Social Media, Marketing y Business Intelligence

Stratebi. Open Business Intelligence

CONTENIDO

DATOS EN LAS REDES SOCIALES

Actualmente gracias a la tecnología, las empresas se ven desbordadas por grandes volúmenes de datos que representan el negocio y todo lo que hay en su entorno y le afecta directa o indirectamente. Gracias a herramientas poderosas de análisis, como Pentaho, conseguimos transformar esos datos en información, y la información en conocimiento. No nos basta con tener los datos e indicadores creados en base a estos datos, sino necesitamos saber sacar provecho de los mismos para tomar decisiones para mejorar competitivamente y obtener una ventaja de cara a otros competidores.

El mundo de la tecnología y la forma de que los usuarios la usen ha evolucionado mucho. Actualmente, casi todas las personas conocen internet y la usan día a día. Desde la aparición de las redes sociales, estos tienden a subir contenido personal: opiniones, datos personales, fotos, videos…etcs. Generando cada día un volumen de datos muy elevado que actualmente no es analizado y que si se recopila, procesa y se generan indicadores las empresas pueden mejorar su estrategia.

Existen millones de blogs hablando de infinidad de cosas que nos rodean, cientos de redes sociales, cientos de comunidades dedicadas a algo en concreto, miles de medios de comunicación…etcs. No solamente tenemos las redes sociales tan conocidas como Facebook o Twitter, sino que tenemos otras como Youtube, Flicker, Blogspot, Linkedin, Myspace… y muchas más donde extraer información. La pregunta es, ¿por qué nos detenemos solamente en Facebook o Twitter?

QUÉ GANAMOS CON EL ANÁLISIS DE LOS DATOS SOCIALES

Gracias a la recopilación de los datos que están dispersados por diferentes fuentes de datos comentadas anteriormente, una empresa puede obtener unos beneficios a corto y medio plazo:

Optimización del marketing. Las herramientas de hoy en día apenas muestran "el qué" pero no el " por qué". Conseguiremos saber a qué sector enfocar una campaña de marketing y saber por qué tenemos que focalizar esta estrategia en ese sector.

Por ejemplo, Lanzando una foto de un nuevo producto de la compañía, a través de los "me gustas" se puede obtener la información de esta gente y se podrá lanzar campañas futuras orientadas a un tipo determinado de persona.

Capturar ideas y clientes insatisfechos. Identificaremos lo que el cliente piensa o desea de nuestros productos. Conseguiremos ver las lagunas en nuestros productos y servicios de la organización a través de los usuarios.

Consciencia de la situación. Conseguiremos identificar y controlar las principales tendencias, comprender cuándo surgen críticas para proteger la

experiencia del cliente o marca.

Mediante los comentarios en Facebook o Twitter se pueden obtener las ideas y problemas que han tenido los clientes y así poder mejorar el producto actuando de forma inmediata. Además, se podrá identificar mediante estos futuras tendencias a las cuales se puede adaptar los productos y así acortar el periodo de adaptación a las mismas.

Análisis de los sentimientos producidos. "¿Qué opinan de nosotros?" o "¿qué opinan de nuestro nuevo producto?" son solo dos ejemplos que difícilmente se pueden responder sin analizar toda la información en las redes sociales. Usando Twitter como ejemplo, si sale una campaña en un país se tendrá miles y miles de opiniones en forma "tweet" que habrá que procesar de alguna forma. Gracias a los algoritmos de Minería de Textos se puede extraer el sentimiento de estos "tweets" para poder saber qué están opinando y poder filtrar por criterios personales y así poder extraer el todo el conocimiento.

En la actualidad, muchas compañías de entretenimiento hacen un seguimiento de la opinión de sus series desde el lanzamiento para así poder retirar este tipo de series y economizar en la compra del espacio televisivo. Ya pueden retirar de la parrilla de programación con más agilidad que antaño ahorrando elevados costes de emisión.

ANALIZANDO CAMPAÑAS DE MARKETING

Toda empresa dispone de una herramienta para la gestión de relaciones con los clientes (Customer Relationship Management, CRM) con la cual podrá encontrar, atraer y ganar nuevos clientes, retener clientes y atraer los antiguos clientes mediante funciones de planificación y comercialización de servicios y/o productos. Estas funciones serán realizadas a través de campañas de marketing mediante diferentes canales de publicidad que dispone la empresa: redes sociales, mail marketing, anuncios web…

Desde Stratebi, se propone abrazar todos estos canales con un objetivo común: optimizar las acciones de publicidad en los canales usados en las acciones publicitarias para obtener el mayor retorno posible de las mismas. Se busca obtener los datos de una sola campaña que se realiza a través de diferentes canales para analizar la situación de la misma y el retorno que nos produce.

EJEMPLO DE ANÁLISIS DE CAMPAÑA MARKETING

Supongamos que se realiza una campaña publicitaria de un evento en la cual se tiene un espacio en una página web con toda la información. Este espacio tendrá una URL asociada clave con la cual se trabajará en diferentes canales. La estrategia es publicitar este evento y obtener todo el conocimiento posible. Para ello se usarán 3 canales y en todas ellas un nexo común que es la URL:

- Mailing: se mandará un e-mail a todos nuestros contactos de nuestro CRM. En un primer momento no se hará ninguna orientación a ningún tipo de contacto. Dispararemos a todos en un email informativo general.

- Twitter: se publicarán una serie de "tweets" informativos con el link de nuestro evento.
- Facebook: se publicarán una serie de noticias en el timeline de la empresa.

Una vez hecho la publicidad en diferentes canales de marketing, se buscará recoger todos los datos que se obtienen de cada uno de los canales para analizar cómo ha sido la campaña y mejorar en una segunda etapa la misma, es decir, obtener los datos y transformarlo en conocimiento y posteriores acciones:

- Mailing: se obtienen estadísticas de la interacción de los contactos con este e-mail. Se obtendrán datos como por ejemplo: nº de contactos que han abierto el e-mail, nº de rebotes, nº de clicks en la URL…
- Twitter: se obtienen estadísticas de los followers y otros usuarios que han visto este link: nº de Re"""tweet"s"", nº de respuestas, perfil de los usuarios, sentimiento de las respuestas, alcance de la campaña.
- Facebook: se obtienen estadísticas de los fans que han visto esa publicidad: nº de "me gustas", nº de veces que ha sido compartida esa publicidad, nº de comentarios, sentimiento de los comentarios, perfil de los usuarios que han interactuado con la campaña.

Una vez obtenidos todos los datos y puestos todos ellos en común, se podrá observar las campañas de publicidad de una forma más profunda que si se analizaran estos datos separadamente por canales como es hecho tradicionalmente. Además, se posibilita el análisis de estos datos en una sola herramienta facilitando así este análisis.

MEJORANDO EL MARKETING DE LA EMPRESA

Como se ha visto en el ejemplo anterior, la empresa, al juntar todos los datos de diferentes canales de publicidad (Mail Marketing, medios digitales y redes sociales) conseguirá principalmente poder modificar la campaña en curso sin tener que esperar a resultados finales.

Hoy en día, las empresas lanzan campañas de publicidad en diferentes canales mediante el departamento de marketing. Estos actúan según la planificación y no obtienen *feedback* de cómo ha ido la/s campaña/s de marketing que han lanzado en base a esta planificación hasta que no acaba el periodo de la/s misma/s. Por ejemplo, si lanzamos una publicidad de un curso y tenemos el periodo de matriculación de 1 mes, este departamento solo podrá ver el número de matriculados solamente cuando acaba el periodo de matriculación, que es cuando se hace un informe de resultados.

Desde Stratebi, se propone facilitar las herramientas de análisis a estos departamentos de marketing para que puedan acompañar la evolución diaria de las campañas emitidas por diferentes canales de publicidad. Haciendo esto, el departamento podrá ver qué está ocurriendo y podrá tomar decisiones de mejora o lanzar nuevas campañas con diferente orientación antes de acabar el periodo establecido.

