es-el-analisis-del-sentimiento/

Torres, Lina. Análisis de sentimientos sobre el posconflicto colombiano utilizando herramientas de minería de texto. (2015). https://repositorio.escuelaing.edu.co/bitstream/001/403/1/Torres%20Samboni%2c%20Lina%20Andrea%20-%202016.pdf

Echevarría, Pedro. Text Mining Aplicado a la Clasificación y Distribución Automática de Correo Electrónico y Detección de Correo SPAM (2009). https://www.dspace.espol.edu.ec/bitstream/123456789/762/1/1412.pdf

Artega Navarrete, Diana. Minería de datos aplicado al marketing. (2018). http://148.215.1.182/bitstream/handle/20.500.11799/99090/Miner%C3%ADa%20de%20Datos%20Aplicado%20al%20Marketing..pdf?sequence=1&isAllowed=y