Esta mejora en la eficiencia de la campaña que se obtendría, repercutiría en los resultados y se obtendría consecuencias e impacto en las

ventas. Por ejemplo, si se está viendo que los primeros días se está vendiendo un producto entre jóvenes de 20 a 27 años, quizás haya que orientar la campaña al sector joven y focalizar esfuerzos en este colectivo, o quizás todo lo contrario. Esto se podría saber a pocos días de empezar la campaña, pudiendo tomar decisiones que optimicen esta campaña de marketing sin tener que esperar a la finalización de la misma.

QUÉ Y CÓMO ANALIZAR

Para poder analizar los datos de estas redes sociales, necesitaremos definir nuestros propios KPIs (Indicadores de desempeño del negocio) y así medir de forma cuantitativa lo que está ocurriendo en las redes sociales.

Actualmente existen muchas herramientas que permiten medir online lo que está pasando en algunas redes sociales pero están limitadas por la propia herramienta que o no dispone de KPIs personalizados o el histórico de datos no posibilita el análisis masivo debido a que está alejado en la nube y por lo tanto no está optimizado al proceso de grandes volúmenes de datos. Hoy en día, están apareciendo herramientas que se instalan en los servidores de las compañías para poder recoger todo este gran volumen de datos y poder así hacer este análisis histórico, pero son herramientas limitadas en el análisis de métricas pues dependen del proveedor y no del propio cliente.

Por ello, desde Stratebi proponemos mendiante las herramientas Open Source de la Suite Pentaho y técnicas que nos proporciona el Business Intelligence, crear un proyecto que busque recopilar todos los datos que se quieran analizar de diferentes fuentes de datos, en este caso las redes sociales, y mediante la creación de recursos visuales para la explotación de los datos analizar los KPIs personales previamente definidos. Con ello, conseguiremos una libertad total para el análisis de estos datos desde la perspectiva del cliente, pudiendo personalizar las soluciones con mucho detalle y huyendo de las limitaciones que poseen otras herramientas al analizar los datos sociales.

Imagen: Proceso de análisis de datos sociales

KPIS A ANALIZAR

Los indicadores de desempeño (KPIs) son los que nos aportarán esa información para poder analizar y saber qué está pasando o si se está cumpliendo lo esperado. Son muchas redes sociales pero en este documento se van a destacar algunos KPIs de las principales redes sociales: Facebook y Twitter.

Facebook.

La red social más importante del mundo y la que más usuarios tiene. Es una fuente de información que si es usada de forma correcta podrá aportarnos mucha información.

Como se sabe, en la red social tenemos lo siguiente para ser analizado:

- **Recursos**. Estos recursos son los elementos con los que los usuarios van a interactuar, pueden ser: fotos, preguntas ,mensajes de TL, links, videos. Todos estos recursos tienen una información en común: comentarios, "likes" y nº de veces que el recurso ha sido compartido por las personas
- **Personas**. Estos son los individuos que van a interactuar con los anteriormente comentados recursos. De las personas podemos tener mucha información pero dependerá de si tenemos o no esa información disponible y permitida por el usuario. Casi siempre suele estar abierta salvo que las personas no lo permitan. Por defecto viene abierto cuando se crea un perfil. Siempre que estudiemos grandes muestras, tendremos una gran aproximación de lo que la mayoría tiene en común.

Imagen: Ejemplo de recurso de fotografía.

Imagen: Ejemplo de persona. Mark Zuckerberg creador de Facebook

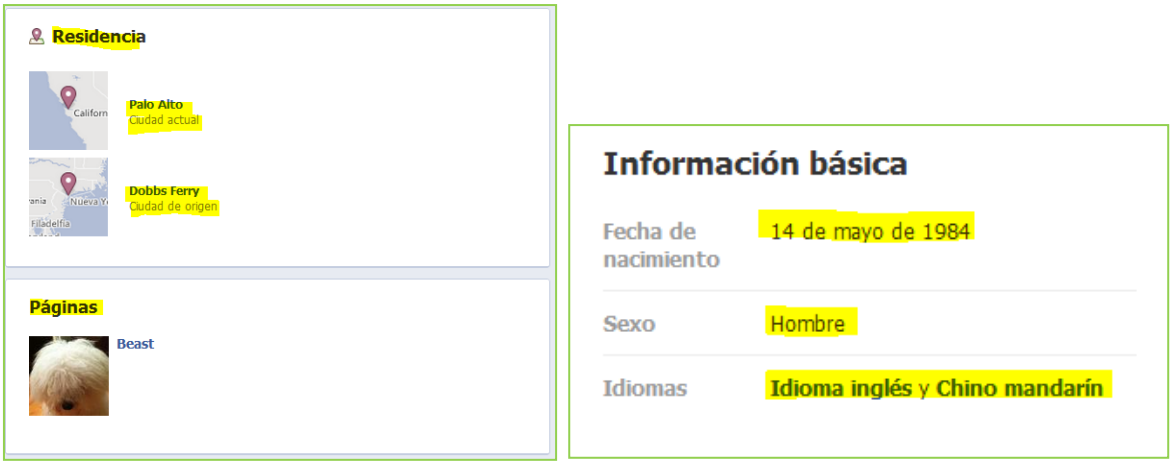

Imagen: Información básica de usuario Facebook

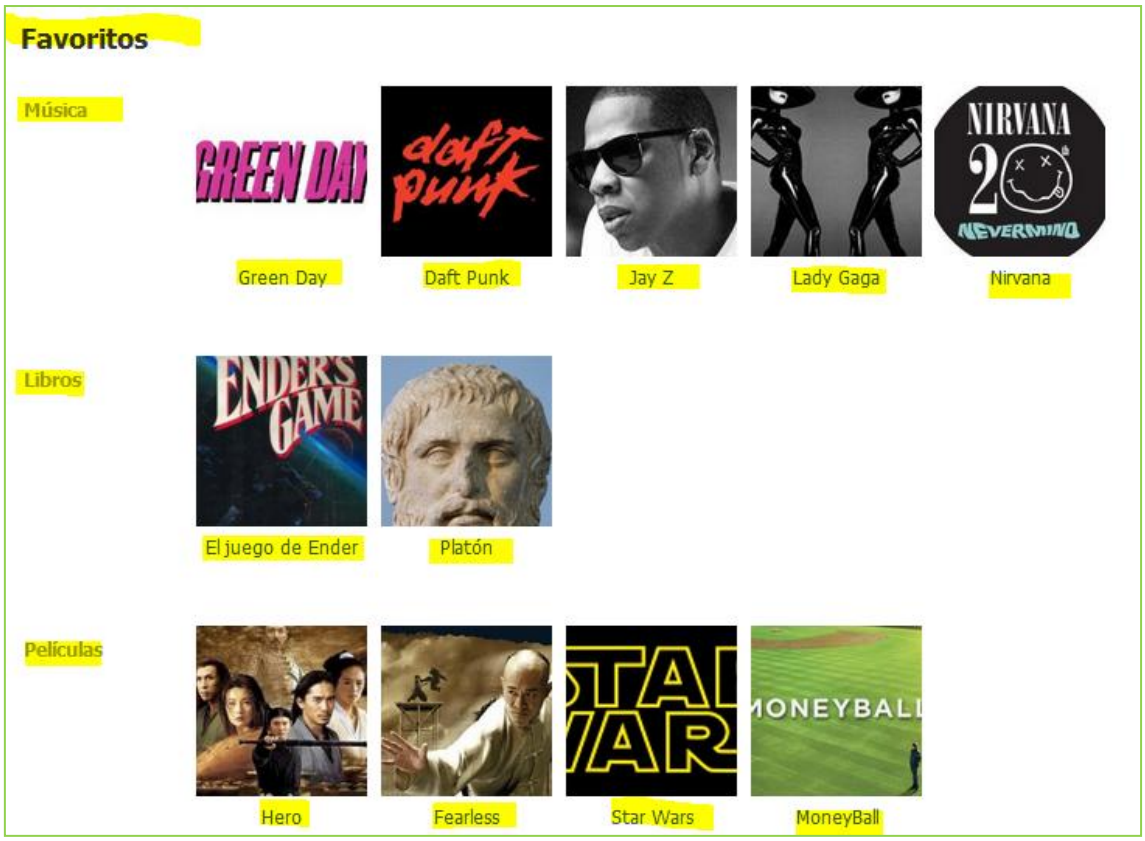

Imagen: Información de gustos de persona de Facebook

Por lo tanto, en base a estos recursos de la red social que podemos analizar, existen algunos KPIs básicos que sirven para monitorizar toda la actividad creada en esta red social:

- Número de veces compartido un recurso
- Recursos que tuvieron más interacciones
- Nº Likes
- Nº Comentarios
- Nº personas que han visto el recurso
- Localización de las personas

Twitter.

La red social basada en mensajes de pocas letras y una de las que más impacto social tiene. Es una gran fuente de información que nos puede ayudar con estrategias de marketing y sobretodo para ver qué opina la gente sobre un determinado asunto u objeto. Cabe destacar que todos los días se vierten millones de opiniones de diferentes asuntos, por lo tanto, aprovechemos esta información para generar conocimiento.

En esta red social se puede analizar los siguientes recursos:

- "Tweets". Son los mensajes que los usuarios escriben en su perfil. Es muy poca información la que tenemos pero podemos, junto con otras informaciones como la temporal o en base a lo que un usuario haya hecho obtener una serie de KPIs que nos será de gran utilidad. Además, cabe

destacar que lo que tiene valor de un "tweet" es la información de opinión que se aplica.

- Personas. Son los que escriben mensajes y crean contenido a través de la red social. Gracias a estos se puede estudiar qué ha pasado con el "tweet". Además, toda persona tiene una información asociada básica que puede ser analizada: localización, descripción, avatar, site y nombre.
- Hashtag. Etiqueta que se escribe en los ""tweet"s". Muy valiosa pues conseguiremos tener los "tweets" de las personas "etiquetados" y será mucho más eficiente el rastreo y búsqueda de opiniones.

Imagen: Información básica de perfil de usuario en Twitter

Imagen: Información de Tweet individual

Imagen: Información de Hashtag

Por lo tanto, aunque no tengamos mucha información a priori, podemos tener una serie de KPIs que nos ayuden a monitorizar esta información básica obtenida de los recursos de esta red social:

- Nº de RT's
- Nº de veces que han marcado favorito el "tweet"
- Nº de privados
- Nº de #FF conseguidos
- Nº de seguidores
- Impacto del "tweet"
- Menciones
- Listas en las que aparecemos

PROCESOS ETLS SOBRE REDES SOCIALES

Los procesos ETLs son aquellos que nos van a ayudar a extraer, normalizar y dejar todos los datos en un almacén de datos o base de datos optimizada para el análisis de grandes volúmenes de datos.

Analizando los datos de redes sociales se manejan muchos datos y que este tipo de almacenes llegan a tener una gran volumetría debido a la carga constante que hay que realizar. Por ejemplo, un comentario en Facebook además de la información natural de este recurso como la tipología, fecha de creación y otras informaciones, tiene información asociada con la interacción de los usuarios que hace que el volumen de datos sea muy grande. Estos recursos tienen X comentarios de respuesta, X veces compartidos y X "likes". Por cada uno de ellos hay una persona, por cada persona hay una serie de información que nos sirve para analizar tanto a la persona como la interacción con el recurso, entonces el grado de volumetría crece mucho conforme vamos recopilando datos periódicamente y debido a esto vemos que hay una gran ventaja en utilizar proyectos BI para el análisis de estos datos.

¿CÓMO Y DE DÓNDE SACAMOS LOS DATOS?

Las redes sociales más conocidas de internet disponen de unos recursos para poder extraer las informaciones de las mismas. El recurso más conocido es la API (Application Programming Interface) con el cual un usuario conocedor de lenguajes de programación puede obtener los datos en diferentes formatos (XML , JSON) y poder manipular en un futuro los mismo. La gran mayoría de redes sociales tienen este tipo de APIs para la extracción de los datos, por lo que el único requisito que tenemos al extraer los datos es tener el conocimiento de esta APIs para poder interactuar y hacer peticiones a las mismas.

Como ejemplo, en este documento solamente se verá la API de Facebook para mostrar cómo se podrían obtener ciertas informaciones de los últimos títulos

API FACEBOOK

En la imagen a continuación se puede observar una captura de la consola de Facebook con la cual se pueden hacer llamadas a la API y poder obtener los datos de la red social. En este ejemplo, se buscaba obtener la información de la página de "La Selección Española". Como se peude ver, esta consola nos devuelve una información básica: categoría, likes, foto, descripción, localización, web, año de fundación, etcs… Además, podemos navegar por diferentes tipos de conexiones y obtener lo último que han publicado en su TimeLine, las fotos, las pregutnas, estatus que ha escrito la página…etcs.

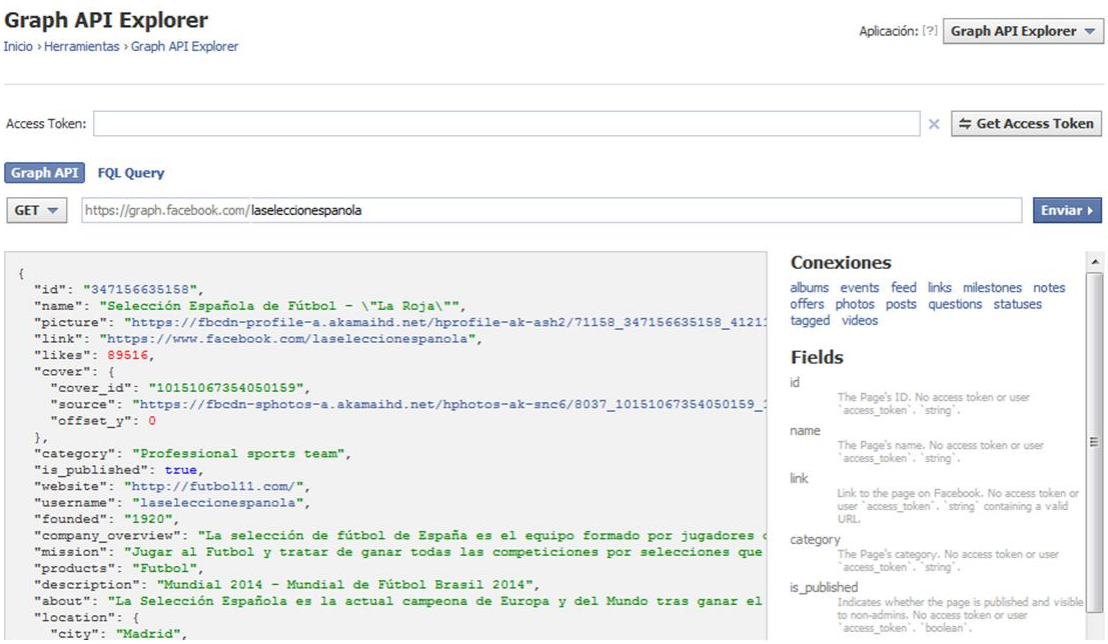

Imagen: Captura de la API de Facebook

Supongamos que queremos obtener la información asociado a los links que se han publicado en esta página pues nos parece importante ver qué respuesta tiene los fans ante el contenido. Para ello, vamos a utilizar Pentaho Data Integration en la extracción de los datos. A continuación se muestran las etapas del proceso de extracción y filtrado de datos que nos proporciona Facebook desde que leemos el

JSON que recibimos de Facebook hasta que devolvemos los resultados de los objetos que queremos.

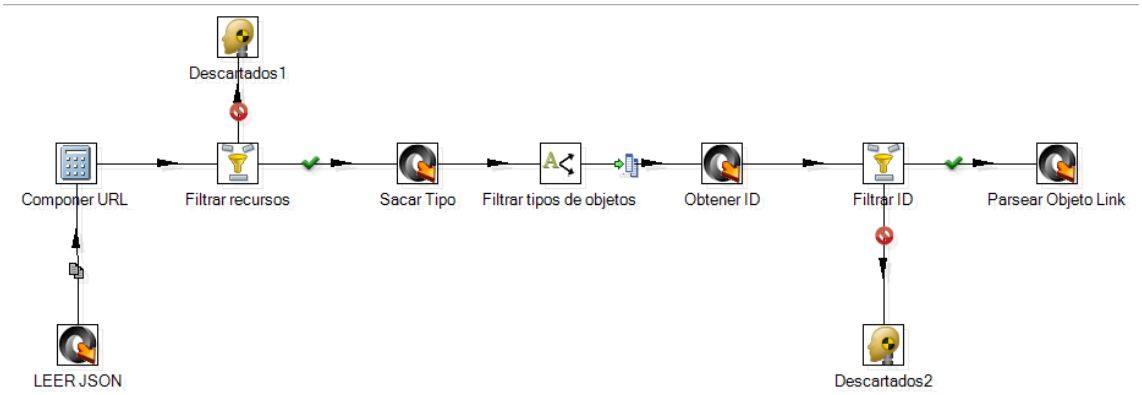

Imagen: Etapas de la ETL de extracción de información de links en fanpage

Resultado obtenido es el de la siguiente imagen. Se ha obtenido información asociada a los últimos recursos que se han creado en el TimeLine de la página de "La Selección Española de Fútbol", y obteniendo en total 10 recursos que son de tipología "Link" con la información de tipo, nº Likes, nº comentarios, título del link y la fecha de creación.

link	59	0	La Vuelta a España	2012-08-19T09:49:34+0000
link	20	0	Levante • El polaco Dariusz Dudka jugara en el levante	2012-08-17T22:38:39+0000
link	14	0	Espanyol • Simao Sabrosa ficha por el Espanyol	2012-08-17T22:38:06+0000
link	15	0	Queens Park Rangers • Bosingwa firma con el QPR	2012-08-17T22:03:10+0000
link	11	0	Fútbol en Portugal • Liga Sagres 2012 / 2013	2012-08-17T18:41:51+0000
link	20	0	Liverpool • Oussama Assaidi, tercer fichaje del Liverpool	2012-08-17T17:35:44+0000
link	15	0	2ª División A • Liga Adelante 2012 / 2013 (Post General)	2012-08-17T17:35:42+0000
link	26	0	Primera Division • Liga BBVA 2012 / 2013 (Post general)	2012-08-17T16:58:19+0000
link	26	1	Primera Division • Liga BBVA 2012 / 2013 (Post general)	2012-08-17T16:26:15+0000
link	9	0	Fútbol Base Madrid • SOLITUD DE ACCESO CATEGORÍAS BASE REAL C.D. CARABANCHEL	2012-08-17T15:50:39+0000

Imagen: Datos resultado de la ETL

Por lo tanto, como se puede ver es bastante sencillo extraer la información de Facebook mediante la herramienta Pentaho Data Integration. Solamente necesitamos tener claro qué queremos extraer y conocer la API de Facebook que nos proporcionará esta información. Cabe destacar que solamente se puede extraer información etiquetada mediante la API de Facebook aunque existen otros métodos más engorrosos para la extracción de datos de esta red social.

EJEMPLOS DE DASHBOARDS DE SOCIAL MEDIA

A continuación se muestran una serie de ejemplos creados con las herramientas Open Source que ofrece la suite Pentaho. Los dashboards han sido creados con el componente para Pentaho llamado STDashboard creado por Stratebi.

Imagen: Panel de login de Pentaho

Imagen: Pantalla de consola de usuario de Pentaho

En la imagen a continuación se muestra unos gráficos de análisis de la cuenta de Twitter. Como se puede observar, este análisis es global y busca mostrar datos básicos sobre la evolución de los usuarios, "retweets", cometarios y menciones en la herramienta durante los meses.

Imagen: Dashboard de análisis de evolución de cuenta Twitter

En la imagen a continuación se busca mostrar un análisis de los "tweets" de una cuenta. Mostrando la distribución de los "retweets" y respuestas por franjas horarias (AM/PM), un resumen de los "tweets" escritos y los comentarios, menciones y "retweets" por cada uno de ellos y una gráfica de evolución sobre los días de la semana de estos últimos.

Imagen: Dashboard de análisis de ""tweet"s"

En la siguiente imagen se busca analizar la evolución de la *fanpage* a lo largo de los meses del año por sexo y en general. Además, se muestra un mapa de las regiones de los fans de la página con el cual se podrá hacer *Drill&Down* y así analizar por sexo esta evolución por país.

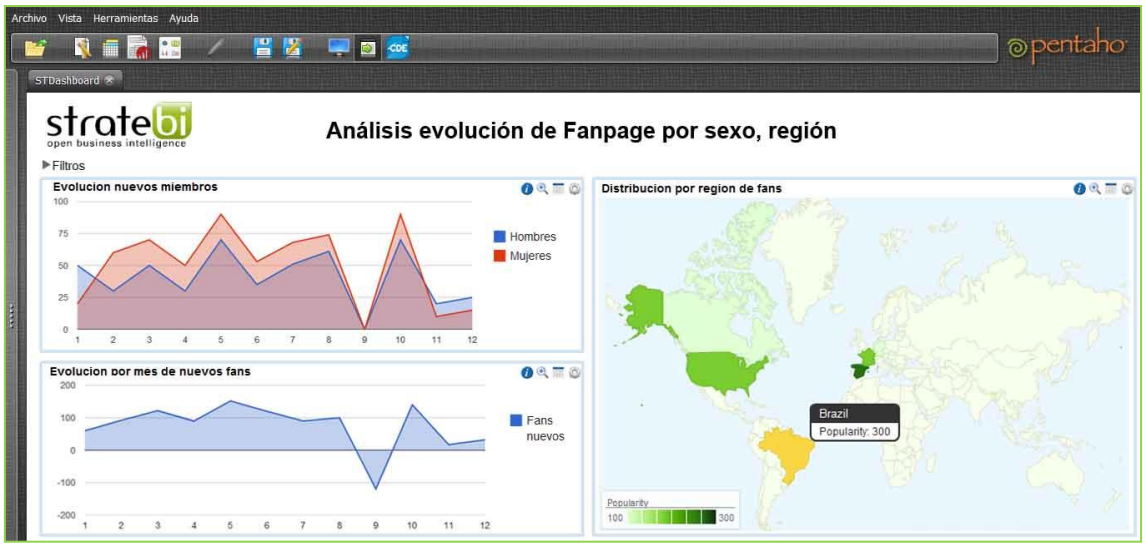

Imagen: Dashboard de evolución de fanpage en Facebook

Por último, en la imagen a continuación se muestra un resumen general de los fans de una página de Facebook. Se muestran gráficos para analizar el sexo, la edad, los estudios, la localización y la situación sentimental.

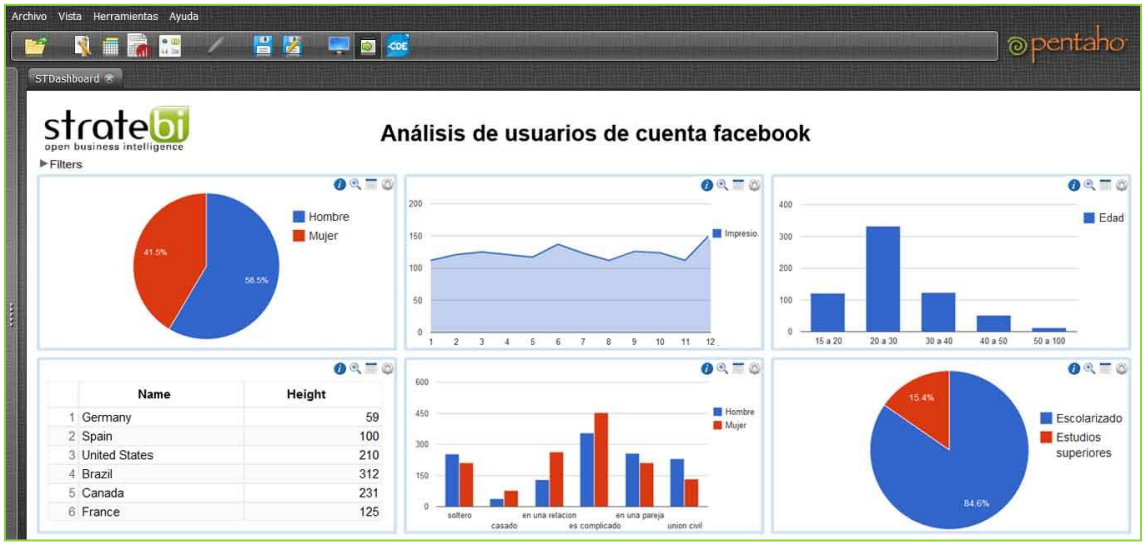

Imagen: Dashboard de análisis de los usuarios en fanpage en Facebook

ANÁLISIS DE SENTIMIENTO EN TWITTER

Gracias al análisis de sentimiento se podrá medir **la repercusión de una marca o producto dentro de las redes sociales**, en los casos de estudio que se ven a continuación se centran los esfuerzos en los términos "Pentaho" y "firefox". Al igual que estas palabras, se puede hacer un estudio para cualquier palabra (nombre de marca, producto, etcs) que desea ser monitorizada por el **departamento de Marketing** de una empresa.

Con el procedimiento que se ha seguido se puede analizar tanto marcas comerciales (ej: Coca Cola, Oracle) como por ejemplo para analizar el impacto del lanzamiento de un nuevo videojuego en el mercado o un estreno de cine. También resultaría adecuado su uso en el ámbito político para observar las opiniones que se tienen sobre los candidatos así como la monitorización de fuentes de información para detectar hostilidades

La técnica que se ha empleado para tal fin es la de Análisis de Sentimientos (conocida también como Minería de Opiniones, Clasificación de Sentimientos o Computación Afectiva). Esta es una técnica que dados unos datos de origen con un formato de texto, en los que aparecen opiniones o sentimientos sobre distintas entidades u objetos, permite extraer las opiniones de los mismos y clasificarlas. Es decir, se podría decir que se trata de un tratamiento computacional de las opiniones, sentimientos y fenómenos subjetivos en los textos.

Esta técnica utiliza el lenguaje natural, ya que es el que utiliza el usuario, y tratar computacionalmente este lenguaje conlleva ciertos problemas como la ambigüedad de las palabras, ya que dependen fuertemente del contexto. Los retos a los que se enfrenta son la extracción de las características sobre las que se está opinando y la clasificación de dichas características.

COMIENZO DEL ESTUDIO

Para el desarrollo de este estudio se ha realizado la búsqueda de la palabra "Pentaho" en la red de microblogging Twitter. Se ha escogido este término puesto que recientemente la compañía dedicada al Business Intelligence, ha lanzado al mercado una nueva versión de su software. En una primera captura se ve cómo podemos obtener la información fuente que vamos a utilizar en este estudio. Se trata de un paso de entrada RSS (versión de PDI utilizada 4.2.1 estable) en el que se especifican que los "tweets" que se van a buscar deben contener la palabra "Pentaho".

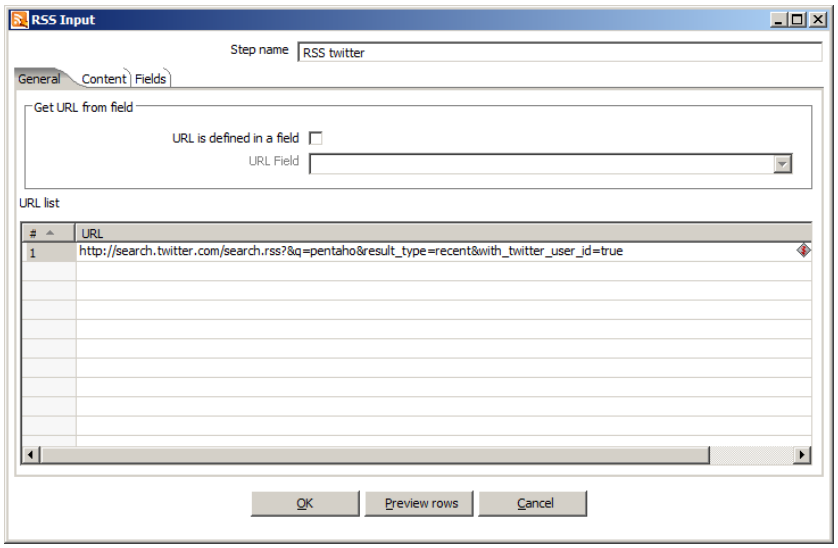

Imagen: Etapa de extracción de tweets en Kettle. URL

En la pestaña Fields se obtienen de manera automática los campos que cada entidad que se van a traer contiene. Es importante destacar que la API de búsqueda de Twitter tiene unas limitaciones para evitar una sobrecarga en sus servidores y que solo permite obtener información de los últimos 10 días o un máximo de 1500 "tweets".

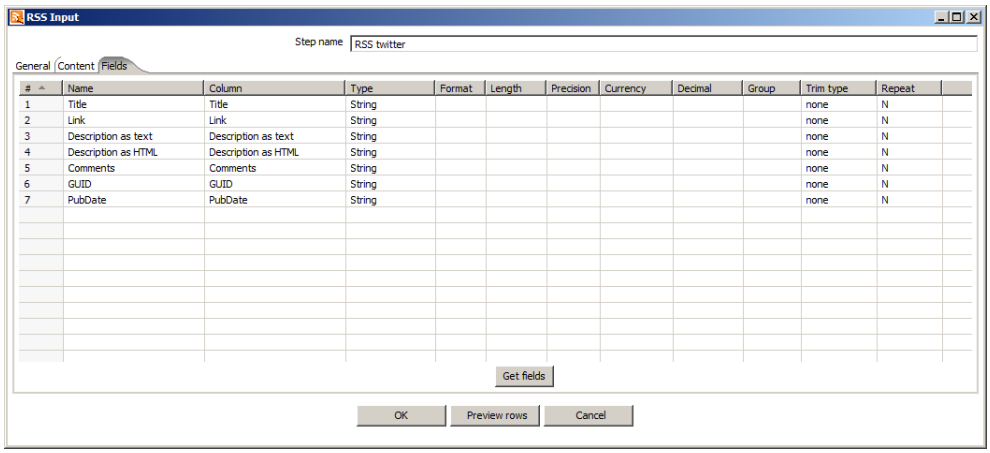

Imagen: Etapa de extracción de tweets en Kettle. Campos

Con todas las entidades procesadas con la herramienta de integración de datos que incluye el propio Pentaho, se realiza una nube de palabras con el texto de los "tweets" para obtener una ligera idea de qué términos tienen relación con la palabra que se está buscando.

Se puede ver esto como un resumen que se podría dar a una persona sin conocimientos del término "Pentaho". Con un simple golpe de vista y analizando las palabras que más resaltan se puede obtener información acerca del término Pentaho.

Imagen: Nube de palabras obtenidas

A continuación se pasará a analizar los "hashtags" de todo la muestra de "tweets" que se ha obtenido en la ETL. Un "hashtag" se trata de una cadena de caracteres precedida de un almohadilla (#), de esta manera los usuarios remarcan un tema o una palabra dotándola de una mayor importancia. Con un hashtag se suele expresar una opinión o sentimiento (ejemplos: #MeGustaPentaho, #BusinessIntelligence, #EleccionesFrancia).

Imagen: Nube de Hashtags obtenidos

En la muestra se puede comprobar cómo el hasthag mayoritario es "Pentaho", y son muy importantes también los términos "bigdata" y "Pentaho45" lo que indica que la nueva versión 4.5 ha sido muy popular en los "tweets" recientes. Además, se puede comprobar que en sus últimas versiones Pentaho incorpora conectividad y posibilidad de trabajar con las tecnologías emergentes de "bigdata". Resulta sorprendente que con un simple análisis automático se pueda obtener tanta información.

En la tercera parte de este estudio se ha tratado de realizar minería de datos con la muestra anterior pero el resultado obtenido no ha sido satisfactorio debido a que se partía de una muestra de sólo 750 "tweets", los cuales tras una minuciosa evaluación humana la gran mayoría expresan una opinión positiva mientras que

sólo 20 muestran rechazo o disgusto con Pentaho. Se ha desestimado este caso de estudio puesto que no se disponen de entidades negativas suficientes para que los algoritmos de minería aprendan con garantías.

Como alternativa, se realiza un análisis de "Firefox" para realizar Data Mining. Con esta búsqueda obtenemos un total de aproximadamente 1500 "tweets" (total de "tweets" que la API de Twitter permite obtener en una consulta). De estas entidades se han evaluado manualmente un total de 419 ""tweet"s", siendo los restantes analizado mediante scripts SQL de búsqueda de palabras o smileys que expresen sentimientos (Ejemplos: like, love, nice,cool, great, hate, crash, crazy, fail, error, ☺ , ☹). Con todos los "tweets" evaluados ya sea de manera manual o automática nos toca generar dos ficheros ARFF que nos servirán de fuente para la herramienta de DM Weka. Se genera un fichero con el total de los datos (1500 "tweets") y otro con el subconjunto de "tweets" que se han analizado de forma manual. Este subconjunto se pasará a llamar "conjunto de entrenamiento" pues en el futuro es lo que va a servir para entrenar a los diferentes algoritmos utilizados.

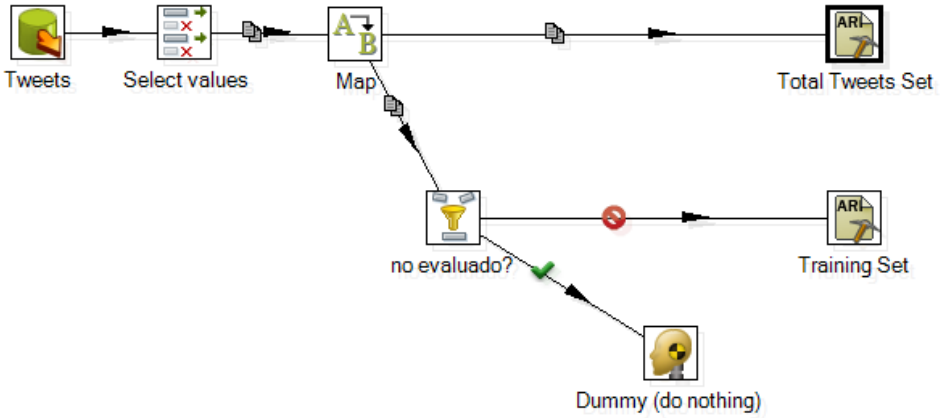

Imagen: ETL de extracción de conjuntos de tweets. Muestra a analizar y muestra entrenamiento.

ANALIZANDO CON EL ALGORITMO K MEDIAS

Se trata de un algoritmo voraz para partir los datos en k clústers. Este procedimiento utiliza la distancia euclídea de cada uno de los puntos al centro de cada clúster para su posterior agrupamiento. Es muy útil como primera aproximación por ser uno de los más veloces y eficientes. Debido a la naturaleza de los datos de muestra, se escogerá como valor de K el 3 puesto que será más fácil ver que existirán 3 conjuntos bien diferenciados de "tweets" (positivos, negativos y neutros).

Utilizaremos la herramienta de la suite Pentaho titulada WEKA explorer y se pasará como fuente el ARFF con los datos del conjunto de entrenamiento (419 "tweets"). Después en la pestaña Cluster elegiremos el algoritmo SimpleKMeans al que manualmente se pondrá el valor 3 en el campo numClusters. El siguiente paso es pasarle el fichero ARFF con el total de los "tweets" como conjunto para realizar los test. Una vez se haya configurado lo anterior se comenzará con el botón Start WEKA comenzará las siguientes acciones:

1) Tomar los datos de entrenamiento como datos maestros para realizar un aprendizaje

2) Aplicar esos conocimientos que ha adquirido al conjunto total de los datos. Realizando una predicción del conjunto final en el que se encontraría cada uno de los "tweets". El proceso de aprendizaje, denominado modelo, puede guardarse en un archivo con extensión .model para su reutilización dentro de Weka o dentro de Pentaho Data Integration a través del paso Weka Scoring.

Datos de la conjunto de entrenamiento (evaluación humana):

Valoración de "tweet"	Número de ""tweet"s"
Positivo	228
Negativo	98
Neutro	93
TOTAL	419

En los resultados que se muestran a continuación se puede ver que el clúster número 0 es el destinado a los "tweets" con sentimientos negativos, dado que los atributos negativos mayor valor medio en él. El clúster 1 por otra parte es el dedicado a las entidades en las que los usuarios de Twitter han proporcionado una opinión positiva puesto que se percibe que en él se alojan los 228 "tweets" que manualmente señalamos como positivos. En el tercer conjunto etiquetado como clúster 2 es en el que se guardan los objetos neutros.

```
Cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                           (419)       (98)      (228)       (93)
=============================================================
positive                  0.6062     0.2551     1.0044          0
                         +/-0.6563  +/-0.5974  +/-0.5268       +/-0

positive_smiley           0.0453          0     0.0833          0
                         +/-0.2083       +/-0   +/-0.277       +/-0

negativo                  0.3031     1.1224     0.0746          0
                         +/-0.5961  +/-0.6464  +/-0.3094       +/-0

negative_smiley           0.0143     0.0612          0          0
                         +/-0.1189   +/-0.241       +/-0       +/-0

h_eval                       Good        Bad       Good    Neutral
  Bad                  98 ( 23%)  98 (100%)   0 (  0%)   0 (  0%)
  Good                228 ( 54%)   0 (  0%) 228 (100%)   0 (  0%)
  Neutral              93 ( 22%)   0 (  0%)   0 (  0%)  93 (100%)
  Not_Classified        0 (  0%)   0 (  0%)   0 (  0%)   0 (  0%)

=== Evaluation on test set ===
Clustered Instances

0       188 ( 12%)
1       395 ( 25%)
2       967 ( 62%)
```

Imagen: Resultado visual del algoritmo.

El algoritmo de "K medias" tras realizar el aprendizaje, clasifica al total de los "tweets" de la siguiente manera: Cluster 0 : 188 "tweets", Cluster 1: 395 "tweets" y Clúster 2.

Datos del conjunto total (evaluación automática WEKA algoritmo 3 medias):

Valoración de "tweet"	Número de ""tweet"s"
Positivo	395
Negativo	188
Neutro	967
TOTAL	1550

ÁRBOL CREADO POR EL USUARIO

El segundo método que se empleará es el de generar manualmente un árbol que servirá como modelo. Para su elaboración se recurrirá a la pestaña superior de Weka y se elegirá User Classifier como algoritmo. Al comenzar, mostrará una ventana con una caja con el total de los datos sobre la que se deberá pulsar para ir a un formato de visualización de eje de coordenadas XY. A través de este eje se puede contrastar el valor de las diferentes variables y formar grupos de valores con un mismo valor.

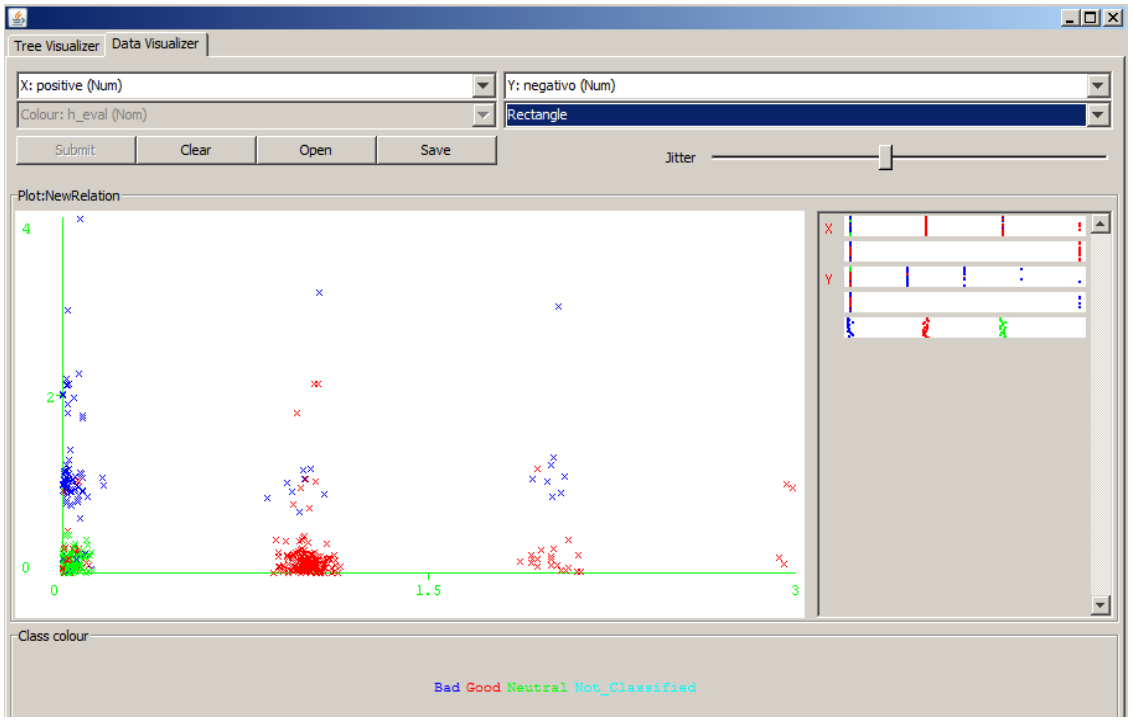

Imagen: visualización de datos de árbol creado manualmente

Una vez clasificados los distintos valores de la manera más homogénea posible se cambiará a la pestaña para visualizar el árbol y se visualizará algo como lo siguiente:

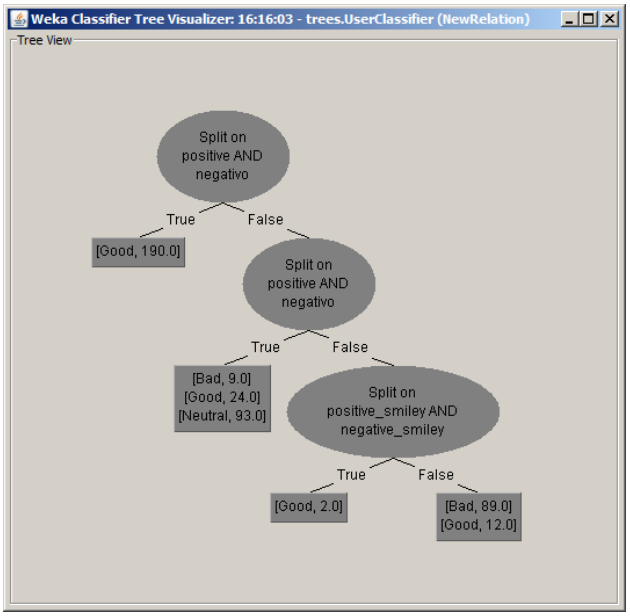

Imagen: visualizador de árbol

Al cerrar la ventana de edición del árbol se podrá ejecutar el procedimiento que se ha creado con el árbol sobre el conjunto de datos total. Hay que recordar especificar el conjunto total como "Supplied Test Set". Los resultados obtenidos son los siguientes

```
=== Confusion Matrix ===

  a   b   c   d   <-- classified as
 89   0   9   0 |   a = Bad
 12 192  24   0 |   b = Good
  0   0  93   0 |   c = Neutral
100 138 893   0 |   d = Not_Classified
```

Imagen: Resultado de algoritmo

En ella se puede ver la distribución de los 1131 "tweets" de la siguiente forma: 100 como negativos, 138 positivos y 893 como neutros. Estos resultados son un resumen de la ejecución global, en un último paso se generará una hoja de cálculo con los "tweets" y su predicción asociada.

ÁRBOL J 48

El algoritmo J48 de WEKA es una implementación del algoritmo C 4.5, uno de los algoritmos de minería de datos más utilizado. Se trata de un refinamiento del modelo generado con OneR (regla mayoritaria sobre un solo atributo).

El parámetro más importante que se debe tener en cuenta es el factor de confianza para la poda, confidence level, que influye en el tamaño y capacidad de predicción del árbol construido.

Para cada operación de poda, se define la probabilidad de error que se permite a la hipótesis de que el empeoramiento debido a esta operación es significativo. A probabilidad menor, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto es del 25%. Según baje este valor, se permiten más operaciones de poda. El árbol que este algoritmo genera automáticamente es el siguiente (recordar pasar como conjunto de test el total de los datos).

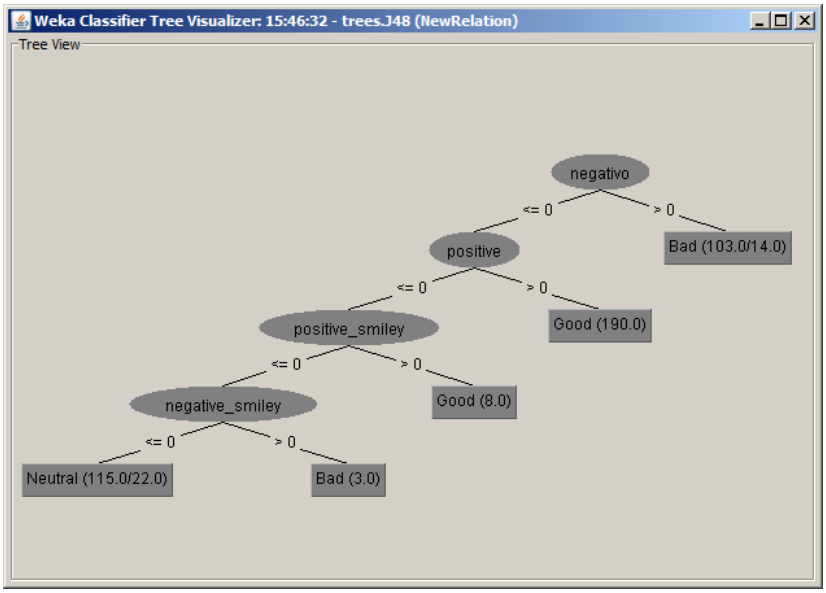

Imagen: visualización de árbol J48

Vemos en ahora el pseudocódigo del árbol.

```
negativo <= 0
|   positive <= 0
|   |   positive_smiley <= 0
|   |   |   negative_smiley <= 0: Neutral (115.0/22.0)
|   |   |   negative_smiley > 0: Bad (3.0)
|   |   positive_smiley > 0: Good (8.0)
|   positive > 0: Good (190.0)
negativo > 0: Bad (103.0/14.0)
```

Los resultados que se obtienen con este algoritmo son los siguientes

```
=== Confusion Matrix ===

  a   b   c   d   <-- classified as
 92   0   6   0 |   a = Bad
 14 198  16   0 |   b = Good
  0   0  93   0 |   c = Neutral
110 147 874   0 |   d = Not_Classified
```

Imagen: resultado de clasificación por árbol j48

En ella se puede ver la distribución de los 1131 "tweets" que no estaban en el conjunto de entrenamiento, los que estaban clasificados como Not_Classified, quedando de la siguiente forma: 110 como negativos, 147 positivos y 874 como neutros.

Con la ejecución de los dos algoritmos que utilizan una forma de árbol se puede realizar una primera comparativa directa, simplemente viendo cómo operan sobre el conjunto de datos de entrenamiento. Los resultados que arroja esta comparativa son que el algoritmo J48 es ligeramente superior a el correspondiente al árbol que se ha creado puesto que tiene un mayor porcentaje de aciertos.

User Tree						
Bad	**Good**	**Neutral**	**Not_classified**	**<--Classified as**		
89	0	9	0	**Bad**	Bien clasificados:	374
12	192	24	0	**Good**	mal clasificados	45
0	0	93	0	**Neutral**	% ✗	10,74 %
100	138	893	0	**Not_Classified**	% ✓	89,26 %

J48						
Bad	**Good**	**Neutral**	**Not_classified**	**<--Classified as**		
92	0	6	0	**Bad**	Bien clasificados:	383
14	198	16	0	**Good**	Mal clasificados	36
0	0	93	0	**Neutral**	% ✗	8,59%
110	147	874	0	**Not_Classified**	% ✓	91,41 %

ALGORITMO DE REGRESIÓN LINEAL

El cuarto algoritmo que se va a utilizar consiste en aplicar una regresión lineal al caso de estudio (método con el que se intenta modelar a través de una recta, la relación entre una variable dependiente Y, las variables independientes X_i y una constante aleatoria k, ecuación $Y= a_1X_1+ a_2X_2 + a_3X_2+\ldots +a_nX_n+ k$).

El primer paso que se debe de recorrer previamente de realizar la regresión propiamente dicha es el de normalizar las variables. Desde la ventana de pre-procesamiento se escogerá los atributos y se aplicará el filtro normalizar. Con esto se logrará los valores originales se distribuyan en el intervalo (0,1) para así de esta manera evitar que atributos con valores elevados metan ruido a la recta.

En el segundo paso se deberá seleccionar en la pestaña "Clasificar" el algoritmo "Linear Regression". En la captura de pantalla bajo estas líneas se puede observar que el propio algoritmo tiene marcada la opción de eliminar los atributos colineales (vectores colineales son aquellos paralelos en el plano) para que se reproduzca la mejor manera posible la recta de predicción.

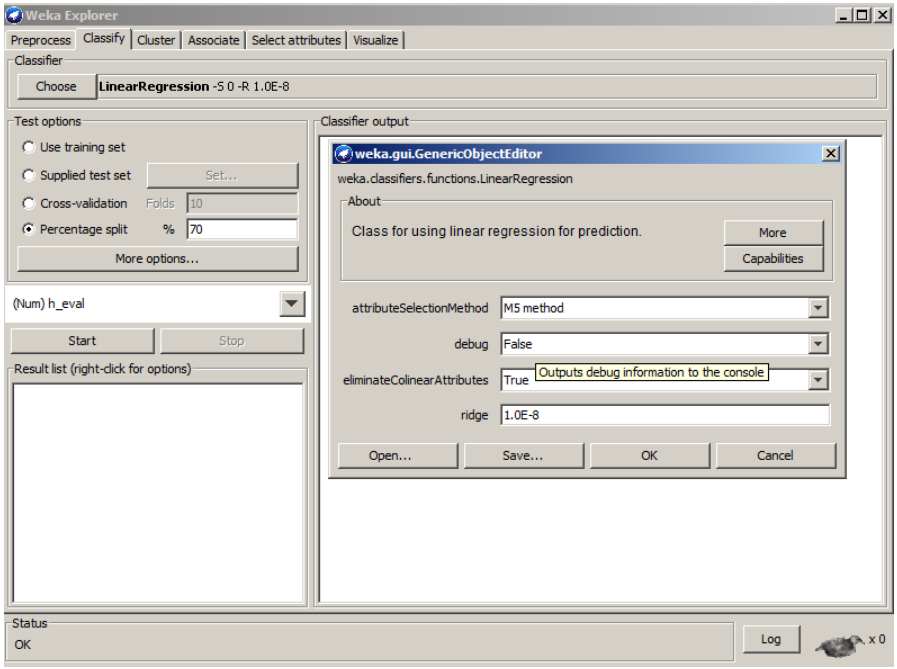

Imagen: visualización de pantalla de configuración.

La recta que este algoritmo genera es la siguiente:

h_eval = 0.257 * positive + 0.3504 * positive_smiley -1.6282 * negative -1.41 * negative_smiley+ 0.6314

COMPARACIÓN DE ALGORITMOS. RESULTADO.

En esta última etapa del estudio, se va a recuperar una transformación de Kettle con la que se obtendrán los "tweets" propiamente junto con su predicción asociada. El formato de las predicciones varía en función del algoritmo empleado: en K medias será un número de clúster, en los de tipo árbol el pronóstico devuelto será un término que describe el "tweet". Por último en la regresión lineal lo que se nos devuelve es un valor entero perteneciente siguiente conjunto {-6,-4,-3,-2,-1, 0, 1} que indica el sentimiento del "tweet". Los valores más altos (1 y 0) del conjunto indican las entidades claramente positivas mientras que los valores iguales o menores a -1 indican aquellos "tweets" que expresan un sentimiento negativo.

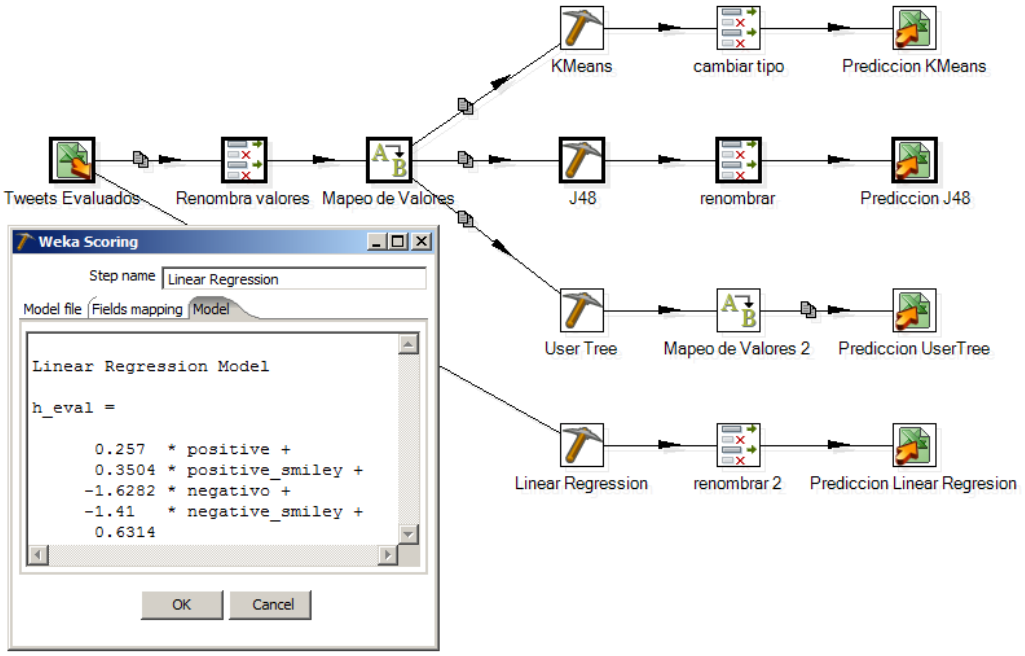

Imagen: Etapa de extracción de tweets en Kettle

En la siguiente tabla se pueden ver los diferentes algoritmos. El único error que se observa es el perteneciente al algoritmo el cual en este documento se ha creado, puesto que esta categorizando como neutro un "tweet" que claramente es negativo. También se observa en cómo la regresión lineal asigna el valor -3 al primer "tweet" expresando que contiene más de un término negativo lo que denota un mayor grado de insatisfacción.

Tweet	Clúster	User Tree	J48	Linear Regression
Is it just I or is Firefox the browser that hangs and crashes the most? :-(	0 (Bad)	Bad	Bad	-3
FastestFox - Browse Faster :: Add-ons for Firefox https://t.co/Khn21Wzo	1 (Good)	Good	Good	1
Icant believe that firefox has a better ftp client than android os has.	1 (Good)	Good	Good	1
FastestFox - Browse Faster :: Add-ons for Firefox https://t.co/uWQ0Mx5F	1 (Good)	Good	Good	1
@gregsidelnikov nice but its dont work on firefox ...	1 (Good)	Bad	Bad	-1
@danmasso Closed out Firefox and started over. :(	0 (Bad)	Neutral	Bad	-1
Honestly Firefox is annoying me now. #memorybloat	0 (Bad)	Bad	Bad	-1
#noscript is the most annoying #firefox addon	0 (Bad)	Bad	Bad	-1
RT @Three_Ninjas: Firefox crashes once a day.	0 (Bad)	Bad	Bad	-1
@misterjaydee thanks for the Firefox <3 ^WR	1 (Good)	Good	Good	1
I hate the new Mozilla Firefox.	0 (Bad)	Bad	Bad	-1
Oh.. #Firefox is cool too #JustSaying	1 (Good)	Good	Good	1
Firefox is getting slower day by day	0 (Bad)	Bad	Bad	-1
Long story short, thanks @firefox	1 (Good)	Good	Good	1
@tfaiso firefox and good show	1 (Good)	Good	Good	1
i hate firefox.	0 (Bad)	Bad	Bad	-1

Imagen: tabla comparativa de algoritmos.